LREC 2026

Joint Workshop on Legal and Ethical Issues in Human Language Technologies and Computational Approaches to Language Data Pseudonymization, Anonymization, De-identification, and Data Privacy (LEGAL2026 and CALD-pseudo 2026) @ LREC 2026

Workshop Proceedings

Editors
Ingo Siegert, Maria Irena Szawerna, Khalid Choukri, Simon Dobnik, Paweł Kamocki, Therese Lindström Tiedemann, Pierre Lison, Ricardo Muñoz Sánchez, Ildikó Pilán, Lisa Södergård, Kossay Talmoudi, Elena Volodina, Xuan-Son Vu

12 May 2026

Proceedings of the Joint Workshop on Legal and Ethical Issues in Human Language Technologies and Computational Approaches to Language Data Pseudonymization, Anonymization, De-identification, and Data Privacy (LEGAL2026 and CALD-pseudo 2026) @ LREC 2026

©ELRA Language Resources Association (ELRA), 2026
These proceedings are licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

ISBN 978-2-493814-86-9

Preface

In recent years, Artificial Intelligence, and in particular Large Language Models (LLMs), have moved from promising innovations to transformative forces shaping research, industry, and society at large. This rapid integration has intensified discussions not only around technological capabilities, but also around the governance of data, data protection, authorship, and accountability on a global scale. As these systems become foundational infrastructures, the legal and ethical implications of their development and deployment have become more complex and more urgent across jurisdictions.

Against this backdrop, a joint workshop is organized at LREC 2026, bringing together two previously distinct but highly complementary venues: the Workshop on Legal and Ethical Issues in Human Language Technologies (LEGAL2026) and the Workshop on Computational Approaches to Language Data Pseudonymization, Anonymization, De-identification, and Data Privacy (CALD-pseudo 2026). By merging these two workshops, this edition explicitly fosters a deeper integration of legal, ethical, and technical perspectives on data governance and privacy in language technologies.

Reconciling innovation with compliance remains a challenge in an increasingly diverse regulatory landscape. While approaches may differ in structure and enforcement, the data protection frameworks of different world regions remain a point of vigilance for the advancement of AI. Japan's Act on the Protection of Personal Information (APPI), South Korea's Personal Information Protection Act (PIPA), Brazil's Lei Geral de Proteção de Dados Pessoais (LGPD), and state-level instruments in the United States such as the California Consumer Privacy Act (CCPA) and the California Privacy Rights Act (CPRA) all reflect this trend. Similarly, countries such as Morocco, through Law 09-08, demonstrate that the establishment of legal frameworks governing personal data is now a widespread and shared priority, even as implementation models and levels of maturity vary.

Within this context, questions surrounding de-identification, pseudonymization, and the assessment of re-identification risk are no longer theoretical concerns but concrete technical and legal challenges, particularly in sensitive domains such as clinical data. Ensuring the accessibility of research data while protecting personal and sensitive information remains a central challenge, requiring approaches that effectively conceal identities while preserving usability and semantic value for downstream research tasks. The detection of indirect personal identifiers and the behavior of embedding representations with respect to personal data further illustrate the complexity of safeguarding privacy in modern AI systems. At the same time, new approaches seek to leverage LLMs themselves to assess privacy sensitivity and to support more robust evaluation of privacy-aware systems, while also accounting for potential biases introduced through pseudonymization processes.

Beyond personal data, the rise of generative and adaptive AI systems also brings to the forefront complex questions of intellectual property rights, liability, and transparency. Issues of authorship, ownership, and derivative works now intersect with algorithmic generation, challenging traditional notions of originality and rights allocation. In response, jurisdictions are progressively enacting AI-specific legal instruments, most notably the European Union's AI Act, which sets harmonized rules for trustworthy and accountable AI.

The increasing use of synthetic data for AI development and fine-tuning introduces additional opportunities, but also raises important legal and ethical questions regarding data provenance, risk mitigation, and compliance. These developments highlight a broader shift from data protection as a constraint toward data governance as an enabling framework for responsible innovation. More broadly, the need to preserve utility and semantic integrity while limiting the

exposure of identifiable entities reflects an ongoing tension between innovation and protection, particularly in light of adversarial re-identification risks.

Held as part of the LREC 2026 Conference, this joint workshop provides a forum for researchers, legal experts, and practitioners to explore the evolving relationship between language technologies and regulatory frameworks worldwide. By explicitly combining the LEGAL and CALD-pseudo communities, the workshop strengthens interdisciplinary exchange and enables a more comprehensive discussion of pseudonymization, data protection, and privacy-preserving technologies.

This year's program reflects a strong focus on privacy-aware methodologies, value-preserving data processing, and compliance-by-design approaches, including the integration of legal requirements into user-oriented infrastructures and consent management tools.

The workshop also addresses emerging concerns around authorship attribution and intellectual responsibility in the age of generative AI, as well as broader governance challenges related to data sharing, contractual frameworks, and the interoperability of infrastructures. The balance between FAIR principles and data protection requirements remains a key issue, particularly in the context of oral and cultural archives.

A notable theme this year is the growing importance of transparency and accountability as system-level properties. The identification of structural compliance gaps and the operationalization of regulatory requirements highlight the need to embed legal considerations directly into the architecture of AI systems.

Ultimately, the workshop underscores the need for a holistic and international approach that integrates technical innovation with legal and ethical responsibility. By fostering dialogue across disciplines and jurisdictions, it aims to contribute to the development of trustworthy, compliant, and socially responsible language technologies in a global context.

This volume presents the proceedings of the Joint Workshop on Legal and Ethical Issues in Human Language Technologies (LEGAL2026) and Computational Approaches to Language Data Pseudonymization, Anonymization, De-identification, and Data Privacy (CALD-pseudo 2026), co-located with LREC 2026. We would like to express our sincere gratitude to all authors for their valuable contributions, and to the Program Committee for their careful and dedicated reviews. Our further thanks go to the generous support from the Swedish Research Council through its funding to the research environment project *Grandma Karl is 27 years old*.

Workshop Organizers, April 2026

Organizing Committee

- **LEGAL 2026:**
  - Ingo Siegert, Otto-von-Guericke Universität Magdeburg, Germany
  - Paweł Kamocki, Leibniz-Institut für Deutsche Sprache, Germany
  - Kossay Talmoudi, ELDA, France
  - Khalid Choukri, ELDA, France
- **CALD-pseudo 2026:**
  - Maria Irena Szawerna, University of Gothenburg, Sweden
  - Simon Dobnik, University of Gothenburg, Sweden
  - Therese Lindström Tiedemann, University of Helsinki, Finland
  - Pierre Lison, Norwegian Computing Center & University of Oslo, Norway
  - Ildikó Pilán, Norwegian Computing Center, Norway
  - Ricardo Muñoz Sánchez, University of Gothenburg, Sweden
  - Lisa Södergård, University of Helsinki, Finland
  - Elena Volodina, University of Gothenburg, Sweden
  - Xuan-Son Vu, Lund University & DeepTensor AB, Sweden

Table of Contents

Workshop Program

Tuesday, May 12, 2026

9:00–10:10 **Welcome Session**

9:00–9:15 *Welcome and basic information from the organizers*
Workshop Organizers

9:15–10:10 *Introductory Lecture*
Paweł Kamocki

10:10–10:30 **Oral Presentations I: LEGAL**

10:10–10:30 *Transparency as Architecture: Structural Compliance Gaps in EU AI Act Article 50 II*
Vera Schmitt, Niklas Kruse, Premtim Sahitaj and Julius Schöning

10:30–11:00 **Coffee Break**

11:00–11:55 **Keynote**

11:00–11:55 *Keynote Speech*
Maja Bogataj Jančič

11:55–13:00 **Oral Presentations II: LEGAL**

11:55–12:15 *Towards Robust Evaluation for Privacy QA Systems*
Anna Leschanowsky, Zahra Kolagar, Erion Çano, Ivan Habernal, Dara Hallinan, Emanuël Habets and Birgit Popp

12:15–12:35 *LDS Contractual Framework: Principles, Status and Implementation*
Penny Labropoulou, Kossay Talmoudi, Dimitrios Gkoumas, Katerina Gkirtzou, Miltos Deligiannis, Leon Voukoutis, Athanasia Kolovou, Khalid Choukri, Stelios Piperidis and Dimitrios Galanis

12:35–12:55 *Authorship Attribution in the Times of LLMs within the Framework of the CRediT Taxonomy*
Pawel Kamocki and Andreas Witt

Tuesday, May 12, 2026 (continued)

13:00–14:00 **Lunch Break**

14:00–14:55 **Invited Talk**

14:00–14:55 *Privacy and anonymization in NLP: Are we barking up the wrong tree?*
Prof. Dr. Ivan Habernal

14:55–16:00 **Oral Presentations III: CALD-pseudo**

14:55–15:15 *DeID-Clinic: A Risk-Aware Pseudonymization Framework for Clinical Text De-identification and Re-identification Risk Assessment*
Angel Paul, Dhivin Shaji, Lifeng Han, Warren Del-Pinto, Goran Nenadic and Suzan Verberne

15:15–15:35 *Distilling Human-Aligned Privacy Sensitivity Assessment from Large Language Models*
Gabriel Loiseau, Damien Sileo, Damien Riquet, Maxime Meyer and Marc Tommasi

15:35–15:55 *Birds of a Feather: Do Embedding Representations of Personal Information Flock Together?*
Maria Irena Szawerna and Simon Dobnik

16:00–16:30 **Coffee Break**

16:30–17:45 **Poster Session**

16:30–17:45 *Modelling Legal Compliance in a Consent Wizard Application as Part of a Research-Centered and User-Oriented Data Infrastructure*
Aliena Strathmann, Marc-Levin Joppek, Maryam Mohammadi, Katja Politt, Paul T. Schrader, Annett B. Jorschick and Hendrik Buschmeier

16:30–17:45 *Balancing FAIR and GDPR: A Governance Framework for Oral Archives*
Elvira Mercatanti, Monica Monachini, Giovanni Abete, Silvia Calamai, Sergio Canazza, Alessandro Casellato, Virginia Niri, Cesarina Vecchia, Giulia Zitelli Conti and Giada Zuccolo

16:30–17:45 *Legal Considerations in the Use of Synthetic Data for AI Development and Finetuning: The Case of LLMs4EU*
Kossay Talmoudi, Khalid Choukri, Amélie Gourgeot and Florine Astruc

Tuesday, May 12, 2026 (continued)

16:30–17:45 *Evaluating Encoder- and LLM-Based Approaches for Robust Indirect Personal Identifier Detection*
Christoph Otto, Ibrahim Baroud, Akiko Aizawa, Sebastian Möller, Roland Roller and Lisa Raithel

16:30–17:45 *VEIL: A Benchmark for Value-Preserving Entity Identification Limitation*
Darina Gold, Shadi Rastegar, Alina Liebel and Alessandra Zarcone

17:45–18:00 Closing Session

17:45–18:00 *Closing Remarks*
Workshop Organizers

Transparency as Architecture: Structural Compliance Gaps in EU AI Act Article 50 II

Vera Schmitt[⋆,1,2,3], **Niklas Kruse**[⋆,4], **Premtim Sahitaj**[1], **Julius Schöning**[4]
Technichal University Berlin XpaliNLP Group[1], Research Center for Artificial Intelligence (DFKI) Berlin[2], Center for European Research on Trusted AI (CERTAIN)[3], Faculty of Engineering and Computer Science, Osnabrück University of Applied Sciences[4]
{vera.schmitt, premtim.sahitaj}@tu-berlin.de, {niklas.kruse,j.schoening}@hs-osnabrueck.de

Abstract

Art. 50 II of the EU Artificial Intelligence Act mandates dual transparency for AI-generated content: outputs must be labeled in both human-understandable and machine-readable form for automated verification. This requirement, entering into force in August 2026, collides with fundamental constraints of current generative AI systems. Using synthetic data generation and automated fact-checking as diagnostic use cases, we show that compliance cannot be reduced to post-hoc labeling. In fact-checking pipelines, provenance tracking is not feasible under iterative editorial workflows and non-deterministic LLM outputs; moreover, the assistive-function exemption does not apply, as such systems actively assign truth values rather than supporting editorial presentation. In synthetic data generation, persistent dual-mode marking is paradoxical: watermarks surviving human inspection risk being learned as spurious features during training, while marks suited for machine verification are fragile under standard data processing. Across both domains, three structural gaps obstruct compliance: (a) absent cross-platform marking formats for interleaved human-AI outputs; (b) misalignment between the regulation's 'reliability' criterion and probabilistic model behavior; and (c) missing guidance for adapting disclosures to heterogeneous user expertise. Closing these gaps requires transparency to be treated as an architectural design requirement, demanding interdisciplinary research across legal semantics, AI engineering, and human-centered design.

Keywords: AI Act, transparency, legal compliance, data governance, synthetic data, fact-checking

1. Introduction

The proliferation of generative AI systems across high-stakes domains such as journalism, scientific research, public administration, and automated decision-making has made the provenance and authenticity of digital content a pressing governance concern. Synthetic text, images, and structured data can now be produced at scale, at low cost, and with a degree of realism that makes human detection increasingly unreliable (Schmitt et al., 2025; Varanasi and Goyal, 2023). In response, the European Union has enacted the Artificial Intelligence Act (AI Act), which represents a decisive shift from *ex post* regulation of digital technologies toward the *ex ante* governance of AI systems (Kruse and Schöning, 2025). Unlike earlier frameworks such as the General Data Protection Regulation (GDPR), which targets data processing and individual rights, the AI Act directly regulates the behavior, outputs, and societal effects of AI systems, most directly through Art. 50 II, which enters into force in August 2026.

Art. 50 II imposes a dual transparency requirement on providers of generative AI: outputs must be labeled in a human-understandable manner *and* in a machine-readable form that enables automated verification. The provision applies across data types and deployment contexts, from text and images to synthetic datasets used in AI training pipelines. Its regulatory intent is clear: restoring epistemic trust in digital content by making AI involvement permanently traceable. Its technical specifications, however, are absent. Art. 50 II mandates that labeling solutions be effective, interoperable, robust, and reliable, while referring to watermarks, metadata labels, and cryptographic methods as candidate approaches. None of these criteria are defined operationally, and the technical standards that would supply such definitions are still under development (European Commission, 2025).

This paper argues that operationalizing Art. 50 II cannot be achieved through labeling solutions appended to existing systems. Generative AI operates under constraints that are structurally at odds with the regulation's requirements. Probabilistic outputs resist deterministic attribution; provenance chains fragment when human and AI contributions are interleaved; and the regulation's demand for reliable and effective transparency provides no guidance on what these properties mean for systems whose outputs vary across identical inputs. These are not implementation difficulties to be engineered away. They reflect conceptual mismatches between legal specification and technical reality that require coordinated research across disciplines.

To expose and structure these mismatches, we analyze two technically distinct but jointly reveal-

⋆Both authors contributed equally to this research.

Proceedings of LEGAL2026 and CALD-pseudo 2026 @ LREC 2026, pages 1–11
12 May 2026. ©ELRA Language Resources Association (ELRA), 2026

ing use cases. In *synthetic data generation*, dual-mode marking introduces a paradox between labeling persistence and training data integrity: watermarks designed to survive human inspection risk being learned as spurious features during model training, while marks suited for machine verification are fragile under standard data processing operations. In *automated fact-checking*, RAG-based multi-source attribution cannot be adequately represented in existing metadata schemas such as Dublin Core or Schema.org, and iterative editorial workflows degrade marking signals to the point where human and AI contributions can no longer be cleanly separated. Beyond these technical limitations, the Art. 50 II assistive-function exemption does not apply to fact-checking systems, as they actively assign truth values and confidence judgments to claims rather than merely supporting editorial presentation, a legal distinction with direct compliance consequences. Together, these cases expose three structural gaps that any compliance pathway must address.

To systematically address these tensions, two research questions are proposed that serve as the foundation for our analysis.

RQ 1: Which requirements of Art. 50 II (human-understandability, machine-readability, effectiveness, interoperability, robustness, and reliability) can be technically implemented within current AI systems, and where do systematic, non-incidental limitations arise?

RQ 2: How do legal concepts such as 'understandability' and 'reliability' diverge from their technical counterparts, explainability, quality assurance, and output consistency, and what are the compliance consequences of these divergences for practitioners and regulators?

Section 2 situates Art. 50 II within the AI Act's regulatory architecture and examines its technical requirements. Section 3 examines both research questions through the lens of two high-stakes use cases, assessing where compliance is feasible and where it breaks down. Section 4 synthesizes the three structural gaps and identifies preliminary compliance pathways grounded in the use case analysis. Section 5 concludes by repositioning transparency from a post-hoc metadata problem to an architectural design requirement, and outlines what a research agenda would need to deliver before Art. 50 II takes effect.

2. Art. 50 II as Technical Requirement

Art. 50 of the AI Act establishes transparency obligations for providers and deployers of certain AI systems, sitting within Chapter IV and entering into force on 2 August 2026. Notably, the EU Digital Omnibus proposal (COM(2025)0836) specifically targets Art. 50(2), proposing a six-month delay until February 2027 for systems already on the market, an extension driven by the same standards vacuum this paper analyses, and one that defers rather than resolves the structural compliance problem. The provision addresses four distinct scenarios: AI systems that interact directly with natural persons (Art. 50 I); systems that generate synthetic audio, image, video or text content (Art. 50 II); emotion recognition and biometric categorisation systems (Art. 50 III); and systems that generate or manipulate deep fake content or text published for public information purposes (Art. 50 IV). This paper focuses on Art. 50 II, which imposes the most technically demanding obligations and applies directly to providers of generative AI systems, including general-purpose AI models.

Art. 50 II requires providers of systems generating synthetic audio, image, video or text content to ensure that outputs are marked in a machine-readable format and detectable as artificially generated or manipulated. Crucially, the obligation is qualified: providers must ensure their technical solutions are effective, interoperable, robust and reliable *as far as this is technically feasible*, taking into account the specificities and limitations of various types of content, the costs of implementation, and the generally acknowledged state of the art as reflected in relevant technical standards. Three exemptions limit the obligation's scope. First, it does not apply where AI systems perform an assistive function for standard editing. Second, it does not apply where the system does not substantially alter the input data or its semantics. Third, it does not apply where use is authorised by law for purposes of detecting, preventing, investigating or prosecuting criminal offences. The interaction between these exemptions and the use cases examined in this paper is non-trivial and is addressed in Section 3.

The regulation is formulated in a deliberately technology-neutral manner: Art. 50 II names no specific technical implementation and provides no operational definition of the four quality criteria. Recital 133 identifies candidate approaches, including watermarks, metadata labels, and cryptographic methods, but does not specify formats, protocols or measurement criteria. This neutrality creates a structural compliance problem. A provider cannot determine from the regulation alone whether a given watermarking scheme satisfies the effectiveness criterion, what interoperability requires across platforms and jurisdictions, how robustness should be measured under realistic processing conditions, or what reliability means for systems whose outputs are probabilistic and non-deterministic. These are not peripheral questions: they determine whether any specific technical implementation is compliant.

The regulation's response to this uncertainty is to defer to technical standards, referenced in Art. 50 II as the intended source of operational guidance. Art. 50 VII further provides that the AI Office shall facilitate codes of practice to support effective implementation, and that the Commission may adopt implementing acts to approve or specify common rules for these obligations. However, the relevant harmonized standards remain under development by international, European and national standardization bodies, and it is unclear whether they will be finalized before the provision enters into force (European Commission, 2025). The standards available to date address adjacent concerns rather than output-level transparency. ISO 42001 governs organizational AI management systems. ISO/IEC 24028 addresses trustworthiness at the system level. ISO/IEC 24027 provides methodologies for bias assessment. None specifies how a label should be designed, embedded, or verified for a given class of AI-generated output, nor how persistence should be maintained once an output passes through downstream processing, editing or format conversion.

This gap between regulatory requirement and available technical guidance is structural rather than incidental. The dual transparency obligation, which this paper reads as requiring both a human-understandable and a machine-readable marking, applies across a heterogeneous class of data types, including text, images, audio and structured datasets, each of which presents distinct technical constraints. The absence of harmonized, output-level standards leaves providers without a clear compliance pathway, and the technically feasible qualification in Art. 50 II, while pragmatic, introduces its own ambiguity: it is unclear who determines feasibility, by what standard, and at what point in the system life-cycle. These unresolved questions set the stage for the specific technical breakdowns examined in the use cases that follow.

3. Use Cases Influenced by Art. 50 II

This section examines the operational challenges of Art. 50 II through two high-stakes use cases: synthetic data generation and automated fact-checking. These cases are technically distinct but jointly detecting similar issues. Each exposes a different facet of the compliance gap identified in Section 2: the synthetic data case reveals how the dual transparency obligation conflicts with the technical requirements of model development pipelines, while the fact-checking case reveals how iterative human-AI workflows undermine provenance tracking and why the assistive-function exemption does not provide support. Together, they ground the three structural gaps described further in Section 4.

3.1. Synthetic Data Generation Systems

The availability of data for developing high-performance AI systems remains a significant challenge across both academic and economic fields (Li et al., 2021; Sambasivan et al., 2021; Malerba and Pasquadibisceglie, 2024). Modern AI systems, particularly those employing transformer architectures (Kaplan et al., 2020; Richter et al., 2022; Halevy et al., 2009), necessitate substantial datasets to attain optimal performance. Conventionally, data has been obtained from the real world, a process frequently associated with high costs and limited availability (Buhrmester et al., 2011). These facts suggest that the development of next-generation AI systems is encountering a significant impediment: insufficient data to enhance their capabilities. A potential solution to this challenge is to augment existing datasets with synthetic data (Mumuni and Mumuni, 2022; Wachter et al., 2025).

As illustrated in Figure 1, the generation of synthetic data encompasses various stakeholders and a range of digital inputs and outputs, including one or several datasets, AI modeling interfaces, and licenses. The synthetic data itself can manifest in various forms, including texts, images, and time series. Although disparate data types are derived from reality, they do not exist outside the synthetic data space. Synthetic data offers a variety of added values. Real-world data can be collected through direct or indirect methods, both of which are derived from reality. These datasets are often subject to real-world biases and can be expensive or time-consuming (Wachter et al., 2024; Gebru et al., 2021). Consequently, existing feature gaps or biases can be addressed by leveraging synthetic data. Another potential application is the pre-training of large AI models, in which synthetic features are used to train fundamental features, thereby improving performance on real datasets. A salient question that remains unanswered is the extent to which synthetic data can completely replace real-world data. In the context of human subjects research, synthetic data often exhibits a high degree of indistinguishability from artificial data, which falls under Art. 50 II if generative AI is used to create it. However, given the inherent variability of image data, particularly in its presentation and design, it is important to note that, e.g., synthetic image data often contains features that are more readily identifiable by an AI system than by humans. Consequently, a discrepancy arises between synthetic and real data, known as the "reality gap" or "domain gap" . (Wachter et al., 2025; Peng et al., 2018). The extent of this discrepancy remains a subject of ongoing research and will be important in determining whether synthetic data needs to be labeled in the context of Art. 50 II.

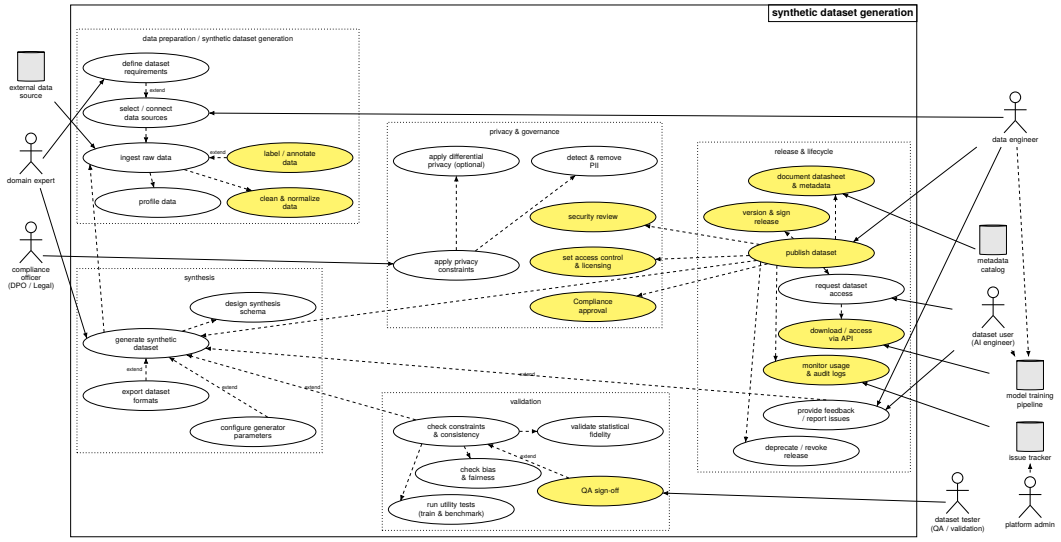


Figure 1: Use case diagram of a synthetic data generation system; the yellow-highlighted use case might be relevant for fulfilling Art. 50 II. Note that, depending on the system design, some use cases may be more or less relevant to Art. 50 II.

3.1.1. Art. 50 II Analysis and Applicability

A system for the generation of synthetic data, as illustrated in Figure 1, constitutes a standard case of Art. 50 II. Such systems overlap with the regulatory aspects of Art. 50 IV, which refers to deepfakes. Whether an AI system is subject to Art. 50 II or IV depends largely on the type of data it generates. According to Recital 134, a deepfake is characterized by the fact that the AI output is intended to falsely convince the recipient that the altered output is actual reality. If synthetic generation is used merely to substitute for the dataset creation process and identical artificial representations of real objects are created, this is likely to constitute a case under Art. 50 IV. This generation process could constitute a deepfake within the meaning of Recital 134. If, instead, data is generated to fill gaps in a dataset that cannot be collected because such data does not exist in a usable form, or to expand an existing dataset, then no deepfake is created; rather, new content is created. Such synthetic data generation systems are regulated under Art. 50 II. As depicted in Figure 1, the use cases within the subsystem data synthesis do not fall under Art. 50 II. However, within the subsystem synthetic dataset generation, three use cases “label & annotate data” clean & normalize” , and “define dataset requirements” , constitute a standard case of Art. 50 II, since the dataset output must be labeled in a way that is both machine-readable and accessible to humans, e.g., using the watermarks mentioned in Recital 133. However, if synthetic datasets are used to train AI systems, the machine-readability of watermarks can become an obstacle to the usability of synthetic data for these purposes. Since in these cases the watermark occurs very frequently, depending on the proportion of synthetic data in the entire dataset, there could be a risk that, due to the distribution of the feature, the AI system treats the watermark as a feature relevant to the training process rather than the actual training content. This links to the subsystem “release & life-cycle,” in which several use cases need to comply with Art. 50 II, e.g., to offer two versions of such synthetic data via an application programming interface (API) so that watermarks visible to an AI can be removed before training, enabling an unbiased training of the AI architecture. The version intended for human recipients contains a human-recognizable watermark instead (Simmons and Winograd, 2024; Bohacek and Vilanova Echavarri, 2025; Kruse and Schöning, 2024). Such watermark methods allow the standard’s purpose to continue to be fulfilled while the whole life-cycle and privacy governance ensure that synthetic data generation systems are usable and compliant with Art. 50 II.

3.1.2. Technical and Operational Challenges

Implementing Art. 50 II’s dual transparency requirements for synthetic data generation systems presents technical and operational challenges within the subsystems “release & life-cycle” , “synthetic dataset generation” , and privacy & governance” . While watermarking techniques, as referenced in Recital 133, appear to offer a straightforward technical solution with permanently linked labels to the data (Militsyna, 2025; Łabuz, 2024), their practical implementation reveals significant complexities and will hinder the use of synthetic datasets.

Firstly, the fundamental tension between human-readable labeling and machine-readability engenders a technical paradox (Militsyna, 2025). Watermarks designed to be discernible to humans, e.g., subtle visual patterns or embedded text, often impede training when synthetic data is used to train AI models. As discussed in Section 3.1.1, watermarks have the potential to evolve into features that AI systems learn to recognize rather than disregard, thereby jeopardizing the integrity of the training data. Jeopardizing the integrity of the training data creates a critical conflict: the regulatory requirement for persistent labeling directly contradicts the technical need for clean, unadulterated data in model training pipelines (Wachter et al., 2025; Chen et al., 2024; Geirhos et al., 2020). The proposed solution of maintaining two versions of synthetic datasets for the subsystem's "release & life-cycle" use cases introduces operational complexity, requiring additional data management infrastructure and potentially increasing data providers' costs.

Secondly, contemporary watermarking technologies are deficient in terms of the robustness required for Art. 50 II's "reliable" standard. Watermarks are often susceptible to compromise when, e.g., subjected to conventional image processing operations such as compression, resizing, cropping, and format conversion. These operations are frequently employed in data pipelines (Chen et al., 2024; Wan et al., 2022). This fragility undermines the "persistence" requirement of Art. 50 II, as watermarks may be inadvertently removed or altered during standard data processing. Furthermore, the efficacy of watermarking varies considerably across different data types and content (Fernandez et al., 2023), with complex or highly unbalanced data posing particular challenges for reliable watermark embedding and detection.

Thirdly, the absence of interoperable technical standards engenders a fragmented implementation landscape (Simmons and Winograd, 2024; Bohacek and Vilanova Echavarri, 2025). Although Recital 133 references watermarks, metadata labels, and cryptographic methods, it does not provide detailed specifications regarding technical formats or protocols. This will result in a proliferation of proprietary watermarking solutions that lack cross-platform compatibility. This is why the use cases "label & annotate data" clean & normalize" , and "define dataset requirements" also need to consider Art. 50 II. To illustrate, a watermark embedded using one provider's technology may not be detectable by another provider's verification system, thereby violating the "interoperable" requirement of Art. 50. This fragmentation engenders substantial impediments for organizations that must integrate synthetic data from multiple sources or utilize data across disparate AI systems.

Fourthly, the scope ambiguity of Art. 50 II exacerbates these technical challenges. The regulatory framework lacks clarity on the necessity of labeling synthetic data used for AI training in a manner consistent with synthetic data distributed to end users. This ambiguity creates operational uncertainty for the system compliance office, which must decide when a watermark is human-recognizable under AI Act Art. 50 and the Accessibility Act. The absence of regulatory guidance on this distinction compels organizations to make potentially costly implementation decisions without clear legal direction.

The tension between regulatory requirements and practical AI system development presents a significant operational challenge. Data scientists and AI engineers generally prioritize data quality and model performance over compliance considerations (Rakova et al., 2021; Varanasi and Goyal, 2023; Sambasivan et al., 2021). The implementation and maintenance of dual labeling systems within the "release & life-cycle" subsystem may be a bottleneck to innovation rather than a necessary compliance measure.

The aforementioned challenges collectively demonstrate that Art. 50 II's transparency requirements cannot be met by simple technical add-ons to existing data-generation pipelines. Instead, a fundamental rethinking of the generation, labeling, and management of synthetic data throughout its life-cycle is necessary.

3.2. Fact-Checking Systems

Fact-checking systems constitute a paradigmatic application domain in which the transparency obligations of Art. 50 II intersect with highly complex technical and editorial workflows. Contemporary automated systems (Guo et al., 2022) increasingly rely on large language models (LLMs) embedded in multi-stage pipelines that typically comprise claim detection (Hassan et al., 2017), evidence retrieval (Sahitaj et al., 2025), veracity assessment, and the retrieval-augmented generation (RAG) (Lewis et al., 2020) of textual justifications or summaries. These pipelines connect content platforms, the actors who use the system's outputs, like journalists or moderators, and the organizations that deploy it, such as newsrooms, turning verified claims into reports that can determine moderation or editorial decisions (Schlichtkrull et al., 2023). When LLMs are used in automated fact-checking pipelines, it is often difficult to understand why the system reached a specific verdict. Explanations can end up sounding like after-the-fact analyses rather than justifying the actual reasoning behind claim-verification decisions (Tan et al., 2025). User studies also suggest that explanations may improve perceived clarity without reliably improving AI-assisted human decision quality, and

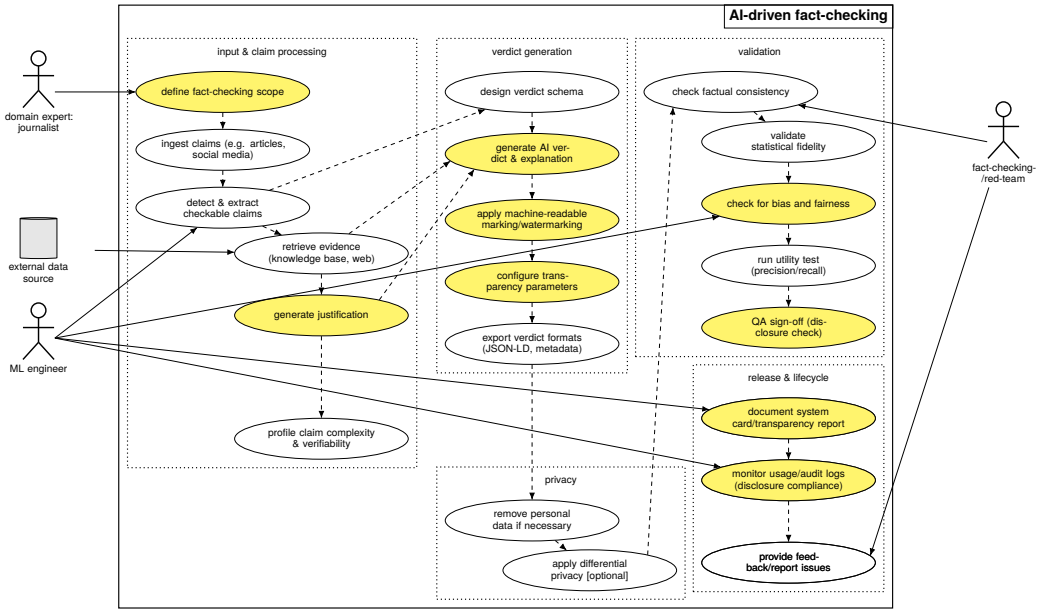


Figure 2: Use case diagram of an AI-driven fact-checking system; the yellow-highlighted use case might be relevant for fulfilling Art. 50 II. Note that, depending on the system design, some use cases may be more or less relevant to Art. 50 II.

can even increase overconfidence (Schmitt et al., 2025), indicating that transparency mechanisms can actively undermine their own regulatory purpose if poorly designed

3.2.1. Art. 50 II Analysis and Applicability

As illustrated in Fig. 2, fact-checking systems fall within the scope of Art. 50 II whenever they generate textual justifications for veracity assessments, synthesize evidence from multiple sources, or disseminate AI-generated evaluations on matters of public interest. The boundary between Art. 50 II and the assistive-function exemption in Art. 50 II sentence 3 turns on semantic transformation. The exemption covers systems that perform standard editing functions or do not substantially alter the semantics of their inputs. Fact-checking systems do neither: they actively transform user-submitted claims by assigning truth values, confidence scores, and evidential justifications. This constitutes a substantive semantic transformation rather than editorial assistance and therefore falls outside the exemption's scope. Unlike the synthetic data case, where the Art. 50 II versus Art. 50 IV boundary depends on intent and realism, the fact-checking case involves a different boundary question: whether the assistive-function exemption applies. The answer is no, and this has direct compliance consequences: providers cannot reduce their marking obligations through the exemption and must implement full dual transparency for all pipeline outputs.

Within the use case architecture depicted in Fig. 2, this obligation maps onto specific subsystems. In the verdict generation subsystem, the use cases *generate AI verdict and explanation* and *apply machine-readable marking and watermarking* constitute the primary compliance locus: outputs must be marked in a machine-readable format and detectable as artificially generated at the point of generation. In the release and lifecycle subsystem, the use cases *monitor usage and audit logs* and *document system card and transparency report* carry secondary compliance obligations, as they must preserve evidence of AI involvement across the output lifecycle. However, a critical complication arises from the interleaving of human and AI contributions. When a journalist reviews, edits, and publishes an AI-generated verdict, the boundary between AI-generated and human-authored content becomes impossible to demarcate cleanly. Unlike the synthetic data case, where a dual-version API can separate watermarked and watermark-free outputs for different recipients, no equivalent architectural solution exists for fact-checking pipelines: the human editorial process itself destroys the provenance chain that machine-readable marking requires.

Current governance standards exhibit the same structural gap identified in Section 2 and Section 3.1. ISO/IEC 42001:2023 addresses organizational AI management rather than output-level transparency. ISO/IEC 24027:2021 provides bias assessment methodologies relevant to source selection and ideological bias in fact-checking, but does not mandate user-facing disclosure of those biases. IEEE P7001 proposes graded transparency

levels that remain misaligned with Art. 50 II's simultaneous human-readable and machine-readable marking requirement. The question of risk classification adds a further compliance dimension. The societal impact of automated fact-checking, particularly in electoral contexts, suggests potential qualification as a high-risk system under Annex III, Art. 8 aa (Schmitt et al., 2024a), which would trigger full Chapter III obligations including conformity assessment, human oversight under Art. 14, and technical documentation under Art. 11. Journalist-facing tools functioning in an advisory rather than decisive capacity may alternatively qualify as limited-risk assistive systems, in which case compliance is largely confined to Art. 50's transparency requirements. This ambiguity is not merely theoretical: the applicable compliance pathway and associated architectural burden differ substantially between these two classifications, and no regulatory guidance or case law currently resolves the question.

3.2.2. Technical and Operational Challenges

Implementing Art. 50 II's dual transparency requirements across the *verdict generation*, *validation*, and *release and lifecycle* subsystems presents four interconnected challenges that parallel those in the synthetic data case but arise from structurally different sources.

First, the fundamental tension between human-readable and machine-readable marking is compounded in fact-checking by the iterative nature of editorial workflows. Text watermarking techniques that could in principle satisfy the machine-readable requirement are highly fragile under the editing, paraphrasing, and summarization practices common in newsrooms (Kirchenbauer et al., 2023). A watermark embedded in an AI-generated justification is unlikely to survive the revisions a journalist applies before publication. Unlike the synthetic data case, where a dual-version solution can separate marked and unmarked outputs at distribution time, no equivalent separation point exists in editorial workflows: the human review process intervenes between AI generation and publication, and it is precisely this intervention that destroys the mark. The regulatory requirement for persistent marking directly contradicts the operational reality of human-AI collaboration in journalism.

Second, the robustness of marking is structurally undermined by RAG-based multi-source attribution. When a justification draws on multiple retrieved sources, weighted by retrieval confidence and filtered by a veracity classifier, no current metadata schema provides a machine-readable format adequate to capture that provenance chain. Established schemas such as Dublin Core or Schema.org lack the semantics for expressing confidence-weighted, multi-step attribution (Kirchenbauer et al., 2023). This is analogous to the watermark fragility problem in synthetic data pipelines, but arises from semantic complexity rather than signal degradation: the provenance information exists but cannot be represented in any interoperable, machine-readable form. The real-time demands of breaking-news scenarios compound this further: cryptographic marking schemes that could support fine-grained provenance are computationally expensive and create latency trade-offs incompatible with live editorial workflows.

Third, the absence of interoperable standards produces the same fragmented implementation landscape identified in the synthetic data case, but with an additional cross-modal dimension. No unified marking approach currently exists for fact-checking outputs that combine text, images, and video, despite Art. 50 II's explicit multimodal scope. Existing provenance frameworks, including C2PA (C2PA Steering Committee, 2024), provide insufficient support for deeply interleaved, text-centric outputs in which human and AI contributions cannot be cleanly separated. A marking solution developed for text-based verdicts will not extend to video fact-checks without significant additional engineering, and no cross-modal standard currently fills this gap.

Fourth, the scope of Art. 50 II is ambiguous with respect to explanation quality in ways that have no direct parallel in the synthetic data case. Art. 50 V requires that disclosures be provided in a clear and distinguishable manner conforming to applicable accessibility requirements, but provides no guidance on calibrating explanation depth or format to different user groups. Post-hoc explainability methods such as LIME or SHAP generate technically accurate feature attributions but are inaccessible to non-technical users including most journalists and content moderators (Schmitt et al., 2024b). Natural-language explanations generated by LLMs are more accessible but introduce explanation-induced overreliance: fluent but incorrect rationales increase user trust without improving decision quality (Bansal et al., 2021; Schmitt et al., 2024b). The result is a one-size-fits-all transparency requirement that is likely to under-serve both technical and non-technical user groups simultaneously, and whose interaction with the hallucination and non-determinism properties of LLMs (Ji et al., 2023) means that legally verifiable quality guarantees cannot be provided under current technology.

Finally, human-AI collaboration introduces additional operational risks. Newsroom environments characterized by time pressure are particularly susceptible to automation bias, potentially eroding independent verification practices and diffusing ac-

countability when AI-assisted fact-checks prove incorrect (Liu et al., 2024). Although Art. 14's human oversight requirements and Art. 50's transparency obligations are conceptually complementary, concrete guidance on their coordinated implementation in fact-checking contexts remains largely absent. Taken together, these four challenges demonstrate that Art. 50 II compliance for automated fact-checking cannot be achieved through marking solutions appended to existing pipelines. The provenance chain that machine-readable marking requires is broken by the editorial workflows that human oversight demands, the semantic complexity of RAG-based attribution exceeds what current standards can represent, and the explanation requirements of Art. 50 V cannot be met without user-group-specific transparency designs that the regulation does not specify. As in the synthetic data case, compliance must be treated as an architectural requirement integrated from the outset, not a post-hoc addition. The next section synthesizes the structural gaps common to both cases and identifies the research directions needed to close them.

4. Structural Gaps and Research Agenda for Art. 50 II Compliance

The use case analyses show three structural gaps that any compliance pathway must address: the absence of cross-platform marking formats for interleaved human-AI outputs (RQ1), the misalignment between the regulation's reliability criterion and probabilistic model behavior (RQ2), and missing guidance for adapting transparency to heterogeneous user expertise (RQ2).

Pathway 1: Interoperable provenance standards. No current standard can represent provenance in outputs where human and AI contributions are interleaved. In synthetic data pipelines, the dual-version API approach partially addresses distribution-time separation but lacks cross-platform verification. In fact-checking pipelines, RAG-based attribution cannot be represented in schemas such as Dublin Core or Schema.org, and editorial workflows destroy provenance chains at the point of human review. Future work should extend frameworks such as C2PA (C2PA Steering Committee, 2024) to interleaved text-centric outputs and establish robustness benchmarks for marking methods under realistic processing conditions.

Pathway 2: Operational feasibility criteria. The undefined quality criteria in Art. 50 II leave providers without a clear compliance pathway. In synthetic data contexts, it is unclear whether training-time watermark removal violates the reliability criterion. In fact-checking contexts, no guidance exists on whether marks destroyed by editorial revision satisfy the regulation's persistence requirement. The Commission's own Code of Practice (European Commission, 2025) drafting process substantiates rather than resolves this gap: the first draft explicitly acknowledges that no single active marking technique currently meets the four quality criteria of Art. 50(2) (European Commission, 2025), and the second draft (March 2026) introduces flexibility without supplying the operational definitions whose absence constitutes the core of this pathway. Future work should develop measurable criteria differentiated by data type, deployment context, and recipient, including the unaddressed distinction between synthetic data for human recipients and for model training pipelines.

Pathway 3: User-group-specific transparency designs. Art. 50 V requires clear and distinguishable disclosures but provides no guidance on calibrating explanations to different user groups. Post-hoc methods such as LIME or SHAP are inaccessible to non-technical users, while natural-language LLM explanations introduce explanation-induced overreliance (Bansal et al., 2021; Schmitt et al., 2024b). Future work should develop and validate disclosure frameworks that adjust explanation depth and format to the recipient's role and expertise, including in time-pressured editorial environments where automation bias poses particular risks (Liu et al., 2024). These pathways are interdependent: interoperable standards are a prerequisite for machine-readable compliance, operational criteria are necessary to evaluate conformity, and user-centered designs determine whether the human-understandable dimension of dual transparency is achieved in practice. Together they define the minimum research agenda needed before Art. 50 II enters into force in August 2026.

5. Conclusion

Using synthetic data generation and automated fact-checking as diagnostic use cases, this paper shows that Art. 50 II compliance cannot be reduced to post-hoc labeling. In synthetic data pipelines, persistent marking conflicts with model training integrity; in fact-checking workflows, human editorial intervention destroys the provenance chains that machine-readable marking requires, and no equivalent architectural solution exists.

For RQ1, some Art. 50 II requirements are achievable where outputs can be accompanied by structured provenance artefacts that support cross-system verification. Systematic limitations arise wherever the regulation presupposes mark persistence and robustness under realistic transformations: precisely the conditions both use cases expose. For RQ2, legal understandability diverges from technical explainability in that explainabil-

ity methods target model behavior rather than informed use at the point of access. Legal reliability diverges from quality assurance in that probabilistic, non-deterministic LLM outputs cannot provide the verifiable consistency guarantees the regulation implies.

The compliance consequence is concrete: treating fluent model-generated rationales as evidence of reliability risks explanation-induced overreliance and weakens the regulatory objective of Art. 50 II. These findings support the paper's central argument: dual transparency must be treated as an architectural design requirement integrated across the full AI lifecycle, not a labeling add-on. Closing the three structural gaps identified in Section 4 requires coordinated action across legal semantics, AI engineering, and human-centered design before Art. 50 II enters into force in August 2026.

Acknowledgements

This research is funded by the Federal Ministry of Research, Technology, and Space (BMFTR) in the scope of the research projects news-polygraph (reference: 03RU2U151C), VeraXtract (reference: 16IS24066), and BIFOLD project FakeXplain. Moreover, this paper includes work carried out within the AgrifoodTEF-DE project. AgrifoodTEF-DE (reference: 28DZI04A23) is supported by funds of the Federal Ministry of Agriculture, Food and Regional Identity (BMLEH) based on a decision of the Parliament of the Federal Republic of Germany via the Federal Office for Agriculture and Food (BLE) under the research and innovation program 'Climate Protection in Agriculture'.

Disclosure: Grammarly, an AI writing assistance tool was used to improve the orthography and grammar of several paragraphs of text.

6. Bibliographical References

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *CHI Conference on Human Factors in Computing Systems*. ACM.

Matyas Bohacek and Ignacio Vilanova Echavarri. 2025. Compliance rating scheme: A data provenance framework for generative ai datasets. In *International Conference on Multimedia*. ACM.

Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. 2011. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *PERSPECT PSYCHOL SCI*, 6(1).

C2PA Steering Committee. 2024. Coalition for content provenance and authenticity (C2PA) technical specification.

Huajie Chen, Chi Liu, Tianqing Zhu, and Wanlei Zhou. 2024. When deep learning meets watermarking: A survey of application, attacks and defenses. *COMP STAND INTER*, 89.

European Commission. 2025. Code of practice on transparency of ai-generated content. Draft version, December 2025.

Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. 2023. The stable signature: Rooting watermarks in latent diffusion models. In *International Conference on Computer Vision (ICCV)*. IEEE.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *COMMUN ACM*, 64(12):86–92.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11).

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10.

Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data. *IEEE INTELL SYST*, 24(2).

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 1803–1812. ACM.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM COMPUT SURV*, 55(12).

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models.

Niklas Kruse and Julius Schöning. 2024. Legal conform data sets for yard tractors and robots. *COMPUT ELECTRON AGR*, 223:109106.

Niklas Kruse and Julius Schöning. 2025. Explainable and trustworthy ai compliance for farms.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *ADV NEUR IN*, 33:9459–9474.

Yifan Li, Xiaohui Yu, and Nick Koudas. 2021. Data acquisition for improving machine learning models. *VLDB Endowment*, 14(10).

Houjiang Liu, Anubrata Das, Alexander Boltz, Didi Zhou, Daisy Pinaroc, Matthew Lease, and Min Kyung Lee. 2024. Human-centered nlp fact-checking: Co-designing with fact-checkers using matchmaking for ai. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW2):1–44.

Donato Malerba and Vincenzo Pasquadibisceglie. 2024. Data-centric ai. *J INTELL INF SYST*, 62(6).

Kateryna Militsyna. 2025. Can copyright law benefit from the marking requirement of the ai act? *IIC-INT REV INTELL P*, 56(9):1734–1751.

Alhassan Mumuni and Fuseini Mumuni. 2022. Data augmentation: A comprehensive survey of modern approaches. *Array*, 16:100258.

Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. 2018. Sim-to-real transfer of robotic control with dynamics randomization. In *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE.

Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where responsible ai meets reality: Practitioner perspectives on enablers for shifting organizational practices. *ACM on Human-Computer Interaction*, 5(CSCW1).

Mats L. Richter, Julius Schöning, Anna Wiedenroth, and Ulf Krumnack. 2022. *Receptive Field Analysis for Optimizing Convolutional Neural Network Architectures Without Training*, pages 235–261. Springer Nature Singapore.

Premtim Sahitaj, Iffat Maab, Junichi Yamagishi, Jawan Kolanowski, Sebastian Möller, and Vera Schmitt. 2025. Towards Automated Fact-Checking of Real-World Claims: Exploring Task Formulation and Assessment with LLMs.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *21 CHI Conference on Human Factors in Computing Systems*, CHI '21. ACM.

Michael Schlichtkrull, Nedjma Ousidhoum, and Andreas Vlachos. 2023. The Intended Uses of Automated Fact-Checking Artefacts: Why, How and Who.

Vera Schmitt, Isabel Bezzaoui, Charlott Jakob, Premtim Sahitaj, Qianli Wang, Arthur Hilbert, Max Upravitelev, Jonas Fegert, Sebastian Möller, and Veronika Solopova. 2025. Beyond Transparency: Evaluating Explainability in AI-Supported Fact-Checking. pages 63–72.

Vera Schmitt, Aljoscha Burchardt, Jakob Tesch, Eva Lopez, Salar Mohtaj, Konstanze Neumann, Tim Polzehl, and Sebastian Möller. 2024a. Implications of regulations on large generative ai models in the super-election year and the impact on disinformation. In *Legal Workshop at LREC-COLING 2024*, LREC-COLING, pages 28–38. ELRA.

Vera Schmitt, Luis-Felipe Villa-Arenas, Nils Feldhus, Joachim Meyer, Robert P. Spang, and Sebastian Möller. 2024b. The role of explainability in collaborative human-ai disinformation detection. In *24 ACM Conference on Fairness, Accountability, and Transparency*, pages 2157–2174. ACM.

John C. Simmons and Joseph M. Winograd. 2024. Interoperable provenance authentication of broadcast media using open standards-based metadata, watermarking and cryptography.

Xin Tan, Bowei Zou, and Ai Ti Aw. 2025. Improving Explainable Fact-Checking with Claim-Evidence Correlations. In *31st International Conference on Computational Linguistics*, pages 1600–1612, Abu Dhabi, UAE. Association for Computational Linguistics.

Rama Adithya Varanasi and Nitesh Goyal. 2023. "it is currently hodgepodge": Examining ai/ml practitioners' challenges during co-production of responsible ai values. In *23 CHI Conference on Human Factors in Computing Systems*. ACM.

Paul Wachter, Niklas Kruse, and Julius Schöning. 2024. Synthetic fields, real gains.

Paul Wachter, Lukas Niehaus, and Julius Schöning. 2025. *Development of Hybrid Artificial Intelligence Training on Real and Synthetic Data: Benchmark on Two Mixed Training Strategies*, pages 175–189. Springer Nature Switzerland.

Wenbo Wan, Jun Wang, Yunming Zhang, Jing Li, Hui Yu, and Jiande Sun. 2022. A comprehensive survey on robust image watermarking. *Neurocomputing*, 488:226–247.

Mateusz Łabuz. 2024. Deep fakes and the artificial intelligence act—an important signal or a missed opportunity? *Policy & Internet*, 16(4).

Towards Robust Evaluation for Privacy QA Systems

Anna Leschanowsky[1], **Zahra Kolagar**[1], **Erion Çano**[2], **Ivan Habernal**[2], **Dara Hallinan**[3], **Emanuël A. P. Habets**[4], **Birgit Popp**[1]
[1]Fraunhofer Institute for Integrated Circuits (IIS), Erlangen, Germany
[2]Research Center for Trustworthy Data Science and Security, Ruhr University Bochum, Germany
[3]FIZ Karlsruhe - Leibniz Institute for Information Infrastructure, Karlsruhe, Germany
[4]International Audio Laboratories Erlangen, Erlangen, Germany

Abstract

The transparency principle of the General Data Protection Regulation requires data-processing information to be clear, precise, and accessible. While Large Language Models (LLMs) show promise in this context, their probabilistic nature raises challenges for ensuring truthfulness and comprehensibility. This paper presents an exploratory evaluation of eight Privacy Question Answering (QA) systems – including LLMs, retrieval-augmented generation, and alignment-based approaches – on two datasets. We propose an evaluation framework that maps both traditional NLP and LLM-as-a-judge metrics to the legal requirements of comprehensibility and precision. Results show that no single system consistently excels across all metrics, and that system rankings can vary depending on the choice of metric and thresholding. We highlight open questions and emphasize the need to translate legal requirements into technical evaluation criteria. Our work provides a foundation for a more robust evaluation of Privacy QA systems.

Keywords: Privacy QA, Data Protection Regulation, Large Language Models

1. Introduction

User privacy is a central concern when interacting with Large Language Models (LLMs). As these systems may process personal data, ensuring transparency in data processing to enable informed decisions and regulatory compliance is essential. The General Data Protection Regulation (GDPR) (European Union, 2016) emphasizes the transparency principle (Art. 5, 12) with three sub-requirements: **accessibility** (everyone should be able to inform themselves about how their data is used), **comprehensibility** (language used should be easy to understand), and **precision** (data subjects should anticipate how their data will be processed) (Article 29 Data Protection Working Party, 2018). Privacy notices are the standard way of providing this information, but their length and complexity often hinder transparency. To address this, Privacy QA systems and personalized privacy assistants have been proposed (Harkous et al., 2016; Morel et al., 2025), and recent work highlights LLMs' potential to answer privacy-related questions (Freiberger et al., 2025; Hamid et al., 2024).

LLMs may aid compliance with the transparency principle by providing comprehensible and precise responses about how personal data is processed. This interactive approach supports accessibility, as users do not need to switch modalities or read static privacy notices (Article 29 Data Protection Working Party, 2018). LLM-based agents assisting with privacy notices can improve comprehension and reduce time spent on privacy management (Sun et al., 2024). However, state-of-the-art LLM applications may provide inaccurate and hallucinated answers (Hamid et al., 2024). Despite their promise, evaluating their performance remains particularly challenging. Prior work has predominantly relied on manual and time-consuming quality annotation (Hamid et al., 2024; Freiberger et al., 2025), making systematic benchmarking difficult. While standard evaluation metrics have been explored in legal NLP (Kelsall et al., 2025), a comprehensive evaluation and comparison of LLM-based Privacy QA systems is missing, and there is limited understanding of how existing metrics map to legal constructs such as precision and comprehensibility. Thus, this work presents an exploratory evaluation aimed at improving the robustness of Privacy QA assessment, with the main contributions summarized as follows:

1. We introduce an evaluation framework that maps 12 state-of-the-art NLP metrics to legal constructs of precision and comprehensibility and analyze their interrelationships (see Section 5).
2. We conduct a comparative assessment of eight LLM-based Privacy QA systems, including baseline LLM, Retrieval Augmented Generation (RAG), and alignment-based approaches, using an expert-generated dataset of data processing questions and an evaluation dataset derived from PolicyQA (Ahmad et al., 2020) (see Section 3).
3. We present MultiRAIN, a multidimensional extension of Rewindable Auto-regressive Inference (RAIN) to jointly optimize for legal precision and comprehensibility, and benchmark its impact within our evaluation framework (see Section 4).

Proceedings of LEGAL2026 and CALD-pseudo 2026 @ LREC 2026, pages 12–25
12 May 2026. ©ELRA Language Resources Association (ELRA), 2026

4. We critically discuss the technical and legal implications of current evaluation practices and Privacy QA system implementations, identifying major challenges and open questions to guide future work (see Section 7).

2. Related work

2.1. Privacy QA systems

Pioneering work introduced Pribots (Harkous et al., 2016), showing that conversational systems can respond to users' questions about personal data processing. Since then, retrieval-based approaches have been tested (Mysore Sathyendra et al., 2017; Ravichander et al., 2019; Ahmad et al., 2020), mostly using classical NLP methods rather than LLMs. Notably, Pribots' output was based on the extraction of legal texts, which can hinder transparency, as these texts can be difficult to understand, even for experts (Martínez et al., 2023; Article 29 Data Protection Working Party, 2018). Recent approaches using LLMs and simple prompting techniques show promise, but can suffer from incorrect or outdated information (Hamid et al., 2024; Freiberger et al., 2025). RAG systems combine LLMs with a document database (e.g., privacy notices, FAQs) to improve accuracy, yet because they are built on top of LLMs, they can still produce hallucinations, raising transparency concerns. Alignment approaches are commonly used to mitigate the risk of hallucinations (Askell et al., 2021; Huang et al., 2024). In this work, we evaluate plain LLM, RAG, and alignment-based techniques. In particular, we experiment with Rewindable Auto-regressive Inference (RAIN), as it does not require costly training, can be integrated into existing language models and performs comparably to other state-of-the-art alignment methods (Li et al., 2024b), making it an ideal candidate for addressing legal transparency in NLP systems.

2.2. Evaluation of QA Systems

QA evaluation has evolved from traditional benchmarks towards frameworks that cover both factual and complex reasoning tasks (e.g., Holistic Evaluation of Language Models (HELM) (Bommasani et al., 2023)). Metrics range from reference-based (e.g., BLEU (Papineni et al., 2002), ROUGE (Lin, 2004)), to embedding-based (e.g., BERTScore (Zhang et al., 2019)) and reference-free approaches (Li et al., 2024a). Recently, LLM-as-a-judge metrics have been proposed to reduce reliance on ground-truth annotations (e.g., RAGAS (Es et al., 2024), ARES (Saad-Falcon et al., 2024)). For Privacy QA, prior work has often relied on standard metrics, such as ROUGE (Sun et al., 2024), or on human evaluation (Hamid et al., 2024). While traditional NLP evaluation and LLM-as-a-judge approaches were found unreliable for court decision predictions in the legal domain (Ammar et al., 2024), a similar comprehensive evaluation is missing for Privacy QA.

3. Evaluation Datasets

3.1. Expert-Generated Dataset

We used a dataset built with legal and linguistic experts (Leschanowsky et al., 2025). The authors presented experts with Alexa's privacy notice and FAQ pages, together with 42 questions. Both legal and linguistic experts generated answers to these questions, taking turns to ensure both legal precision and linguistic simplicity. Questions cover nine information types, e.g., contact information, location, and voice recordings, and are categorized into six data practice categories, e.g., First Party Collection/Use, User Rights, and Choice/Control. As the answers are expert-generated, this dataset allows for evaluating system-generated answers with respect to human expert-generated answers.

3.2. PolicyQA Subset

To assess generalizability, we evaluated on a PolicyQA (Ahmad et al., 2020) subset. We chose PolicyQA over other Privacy QA corpora as it provides context information for each question, allowing assessment reference-based metrics such as context adherence (Ahmad et al., 2020). PolicyQA consists of 115 website privacy policies from the OPP-115 corpus, annotated by OPP-115 categories. To limit computation time, we only use the *internetbrands.com* notice from the development set of PolicyQA, as it contained the most associated contexts. We extracted 47 out of the 429 questions linked to this policy. To select a diverse subset, we computed pairwise semantic textual similarity via SentenceBERT (Reimers and Gurevych, 2019) and chose the most dissimilar questions within each category. We ensured at least two questions per privacy practice category (except "Do Not Track"). As PolicyQA contexts are scattered paragraphs of the privacy notice, we reconstructed a complete notice by consolidating these contexts and using the current *internetbrands.com* notice as a formatting reference. This reconstructed notice was used as a document database for the RAG-based systems.

4. Privacy QA Systems

We evaluated eight LLM-based Privacy QA systems, including plain LLMs, RAG, and alignment-based approaches, specifically Rewindable Auto-regressive Inference (RAIN) (Li et al., 2024b). Fur-

thermore, we extend RAIN to optimize two criteria: precision and comprehensibility. To the best of our knowledge, our evaluation is the first to systematically compare these approaches for Privacy QA.

RAIN and MultiRAIN systems included real-time evaluation modules that monitored generation and rewound when quality criteria were unmet. Due to computational constraints, optimization was limited to one or two metrics and differs from the comprehensive post-generation evaluation (see Section 5). For system implementation, we operationalized precision and comprehensibility using both LLM-as-a-judge and traditional metrics.

Baseline - LLM Plain LLM answering one question at a time with the privacy notice as context.

Retrieval Augmented Generation (RAG) RAG retrieves the top three relevant policy excerpts and generates answers conditioned on this context (prompt in Appendix 12.4).

Rewindable Auto-regressive Inference (RAIN) Four systems use RAIN (Li et al., 2024b) as an alignment method. RAIN operates as a rewindable tree search, in which generated tokens are evaluated for precision and comprehensibility, and the response is revised when criteria are not met. Importantly, since RAIN optimizes based on retrieved responses, it cannot correct an incorrect retrieval. However, we focused on investigating ways to jointly optimize for multiple features, such as precision and comprehensibility, in LLM generation and thus kept the retrieval module the same across the tested systems. We instantiated RAIN with two metric choices:

- LLM-as-a-judge metrics: Correctness and Readability as implemented by Trott and Rivière (2024) (see Appendix 12.4 for the prompt templates), resulting in **RAIN Correctness** and **RAIN Readability**. We prompted only for scores, without providing additional examples, and applied thresholds of 78.64 (correctness) and 90.74 (readability) derived from the mean scores of expert-generated answers.
- Traditional NLP metrics: BERTScore (Reimers and Gurevych, 2019) and Flesch–Kincaid Readability (Kincaid et al., 1975), resulting in **RAIN BERTScore** and **RAIN Flesch–Kincaid Readability**, with thresholds of 0.312 and 62.69, respectively.

Multi Rewindable Auto-regressive Inference (MultiRAIN) Two systems used MultiRAIN, a multidimensional adaptation of RAIN that jointly optimizes multiple criteria (mathematical formulation and pseudocode in Appendix 12.1). We instantiated:

- **MultiRAIN (LLM)** based on LLM-as-a-judge metrics of readability and correctness using the same thresholds as **RAIN Readability** and **RAIN Correctness**.
- **MultiRAIN (Traditional)** based on traditional NLP metrics, BERTScore, and Flesch-Kincaid Readability using the same thresholds as RAIN BERTScore and RAIN Flesch–Kincaid Readability, where BERTScore is multiplied by 100 before averaging due to scale differences.

For text generation across all system variations, we used Mistral-7B-Instruct-v0.2[1] as it is openly available and its moderate model size enables both reproducibility and efficient experimentation. For RAG, we used OpenAI's text-embedding-3-small model[2] for embedding documents.

5. Evaluation Framework

We combine traditional NLP metrics and LLM-as-a-judge metrics and map them to legal constructs of precision and comprehensibility. Evaluation code is provided on GitHub (https://github.com/audiolabs/transparentnlp).

5.1. Automated Evaluation Metrics

To our knowledge, no established mapping of legal constructs to technical metrics exists. Thus, our mapping (see Table 1) represents a preliminary and principled attempt. By mapping up to eight technical metrics to each legal construct, we can assess how well the evaluation metrics satisfy legal requirements. We use both LLM-as-a-judge metrics and traditional NLP metrics, as well as reference-based and reference-free metrics. As prompt variation can impact evaluation of LLM-as-a-judge metrics, we relied on well-established and previously used prompts to ensure comparability and reproducibility (Trott and Rivière, 2024; Galileo AI, 2024; Friel and Sanyal, 2023; Dale and Chall, 1949; Kincaid et al., 1975; Zenker and Kyle, 2021; Mehrpour and Riazi, 2004). For reference-based metrics, we used excerpts from privacy notices as ground truth.

5.1.1. Measuring with LLM-as-a-judge

We evaluated with OpenAI's GPT-4 (used here solely as an evaluator, distinct from the Mistral model used for text generation (Achiam et al., 2023)) and adopted best practices for prompt design, such as chain-of-thought prompting.

Measuring precision. Metrics assessing LLM response precision lack standard terminology, with terms like "correctness" and "faithfulness" often used interchangeably. To address this, we reference Galileo.ai's LLM-as-a-judge metrics without

[1] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

[2] https://platform.openai.com/docs/models/text-embedding-3-small

	With LLM-as-a-judge	Without LLM-as-a-judge
Precision	Context Adherence Completeness Correctness Answer Relevancy	BLEU ROUGE-1 BERTScore STS
Comprehensibility	Readability as implemented by Trott and Rivière (2024)	Flesch-Kincaid Readability Lexical Diversity Sentence Length

Table 1: Overview of evaluation metrics with and without LLM-as-a-judge. Details on the metrics, including references, are provided in Section 5. For Context Adherence, Completeness, BLEU, ROUGE-1, BERTScore, and STS, which require a ground truth for comparison, we used excerpts from the privacy notices as a reference.

endorsing their platform:[3]

- **Context Adherence (Faithfulness)**: Measures whether responses align with the provided context (Friel and Sanyal, 2023), akin to "Faithfulness" in LLamaIndex (LlamaIndex, 2024).
- **Completeness**: Evaluates whether all relevant context information is included (Galileo AI, 2024).
- **Correctness**: Detects open-domain hallucinations or factual inaccuracies unrelated to specific documents (Friel and Sanyal, 2023).
- **Answer Relevance (Relevancy)**: Assesses the relevance of generated answers to user queries (Galileo AI, 2024).

Measuring comprehensibility. We rely on Trott and Rivière (2024), as they found significant correlations between LLM and human readability assessments using GPT-4 Turbo with the CLEAR corpus.

5.1.2. Measuring without LLM-as-a-judge

Measuring precision. Traditional metrics like BLEU and ROUGE-1 assess n-gram overlap and response similarity to reference texts, as used in prior work (Huang et al., 2024; Friel and Sanyal, 2023; Forbes et al., 2023). BERTScore measures token-level similarity using contextual embeddings, while Semantic Textual Similarity (STS) quantifies semantic similarity (Cer et al., 2017).

Measuring comprehensibility. We evaluated comprehensibility through readability, lexical diversity, and sentence length:

- **Readability**: Readability formulas, like Flesch-Kincaid, evaluate ease of comprehension (Dale and Chall, 1949; Kincaid et al., 1975) and are used in privacy notice research (Cadogan, 2004; Fabian et al., 2017).
- **Lexical Diversity**: We rely on the Measure of Textual Lexical Diversity (MTLD) to assess vocabulary richness, as it is better suited for varying text lengths (Zenker and Kyle, 2021).
- **Sentence Length**: Shorter sentences can improve comprehension, though results may vary (Mehrpour and Riazi, 2004).

5.2. Evaluation Procedure and Thresholding

The availability of expert answers enables the definition of metric-specific thresholds (Leschanowsky et al., 2025) to assess whether generated answers are "at least as good" as, or better than, expert references. Both sets of answers designed in (Leschanowsky et al., 2025) were used for thresholding. We compared four thresholding methods to illustrate how threshold selection can affect system rankings and evaluation robustness:

Min: We took the minimum value over the designed answers to compute a lower bound. For interval-based metrics such as sentence length, we used the minimum and maximum as the acceptable range. This method counts answers as acceptable if they fall within the observed range, but the method is highly sensitive to outliers.

Mean: We used the arithmetic mean as the threshold. For interval-based metrics, we used the range defined by mean $\pm$ one standard deviation. This threshold is easy to compute, but it sets the bar high, as designed answers that fall below the mean may be deemed insufficient.

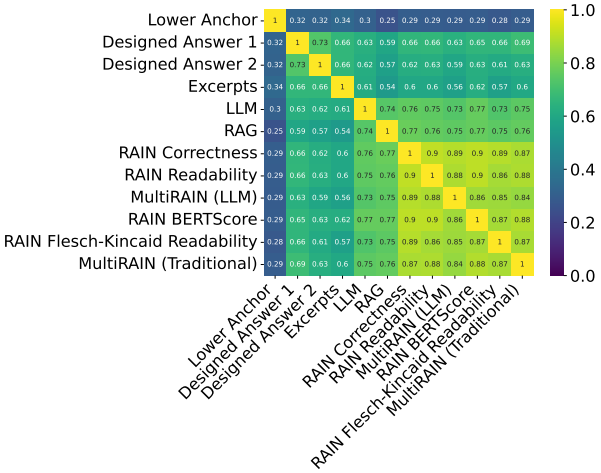
Percentiles: We computed two percentile-based thresholds. For interval metrics, we used the 10th and 90th percentiles as outer bounds and the 25th and 75th percentiles as the interquartile range. For other metrics, we used either the 10th or 25th percentile as a lower bound. The 10th percentile excludes outliers while considering 90% of designed answers as sufficient; the 25th percentile is stricter, but still captures most designed answers.

[3] https://docs.galileo.ai/galileo/gen-ai-studio-products/galileo-guardrail-metrics

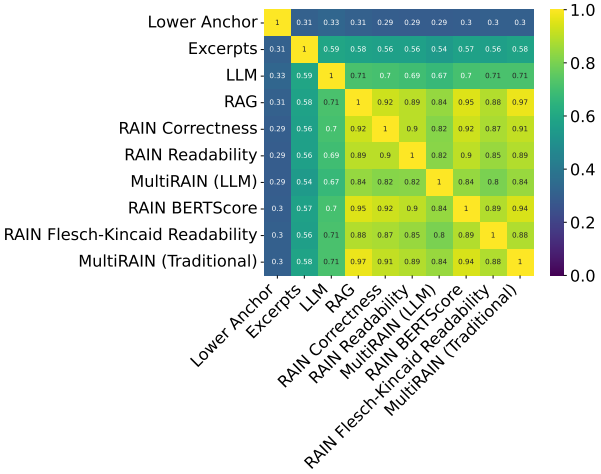
While expert-generated answers serve as an upper bound, a lower anchor is provided by random-word answers to all questions, with words drawn from the vocabulary of system-generated answers. We used an average sentence length of 18 words for the lower-anchor answers, matching the average sentence length of expert-generated answers. Since these answers lack semantic structure, we expected them to perform poorly.

6. Results

6.1. Answer Similarity



(a) Answer Similarity of the expert-generated Dataset.



(b) Answer Similarity of the PolicyQA Subset.

Figure 1: Pairwise answer similarity between different Privacy QA system realizations.

Figure 1 shows averaged pairwise answer similarity, computed with SentenceBERT (Reimers and Gurevych, 2019). In the expert-generated dataset, excerpts and designed answers are most dissimilar (approx. 0.60 and 0.66). This is expected, as the designed answers were selected to maximize dissimilarity, thereby illustrating the range of variability possible in expert-generated responses (Leschanowsky et al., 2025). LLM and RAG system answers exhibit moderate pairwise similarity scores of approximately 0.75. While similarity scores do not provide direct insights into answer quality, they highlight that exploring alternative system implementations beyond plain LLMs can be valuable for Privacy QA. Alignment-based methods (RAIN and MultiRAIN) yield the highest mutual similarity, suggesting that optimization induces only modest changes. Nevertheless, these answers diverge from the base RAG outputs, confirming that alignment can alter responses.

For the PolicyQA subset, similar patterns emerge, with excerpts being the least similar and LLMs following. However, RAG demonstrates high similarity with answers from RAIN and MultiRAIN. This suggests that, in the context of PolicyQA, alignment methods resulted in only minor modifications when compared to RAG answers. Qualitative inspection of the top 10 answers with the lowest similarity in the expert-generated dataset reveals that RAG answers often express uncertainty (e.g., *"the sources do not mention contacts specifically"*). In contrast, optimized answers tend to be more assertive (e.g., *"Yes, we use [...]"*). In PolicyQA, most RAG outputs do not express uncertainty, possibly because questions are framed around general data rather than specific information types, which are only subtly covered in the privacy notice.

6.2. Evaluation Metric Behavior and Threshold Influence

Evaluation of raw metric scores for the expert-generated dataset (see Figures 5 and 6 in the Appendix) reveals several key trends across Privacy QA systems. The lower anchor behaves as expected, scoring low on all metrics and setting a lower bound with few outliers for completeness. Only Flesch-Kincaid Readability scores vary from around 10 to 100, as it depends on the number of words, syllables, and sentences, so even nonsensical sentences can appear readable. In contrast, the LLM-as-a-judge readability metric shows a sharp decrease for the lower anchor, but exhibits a ceiling effect for all Privacy QA systems. Here, Flesch-Kincaid Readability provides more nuanced differentiation, with average system scores around 50%, highlighting the need to include both traditional and LLM-as-a-judge metrics for a more comprehensive evaluation. Correctness also shows a ceiling effect, while BLEU shows a floor effect, making system differentiation challenging. Context adherence varies widely (0-100), and this variation also holds for expert answers, indicating imperfect alignment between the automated evalu-

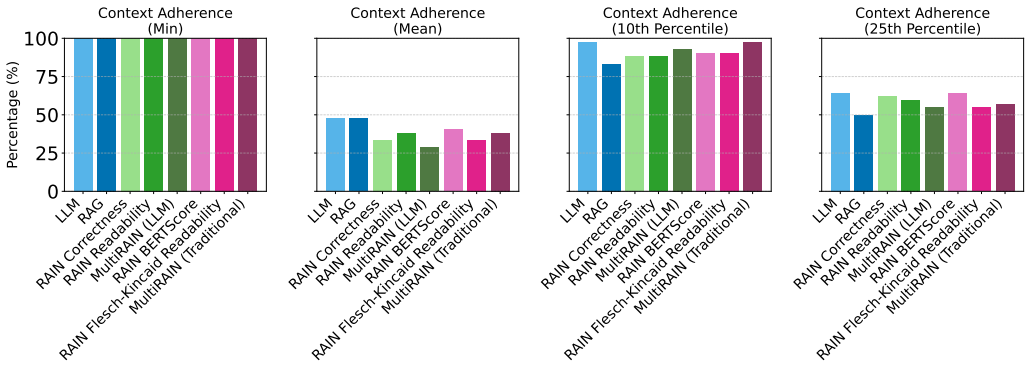


Figure 2: Comparison of "at least as good" as human expert answers under different thresholds in the expert-generated dataset.

ation and expert opinion. Traditional NLP metrics (BLEU, ROUGE, BERTScore, STS) tend to yield lower scores than LLM-based metrics (correctness, completeness, answer relevance). Although expert-designed answers often rank highest on traditional metrics, LLM-generated answers surpass them in completeness and answer relevance. This discrepancy may suggest that LLM-based evaluation metrics are not perfectly aligned with human expert annotations and may, in some cases, favor system-generated outputs over human-written gold standards. Future work should invest in expert-generated datasets to explore alignment between human annotation and LLM-based metrics.

The availability of expert answers enables the definition of metric-specific thresholds to determine whether a generated answer is "at least as good" as the reference. However, threshold selection is non-trivial. Figure 2 illustrates how four thresholding methods affect the percentage of responses meeting expert performance for context adherence. This can substantially affect system comparison. Using extreme thresholds (minimum or maximum) leads to extreme outcomes. For example, since context adherence and completeness are 0 for expert-generated answers, all generated responses would meet a minimum-based threshold. In contrast, mean-based thresholds set a high bar that even expert answers fail to meet. Importantly, threshold choice affects both absolute performance and system ranking (see Figure 2). For example, using the mean results in approximately a 40-percentage-point drop in context-adherence performance compared to the 10th percentile, and systems equivalent under one threshold (e.g., LLM and RAG under the mean) diverge under others (e.g., LLM and RAG under the 25th percentile). Thus, threshold selection is not arbitrary and influences which systems are preferred and which answers are considered sufficient.

6.3. Variation Across Datasets, Information Type and Data Practice

Figure 3 shows the percentage of responses considered at least as good as expert answers across both datasets and thresholds. Overall, performance is similar across datasets. Correlation analysis supports this finding, with moderate positive correlations for completeness, context adherence, answer relevance, ROUGE, and readability (e.g., Pearson 0.74 for completeness and 0.69 for context adherence). At the same time, BERTScore and lexical diversity show weak or negative correlations. Although RAIN and MultiRAIN were aligned using thresholds derived from the expert-generated dataset, systems optimized on that dataset do not consistently outperform those evaluated on the PolicyQA subset, suggesting that results may generalize to other Privacy QA datasets. While system rankings remain sensitive to threshold (Section 6.2), dataset comparison shows that no system consistently dominates across metrics and thresholds. Only RAIN Flesch–Kincaid Readability shows consistent improvements on its optimized metric across datasets, indicating that targeted alignment can be effective.

We further investigated performance variation across information types and data practices. In the expert dataset, metrics with high variance (e.g., context adherence and completeness) show differences across information types and systems, but no system consistently outperforms others on a specific type, and no type appears especially difficult. Grouping by data practice reveals clearer patterns. In the expert-generated dataset, context adherence appears higher for categories such as *Privacy Policy* and *First Party Collection/Use - Information*, and lower for *Third-Party Collection/Use*. In the PolicyQA subset, categories such as *Do Not Track* and *User Access/Edit and Deletion* achieve higher scores than *Third Party Sharing/Collection* and *International and Specific Audiences*. However, the number of questions per category varies,

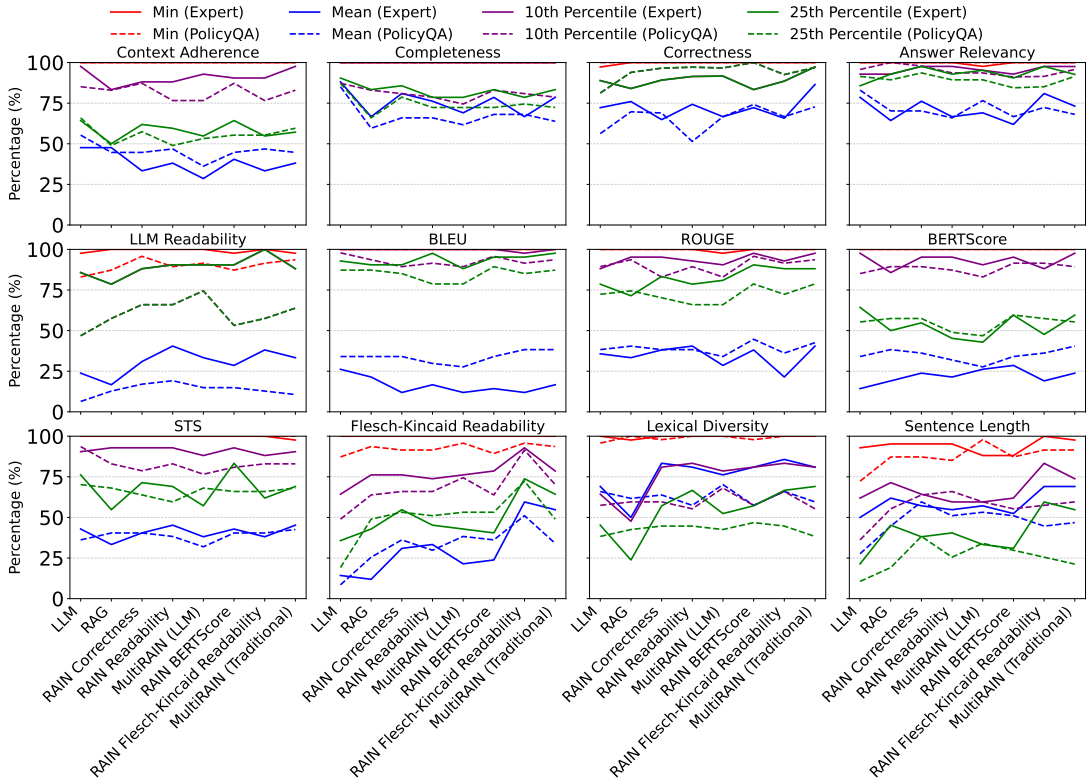


Figure 3: Performance trends for both evaluation datasets across systems for each evaluation metric. The mean is chosen as a threshold for exemplary purposes.

and analysis shows high variance across questions in categories with low average performance. This suggests that differences between categories are driven by the number and diversity of questions rather than category difficulty.

6.4. Principal Component Analysis

We conducted an exploratory Principal Component Analysis (PCA) on the results from both datasets to examine whether the metrics cluster into constructs for precision and comprehensibility. Figure 4 shows PCA projections of the two main components. We observe that comprehensibility and precision metrics generally separate along Principal Component 1 (PC1), with most metrics exhibiting loadings above 0.3 on this axis, indicating an association with the primary dimension of variation. Only correctness and context adherence have minimal loadings, suggesting they capture distinct aspects not aligned with the main axes of variance. This may imply that these two metrics capture unique aspects of answer quality and raise questions about our categorization of precision measures in Table 1.

Correctness and answer relevancy differ in that they assess factual correctness or relevance without a ground truth. Notably, both reference-based LLM- and traditional NLP metrics cluster together, suggesting that precision metrics may be further differentiated into reference-based and reference-free metrics. For comprehensibility metrics, loadings on PC2 exceed 0.4 for readability metrics and fall below -0.4 for interval-based metrics, likely indicating two distinct dimensions of comprehensibility. While the first two PCs explain about 50% of the variance (see Figure 7), future work could explore additional components to uncover additional latent structure.

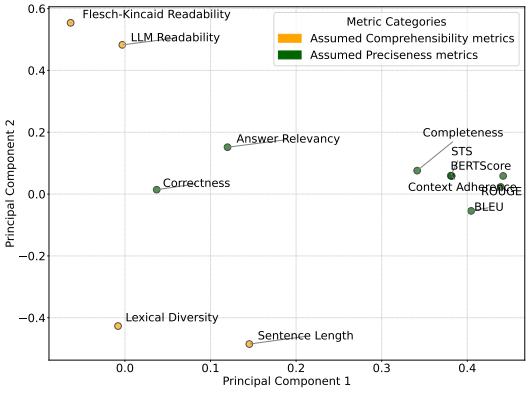


Figure 4: The 2D PCA projection shows relationships between text evaluation metrics. Metrics are colored based on their assumed relationship, i.e., precision (dark green) and comprehensibility (yellow).

7. Discussion and Open Questions

Our study reveals challenges in robust Privacy QA evaluation and system comparison. We evaluated eight systems on 12 state-of-the-art NLP metrics spanning legal precision and comprehensibility. No single system outperforms the others across all metrics. Yet alignment can improve performance on targeted metrics, such as Flesch-Kincaid Readability. We demonstrate that comparisons to expert answers and thresholding influence performance and rankings. A comparison of two datasets suggests some generalizability, but performance varies with the diversity and number of questions per category. A key focus of our study is mapping legal constructs, i.e., comprehensibility and precision, to technical metrics. Our initial categorization of evaluation metrics (see Table 1) was not fully supported by the PCA. We discuss open questions for robust Privacy QA evaluation in light of these findings.

7.1. What Makes a Robust Evaluation Dataset for Privacy QA?

Our analysis of answer similarity and dataset variation shows that some answers seem easily extractable from the privacy notice, while others are affected by vagueness or ambiguity in the underlying privacy notice. Further, the number and diversity of questions affect the results. A robust evaluation dataset should therefore include a spectrum of question types, e.g., fully answerable questions, clearly unanswerable questions, and questions that are potentially answerable, but subject to vague or ambiguous information, possibly drawing on previous work (Ravichander et al., 2019). This requires legal expert input and systematic labeling of vagueness to categorize question difficulty. Paraphrased variants of curated questions (Hamid et al., 2024) can increase diversity, and diversity across data practices per established annotation schemes (Wilson et al., 2016) can support robust evaluation.

7.2. How to Choose Evaluation Metrics That Align with Legal Concepts?

Our work maps metrics to legal constructs of comprehensibility and precision, including traditional NLP metrics and LLM-as-a-judge metrics. None of the evaluated metrics demonstrated clearly superior performance, due to misalignment with expert-generated answers, poor separation from the lower anchor, or ceiling and floor effects. PCA results indicate sub-clusters rather than a clean two-factor structure, with additional distinction between reference-based and reference-free metrics. We therefore recommend using both LLM-based and traditional NLP metrics, as well as both reference-based and reference-free metrics, for comprehensive evaluation. Developing a joint metric that balances precision and comprehensibility could help streamline evaluation and further improve system alignment. As LLM-as-a-judge metrics are sensitive to prompting (Li et al., 2024a), future work should evaluate robustness to prompt variation. Translating legal concepts into concrete technical metrics remains a challenge, but presents significant opportunities. Our work takes a first step and highlights the importance of interdisciplinary collaboration in defining, translating, and testing legal constructs as evaluation metrics.

7.3. How Do We Define "Good Enough" for Privacy QA System Comparison?

Our analysis used two reference points to assess whether answers are "good enough": i) expert-generated answers as an upper bound and ii) randomly concatenating words from system outputs as a lower bound. The lower anchor illustrates metric limitations, e.g., its high Flesch-Kincaid score shows the metric alone does not guarantee comprehensibility or distinguish expert from non-expert quality. Expert-generated answers showed high variance on metrics like context adherence, so setting a threshold at the minimum expert score can yield 0 for context adherence, making all outputs appear sufficient. Therefore, we suggest that thresholds should meaningfully separate expert and anchor signals. For some metrics, the 10th percentile suffices, while for others (e.g., completeness and ROUGE), the mean may be needed. Given the legal requirements, a higher threshold, such as the mean, may be justified, though the appropriate choice of threshold and the responsibility for setting it remain open questions. An alternative is to vary thresholds to compare system rankings in a threshold-robust way.

7.4. What are Legal Implications for Privacy QA?

Despite technological developments, there has been little focused legal analysis of Privacy QA, which is necessary to advance the field. Several lines of legal research seem most pertinent: (1) analyze the scope and content of the relevant legal transparency obligations in relation to Privacy QA; (2) assess the degree to which Privacy QA systems can meet those obligations, and, in particular, the degree to which an inaccurate system for providing legal information can be legally permissible. Our experiments show that, depending on the chosen threshold, current approaches often fail to meet expert-generated standards, potentially failing to fulfill legal requirements or constituting misleading information. This highlights the need to define what counts as "good enough" and, if thresholds cannot

be consistently met, to consider safeguards such as disclaimers, layered-information provisions, and inclusion of reference texts to address legal issues of inaccuracy. (3) analyze other legal obligations relevant to developers and users, specifically considering legal frameworks governing AI.

8. Conclusion

In an exploratory study, we evaluated eight Privacy QA systems (LLM, RAG, and alignment-based methods) across two datasets, using a framework that maps traditional NLP and LLM-as-a-judge metrics to legal constructs of comprehensibility and precision. No single system dominates, and rankings vary with the choice of threshold. We identify limitations of current metrics, including poor separation between high- and low-quality answers, and discuss open questions regarding the translation of legal requirements into technical evaluation criteria.

9. Limitations

Our study presents a first attempt to systematically compare Privacy QA systems with varying architectures using 12 metrics to approximate precision and comprehensibility. However, this selection is a snapshot of the large space of possible systems and metrics. For example, varying prompts for LLM-as-a-judge metrics can alter measurements and outcomes, underscoring the complexity of defining these constructs. Further, our implementations of MultiRAIN are limited by algorithmic efficiency, as generating 42 answers using alignment modules took 20–58 hours on one GPU (NVIDIA A100 SXM4); practical applications require answers in seconds. The dataset scope is restricted to privacy notices from two providers and a small, diverse question set, limiting generalizability. Results of the PCA depend on the examined datasets, and assessments of the metrics may change when additional data are incorporated. Together, these factors mean our findings should be viewed as exploratory and may not generalize to systems using different LLMs or datasets.

10. Acknowledgments

This work has been partly supported by the Research Center Trustworthy Data Science and Security (https://rc-trust.ai), one of the Research Alliance centers within the https://uaruhr.de.

11. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Wasi Ahmad, Jianfeng Chi, Yuan Tian, and Kai-Wei Chang. 2020. PolicyQA: A reading comprehension dataset for privacy policies. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 743–749, Online. Association for Computational Linguistics.

Adel Ammar, Anis Koubaa, Bilel Benjdira, Omer Nacar, and Serry Sibaee. 2024. Prediction of Arabic legal rulings using large language models. *Electronics*, 13(4):764.

Article 29 Data Protection Working Party. 2018. Guidelines on Transparency under Regulation 2016/679. Accessed: 2024-11-22.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Benjamin Mann, Nova Dassarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, John Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Christopher Olah, and Jared Kaplan. 2021. A general language assistant as a laboratory for alignment. *ArXiv*, abs/2112.00861.

Rishi Bommasani, Percy Liang, and Tony Lee. 2023. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1):140–146.

Rochelle A. Cadogan. 2004. An imbalance of power: the readability of internet privacy policies. *Journal of Business & Economics Research (JBER)*, 2.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Edgar Dale and Jeanne S. Chall. 1949. The concept of readability. *Elementary English*, 26(1):19–26.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation.

In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

European Union. 2016. Regulation (EU) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (general data protection regulation). Official Journal of the European Union, L 119/1. Accessed: 2024-11-21.

Benjamin Fabian, Tatiana Ermakova, and Tino Lentz. 2017. Large-scale readability analysis of privacy policies. In *Proceedings of the international conference on web intelligence*, pages 18–25.

Grant C. Forbes, Parth Katlana, and Zeydy Ortiz. 2023. Metric ensembles for hallucination detection. *arXiv preprint arXiv:2310.10495*.

Vincent Freiberger, Arthur Fleig, and Erik Buchmann. 2025. "you don't need a university degree to comprehend data protection this way": LLM-powered interactive privacy policy assessment. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, New York, NY, USA. Association for Computing Machinery.

Robert Friel and Atindriyo Sanyal. 2023. Chainpoll: A high efficacy method for llm hallucination detection. *arXiv preprint arXiv:2310.18344*.

Galileo AI. 2024. Guardrail metrics. Accessed: 2024-11-25.

Aamir Hamid, Hemanth Reddy Samidi, Primal Pappachan, Tim Finin, and Roberto Yus. 2024. Genaipabench: A benchmark for generative ai-based privacy assistants. *Proceedings on Privacy Enhancing Technologies*.

Hamza Harkous, Kassem Fawaz, Kang G. Shin, and Karl Aberer. 2016. PriBots: Conversational privacy with chatbots. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, Denver, CO. USENIX Association.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Hanchi Sun, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li, Eric P. Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, Joaquin Vanschoren, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi Zhou, William Yang Wang, Xiang Li, Xiangliang Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, Yong Chen, and Yue Zhao. 2024. Position: TrustLLM: Trustworthiness in large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 20166–20270. PMLR.

Joshua Kelsall, Xingwei Tan, Aislinn Bergin, Jiahong Chen, Maria Waheed, Tom Sorell, Rob Procter, Maria Liakata, Jenny Chim, and Serene Chi. 2025. A rapid evidence review of evaluation techniques for large language models in legal use cases: trends, gaps, and recommendations for future research. *AI & SOCIETY*, pages 1–19.

J. Peter Kincaid, Robert P. Fishburne Jr. Fishburne, Richard L. Rogers Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas: (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Chief of Naval Technical Training, Naval Air Station Memphis, Millington, Springfield. Distributed by NTIS.

Anna Leschanowsky, Farnaz Salamatjoo, Zahra Kolagar, and Birgit Popp. 2025. Expert-generated privacy q&a dataset for conversational ai and user study insights. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, New York, NY, USA. Association for Computing Machinery.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024a. LLMs-as-Judges: a comprehensive survey on LLM-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Y. Li, FangyunWei, J. Zhao, C. Zhang, and H. Zhang. 2024b. RAIN: Your language models can align themselves without finetuning. In *International Conference on Learning Representations*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

LlamaIndex. 2024. Evaluating with LlamaIndex. Accessed: 2024-11-25.

Eric Martínez, Francis Mollica, and Edward Gibson. 2023. Even lawyers do not like legalese. *Proceedings of the National Academy of Sciences*, 120(23):e2302672120.

Saeed Mehrpour and Abdolmehdi Riazi. 2004. The impact of text length on reading comprehension in English as a second language. *Asian EFL Journal*, 3(6):1–14.

Victor Morel, Leonardo Horn Iwaya, and Simone Fischer-Hübner. 2025. AI-driven personalized privacy assistants: A systematic literature review. *IEEE Access*, 13:160982–161002.

Kanthashree Mysore Sathyendra, Shomir Wilson, Florian Schaub, Sebastian Zimmeck, and Norman Sadeh. 2017. Identifying the provision of choices in privacy policy text. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2774–2779, Copenhagen, Denmark. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question Answering for Privacy Policies: Combining Computational and Legal Perspectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4946–4957, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354, Mexico City, Mexico. Association for Computational Linguistics.

Bolun Sun, Yifan Zhou, and Haiyun Jiang. 2024. Empowering users in digital privacy management through interactive llm-based agents. *arXiv preprint arXiv:2410.11906*.

Sean Trott and Pamela Rivière. 2024. Measuring and modifying the readability of English texts with GPT-4. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 126–134, Miami, Florida, USA. Association for Computational Linguistics.

Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, Thomas B. Norton, Eduard Hovy, Joel Reidenberg, and Norman Sadeh. 2016. The Creation and Analysis of a Website Privacy Policy Corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1340, Berlin, Germany. Association for Computational Linguistics.

Fred Zenker and Kristopher Kyle. 2021. Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, 47:100505.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.

12. Appendix

12.1. MultiRAIN Formulation and Algorithm

RAIN was designed for unidimensional optimization problems, but we aim to optimize two criteria: precision and comprehensibility. To address this, we propose MultiRAIN, a multidimensional adaptation of the RAIN algorithm.

To explain MultiRAIN, we use the notation introduced by Li et al. (2024b). We refer to Li et al. (2024b) for further details on unchanged processing components. We denote individual tokens or values by lowercase letters, such as y, and represent sequences of tokens or values by uppercase letters, such as Y. In particular, $Y_{i:j}$ refers to the token set $(y_i, y_{i+1}, y_{i+2}, \ldots, y_j)$. The RAIN algorithm starts from the root node (the user query) and selects the next token set based on the formula:

$$Y' = \arg\max_{Y_{i:j}} \left(f(V_{\alpha:\beta}(Y_{i:j}; Y_{1:i-1}), \theta_{\alpha:\beta}) + c \cdot u(Y_{i:j}; Y_{1:i-1}) \right) \quad (1)$$

where f is a function to combine multiple metrics, $V_{\alpha:\beta}$ is a set of values of metrics, $Y_{i:j}$ are tokens that are being generated, $Y_{1:i-1}$ are all the tokens that have been previously generated, c is a regularization hyper-parameter balancing exploitation and exploration of the optimization search, $u(Y_{i:j}; Y_{1:i-1})$ indicates the extent to which a token set has been explored. The value $u(Y_{i:j}; Y_{1:i-1})$ increases when rarely visited branches are explored (see Li et al. (2024b, pp. 5–6) for a detailed description). If V represents values of a single metric, the equation aligns with RAIN.

Function f: Combining Multiple Metrics The function $f(V_{\alpha:\beta}(Y_{i:j}; Y_{1:i-1}), \theta_{\alpha:\beta})$ reflects a method to combine the values $V_{\alpha:\beta}$, e.g., via a sum or an average. Moreover, $f(V_{\alpha:\beta}(Y_{i:j}; Y_{1:i-1}), \theta_{\alpha:\beta})$ penalizes any individual value below a threshold θ, guaranteeing that a minimum level of desired metrics (e.g., precision and comprehensibility) is reached. A general formulation of f is:

$$f(V_{\alpha:\beta}(Y_{i:j}; Y_{1:i-1}), \theta_{\alpha:\beta}) = g(V_{\alpha:\beta}(Y_{i:j}; Y_{1:i-1})) \cdot p \quad (2)$$

where g is a combination function like an average or a sum, and p is a penalty factor.

Specifically, we implement the following for the Privacy QA system. Let $V_{\alpha:\beta}$ be the set $\{v_{\text{precision}}, v_{\text{comprehensibility}}\}$ and $\theta_{\alpha:\beta}$ be the set $\{\theta_{\text{precision}}, \theta_{\text{comprehensibility}}\}$.

First, we define a penalty factor p, where $0 \leq p \leq 1$. If none of the values $\{v_{\text{precision}}, v_{\text{comprehensibility}}\}$ falls below their corresponding thresholds $\{\theta_{\text{precision}}, \theta_{\text{comprehensibility}}\}$, then $p = 1$ (no penalty), otherwise $p = 0$.

Second, we define the combination function as the average across $v_{\text{precision}}$ and $v_{\text{comprehensibility}}$ (abbreviated as *prec* and *comp*, respectively) and combine to:

$$f(V_{\text{prec,comp}}(Y_{i:j}; Y_{1:i-1}), \theta_{\text{prec,comp}}) = \frac{v_{\text{precision}} + v_{\text{comprehensibility}}}{2} \times p.$$

Backward process After reaching the leaf node $Y_{i:j}$, a multidimensional evaluation is performed that computes scores $s_{\alpha:\beta}(Y_{1:j})$. This self-evaluation initiates the "backward process" as described by Li et al. (2024b, pp. 6–7). Scores s are the basis for values $V_{\alpha:\beta}(Y_{i:j}; Y_{1:i-1})$ in that the value $v_\alpha(Y_{i:j}; Y_{1:i-1})$ represents the mean scores of the token sequences that take $Y_{1:j}$ as their prefix (Li et al., 2024b, p. 6).

MultiRAIN Algorithm Note that Algorithm A is based on Algorithm 1 as presented by Li et al. (2024b). However, we made four changes to generalize for multidimensional optimization and to maintain clarity. We highlight these changes in purple. Firstly, in the presented algorithm, we refer to our Equation 1, which is generalized for multi-dimensional optimization. Secondly, we changed the algorithm to use the output of the function f, see Equation 1, as this function combines multiple metrics. Thirdly, we added the option to evaluate answers not only through the language model's self-evaluation, but also through rule-based evaluation. Finally, we changed the notation for the language model from "f" to "L" to avoid confusion with the function f as used in Eq. 1.

Algorithm 1: Multi Rewindable Auto-regressive Inference

```
Input: Language model L, current token sequence X, maximum number of search iterations T, minimum number of search iterations Tm, value threshold V, output Ω of function f as used in Eq. (1);

Output: Next token set Y;

1: t ← 0, root node ← X, current node ← X;
2: for t ≤ T do
3:     while the current node is not a leaf node do
4:         current node ← child node of current node according to Equation (1);
5:     end while
6:     Score s_α ← self-evaluation (current node and its context);
7:  if rule-based evaluation exists then
8:  Score s_β = rule-based evaluation (current node and its context);
9:  end if
10:    Querying L to sample q candidate token sets and appending them to the current node
11:    Rewind to the root node and update according to Equation (2) as in Li(2023);
12:    t ← t + 1;
13:    if t ≥ Tm & Ω of the values of the most-visited child node from the root ≥ V then
14:        break;
15:    end if
16: end for
17: Y ← the most-visited child node from the root;
```

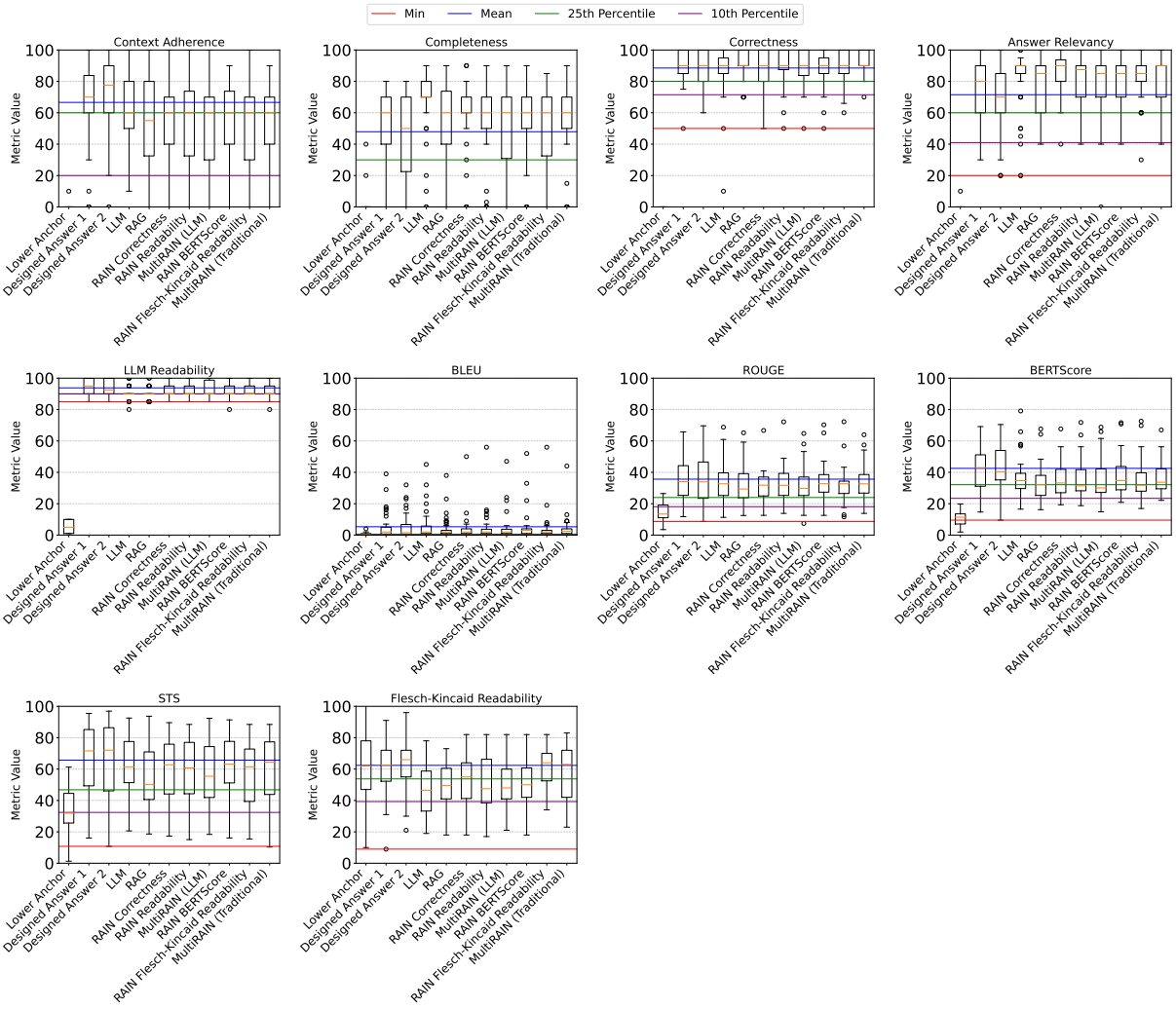


Figure 5: Raw metric scores and thresholds evaluated on the expert-generated dataset for all metrics where “bigger is better”. Threshold computation depends on the metric implementation (see Section 5.2).

12.2. Raw Metric Scores and Thresholds

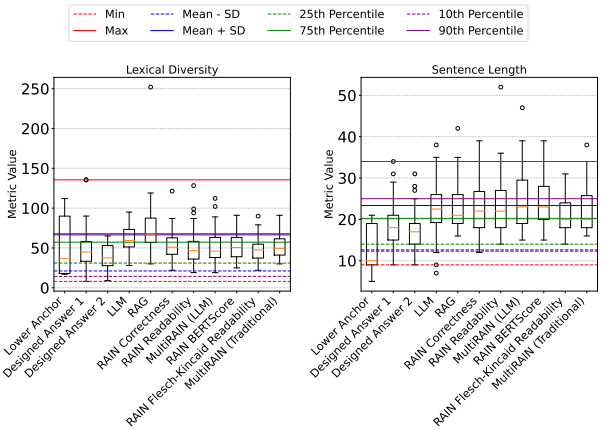


Figure 6: Raw metric scores and thresholds evaluated on the expert-generated dataset for all interval-based metrics (see Section 5.2 for information on threshold computation).

12.3. PCA

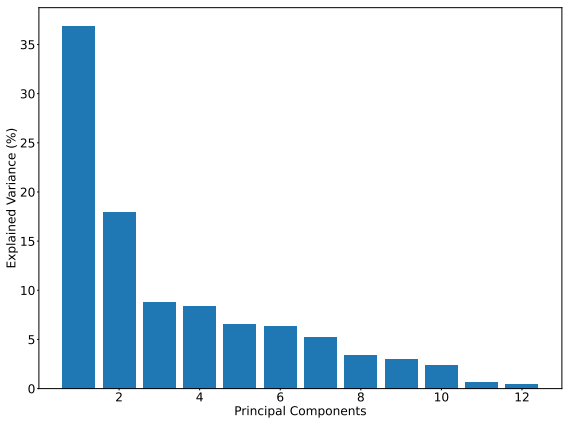


Figure 7: Explained variance by principal components. The first and second principal components are projected as x- and y-axes, respectively, in Figure 4.

Figure 7 shows explained variance across components, with a steep drop after the second, justifying using 2D PCA projections as in Figure 4.

12.4. Prompt Templates

The prompt template used for RAG is shown in Prompt (A). In the first realization of RAIN and MultiRAIN, we apply LLM-as-a-judge metrics for alignment, but use continuous self-evaluation strategies without providing examples. Prompts (B) and (C) show the prompt templates to assess correctness and readability.

Prompt (A)
You are an assistant who answers questions about data protection. Only the following knowledge is available for answering: **<Documents>**
Do not use knowledge that does not appear in the sources. Not all sources need to be used.
User: <*Query*>
Assistant:

Prompt (B)
Correctness measures whether a given model response is factual or not. Correctness (f.k.a. Factuality) is a good way of uncovering open-domain hallucinations: factual errors that don't relate to any specific documents or context. A high Correctness score means the response is more likely to be accurate vs a low response indicates a high probability for hallucination. Evaluate the correctness of the assistant's response: {text}. The Correctness should be given as a score from 0 to 100, where 100 is perfect correctness and 0 is no correctness. Think step by step, and present your reasoning before giving the answer. After reasoning, provide an overall score in the following format: 'Overall score: number'. The overall score can be an average of scores that you come up with during the reasoning. If no sensible overall score can be provided, because the metric does not apply then you can provide 'Overall score: NA'.

Prompt (C)
Read the text below. Then, indicate the readability of the assistant's response, on a scale from 1 (extremely challenging to understand) to 100 (very easy to read and understand). In your assessment, consider factors such as sentence structure, vocabulary complexity, and overall clarity. Text: {text} Think step by step, and present your reasoning before giving the answer. After reasoning, provide an overall score in the following format: 'Overall score: number'. The overall score can be an average of scores that you come up with during the reasoning. If no sensible overall score can be provided, because the metric does not apply then you can provide 'Overall score: NA'.

LDS Contractual Framework: Principles, Status and Implementation

Penny Labropoulou[1], Kossay Talmoudi[2], Dimitris Gkoumas[1], Katerina Gkirtzou[1], Miltos Deligiannis[1], Leon Voukoutis[1], Athanasia Kolovou[1], Khalid Choukri[2], Stelios Piperidis[1], Dimitrios Galanis[1]

[1] "Athena" R. C., Institute for Language and Speech Processing (ILSP), Greece
[2]Evaluations and Language Resources Distribution Agency (ELDA), France
Corresponding author: penny@athenarc.gr

Abstract

To strengthen competitiveness and digital sovereignty, the European Union has promoted the development of Common European Data Spaces to enable secure and interoperable data sharing between participants for various sectors. Data spaces combine technical infrastructure with governance mechanisms to ensure trust, transparency, data sovereignty and interoperability. Their operation must comply with the evolving European regulatory framework as well as contractual law. This paper presents the strategy adopted in the Language Data Space (LDS) to operationalise these requirements, focusing on its contractual framework and supporting instruments. It outlines the governing principles designed to ensure lawful, transparent, and fair data transactions while safeguarding the rights and obligations of data providers and consumers alike. It further describes the actual framework, and the recommended data sharing licences, with a particular emphasis on the LDS standard licence. Finally, it presents the automation tools designed and developed to support the relevant workflows while serving a wide range of users that have little or no knowledge of technical and legal complexities.

Keywords: data space, contracts, licences, semantic representation, ODRL

1. Introduction

Recognising the strategic value of data for competitiveness and digital sovereignty, the European Union has set out an ambitious vision to establish a genuine single market for data[1]. At the heart of this vision lie the Common European Data Spaces[2] (CEDS), designed to facilitate the availability and reuse of data across key economic sectors while ensuring that data holders retain control over their assets. As a critical driver of economic growth, innovation, job creation, and societal progress, data require a robust, coordinated, and trustworthy ecosystem capable of unlocking their full potential.

In a rapidly evolving digital environment, **data spaces** have emerged as a foundational framework for enabling secure, transparent, and reliable data sharing among multiple stakeholders. By combining technical infrastructure with the required governance mechanisms, data spaces strengthen trust and interoperability across sectors and borders, thereby reinforcing Europe's data economy. According to the DSSC glossary, a data space is defined as an "interoperable framework, based on common governance principles, standards, practices and enabling services, that enables trusted data transactions between participants"[3]. The Data Act Article 33(1) further states: "[...] common European data spaces [..] are purpose- or sector-specific or cross-sectoral interoperable frameworks for common standards and practices to share or jointly process data for, inter alia, the development of new products and services, scientific research or civil society initiatives". Both definitions highlight that data spaces are not merely technical platforms, but structured environments built on shared rules, common and mutual accountability.

Trust and data sovereignty can only be achieved if all transactions within data spaces are legally transparent and fully compliant with applicable regulatory and contractual frameworks. Both EU and national legislation, covering areas such as data protection, competition law, and intellectual property rights, must be carefully observed. Key instruments including the Data Act[4], the Data Governance Act[5], the Artificial Intelligence Act[6], and the General Data Protection Regulation (GDPR)[7]

[1] https://ec.europa.eu/digital-single-market/en/policies/building-european-data-economy

[2] https://digital-strategy.ec.europa.eu/en/policies/data-spaces

[3] https://blueprint.dssc.eu/?pane=glossary&glossary=1-key-concept-definitions

[4] https://digital-strategy.ec.europa.eu/en/policies/data-act

[5] https://digital-strategy.ec.europa.eu/en/policies/data-governance-act

[6] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[7] https://eur-lex.europa.eu/eli/reg/

Proceedings of LEGAL2026 and CALD-pseudo 2026 @ LREC 2026, pages 26–34
12 May 2026. ©ELRA Language Resources Association (ELRA), 2026

play a central role in shaping this legal landscape. In addition, strict adherence to competition rules is essential to ensure that data spaces operate on fair and non-discriminatory terms, granting all eligible participants equal access to documentation, governance structures, and support necessary to conduct lawful data transactions.

At the same time, *data providers* must retain the ability to define the terms/conditions under which their data are accessed and reused, thereby preserving control over their assets. Conversely, *data users* must be clearly informed of, and fully understand, the contractual obligations they assume before agreeing to such terms, ensuring accountability in the event of non-compliance. Where feasible, both parties should be supported by automation tools and standardised contractual mechanisms to streamline negotiations, enhance legal certainty, and reduce transaction costs within data spaces.

This paper presents the strategy adopted for the implementation of the above principles in the context of the Language Data Space (LDS), focusing on the contractual framework, as well as the instruments designed and developed to support it.

The following section outlines the topic of the paper by presenting the background that guides the legal work in LDS. Section 3 takes a closer look at the LDS contractual framework, its governing principles and recommendations, while the next section is devoted to the tools/mechanisms that support the contractual specifications in the various LDS workflows. Section 5 describes relevant work and, finally, Section 6 concludes with the current status and an outlook into the next steps to be undertaken.

2. Background and Requirements

The sharing of language data has a long-standing tradition within the Language Resources and Technology (LRT) community. From an early stage, it became clear that sustainable data sharing requires a coherent and legally sound contractual framework governing both distribution and reuse. Assigning a licence, i.e., a formal legal document specifying the rights granted and the restrictions imposed on the use of an asset, is essential whenever resources are made available beyond their original creators. Without clear licensing terms, even high-quality datasets cannot be confidently (re)used, integrated, or (re)distributed.

In response to this need, the community has increasingly adopted standardised open licences over the years to enable clarity and interoperability. Widely recognised frameworks, such as the Creative Commons (CC) family of licences[8], the Apache License 2.0[9], etc., are commonly used for datasets produced through national and EU-funded projects. At the same time, community-driven licensing schemes have been developed to address more specific research and commercial requirements. Notable examples include the licences used by the European Language Resources Association (ELRA)[10] and those developed within META-SHARE[11] (Piperidis, 2012), which offer tailored solutions aligned with the particularities of language data.

Despite this progress, a substantial volume of legacy data continues to circulate without a clearly defined licence. In many cases, such resources are accompanied only by a brief informal statement (often referred to as "access statement") that outlines general conditions of use (e.g., "free for research purposes"). While these statements may signal an intention to allow reuse, they typically lack the legal precision and enforceability of a formal licence. In other cases, even such minimal guidance is absent. This legal ambiguity creates uncertainty for potential users, discourages responsible reuse, and may ultimately render valuable data effectively unusable.

In today's data and AI ecosystem, new requirements and new opportunities emerge.

In this setting, the demand for a comprehensive and transparent legal framework becomes even more pressing as data are no longer consumed solely by humans, but are also processed directly by machines (e.g., AI agents), often with little or no human intervention. In such environments, ambiguities in usage conditions can no longer be resolved through manual interpretation; consequently, licensing frameworks must evolve beyond human-readable legal texts. While licences are indispensable for ensuring legal compliance, especially in the case of legal disputes, acting as evidence, the terms included in them must also be machine-readable and, ideally, machine-understandable and machine-actionable, enabling automated compliance checks and dynamic access control. Achieving this objective requires semantic accuracy grounded in shared, standardised vocabularies capable of encoding the terms (permissions, obligations and restrictions) included in licences. Such vocabularies must be commonly agreed upon and interoperable, ensuring that both machines and human actors interpret the same legal concepts consistently and act upon them accordingly.

2016/679/oj

[8] https://creativecommons.org/share-your-work/

[9] https://www.apache.org/licenses/LICENSE-2.0

[10] https://www.elra.info/

[11] http://www.meta-share.org/

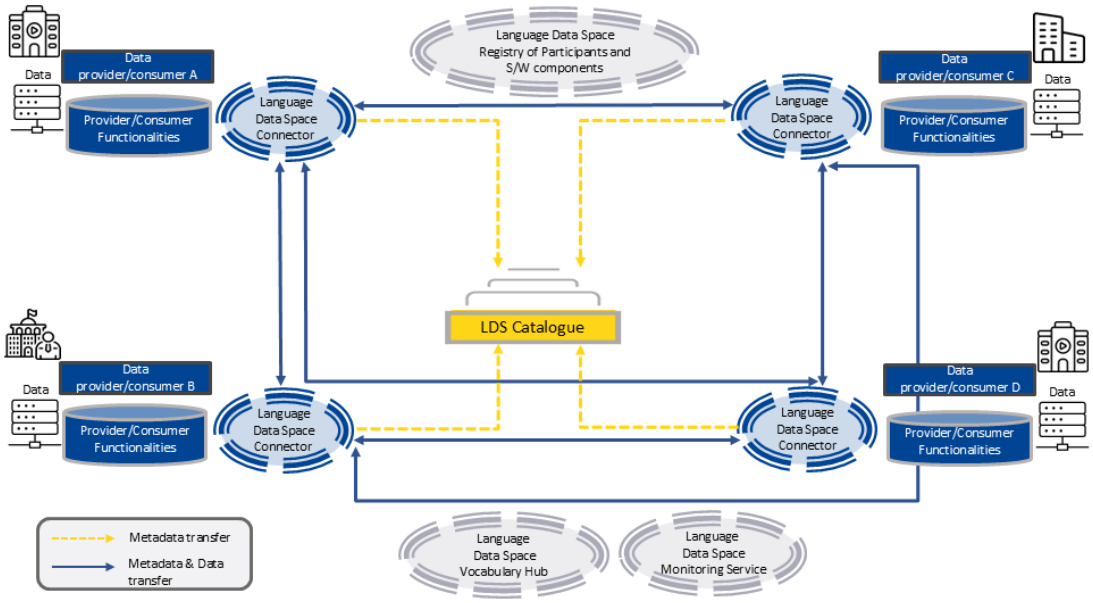


Figure 1: LDS architecture

To serve the principles of data sovereignty and interoperability, a **data space** is conceived of and implemented as a set of separate technical components ("participant agents", aka "Connectors") inside which their owners/operators perform all necessary actions related to their own assets and interact with other components (central or other participants' components) through secure communication channels following specified technical protocols when specific criteria are met. One of the most important protocols is the *Dataspace Protocol (DSP)*[12], a recently developed standard that lays the foundations for technical and semantic interoperability. More specifically, the DSP regulates data sharing transactions between participants in data spaces. Metadata descriptions of assets are exchanged between participants in the form of **DCAT**[13] catalogues; DCAT is an RDF vocabulary designed to facilitate interoperability among catalogues published in the web, catering for the description of datasets and data services. Data access and usage conditions are expressed as *policies*, encoded as formal statements with the **Open Digital Rights Language (ODRL)**[14] vocabulary; ODRL is a policy expression language that provides a flexible and interoperable information model, vocabulary, and encoding mechanisms for representing statements about the usage of content and services. Data (assets) are automatically negotiated and accessed using the respective data transfer APIs/protocols. Finally, logs are generated for each transaction, enabling monitoring and accounting thereof.

The LDS complies with data space principles, adheres to the recommended standards, such as the DSP, and is built based on, extending and customising relevant technologies to language data requirements in view of developing AI trustworthy systems.

Figure 1 shows the architecture of LDS. The LDS is framed as a decentralised network of organisations that install and operate the LDS Connector, which offers functionalities supporting all operations centered around data exchange, from publication of language data to their discovery and actual transfer, as well as the logging of such transactions. Transactions, such as metadata discovery and data flows, are performed Connector-to-Connector (peer-to-peer). The picture is complemented by the four LDS Central Components depicted in Figure 1, which are controlled by the LDS Governance Board (GB) and aim to facilitate the interactions between participants at different dimensions.

3. LDS Contractual Framework

3.1. Principles and Considerations

With regard to the contractual framework, it is important to take into account that LDS end users come from a variety of backgrounds with different expertise and, most important, varying levels of knowledge both on legal and technical issues. Intricate legal concepts, complex and ambiguous terms in legal contracts can be the source of misunderstandings and potential disputes for the usage of

[12] https://eclipse-dataspace-protocol-base.github.io/DataspaceProtocol/2025-1/

[13] https://www.w3.org/TR/vocab-dcat-3/

[14] https://www.w3.org/TR/odrl-model/

data assets in a legal and legitimate way. Moreover, familiarity with writing and understanding ODRL statements is not to be expected from the majority of LDS users. It is, therefore, important that the contractual framework in the LDS ecosystem is developed and presented to users in an easy to understand and apply way, while the technical implementation provides sufficient, flexible and user-friendly mechanisms for the description and selection of policies. The LDS, therefore, employs a set of principles and instruments that aim to cover the requirements described in Section 2:

- all assets offered through the LDS should be assigned a licence, preferably from among the ones recommended by the LDS GB (see Section 3.2); yet, respecting the data sovereignty principle, providers are also allowed to use their own licences provided that they are compliant with the LDS governance framework; for such cases, providers are advised to contact the LDS GB;
- all assets offered through the LDS must be assigned a "policy" in the form of an ODRL statement; in the case of an assigned licence, the policy terms must be aligned with the terms included in the legal text;
- to support users that have little or no technical knowledge of ODRL, the LDS technical platform offers a suite of tools described in Section 4;
- to support users with little legal knowledge, the LDS GB recommends the use of specific licences (see Section 3.2), while the LDS helpdesk, offers general consultation services and collects FAQs and useful documents and publishes them at the official LDS website.

3.2. Recommended Data Sharing Licences

The LDS GB recommends specific licences that can be used by all data providers and which are compliant with the LDS principles. During the first period, these were mainly the standard open licences that are most frequently used in the LRT community.

In the most recent release of LDS (v3.0.0), the **LDS Standard Licence** has been introduced and is proposed to data participants. This standard licence implements a modular framework for the provision of data. The licence distinguishes between the roles of data provider and data recipient (consumer) and enables to configure through annexes the “acceptable purposes” that are allowed by a data provider in regard to the data. The licence presents a common list of definitions that are compatible with common practices in the data sharing ecosystem (e.g., internal use, non-commercial use, commercial use, data derivation, product development, public release).

The licence adopts a liability capping mechanism in the sense that liability is aggregated per breach of the licence and is calculated on the basis of the economic value of the dataset. However, special mechanisms are set for zero-fee datasets (fixed per-breach caps), in order to avoid uncapped exposure for open data providers while at the same time maintaining a deterrent effect for serious non-compliance.

The licence is aligned with European Union instruments relevant to data sharing (notably the Data Governance Act and the Data Act) and is complemented by specific data protection clauses that allocate controllership between the data provider and data recipient and clarify the obligations of compliance with GDPR principles. Intellectual property compliance is treated with care as well, in the sense that the data provider vouches for their ownership of rights on the shared data.

4. Implementation Mechanisms

Depending on the LDS workflow, various types of tools and mechanisms are required for the implementation of features related to the contractual framework. More specifically, the following functionalities need to be supported:

- data owners and providers need to *describe in a formal manner* the access and/or usage policies (e.g. free vs. on-a-fee base, for a limited time duration, for usage calculated by times or volume used) under which they wish to share their data, as well as to give or revoke their authorisation to the usage of their data, and, upon conditions, change the access rights for their data;
- data consumers need to *view and understand* the policies under which the data are offered before requesting to acquire them, *get access to them in compliance with their access rights*, as well as *perform any actions required* for getting access to them in a lawful manner (e.g., pay the required fee);
- LDS operators and stakeholders must implement *technical enforcement mechanisms* that control, to the extent possible, data access, and authorise such access only in compliance with the designated policies, as well as *monitoring mechanisms* that keep track of the transactions in the LDS ecosystem and the flow(s) of datasets after leaving the providers’ trusted

boundaries, which can be used as evidence in the case of breach of usage policies.

Policies in data spaces are distinguished between *publication policies*, that define the access to (aka. visibility of) the metadata descriptions of offers, and *contract policies*, that regulate access (i.e., "who can access data and under what conditions") and usage (i.e., "what actions can be performed and which obligations are provided according to the policy once accessed) of the actual data.

Furthermore, in data spaces, the offering of assets is performed in separate steps: providers create the metadata descriptions for their assets and define the policies they would like to share them with, independently of each other. They can then combine them together assets and policies in order to create "offers". The same asset may be offered with different policies under different conditions, e.g., for free for research purposes and under a fee for commercial applications.

4.1. Defining policies

For the definition of policies data providers/owners have at their portfolio a range of tools covering the needs and preferences of a varied range of users.

4.1.1. Pre-population with standard licences

The LDS GB has selected and recommends a number of open standard licences (cf. Section 3.2). These have been transformed into ODRL policies, and are included in the LDS Connector, so that they can be used out of the box during the selection process (cf. Section 4.2), thus alleviating providers from the burden of creating them from scratch.

In this endeavour, we take benefit of existing resources, namely the DALICC License Library[15] and the library of RDF representations of public licences offered by the Universidad Politécnica de Madrid (UPM)[16]. DALICC is a software framework that supports the automated clearance of rights and provides various APIs granting access to their licence database, including the ODRL representations of standard licences. The UPM library exposes links to RDF representations of licences popular in the language data and services community; these include an ODRL representation for them, as well as an RDF representation of information derived from the SPDX License List[17]. For our purposes, we have defined a workflow whereby we retrieve information from these two sources, combine them and transform them, as needed, in order to create the final representations that are imported in the LDS database of policies. Given that the EDC connector[18], which is the foundation of the LDS connector, does not support the full ODRL, these transformations aim to ensure that the imported policies are valid and compliant with the limitations imposed by the EDC. In addition, a script automating this transformation has been developed and can be adapted to retrieve information from other sources if found.

Out of this process, 19 contract policies and 1 publication policy (imposing no restrictions on the visibility) are included in the LDS connector that is installed by all users.

4.1.2. LDS standard licence editor

As depicted in Figure 2, the LDS licence editor guides the user into customising and consolidating the LDS standard licence template through the selection of a set of predefined options for the access and usage of their assets (e.g., scope of use, permission/prohibition of derived works, charge of fees, etc.). The outputs of this process are (a) a human-readable licence, combining the legal text from the template and the annex with all the options selected by the user, and (b) the ODRL representation of the selected options combined into a policy intended mainly for machine consumption. To avoid creating duplicates, the licence automatically takes a title with an aggregation of the acronyms used for each of the options.

4.1.3. Generic policy editor

The LDS generic policy editor exploits the fact that licences/policies consist of terms/conditions, in the form of permissions, prohibitions, obligations and restrictions imposed on them. Each of these can be defined in a generic way as a "policy class", i.e., an atomic policy template referring to a specific rule governing data access and/or usage. For instance, restrictions can be based on geographic criteria (e.g., data can only be accessed by participants registered in certain geographic areas), time criteria (e.g., data can only be accessed or used for a certain time period), related to purpose of use (e.g., data to be used only for research purposes), type of recipient (e.g., data to be used only by SMEs), following a financial transaction (e.g., data to be accessed with a certain fee), etc.

Such restrictions can be represented in the form of a ready-to-use ODRL abstract statement. The generic policy editor offers these statements to data providers like building blocks that they can easily select, instantiate with their desired values (e.g., adding an amount to the fee for a data asset, or selecting a country where an asset can only be distributed), bundle together and, thus, create the

[15] https://www.dalicc.net/

[16] https://rdflicense.linkeddata.es/

[17] https://spdx.org/licenses/

[18] https://projects.eclipse.org/projects/technology.etc

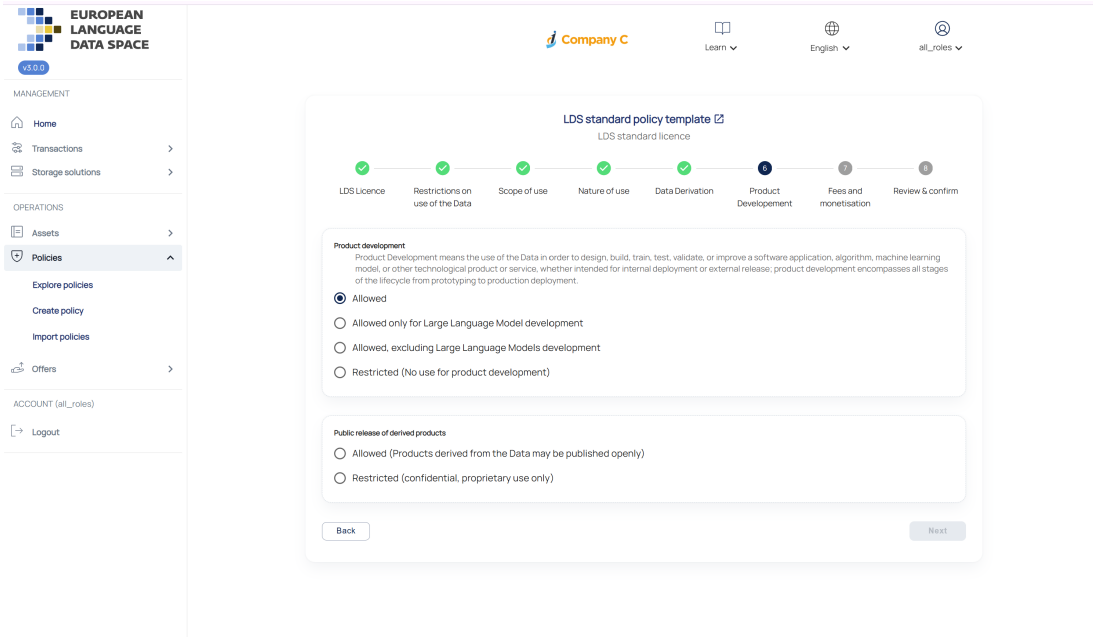


Figure 2: LDS standard licence editor: Selection of attributes (derivatives)

policies they wish to assign to their data assets, as shown in Figure 3.

The LDS data governance framework defines the set of policy classes that can be used by LDS participants. The first set has been selected through (a) the extraction of patterns of conditions that are common among the most popular licences used in LRT catalogues and (b) a subset of the policy classes identified in data spaces (Steinbuss et al., 2021). The current LDS version includes a set of 11 policy classes for contract policies and 2 for the publication policies.

4.1.4. Import of user-created policies

The import option, with the upload of a JSON-LD file with the policy expressed in ODRL, is meant for (a) advanced users with experience in ODRL and (b) cases of metadata descriptions imported from other infrastructures. To avoid unnecessary complexities, it can only be used for contract policies. The imported file is validated with the built-in validator of the EDC connector, which means that it must comply with the ODRL schema as implemented by EDC. Finally, users of this functionality are warned that terms included in the policy that don't belong to the LDS preset policy classes cannot be enforced, even if technically feasible, as the software code that implements the enforcement is specific to each policy class and cannot automatically be generated and integrated into the LDS. In general, users are advised to contact the LDS Governance Board and technical team for this option.

4.2. Viewing and assigning policies

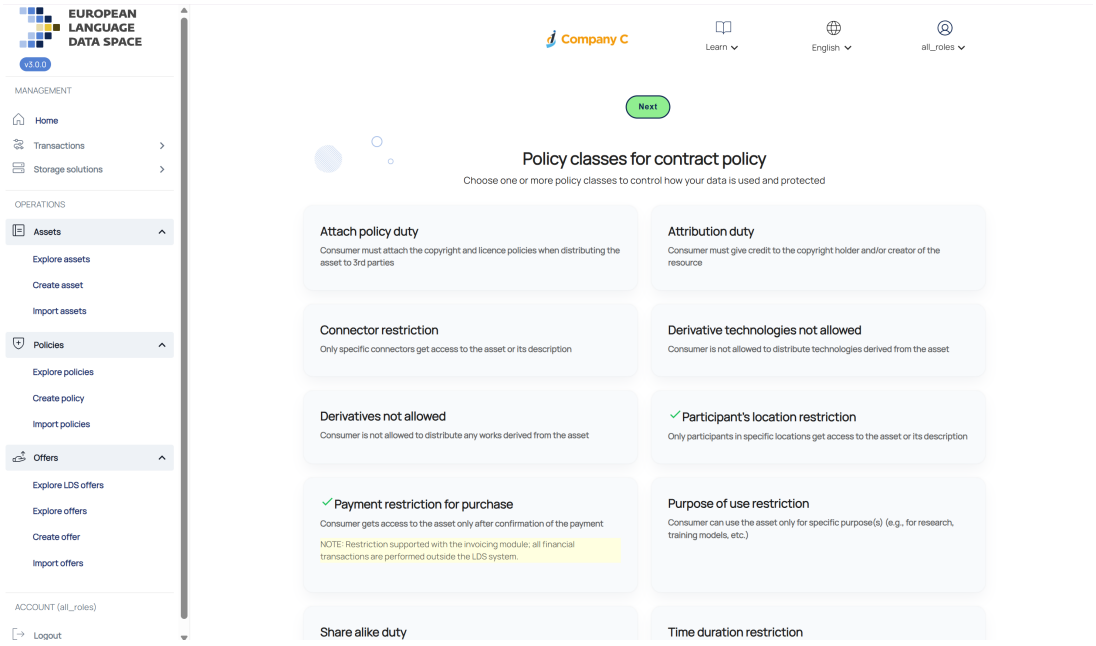
During the creation of offers, providers are prompted to select the appropriate publication policy and contract policy from the predefined list. To facilitate the selection, the list displays the title, a description, a link to the URL with the legal code (if added by the user) and, in the case of policies charging fees, the respective amount. They can also use the free text search to narrow the selection.

It should be noted that the EDC connector library for the display of policies/licences does not support the view described above. Although the EDC connector supports adding and storing additional metadata for policies (besides those of the ODRL vocabulary), such as title and description, it does not support their retrieval and hence their display. To overcome this, an LDS customised extension was developed.

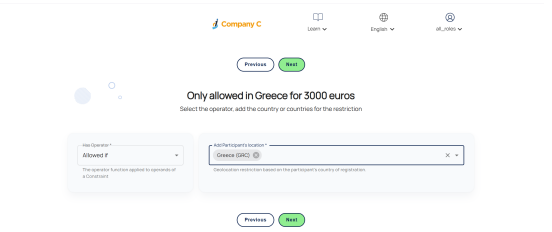
The same extension is utilised to display the full legal text alongside its ODRL representation to consumers (Figure 4), thus helping them understand the terms and make an informed decision.

4.3. Enforcing policies

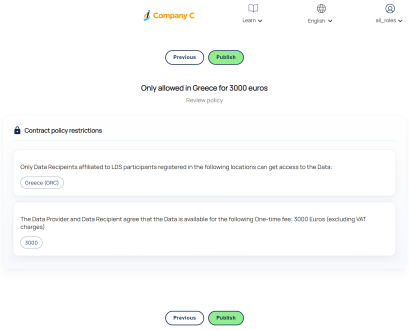
Requesting and getting grant to access an offer is electronically negotiated and concluded in data spaces. Policy enforcement is performed automatically to the extent that this is technically implementable. Thus, the fulfilment of access prerequisites must be evaluated together with the application of access restrictions before granting access to an offer. Policy evaluation takes place at specific stages of the respective implemented workflows.



(a) Step1: Selecting policy classes



(b) Step2: Instantiating policy classes (location restriction)



(c) Step3: Reviewing and publishing policy

Figure 3: LDS generic policy editor

The first stage is when a consumer requests *access to the catalogue* of another participant. At this point, the evaluation looks into the publication policies which determine access to the offer, i.e., blocking the visibility of an offer in a consumer's catalogue. For instance, a provider's offers may be visible only by participants registered in the European Union. The checks are implemented via the respective policy evaluation functions in EDC.

When the consumer finds a specific offering of interest in the respective catalogue from other participants, he/she can initiate the so-called *"contract negotiation"* process. The consumer's connector contacts the provider's connector and sends the request. The provider's connector checks the validity of the request. Firstly, it checks the identity of the connector, and, if valid, further processes the request. Then, it checks whether the consumer fulfills the required terms encoded in the contract policy offer via the respective EDC policy evaluation functions to decide whether to accept or reject it.

The same policy evaluation functions and checks may be bound to the *data transfer* process, thus blocking the actual access to the dataset. This is the third policy evaluation point, and it has been used for the implementation of the exchange workflow of on-a-fee datasets, together with the invoicing module that is integrated in the LDS connector.

More specifically, in LDS a policy may define charges for granting access to a dataset. In this workflow, consumers negotiate successfully assets that are offered on a charge basis but cannot get access to the actual asset immediately. Instead, an invoice is automatically created at the provider's side and transferred (on demand) to the consumer's side. The consumer can download the invoice and perform the payment (outside the LDS). Once the provider is notified that the payment has been completed, he/she updates the status of the invoice in the invoicing module. The new invoice (i.e., the one marked as "paid") is transferred again to the

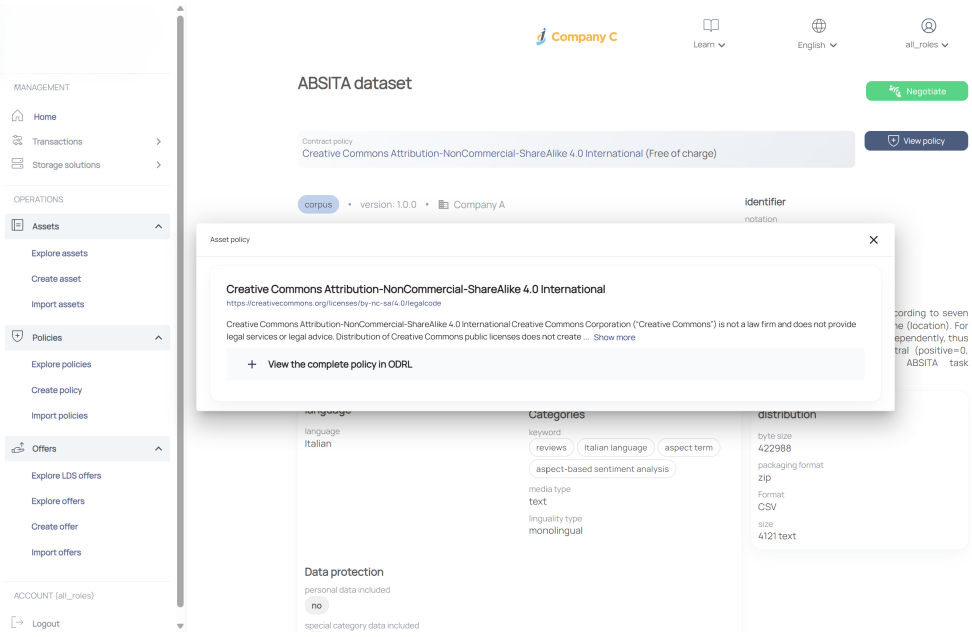


Figure 4: View of licence

consumer's connector. It is only after this step, that consumers can get access to the data and transfer the files to their own connector, since the respective evaluation function at the provider side checks whether the invoice is paid.

For the policy enforcement through the aforementioned evaluation functions, we rely on the EDC policy engine. The engine has access to the (access) token that is included in the request (e.g., start transfer) of a consumer's connector. The function parses the ODRL policy and checks whether the required conditions are met. For example, if the ODRL policy requires that the consumer's organisation must be active in EU, then it checks whether the respective (predefined within LDS) claim included in the token (e.g., "location") has a specific value (e.g., "EU"). If yes, the function returns "true" and the transfer starts; otherwise, it returns "false", and transfer is not allowed. Policy functions are added/registered to the policy engine of the connector as an EDC extension which is the standard/recommended way for adding functionalities to EDC. To the best of our knowledge, the version of EDC that we use (0.7.0) contains only one built-in policy function which controls data transfers for assets offered within a restricted time interval. All other policy restrictions have been implemented by the LDS technical team.

4.4. Monitoring contracts

Data sharing transactions are governed by contracts, thus mitigating potential disputes and ensuring more efficient and transparent data management. Monitoring and recording of all transactions related to the exchange of data offerings upon the conclusion of relevant contracts is of utmost importance. LDS supports the logging of each contract concluded among two participants, i.e., the details of the participants, the metadata descriptions of the offer, the licence/policy under which it has been acquired, as well as the data transfer requests and exchanges following the contract, locally at both involved connectors as well as at the Central Logging Component. Both participants as well as the GB members (in the Central Monitoring Component) have full access to the text and ODRL statement representing it.

5. Related Work

In the context of data spaces, the Data Spaces Support Centre (DSSC)[19], in their capacity to advise and assist CEDS setting up their infrastructures, provides recommendations and relevant information on the contractual framework and promotes legal and technical interoperability across data spaces. To achieve this objective, it collaborates with legal as well as technical and business experts and issues the DSSC Blueprint (Data Spaces Support Centre, 2026), which crystallises the current state-of-the-art and recommendations. The Contractual Framework building block[20] describes the legally enforceable agreements that underlie the operation of a data space, as entered into by different parties in a relationship with the data space,

[19] https://dssc.eu/

[20] https://blueprint.dssc.eu/?pane=business&business=contractual-framework

i.e., not only the data sharing agreements that are discussed in this paper, but also institutional and services agreements.

Data spaces are free to define their governance frameworks in which the selection and recommendation of data sharing policies constitutes an essential part. However, to the best of our knowledge, no other data space has devised an official legal document, like the LDS standard licence template, that can be used by the respective participants.

The concept of "policy classes" is not entirely novel in the licensing landscape. It shares commonalities with licensing terms and conditions, such as the "attribution", "no derivatives", etc. attributes used in the Creative Commons family of licences or the "access statements" mentioned in Section 2. Such terms are usually rendered as metadata elements accompanying licences, represented as simple labels/tags or in a more formal way in the metadata description of the licence (Rodriguez-Doncel and Labropoulou, 2015). In data spaces, "policy classes" and their enforcement mechanisms have been introduced by IDSA (Steinbuss et al., 2021).

Finally, for the implementation of the LDS generic policy editor, we have drawn inspiration from the PAP editor[21], which supports data providers to specify their Usage Control policies in ODRL and IDS formats, with a set of predifined policy classes.

6. Summary and Next Steps

The principles and foundations of the LDS contractual framework have already been defined, while new data sharing licences, especially those submitted by data providers, will continue to be reviewed, keeping up-to-date with changes and additions in the licensing domain.

The tools offered through the LDS connector will implement emerging requirements (e.g., representation and enforcement of new policy classes in the generic policy editor, addition of new recommended licences, etc.). Among the planned enhancements of the tools, the compatibility of policy classes combined together to create valid policies is one of the priorities.

The latest release of the LDS infrastructure is offered in all EU official languages, combining automatic translation (exploiting the eTranslation service[22], that is offered by the European Commission) and, where available, human curation. Given that the comprehension of legal texts is crucial for the contract conclusion and avoidance of contract breaches and misuses, this release does not include translations of licences. For upcoming releases, we plan to re-use existing translations/adaptations of standard licences by legal experts, where available, and further investigate the workflow, always in respect of the intended quality.

Finally, in order to assess and enhance both the contractual framework and the implementation mechanisms, we intend to conduct a survey among the LDS users and the LDS Interest Group [23] that will help us determine their usability and ways of improving them.

7. Acknowledgements

The Common European Language Data Space is funded by the European Union through the contract LC-01936389.

8. Bibliographical References

Data Spaces Support Centre. 2026. Dssc blueprint.

Stelios Piperidis. 2012. The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Víctor Rodriguez-Doncel and Penny Labropoulou. 2015. Digital Representation of Licenses for Language Resources. In *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, pages 49–58, Beijing, China. Association for Computational Linguistics.

Sebastian Steinbuss, Andreas Eitel, Christian Jung, Robin Brandstädter, Arghavan Hosseinzadeh, Sebastian Bader, CHristian Kühnle, Pascal Birnstill, Gerd Brost, Gall, Fabian Bruckner, Norbert Weißenberg, and Benjamin Korth. 2021. Usage control in the international data spaces.

[21] https://odrl-pap.mydata-control.de/

[22] https://cor.europa.eu/en/etranslation

[23] The LDS Interest Group consists of European stakeholders from mainly industry, representing all market segments, but also public administration and academia, that wish to contribute to and take advantage of the LDS, bringing in their own requirements and validating the emerging LDS infrastructure.

Authorship Attribution in the Times of LLMs within the Framework of the CRediT Taxonomy

Paweł Kamocki, Andreas Witt

Leibniz Institute for the German Language
R5, 6-13, 68161 Mannheim, Germany
{kamocki, witt}@ids-mannheim.de

Abstract

This article examines the concept of authorship in the context of generative language models and other uses of Artificial Intelligence, and how this new 'authorshipness' can be represented in metadata. It analyses authorship under copyright law and proposes a metadata-based approach to disclosing the use of AI in publications, drawing on the widely adopted CRediT taxonomy developed by the National Information Standards Organization (NISO), and informed by guidance from the United States Copyright Office (USCO) and the International Association of Scientific, Technical and Medical Publishers (STM).

Keywords: metadata, authorship, AI

1. Introduction

AI tools have become an integral part of research. A recent report by Wiley (2025) indicates that 84% of researchers use AI tools in their work and that demand for expanded use is growing. Writing assistance is among the most common use cases: 74% of researchers report interest in using AI for this purpose, and 59% believe that AI already outperforms humans in this task. At the same time, researchers call for transparency: according to the same report, 66% consider it highly important for authors to disclose their use of AI in drafting and editing.

These developments are likely to drive changes in the concept of authorship in academia and beyond. Authorship is a foundational concept in copyright law and can be regarded as central to contemporary economic and cultural production.

This article proposes a mechanism to address the need to disclose the use of AI (Section 5), based on the well-established CRediT taxonomy. Before presenting this solution, Section 2 examines the concept of authorship and its implications in copyright law; Section 3 introduces the CRediT taxonomy; and Section 4 reviews recent developments concerning AI and authorship from both copyright law and publishing sector perspectives.

2. Authorship in Copyright Law

The author is placed at the centre of copyright law (called "author's right" in many languages) as the initial holder of exclusive rights in a work.

In EU law, the sufficient and necessary condition for a work to be protected by copyright is its originality, understood as "author's own intellectual creation". According to the Court of Justice of the European Union (e.g., C-469/17 *Funke Medien*), this means that the work must reflect the author's personality, which is the case when *"the author was able to express his creative abilities in the production of the work by making free and creative choices"*. A logical consequence of this approach to originality is that only a work created by a human author can be protected by copyright, since only humans have a personality and are capable of making creative choices. The possibility of corporate authorship (like in a *work for hire*, where copyright is initially held by a legal entity that employs the human creator), although admitted in some (mostly common law) jurisdictions, remains an exception from the general principle of human authorship.

In many national jurisdictions copyright is transferrable, and in practice it is often transferred by the initial holder e.g. to a publisher. However, the author remains important throughout the lifecycle of a copyright-protected work. For example, the term of copyright protection (in most jurisdictions, life+70) is determined by the death of the author; the author also retains, even after the transfer of copyright, the right to claim authorship of his or her work (Article 6bis of the Berne Convention). *Author* metadata should be treated with great care also for another reason: the presumption of authorship of Article 15 of the Berne Convention. According to this provision, briefly put, in order to file an infringement case, it is sufficient that the claimant's name appears on the work *"in the usual manner."* Arguably, a "usual manner" to indicate the name of an author of a born-digital publication is in the metadata. Therefore, the person whose name appears as author in the metadata can sue for copyright infringement, even if the publication was in fact AI-generated and as such it is not protected by copyright. The burden of proving that the work was not in fact created by the person identified as the author would then rest on the defendant, and such proof is becoming increasingly difficult. This is one of the reasons why the AI Act (Article 50) requires that outputs of generative AI systems be clearly identifiable as such, but this requirement is particularly difficult to enforce in case of plain text outputs. Regardless of this legal obligation

Proceedings of LEGAL2026 and CALD-pseudo 2026 @ LREC 2026, pages 35–39
12 May 2026. ©ELRA Language Resources Association (ELRA), 2026

(incumbent on the providers of AI systems, not on the users), disclosing the AI-generated nature of contents should be considered an ethical obligation of paramount importance (cf. Kamocki, Witt, 2022 and 2024).

Copyright law does not, on the other hand, recognise "Contributor" as an autonomous concept: in works created by multiple individuals (synchronically or asynchronically), each of them is granted the status of an author, as long as his or her contribution is original (i.e., in other words, as long as he or she left his or her "personal stamp" in the work). This is of practical significance, e.g. for the term of copyright protection, which for works of joint authorship is determined by the death of the last surviving co-author. Hypothetically, if a much younger assistant is to be considered a co-author of a work, the work is likely to remain in copyright for a much longer period. However, contributors who cannot be considered co-authors of a work, i.e. those whose contribution was not original, are not protected by copyright law in any way. Classically, a PhD supervisor who contributes ideas and provides guidance to a student, but does not participate in the drafting process, is not regarded as a co-author and therefore does not hold any copyright in the thesis. The same holds, e.g, for a technical group that contributes the data analyses in a life science article – despite the fact that the findings presented in such a paper depend on the data analysis.

This may contrast with the established academic practice described in the following section.

3. Authors/Contributors in Academic Contexts according to the CRediT system

The academic community tends, for good reason, to recognize a wide range of contributions, also with 'faux' authorship. In a widely quoted example, one physics paper had 5,154 authors (Aad et al., 2015). Obviously, such "authorship" often does not meet the standards of copyright law discussed in the previous section.

This phenomenon is partly addressed by the *Contributor* metadata term (present e.g. in the Dublin Core Metadata Terms, DCMI 2020), defined as "an entity responsible for making contributions to the resource" and distinct from the *Creator* term, defined as "an entity responsible for making the resource" (DCMI 2020; note that the Dublin Core Metadata Terms diplomatically avoid the term Author altogether).

In order to allow the community to distinguish between the various contributions to a published work, the CRediT (Contributor Roles Taxonomy, https://credit.niso.org) was introduced by the National Information Standards Organisation (NISO) (Hosseini et al., 2026). Approved in 2022 as an ANSI/NISO stadard, the Taxonomy recognises the following roles, which can be attributed to every author of a published paper :

- **Conceptualisation** (formulation or evolution of overarching research goals and aims),
- **Data Curation** (management activities to prepare data for initial use and later re-use),
- **Formal Analysis** (application of formal techniques to analyse or synthesise the data),
- **Funding acquisition**
- **Investigation** (condctuing a research and investigation process, specifically performing experiments or data collection),
- **Methodology** (development or design of methodology, creation of models),
- **Project Administration** (management and coordination of the research activity planning and execution),
- **Resources** (provision of metarials, samples, instrumentation, computing resources or other analysis tools),
- **Software** (programming, implementation of existing code, testing of existing code),
- **Supervision** (oversight and leadership, including mentorship),
- **Validation** (verification of replication/reproducibility of results),
- **Visualisation** (preparation of the published work, specifically visualisation/data presentation),
- **Writing – original draft** (drafting original text, including substantive translation),
- **Writing – review & editing** (specifically critical review, commentary or revision, including pre- or post-publication stages).

In the era of generative AI, most of these contributions (arguably, all of them, apart from Funding acquisition and Supervision) can be made with AI assitance. Nevertheless, this article focuses on the roles that are decisive for authorship as it is understood in copyright law (cf. Section 2 above), that is Writing, both the "original draft" and "review & editing". Incidentally, these are also the roles (alongside Visualisation) where there is the greatest demand for disclosure of the use of AI tools according to the abovementioned report (Wiley, 2025).

Metadata are the right place for such a disclosure. Before the specific proposal for how to incorporate this information within the CRediT taxonomy is made in Section 5, Section 4 explores the various ways in which AI can be used in the writing process.

4. AI and Authorship: Copyright and Publishing Sector Perspectives

Widespread use of AI in the writing process is very likely to have a disruptive effect on copyright law and the publishing sector.

The issue of copyrightability of AI-assisted works is one of the main challenges copyright law will have to face in the near future. On the one hand, it is rather undisputed that AI-generated works should not be protected by copyright (cf. above in Section 2); on the other hand, some degree of AI-assistance in the creative process should not bar copyrightability of the output.

In Part 2 of its Report on Copyright and Artificial Intelligence, The United States Copyright Office (USCO, 2025) presented the conclusions from a large public consultation on copyrightability of AI outputs. The Report distinguishes between five types of human interactions with generative AI systems in the creative process:

- **Assistive uses** such as e.g. error correction or "brainstorming". These uses do not affect copyrightability of the resulting works;
- **Prompting**; according to USCO, prompting alone, even very detailed and repeated, *"[does] not provide sufficient human control to make users of an AI system the authors of the output".* USCO, however, leaves open the possibility that this could change with technological progress, if AI tools give users greater control over the final shape of the outputs;
- **Expressive Inputs** that are intended to be perceptible in the output (e.g. where AI is used for translation); according to USCO, in such cases the user retains at least partial authorship of the output which, however, does not extend to the AI-generated components, in a manner analogous to a derivative work;
- **Modifying or arranging AI-generated content**. In cases where the modifications are sufficiently creative, human authors can claim copyright in the final result, which, however, does not extend to individual AI-generated elements. Since this particular use of AI is of little relevance for the writing process (as opposed to, e.g., data compilation), it will not be elaborated upon in this article;
- **Inclusion of AI-Generated Content** in a larger human-authored work (e.g., special effects in a movie) should not affect copyrightability of the larger work.

Also in 2025, the International Association of Scientific, Technical & Medical Publishers (STM) issued its Recommendations for a Classification of AI Use in Academic Manuscript Preparation (2025). The document contains 9 categories of AI uses:

1. *Refinement*, correction, editing and formatting the manuscript to improve clarity of language, e.g. via the use of spell checkers, grammar checkers and similar tools – according to STM, this is the only use of AI that should not be necessary to disclose;
2. *Writing* or drafting (parts of) manuscript content (either from prompts, or by asking to substantially expand or rewrite the input);
3. *Translation* of manuscript text for the purpose of publishing (as distinct from translation of source texts in the research process);
4. Refining or *formatting of data* reported in the manuscript;
5. *Generation*, refinement, correction, editing or formatting of images, diagrams or other *figures* for illustrative purposes only;
6. *Generation*, refinement, correction, editing or formatting of *visualisations* of research *data* or results;
7. *Refinement* or formatting of *code* reported in the submitted manuscript;
8. Assisting with *gathering references*;
9. Presentation of any kind of content generated by AI tools as though it were original research data/results from non-machine sources.

According to STM, the use described in point '9' should be prohibited. All the other uses are allowed, but should be disclosed (apart from use 1, for which disclosure is not mandatory).

Due to its level of detail, the STM classification can be tentatively mapped onto the classification proposed by USCO, as shown in Table 1.

USCO	STM
Assistive use	Refinement (1), Gathering references (8), Formatting of data (4), Refinement of code (7)
Prompting	Writing (2) [from prompt]
Expressive Inputs	Writing (2) [from expressive input], Translation (3) [of manuscript]
Inclusion of AI-Generated Content	Generation of figures (5), Generation of data visualisations (6)

Table 1: The relation between the categories of AI uses proposed by USCO and STM

5. Disclosure of AI Use in *Author* Metadata: Proposal for Extension of CRediT

Author metadata is an appropriate place to disclose the use of AI in the writing process. This disclosure should on the one hand ensure a high level of transparency and by doing so promote trust; on the other hand, the adopted solution should not create unnecessary burden or stigma that could discourage users from disclosing the use of AI or, worse, prevent legitimate uses of AI.

The authors of this article propose to achieve this by extending the existing CRediT taxonomy (Section 3). Contributions consisting of Writing (be it "original draft" or "review & editing") should come with an additional information concerning the use of AI. For the sake of systematisation, the authors of this article propose to distinguish between five types of AI use (Table 2), based on the USCO and STM classifications presented in Section 4.

AI use category	Examples
Assistive use	Idea generation, topic discovery, argument development and critique, gap identification, summarising and suggesting sources
Editing	Grammar and syntax correction, spellchecking, flow improvement, style and tone adaptation
Generation of minor elements	Generation of introduction/conclusion/ abstract, finding examples (including e.g. linguistic structures)
Translation	Translation of the entire manuscript, translation of source texts/quotations
Compression/expansion	Using AI to shorten or expand existing human-generated input to meet word/page limit
Generation from prompt	Generation of an entire article from a prompt with a human assuming editorial control over the output.

Table 2: AI use categories to enrich the CRediT taxonomy

This classification contains two essential modifications compared to the USCO and STM classifications.

First, the authors of this article believe that "editing", as a relatively simple and mechanical task in which LLMs clearly outperform humans should be distinguished from other assistive uses, in which the gap in performance between humans and AI is much more debatable.

Second, in the authors' view translation deserves to be distinguished from both the "assistive uses" and from "expressive inputs", and constitute a class of its own. The use of AI to take down the language barrier should be viewed differently from other uses of AI, as it assists rather than substitutes the human thinking process. At the same time, translation seems too substantial to fall within the "assistive use" category.

The proposed classification is also meaningful from the point of view of copyright law: "generation from prompt" would most likely exclude the output from copyright protection, whereas all the other uses should have little impact on the output's copyright status

As a final step, the specific tool used in the process and its version should also be disclosed in the metadata. In short, the revised *Author* metadata can be represented as in Table 3:

Author	*Person*
CRediT role	Writing – original draft OR Writing – review & editing
AI use category	None Assistive use Editing Generation of minor elements Translation Compression/expansion Generation from prompt
AI tool	*Tool name and version*

Table 3: Author metadata disclosing the use of AI

6. Conclusion and Next Steps

The article proposes a clear and structured approach to disclosing the use of AI in author metadata, based on the well-established CRediT taxonomy. It addresses one of the most significant challenges faced by researchers, publishers, and legal scholars in the age of AI. It supports transparency and trust, reduces the risk of misattribution (thereby safeguarding the author's moral right to claim authorship), and helps to align research and publishing practices with the copyright system. It preserves the centrality of

human authorship while acknowledging the growing role of AI in the writing process.

The authors further contend that the proposed approach can be extended to other roles within the CRediT taxonomy that may involve AI assistance.

The authors intend to engage with the National Information Standards Organization (NISO) and its CRediT Standing Committee to advocate for the incorporation of this approach into a future version of the taxonomy. They will report on the outcome at a next conference.

7. Bibliographical References

Aad, G. et al. (2015), Combined Measurement of the Higgs Boson Mass in pp Collisions at √ s = 7 and 8 TeV with the ATLAS and CMS Experiments, *Physical Review Letters*, 114, 191803

DCMI (2020). DCMI Metadata Terms. Available at: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/ (last visit: 20.02.2026).

Hosseini, M., Kerridge, S., Allen, L., Kiermer, V. and Holmes, K. (2026). CRediT Roles and Example Research Tasks That Could be Attributed to Them. Available at: https://doi.org/10.5281/zenodo.18421448 (last visit: 20.02.2026).

Hosseini, M., S. Kerridge, L. Allen, V.K iermer, and K. Holmes (2026). Enhancing, Understanding and Adoption of the Contributor Roles Taxonomy (CRediT). *Learned Publishing* 39, no. 2: e2048. https://doi.org/10.1002/leap.2048.

Kamocki, P. and A. Witt (2022). Ethical Issues in Language Resources and Language Technology – Tentative Categorisation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 559–563, Marseille, France. European Language Resources Association.

Kamocki, P. and A. Witt (2024). Ethical Issues in Language Resources and Language Technology – New Challenges, New Perspectives. In *Proceedings of the Workshop on Legal and Ethical Issues in Human Language Technologies @ LREC-COLING 2024*, pages 19–23, Torino, Italia. ELRA and ICCL.

STM (International Association of Scientific, Technical & Medical Publishers) (2025). Recommendations for a Classification of AI Use in Academic Manuscript Preparation. Available at: https://s3.eu-west-2.amazonaws.com/stm.offloadmedia/wp-content/uploads/2025/04/23020709/STM_AI_Classification_Recs_19_Sept2025-1.pdf (last visit: 20.02.2026)

USCO (United States Copyright Office) (2025). *Copyright and Artificial Intelligence, Part 2: Copyrightability*. January 2025. Available at: https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-2-Copyrightability-Report.pdf (last visit: 20.02.2026).

Wiley (2025). Explanations 2025: The Evolution of AI in Research. Available at: https://www.wiley.com/en-de/about-us/ai-resources/ai-study/ (last visit: 26.03.2026)

DeID-Clinic: A Risk-Aware Pseudonymization Framework for Clinical Text De-identification and Re-identification Risk Assessment

Angel Paul[1,†], **Dhivin Shaji**[1,†], **Lifeng Han**[2,3,*]
Warren Del-Pinto[1], **Goran Nenadic**[1], **Suzan Verberne**[2]
[1]University of Manchester, Manchester, UK
[2]Leiden Institute of Advanced Computer Science (LIACS), Leiden University, NL
[3]Biomedical Data Sciences, Leiden University Medical Center, Leiden, NL
[†]co-first [*]*corresponding*: {l.han, s.verberne}@liacs.leidenuniv.nl

Abstract

The increasing availability of sensitive textual data has created an urgent need for robust de-identification methods that enable compliant data sharing while preserving downstream utility. This paper presents DeID-Clinic, a multi-layered framework for automated pseudonymization and re-identification risk assessment of clinical free-text data. Our approach integrates domain-adapted transformer models, including BioBERT and ClinicalBERT, into the MASK de-identification framework to improve the detection and masking of protected health information (PHI). Beyond entity recognition, we introduce a novel document-level risk assessment module that quantifies residual re-identification risk using a combination of k-anonymity, l-diversity, t-closeness, contextual similarity, and entity co-occurrence analysis. Experiments conducted on the i2b2 2014 de-identification dataset demonstrate strong performance, achieving macro-level F1 scores above 0.96 for several entity categories, while enabling quantitative prioritization of high-risk documents for further review. Our results highlight the effectiveness of combining neural de-identification with explicit risk modeling, supporting privacy-preserving data sharing in sensitive domains. Although evaluated on clinical text, the proposed framework is generalizable to other privacy-critical domains such as legal and administrative documents, where reliable pseudonymization and risk-aware anonymization are essential.

Keywords: Automated De-Identification, Risk Assessment, Patient Privacy, Pseudonymization, Personal Health Information

1. Introduction

The widespread adoption of electronic health records and other sensitive textual datasets has created an urgent need for reliable de-identification methods that not only remove personally identifiable information but also quantify the residual risk of re-identification (Scaiano et al., 2016; Subramanian et al., 2024; Sarkar et al., 2024). While recent neural approaches have achieved high accuracy in detecting protected health information, most existing systems focus solely on entity masking without assessing whether the resulting text remains vulnerable to re-identification through contextual clues or rare entity combinations (Sondeck and Laurent, 2025). This limitation poses a significant challenge for privacy-preserving data sharing, as effective anonymization requires both accurate pseudonymization and rigorous risk evaluation. Addressing this gap, we propose a *risk-aware de-identification* framework that integrates transformer-based entity recognition with document-level privacy risk assessment, enabling more reliable and accountable anonymization of clinical free-text data.

The need for privacy-preserving text processing is particularly critical in healthcare, where clinical narratives contain sensitive patient information protected under regulations such as GDPR and HIPAA (El Emam et al., 2006; Voigt and Von dem Bussche, 2017; Edemekong et al., 2024). De-identification aims to reduce the risk of re-identification by removing or replacing sensitive information while preserving data utility for research and clinical applications (Sweeney, 2002a; Dankar et al., 2012). For example, a clinical sentence containing a patient name, date, and location may be transformed into a pseudonymized version that maintains clinical meaning but protects individual privacy (Meystre et al., 2010; Stubbs et al., 2015).

Recent advances in neural language models, particularly transformer-based architectures such as BERT and its domain-specific variants, have significantly improved the accuracy of identifying sensitive entities in text. Models such as BioBERT and ClinicalBERT (Lee et al., 2020; Alsentzer et al., 2019) leverage domain-specific pre-training to better capture the linguistic characteristics of clinical narratives. These models have demonstrated strong performance in named entity recognition tasks, making them promising candidates for automated de-identification. However, accurate entity detection alone does not guarantee effective privacy protection (Kovačević et al., 2024). Even after

Proceedings of LEGAL2026 and CALD-pseudo 2026 @ LREC 2026, pages 40–52
12 May 2026. ©ELRA Language Resources Association (ELRA), 2026

pseudonymization, residual information such as rare entity combinations or unique contextual patterns may enable re-identification. Consequently, there is a growing need for methods that not only perform de-identification but also quantify the residual risk associated with anonymized text. Such risk-aware approaches are critical for supporting responsible data sharing and ensuring compliance with privacy regulations (Hara et al., 2018).

To address these challenges, we present **DeID-Clinic**, a multi-layered framework for automated pseudonymization and re-identification risk assessment of clinical free-text data. Our approach integrates domain-adapted transformer models into the open-sourced MASK framework (Milosevic et al., 2020) to improve entity detection and masking and introduces a document-level risk assessment module to quantify residual privacy risks [1].

This work advances privacy-preserving language processing by introducing a risk-aware pseudonymization framework that integrates neural entity recognition with quantitative privacy risk estimation. The key contributions are: 1) Risk-aware pseudonymization framework: We propose DeID-Clinic, a unified framework that combines neural de-identification and document-level re-identification risk assessment, enabling both automated pseudonymization and quantitative privacy evaluation. 2) Document-level privacy risk modeling: We introduce a novel risk scoring method that integrates classical anonymization metrics (k-anonymity, l-diversity, t-closeness) with contextual embedding similarity and entity co-occurrence analysis to estimate residual re-identification risk in free-text documents. 3) Integration of domain-adapted transformer models: We extend the MASK platform by incorporating BioBERT and ClinicalBERT models, improving sensitive entity detection accuracy on clinical text. 4) Comprehensive experimental and case-study evaluation: We evaluate the framework on the i2b2 2014 dataset and demonstrate its effectiveness in both entity detection performance and risk-aware document prioritization.

2. Related Work

Automated de-identification of clinical text has been extensively studied, with approaches evolving from rule-based systems to modern deep learning models. In addition, emerging research has begun to explore methods for assessing re-identification risk after de-identification.

[1] this is an extended work from our 2-page poster paper (Shaji et al., 2025). In this longer paper, we describe the details on methodology design and carry out more experimental evaluations and analysis.

2.1. Rule-based and Traditional ML

Early de-identification systems primarily relied on rule-based approaches, which use manually defined patterns and dictionaries to identify sensitive entities such as names, dates, and locations (Friedlin and McDonald, 2008; Meystre et al., 2014). While effective in structured settings, rule-based systems often lack flexibility and struggle with linguistic variability and ambiguity in clinical narratives.

Machine learning approaches such as Conditional Random Fields (CRFs) were later introduced, allowing models to learn entity patterns directly from annotated data (Yang et al., 2019; Liu et al., 2017). These methods improved adaptability and performance but still faced limitations in capturing long-range dependencies and contextual relationships.

Recurrent neural network architectures, particularly BiLSTM models, further improved performance by modeling sequential dependencies in text (Dernoncourt et al., 2017; Kim et al., 2018). However, these models often require extensive feature engineering and may struggle with complex contextual interactions, e.g., in long clinical documents (Lin, 2020).

2.2. Transformer-based De-identification

The introduction of transformer-based language models has significantly advanced clinical de-identification. Domain-adapted models such as BioBERT (Lee et al., 2020) and ClinicalBERT (Alsentzer et al., 2019) leverage pre-training on biomedical and clinical corpora to improve entity recognition performance. These models have demonstrated strong results across multiple clinical NLP tasks, including PHI detection. Recent systems have successfully applied transformer-based architectures to de-identification tasks, achieving state-of-the-art performance while reducing the need for manual feature engineering (Kraljevic et al., 2023).

2.3. De-identification Frameworks and Systems

The MASK framework (Milosevic et al., 2020) provides a flexible open-sourced platform for clinical text de-identification, supporting multiple named entity recognition models and masking strategies [2]. MASK enables both redaction and pseudonymization and allows integration of custom NER models. Its modular design makes it suitable for deployment in real-world clinical environments.

Another widely used system is Philter, a rule-based de-identification tool designed for large-

[2] https://github.com/icescentral/MASK_public

scale clinical text processing (Hartman et al., 2020). Philter offers high customizability and transparency through manually defined filtering rules, making it particularly suitable for environments where explainability and precise control are required. However, rule-based approaches may require extensive manual tuning and may not generalize well across datasets.

More recently, AnonCAT, integrated within the MedCAT ecosystem, combines transformer-based models with biomedical knowledge graphs to improve de-identification accuracy and contextual understanding (Vakili and Dalianis, 2022; Kraljevic et al., 2023). By leveraging domain knowledge and fine-tuning strategies, AnonCAT provides a flexible and scalable solution for clinical text anonymization.

2.4. Risk Assessment and Privacy Evaluation

While significant progress has been made in entity detection and masking, fewer studies have focused on evaluating residual re-identification risk after de-identification. Privacy models such as k-anonymity (Sweeney, 2002b), l-diversity (Machanavajjhala et al., 2007), and t-closeness (Li et al., 2007) provide formal mechanisms for assessing identifiability in structured data.

These methods have been adapted to evaluate privacy risks in clinical datasets (Dankar et al., 2012; Hara et al., 2018). However, their integration into automated de-identification pipelines for unstructured clinical text remains limited.

In this work, we extend existing de-identification frameworks by incorporating document-level risk assessment alongside neural pseudonymization, enabling both sensitive entity detection and quantitative evaluation of residual privacy risk.

3. Methods and Design

3.1. Architecture Overview

The system architecture, as depicted in the diagram (Figure 1), is divided into three major sections: Entity Recognition, Masking, and Risk Assessment. The clinical letters serve as the input to the system, and they are processed through multiple pipelines before final redacted or replaced documents are produced. The steps are as follows: 1) **Data Ingestion**: Clinical letters are fed into the system via the user interface (UI), allowing users to upload documents in bulk for de-identification. 2) **Entity Recognition**: We applied several techniques to identify sensitive information in the clinical letters, such as names, professions, dates, ages, and locations. To accommodate the complexity of clinical data, we integrated multiple approaches: Dictionary look-up leverages predefined dictionaries to detect common sensitive information categories such as names, professions, and locations, ensuring consistent identification of widely known terms across the dataset. Rule-based search handles entities with structured formats, such as dates and ages, using regular expressions to capture variations in formatting. Additionally, a machine learning model based on pre-trained BioBERT and ClinicalBERT is integrated to enhance entity recognition accuracy by capturing contextual information beyond the capabilities of rule-based and dictionary methods. 3) **Union of Entities**: After we have applied the various methods of entity recognition, the system combines the identified entities into a unified list for further processing, ensuring comprehensive entity recognition. This phase incorporates multiple techniques that complement each other. 4) **Masking Strategies**: The next phase involves applying masking strategies. Users can choose between Redaction and Replacement. Redaction involves replacing the identified entities with placeholders (e.g., "XXX-Name"). Replacement, on the other hand, substitutes sensitive information with synthetically generated or random replacements to maintain the structure of the original document. Replacement is more suitable for contexts where document coherence needs to be preserved, such as for clinical research purposes (Neamatullah et al., 2008). The Replacement mapping is stored and later used for reference in the risk assessment stage. 5) **Risk Assessment**: The probability of re-identification of the de-identified data is evaluated, ensuring that the transformation process sufficiently anonymizes sensitive information (El Emam et al., 2008). The integration of these metrics enables a quantitative assessment of the risk associated with the data after de-identification. 6) **Final Output**: Depending on the chosen masking strategy, the system generates either redacted or replaced documents. These documents are returned to the user via the UI, alongside a risk assessment report.

By integrating advanced language models with robust masking strategies and risk assessment techniques, this architecture enables unified pseudonymization and quantitative risk assessment within a single processing pipeline, supporting privacy-aware text release workflows.

3.2. The Risk Assessment Framework

We implement a set of risk metrics to evaluate the robustness of the de-identification process and mitigate the risk of re-identification.

For entity extraction with context, each entity is paired with a window of surrounding words to form a quasi-identifier, simulating realistic re-identification scenarios where attackers may have access to aux-

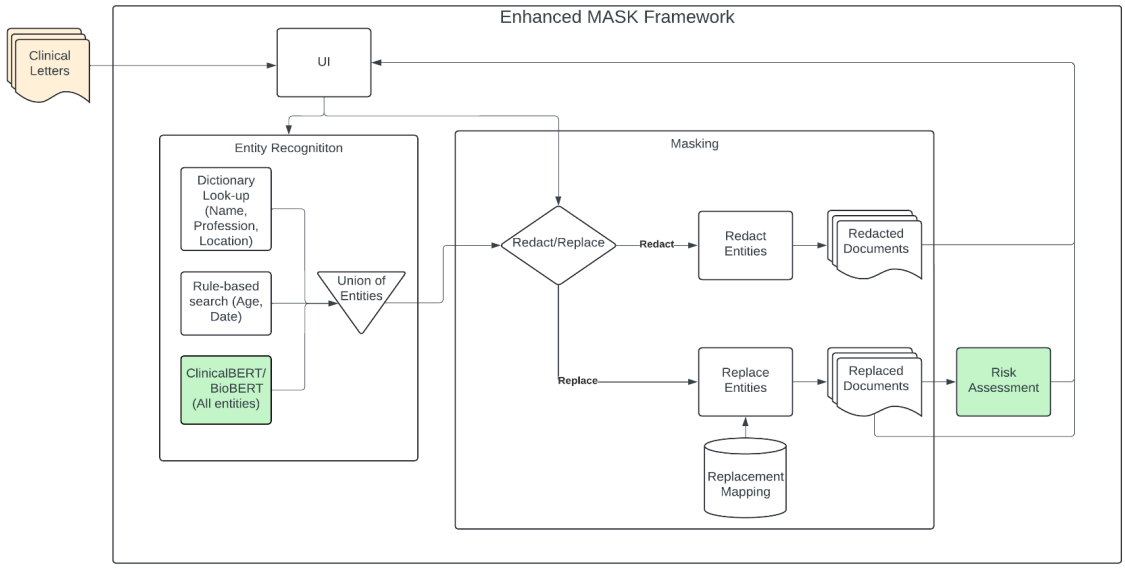


Figure 1: Detailed Architecture of Enhanced MASK framework - DeID-Clinic

iliary information. Text is tokenized using the BertTokenizerFast, and character indices are aligned with token indices to accurately extract contextual spans.

For k-anonymity (Sweeney, 2002b), entities are grouped according to their quasi-identifiers, defined by both entity type and contextual information. Context is represented through high-dimensional contextual embeddings that encode semantic meaning. The number of similar records within each group determines the anonymity level, and the smallest group size defines the overall k-anonymity score.

L-diversity (Machanavajjhala et al., 2007) is computed by measuring the number of distinct sensitive attribute values (e.g., age or profession) within each quasi-identifier group to ensure sufficient diversity.

Unicity (De Montjoye et al., 2013) is measured by counting unique quasi-identifier combinations, where higher uniqueness implies increased re-identification risk.

The quasi-identifier risk likelihood is estimated by assigning each combination a probability inversely proportional to its dataset frequency, and averaging these probabilities to approximate expected re-identification risk (El Emam et al., 2011).

T-closeness (Li et al., 2007) is evaluated by comparing the distribution of sensitive attributes within each group to the global distribution across the dataset using Kullback–Leibler divergence; smaller divergence indicates stronger privacy preservation. Cosine similarity is used to measure contextual similarity between embeddings (Yu et al., 2022). Similar contexts across documents (scores near 1) indicate common entities with lower risk, whereas low similarity (scores near 0) suggests uniqueness and higher potential re-identification risk.

After computing these metrics for both the original and de-identified datasets, a comparative risk assessment is conducted. The framework analyzes entity frequencies and co-occurrences, as combinations of entities increase identifiability. Each co-occurrence is assigned a sensitivity weight, and the document-level risk score (RS) is defined as:

$$\mathrm{RS} = \sum(\mathrm{EF} + \mathrm{CoWeight}) \tag{1}$$

where EF denotes entity frequency and CoWeight denotes co-occurrence weight. The system further counts contexts with cosine similarity below a threshold of 0.5, and computes the proportion of unique contexts within each document relative to all contexts. This proportion is combined with the co-occurrence statistics to obtain the final risk score (FRS):

$$\mathrm{FRS} = \sum(\mathrm{EF}+\mathrm{CoWeight}) \times \left(\frac{\mathrm{Count}}{\mathrm{TotalCount}}\right) \times 100 \tag{2}$$

The resulting score is expressed as a percentage and used to assign documents to three risk categories: low risk (below 25%), moderate risk (25–50%), and high risk (above 50%). Finally, documents are prioritized accordingly, with high-risk documents requiring manual review and stricter de-identification, while low-risk documents require minimal intervention.

Unlike traditional anonymization approaches that focus solely on entity removal, our risk assessment framework explicitly models residual identifiability after pseudonymization. By combining statistical privacy metrics with contextual semantic similarity, the proposed approach provides a practical approx-

imation of real-world re-identification risk, where adversaries may exploit contextual clues rather than isolated identifiers. This enables more informed decisions regarding data release and manual review prioritization.

4. Experimental Evaluation

4.1. Dataset Overview

The dataset used in this work is the i2b2/UTHealth De-identification and Heart Disease Risk Factors dataset, specifically the 2014 PHI Gold Set 1 and 2 (Stubbs and Uzuner, 2014), which is part of the National NLP Clinical Challenges (n2c2) initiative (n2c2 NLP Research Data Sets).[3] This dataset comprises de-identified clinical notes that are extensively annotated for Protected Health Information (PHI) and are intended for evaluating and advancing the performance of de-identification systems in clinical settings. The dataset is sourced from the Research Patient Data Registry (RPDR) at Partners Healthcare and was manually annotated by domain experts. The dataset consists of 790 clinical notes spanning multiple years of patient data. The dataset contains the following de-identification entities: 4,456 names, 7,495 dates, 897 medical identifiers, 2,767 locations, 234 professions, 323 contact details, and 1,424 ages. This annotation process provides a dataset for training and evaluating de-identification models across a diverse range of PHI categories. The dataset is structured to facilitate research in clinical text de-identification, with annotations corresponding to several categories of PHI. The entity categories can be further categorised as direct identifiers (e.g., names and contact information) and quasi-identifiers (e.g., ages, locations). Direct identifiers refer to information that identifies an individual, such as names, contact details, or Social Security numbers. Quasi-identifiers are pieces of information that do not directly identify an individual but can be combined with other data to re-identify someone (Scaiano et al., 2016).

4.2. Model Setup and Finetuning

The BioBERT and ClinicalBERT models are integrated into the MASK framework to identify and classify sensitive entities in clinical text. The models are further fine-tuned using the i2b2 dataset to adapt them for clinical NER tasks.

We implement NER following a common token classification approach, where each token in a sequence is assigned a label. Given the nature of clinical notes, where a single entity may span multiple tokens, the model uses **BIO** tagging (Begin, Inside, Outside tagging) to ensure that multi-token entities are labelled correctly. For the sentence, “John Smith visited the hospital on 12th August 2024.", the BIO tags might be as in Table 1.

Here, the name “John Smith" is recognised as a “NAME" entity, the “hospital" as an “ORG" (Organization) entity, and “12th August 2024" as a “DATE" entity. Each of these entities has appropriate "B" and "I" tags depending on whether the token is at the beginning (B) or inside (I) of the entity.

The finetuning of Bio/ClinicalBERT involves the following key steps: I) Data Preprocessing: 1) The input text is split into sentences using the `sent_tokenize` function from NLTK, ensuring that the model processes manageable text chunks. 2) Each sentence is then tokenized using the BioBERT and ClinicalBERT tokenizer, which handles subword tokens. This is crucial for preserving the granularity of clinical entities. 3) The tokenized sentences are padded to a maximum length of 75 tokens, ensuring uniformity in batch processing.

II) Training Setup: 1) The model is set up to utilise a GPU (T4) with CUDA when available; otherwise, it will default to running on the CPU. 2) BioBERT and ClinicalBERT are finetuned on the i2b2 2014 dataset, specifically focusing on the NER task. 3) The training process uses the AdamW optimizer with a learning rate of 3×10^{-5} and weight decay to prevent overfitting. This value was found to be small enough to ensure stable convergence during training while allowing the model to learn efficiently from the dataset. 4) The model was trained using a batch size of 4 to optimize memory utilisation on GPU hardware. The training process was carried out over 20 epochs, with loss computed at each step.

III) Learning and Evaluation: 1) During the learning process, the model’s parameters were updated iteratively based on the cross-entropy loss between predicted and true labels. 2) After each epoch, the model’s performance was evaluated using validation data, and loss curves were plotted to monitor overfitting or underfitting tendencies. 3) The primary evaluation metrics used were Precision, Recall, F1-score, and Accuracy.

4.3. BioBERT and ClinicalBERT Results

Based on the comparisons in Table 2 of BioBERT and ClinicalBERT in each entity category, we summarise the results as follows. 1), BioBERT wins more precision than ClinicalBERT, e.g. on DATE, AGE, LOCATION, CONTACT. 2), ClinicalBERT wins more recall than BioBERT, e.g. on NAME, DATE, ID, LOCATION, CONTACT, and PROFESSION, except for the only entity category AGE. 3), BioBERT wins four entities on F1, i.e. NAME, AGE, LOCATION, and PROFESSION. 4), ClinicalBERT wins the rest three entity types, i.e. DATE, ID, and CONTACT. 5), in average across all entities,

[3] https://n2c2.dbmi.hms.harvard.edu

Token	John	Smith	visited	the	hospital	on	12th	August	2024
Tag	B-NAME	I-NAME	O	O	B-ORG	O	B-DATE	I-DATE	I-DATE

Table 1: BIO Tagging

Entity	**BioBERT**			**ClinicalBERT**			**items**
	P	**R**	**F1**	**P**	**R**	**F1**	
NAME	0.963	0.960	**0.985**	**0.970**	**0.970**	0.970	622
DATE	**0.974**	0.974	0.974	0.960	**0.982**	**0.971**	655
ID	0.948	0.973	0.988	**0.986**	**0.993**	**0.989**	75
AGE	**0.956**	**0.978**	**0.967**	0.930	0.974	0.951	89
LOCATION	**0.933**	0.903	**0.928**	0.919	**0.922**	0.921	278
CONTACT	**0.955**	0.928	0.941	0.947	**0.960**	**0.954**	69
PROFESSION	0.810	0.630	**0.748**	**0.844**	**0.643**	0.730	27
Micro Avg	**0.959**	**0.964**	**0.965**	0.955	0.963	0.959	1816
Macro Avg	0.817	0.793	0.804	**0.820**	**0.805**	**0.811**	1816
Weighted Avg	**0.958**	**0.961**	**0.964**	0.953	0.963	0.958	1816

Table 2: Evaluations of BioBERT and ClinicalBERT models with higher scores in bold

BioBERT wins micro avg scores and weighted avg scores, while ClinicalBERT wins macro avg scores on P/R/F1.

The F1 scores across all entity types mostly fall between 0.92 (LOCATION) and 0.99 (ID), except for PROFESSION who has the lowest F1 scores 0.748 (BioBERT) and 0.730 (ClinicalBERT). This lower performance suggests that professions are more **ambiguous** and context-dependent, making them harder to identify in comparison to other entities such as IDs or Names; this is a known issue (Uzuner et al., 2007; Dernoncourt et al., 2017).

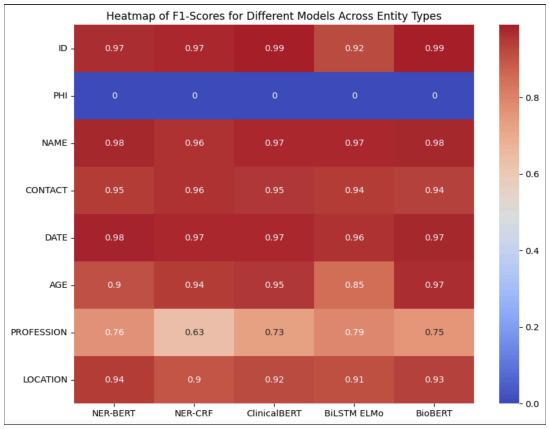


Figure 2: Heat-Map of All Evaluated Models

4.4. Baseline Results

We also list the comparisons on BERT, BiLSTM, and CRF in Table 3, where it shows that the BERT model wins the most Precision scores on 4 entities, versus BiLSTM (1) and CRF (2). In comparison, the CRF model wins the most Recall scores on 4 entities (NAME, ID, LOCATION, PROFESSION), versus BERT (3 including 1 tie) and BiLSTM (1), which indicates that CRFs produce more false positives for the sake of true positives. This is especially true for the PROFESSION entity type, where CRFs give the lowest precision score of 0.470. In contrast, the BERT model has much higher Precision than Recall (0.907 vs 0.680), indicating that it sacrifices potential true outputs by restricting false positives. Interestingly, the BiLSTM model has a more balanced P/R on the PROFESSION category (0.81 vs 0.78). Looking at both Table 2 and 3, we can see that the domain adapted models BioBERT and ClinicalBERT have improved the performance on entity types **NAME**, **ID**, and **AGE** in comparison to BERT model from (0.979, 0.962, 0.893) to (0.985, 0.989, 0.967) on F1 scores.

4.5. Heat-maps from All Models

Figure 2 presents the heat-map comparative results of F1-scores of MASK-BioBERT/ClinicalBERT and other MASK NER models. The results indicate that Biomed-Clinical BERT models consistently perform strongly or match the performance of other models in key entity categories, demonstrating their effectiveness for clinical data de-identification. Notably, MASK-BioBERT demonstrated very high performance for **structured entity types** like IDs, Dates, and Names, where it consistently achieved higher or equal F1-scores compared to ClinicalBERT, NER-BERT, and NER-CRF, all likely benefiting from domain-specific pre-training.

However, while MASK-BioBERT excels in structured entity recognition, its lower performance on context-dependent entities like **Profession** (0.75 F1-score) highlights its limitations in handling ambiguity. This is an area where even ClinicalBERT, which focuses on clinical texts, struggles.

Entity	BERT			BiLSTM			CRF		
	P	R	F1	P	R	F1	P	R	F1
NAME	**0.979**	**0.980**	**0.979**	0.980	0.960	0.970	0.940	**0.980**	0.960
DATE	0.965	**0.988**	**0.977**	0.960	0.970	0.960	**0.960**	0.980	0.970
ID	0.940	0.986	0.962	**0.980**	0.860	0.920	0.950	**0.990**	**0.970**
AGE	0.878	0.908	0.893	0.740	**0.990**	0.850	**0.910**	0.980	**0.940**
LOCATION	**0.943**	0.937	**0.940**	0.890	0.940	0.910	0.850	**0.970**	0.900
CONTACT	**0.953**	**0.994**	**0.973**	0.950	0.930	0.940	0.930	0.990	0.960
PROFESSION	**0.907**	0.680	**0.777**	0.810	0.780	0.790	0.470	**0.930**	0.630
Micro Avg	**0.962**	0.972	**0.967**	0.940	0.950	0.950	0.920	**0.980**	0.950
Macro Avg	**0.821**	0.809	**0.813**	0.790	**0.800**	0.790	0.750	**0.850**	0.790
Weighted Avg	**0.960**	0.972	**0.966**	0.950	0.950	0.950	0.930	**0.980**	0.950

Table 3: Comparison of evaluation metrics for BERT, BiLSTM, and CRF models.

Metric	M-BioBERT	M-ClinicalBERT
TP	171	181
FP	10	12
FN	2	0
Precision	**0.9448**	0.9378
Recall	0.9948	**1.0000**
F1	0.9648	**0.9679**

Table 4: NER evaluation comparisons.

Correct	`<LOCATION id=P13 start=967 end=977 text=Clarkfield TYPE=HOSPITAL>`
False Positives	`2071(1576-1580, Date), US(1602-1604, Location), 2071(1745-1749, Date), Thiel(4026-4031, Name), 4(2138-2139, Age), Clark-field(1317-1327, Name), Thiel(5890-5895, Name)`

Table 5: Sample false positives of MASK-BioBERT.

4.6. Qualitative Case Study

To illustrate the practical behavior of the proposed framework in realistic deployment scenarios, we conducted a qualitative case study on five randomly selected clinical documents from the i2b2 dataset. This case study complements the quantitative evaluation presented earlier by providing insight into system behavior at the document level, including entity detection accuracy and masking effectiveness. These documents originally contained a total of 181 sensitive entities. From the annotations by two fluent English speakers (MSc graduates), the system using MASK-BioBERT and MASK-ClinicalBERT identified a total of 183 and 193 entities, respectively, in Table 4 (*left* and *right*), with the key metrics observed from this test.

Here's a detailed analysis of the metrics observed: 1) On Precision: The systems achieved the precision of 0.944 and 0.937, indicating a high level of accuracy in identifying entities. These values suggest that the majority of identified entities were correct, with only a small number of false positives identified when running the MASK-Bio/ClinicalBERT. 2) On Recall: A recall score of 0.9948 and 1 reflects the system's ability to correctly identify nearly all relevant entities present in the dataset. Out of all potential entities, only **2 false negatives** for MASK-BioBERT were missed, indicating a highly efficient model in terms of capturing the intended entities. In addition, MASK-ClinicalBERT had 0 false negatives on this task. 3) On F1 Score: The F1 score, calculated as the harmonic mean of precision and recall, stood at 0.964 and 0.967 for the two models. The robust F1 score illustrates that the model strikes an effective balance between precision and recall, providing reliable and consistent performance across different entity types. The **10 and 12 false positives** from two systems indicate some over-identification by the models. These might arise from the model's sensitivity in identifying entities, where certain words are incorrectly flagged as entities, likely due to ambiguity in the context or overlaps in entity types. As seen in Table 5 'Clarkfield' is identified as a name, when in reality it's actually a location.

Despite these challenges, the system's high precision, recall, and F1 score suggest that it performs reliably in recognising sensitive information in clinical documents. These metrics highlight the system's potential to be a strong candidate for real-world applications in medical entity de-identification.

The masking process, both *redaction* and *replacement*, was successfully implemented. In redaction mode, sensitive entities such as names, dates, and ages were replaced with their respective entity type placeholder (e.g., "XXX-NAME", "XXX-DATE"). In replacement mode, realistic replacements were used for names and temporal entities from the list of full names and surnames extracted from the i2b2 2014 dataset, ensuring that the structure of the clinical document was maintained while still protecting the patient's identity.

Replacement output
Oakley→Jones; 2065→2063; 3/67→01/65; 2068-12-05→2066-09-09; 37→34; 66→62

Table 6: Example replacements produced.

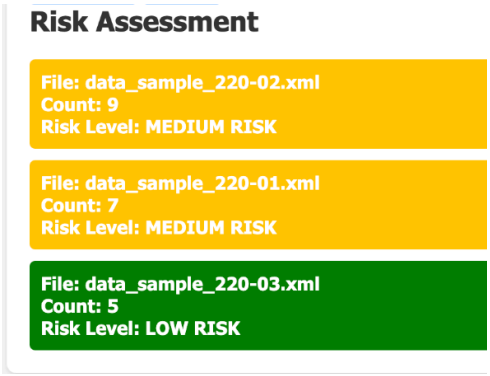


Figure 3: Risk Assessment Results for Batch upload of documents

4.7. Risk Assessment Results

Figure 3 risk assessment visualisation underscores the system's proficiency in identifying documents that still pose re-identification risks and provides actionable insights for further mitigating potential vulnerabilities in sensitive data. The risk assessment results in Figure 3 effectively categorise documents based on the risk of re-identification. As shown, the system assigns each document a risk level — **high** (red), **medium** (yellow), or **low** (green) — depending on how many instances of unique entity contexts were found, referring to Section 3.2 (Risk Assessment).

For example, `data_sample_220-02.xml` is marked with Medium Risk, having 9 unique contexts, while `data_sample_220-01.xml` similarly displays Medium Risk with 7 occurrences. These files flagged as medium risk suggest that while some sensitive information has been de-identified, there is still a non-negligible possibility of re-identification due to the unique contexts of certain entities.

In contrast, `data_sample_220-03.xml` is classified as Low Risk with only 5 instances of unique contexts, suggesting that most of the entities in this document share common contexts across the dataset, thus significantly lowering the chances of re-identification.

4.8. Error Analysis and Limitations

Through manual inspection of the model outputs, we identified some primary sources of errors. First, *boundary* errors occurred when the model slightly misidentified the start or end of an entity, a common issue in NER tasks, particularly for names that include prefixes or titles (e.g., "Dr. Oakley" in Figure 7, Appendix). Second, the model produced *false positives* for *dates* and *ages* by misclassifying numerical values unrelated to temporal information (e.g., medical measurements such as '2071' in Table 5), which negatively affected precision. Third, *frequent terms* were occasionally over-masked as sensitive entities; for example, "US" was sometimes labeled as a location (Table 5), reducing precision by masking non-sensitive content. Forth, overlapping entities and patterns introduced ambiguity, as the same text segment could be detected as multiple entity types by different methods (e.g., a place name misclassified as a person name). Fifth, the rule-based component occasionally fragmented multi-word entities into separate tokens, particularly for dates, where a single expression was split and treated as multiple independent entities.

From the experimental investigation outcomes of our work, several limitations and areas for improvement remain. First, the model was optimized for a single dataset due to the paucity of readily available data, resulting in dataset-specific performance and limited generalization to diverse clinical settings or real-world hospital deployments. Second, due to limitations of computational facilities, we only tested on domain-specific BERT models for integration into the Mask framework, without using LLMs. [4]

5. Conclusions and Future Work

This work presents DeID-Clinic, a risk-aware pseudonymization framework for privacy-preserving processing of clinical free-text data. By integrating domain-adapted transformer models with document-level privacy risk assessment, the proposed system extends traditional de-identification pipelines beyond entity masking toward quantitative privacy evaluation.

Experimental results on the i2b2 dataset demonstrate strong performance in sensitive entity detection, while the proposed risk scoring framework enables identification of documents with elevated re-identification risk. This capability is particularly important in real-world privacy-sensitive applications, where automated de-identification alone may not fully eliminate privacy threats. More broadly, this work highlights the importance of combining neural language models with explicit privacy risk modeling to support responsible data sharing. While evaluated on clinical data, the proposed framework is applicable to other privacy-critical domains, including legal and administrative text, aligning with emerging requirements for privacy-preserving language technologies.

[4] Recent work using LLMs for biomedical NER includes (Mazzucato et al., 2026) for Dutch and Italian languages (preprint).

Future work will focus on extending the proposed framework in several directions: 1) evaluate the risk assessment module across multiple datasets and domains to further validate its effectiveness in estimating re-identification risk; 2) integrate newer LLMs, which may improve performance on context-dependent entity types such as professions and organizations; 3) explore adaptive risk thresholds and user-configurable privacy settings to support more flexible deployment in real-world privacy-sensitive environments.

6. Ethical Statement

The data we used for this work is already de-identified and anonymized by the shared task organisers who released the data for research purposes only. We did not use any third party commercial platforms to disclose the data.

7. Acknowledgement

We are grateful to the reviewers for valuable comments. Funded by the European Union under Horizon Europe Work Programme 101057332, views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. The UK team are funded under the Innovate UK Horizon Europe Guarantee Programme, UKRI Reference Number: 10041120. WDP and GN are grateful for the support from the grant "Assembling the Data Jigsaw: Powering Robust Research on the Causes, Determinants and Outcomes of MSK Disease", and the grant "Integrating hospital outpatient letters into the healthcare data space" (EP/V047949/1; funder: UKRI/EPSRC).

8. Bibliographical References

Emily Alsentzer, John R Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Fida K Dankar, Khaled El Emam, Angelica Neisa, and Tyson Roffey. 2012. Estimating the re-identification risk of clinical data sets. *BMC Medical Informatics and Decision Making*, 12(1):1–13.

Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. 2013. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3(1):1376.

Franck Dernoncourt, Ji Young Lee, Özlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Cynthia Dwork. 2006. Differential privacy. In *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP)*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer-Verlag.

Peter Edemekong, Pavan Annamaraju, Muriam Afzal, and Micelle Haydel. 2024. Health insurance portability and accountability act (hipaa) compliance. *StatPearls*.

Khaled El Emam, Fida K Dankar, Romain Issa, Elizabeth Jonker, Daniel Amyot, Elise Cogo, Jean-Pierre Corriveau, Mark Walker, Sadrul Chowdhury, Regis Vaillancourt, Tyson Roffey, and Jim Bottomley. 2006. A globally optimal k-anonymity method for the de-identification of health data. *Journal of the American Medical Informatics Association*, 13(5):556–569.

Khaled El Emam, Fida K Dankar, Regis Vaillancourt, Tyson Roffey, and Mark Lysyk. 2008. Evaluating the risk of re-identification of patients from hospital prescription records. *Canadian Journal of Hospital Pharmacy*, 61(3):191–198.

Khaled El Emam, Elizabeth Jonker, Luk Arbuckle, and Bradley Malin. 2011. A systematic review of re-identification attacks on health data. *PloS one*, 6(12):e28071.

F Jeff Friedlin and Clement J McDonald. 2008. A software tool for removing patient identifying information from clinical documents. *Journal of the American Medical Informatics Association*, 15(5):601–610.

Kazuma Hara, Takuya Matsuzaki, and Yusuke Miyao. 2018. Risk analysis of de-identification methods for anonymizing biomedical text data. *Journal of Biomedical Informatics*, 84:136–146.

Tal Hartman, Michael D Howell, Jeffrey Dean, Shahar Hoory, Ronit Slyper, Irit Laish, Oren Gilon, and Yossi Matias. 2020. Customization scenarios for de-identification of clinical notes. *BMC Medical Informatics and Decision Making*, 20(1):1–9.

Youngjun Kim, Patricia Heider, and Stéphane Meystre. 2018. Ensemble-based methods to improve de-identification of electronic health record narratives. In *AMIA Annual Symposium Proceedings*, pages 663–672.

Aleksandar Kovačević, Bojana Bašaragin, Nikola Milošević, and Goran Nenadić. 2024. De-identification of clinical free text using natural language processing: A systematic review of current approaches. *Artificial Intelligence in Medicine*, 151:102845.

Zeljko Kraljevic, Anthony Shek, Joshua Au-Yeung, Ewart Sheldon, Mohammad Al-Agil, Haris Shuaib, Bai Xi, Kawsar Noor, Anoop Shah, Richard Dobson, and James Teo. 2023. Deploying transformers for redaction of text from electronic health records in real world healthcare.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. 2007. T-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115. IEEE.

Jennifer Lin. 2020. A comparison of recurrent neural networks and conditional random fields for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2324–2335.

Zengjian Liu, Buzhou Tang, Xiaolong Wang, and Qingcai Chen. 2017. De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of Biomedical Informatics*, 75:S34–S42.

Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3–es.

Sara Mazzucato, Tom M Seinen, Sara Moccia, Silvestro Micera, Andrea Bandini, and Erik M van Mulligen. 2026. Advancements in multilingual biomedical natural language processing: exploring large language models for named entity recognition and linking. *medRxiv*, pages 2026–01.

Stcaféphane M Meystre, Óscar Ferrcafández, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. 2014. Text de-identification for privacy protection: A study of its impact on clinical text information content. *Journal of Biomedical Informatics*, 50:142–150.

Stcaféphane M Meystre, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. 2010. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Medical Research Methodology*, 10(1):1–16.

Nikola Milosevic, Gangamma Kalappa, Hesam Dadafarin, Mahmoud Azimaee, and Goran Nenadic. 2020. Mask: A flexible framework to facilitate de-identification of clinical texts. *arXiv preprint arXiv:2005.11687*.

NCA NHS Foundation Trust. 2021. Mask api. https://github.com/NCA-NHS-Foundation-Trust/MASK_API_Copy/tree/master. Accessed: 2024-10-15.

Ishna Neamatullah, Margaret M Douglass, Li-wei H Lehman, Andrew T Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George B Moody, Roger G Mark, and Gari D Clifford. 2008. Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, 8(1):32.

Atiquer Rahman Sarkar, Yao-Shun Chuang, Noman Mohammed, and Xiaoqian Jiang. 2024. De-identification is not enough: a comparison between de-identified and synthetic clinical notes. *Scientific reports*, 14(1):29669.

Martin Scaiano, Grant Middleton, Luk Arbuckle, Varada Kolhatkar, Liam Peyton, Moira Dowling, Debbie S Gipson, and Khaled El Emam. 2016. A unified framework for evaluating the risk of re-identification of text de-identification tools. *Journal of biomedical informatics*, 63:174–183.

Dhivin Shaji, Angel Paul, Lifeng Han, Warren Del-Pinto, Goran Nenadic, and Suzan Verberne. 2025. De-identifying clinical texts using biomed-clinical berts and comprehensive risk assessment. In *2025 IEEE 13th International Conference on Healthcare Informatics (ICHI)*, pages 683–684. IEEE.

Louis Philippe Sondeck and Maryline Laurent. 2025. Practical and ready-to-use methodology to assess the re-identification risk in anonymized datasets. *Scientific Reports*, 15(1):23223.

Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. 2015. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1. *Journal of Biomedical Informatics*, 58:S11–S19.

Amber Stubbs and Özlem Uzuner. 2014. Annotation guidelines for de-identification of medical records. Version 1.0, i2b2/UTHealth.

Hemang Subramanian, Arijit Sengupta, and Yilin Xu. 2024. Patient health record protection beyond the health insurance portability and accountability act: mixed methods study. *Journal of Medical Internet Research*, 26:e59674.

Latanya Sweeney. 2002a. Achieving k-anonymity privacy protection using generalisation and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):571–588.

Latanya Sweeney. 2002b. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570.

Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 14(5):550–563.

Taher Vakili and Hercules Dalianis. 2022. Utility preservation of clinical text after de-identification. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 383–388, Dublin, Ireland.

Paul Voigt and Axel Von dem Bussche. 2017. *The EU General Data Protection Regulation (GDPR). A Practical Guide*, 1 edition. Springer International Publishing.

Xiaofeng Yang, Tian Lyu, Qing Li, Chih-Yin Lee, Jiang Bian, William R Hogan, and Yonghui Wu. 2019. A study of deep learning methods for de-identification of clinical notes in cross-institute settings. *BMC Medical Informatics and Decision Making*, 19(S5).

Wenguang Yu, Yu Weng, Ronghua Lin, and Yong Tang. 2022. Cosbert: A cosine-based siamese bert-networks using for semantic textual similarity. In *CCF Conference on Computer Supported Cooperative Work and Social Computing*, pages 376–389. Springer.

9. Appendix

Data Statistics and Training

Figure 4 shows the entity occurrence distribution in the i2b2 dataset. Figure 5 is an example of how loss evolved during training, which is indicative of the model's learning trajectory.

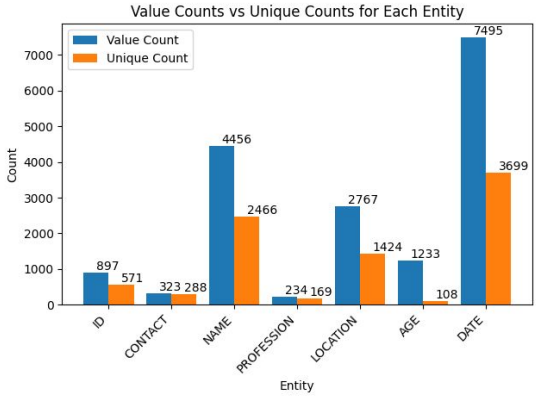


Figure 4: Entity occurrence distribution in the i2b2 dataset, showing value counts and unique counts for each entity type

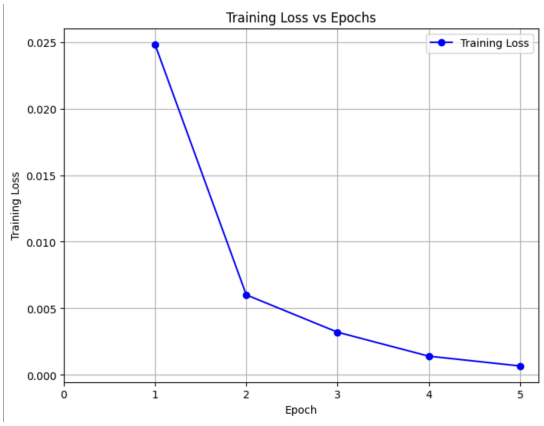


Figure 5: Training curve visualisation

Tools and Technologies in Detail

Some technical details are listed below:

- This work is built on the MASK API developed by the Northern Care Alliance NHS Foundation Trust, available publicly on GitHub (Milosevic et al., 2020; NCA NHS Foundation Trust, 2021).
- Risk Assessment Metrics: Libraries such as Scikit-learn and SciPy were used for calculating re-identification risk metrics (Dwork, 2006).
- Google Colab provides the necessary T4-GPU resources for fine-tuning.

Platform Interface: Human-in-the-loop

The de-identification platform, as demonstrated in Figure 6, can support both single file and multiple files processing. The De-identificaiton procedure is:

- Load Finetuned (saved) models, e.g. MASK-BioBERT/ClinicalBERT
- Run De-identification using the loaded model
- Mark/Remove Entities Option, human-in-the-loop, editable results
- Store/Download final output

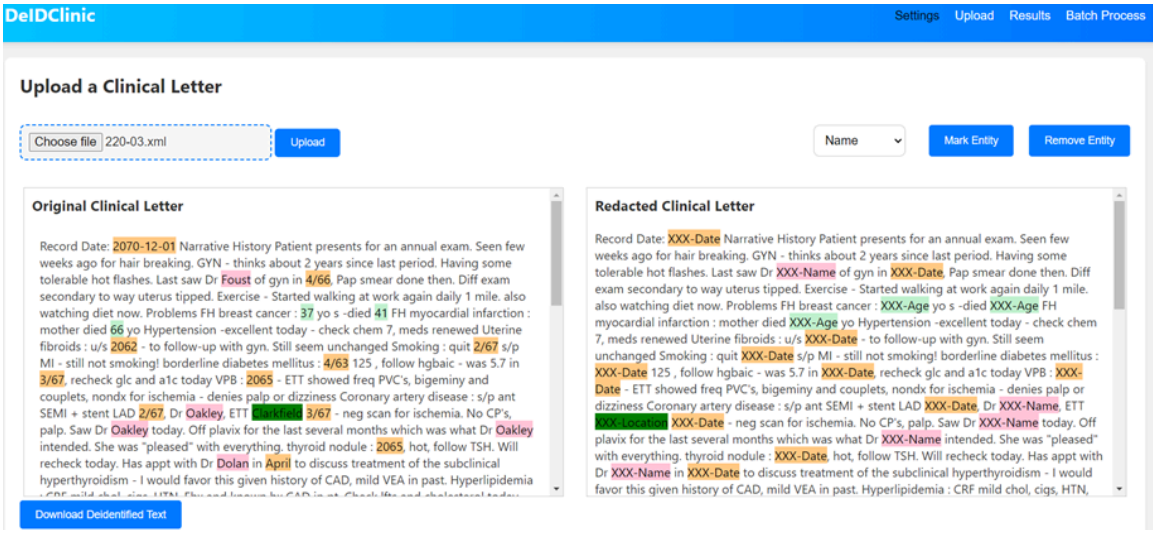


Figure 6: **Interface Demo** with De-identification output using uploaded letter (**cancer** domain text)

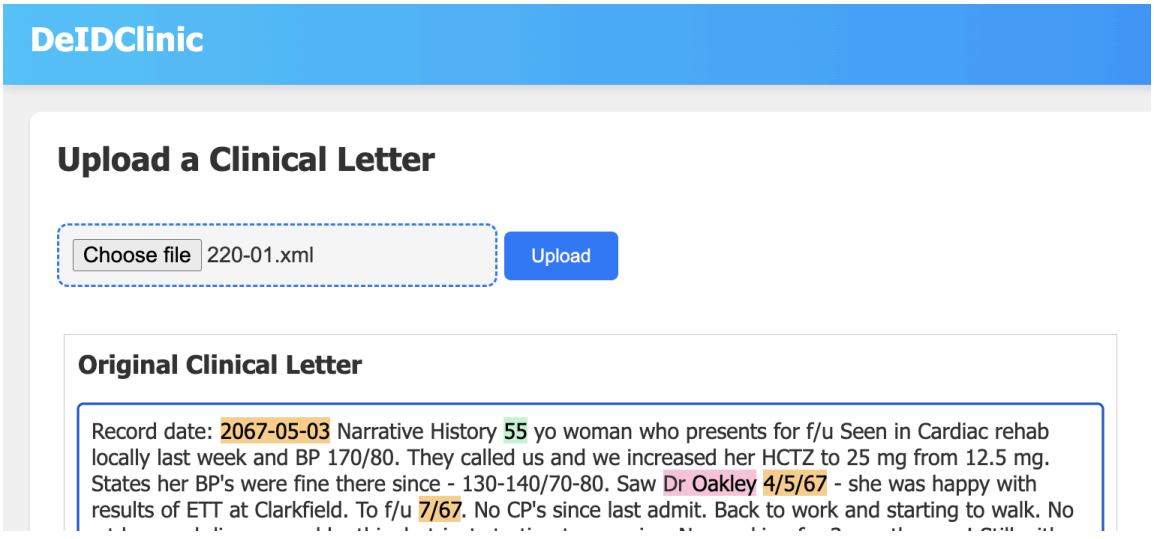


Figure 7: Boundary Error Example in MASK-BioBERT (**cardiac** domain text)

Distilling Human-Aligned Privacy Sensitivity Assessment from Large Language Models

Gabriel Loiseau[1,2], **Damien Sileo**[2], **Damien Riquet**[1], **Maxime Meyer**[1], **Marc Tommasi**[2]
[1]Hornetsecurity, Hem, France
[2]Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France
gabriel.loiseau@inria.fr

Abstract

Accurate privacy evaluation of textual data remains a critical challenge in privacy-preserving natural language processing. Recent work has shown that large language models (LLMs) can serve as reliable privacy evaluators, achieving strong agreement with human judgments; however, their computational cost and impracticality for processing sensitive data at scale limit real-world deployment. We address this gap by distilling the privacy assessment capabilities of Mistral Large 3 (675B) into lightweight encoder models with as few as 150M parameters. Leveraging a large-scale dataset of privacy-annotated texts spanning 10 diverse domains, we train efficient classifiers that preserve strong agreement with human annotations while dramatically reducing computational requirements. We validate our approach on human-annotated test data and demonstrate its practical utility as an evaluation metric for de-identification systems.

Keywords: privacy evaluation, knowledge distillation, de-identification

1. Introduction

Quantifying privacy in textual data remains an open challenge due to the absence of a unified definition and the inherently contextual nature of privacy (Bambauer et al., 2022; Tesfay et al., 2016). Formal frameworks such as differential privacy (Dwork, 2006) provide rigorous guarantees, and proxy-based evaluation through attack success rates or information-theoretic measures are well-established in practice (Ren et al., 2025). However, these approaches capture specific, well-defined threat models rather than the broader, human-perceived notion of what constitutes sensitive content. Large language models (LLMs), with their capacity for nuanced language understanding, have emerged as promising candidates for human-aligned evaluation, demonstrating strong agreement with human judgments across a variety of NLP tasks (Zheng et al., 2023; Li et al., 2024).

Recent work by Meisenbacher et al. (2025) represents a significant step toward closing this gap in privacy evaluation by applying the *LLM-as-a-Judge* paradigm to this domain. Across 10 datasets and 677 human annotators, they show that LLMs can approximate a "global human privacy perspective" with strong agreement to aggregated human ratings, even exceeding inter-human agreement. These findings suggest that LLMs can serve as practical, human-aligned privacy evaluators.

Yet, deploying frontier LLMs for privacy assessment poses two central challenges. First, their computational and financial costs limit large-scale use. Second, evaluating sensitive text through third-party APIs introduces additional privacy concerns, as the very data being assessed may not be shareable. This creates a paradox: using powerful external LLMs to evaluate privacy may itself compromise privacy constraints.

In this work, we address this deployment gap through *knowledge distillation*. Using Mistral Large 3 (Mistral AI, 2025) as a teacher model, we annotate 200,000 user-written texts with privacy sensitivity scores following the structured Likert-scale methodology of Meisenbacher et al. (2025). We then distill these judgments into lightweight encoder-based classifiers, enabling fast, local, and privacy-preserving inference. Our central research question is whether the privacy reasoning capabilities of LLMs can be transferred to smaller models without sacrificing alignment with human judgments.

We validate the distilled models on human-annotated test data and show that they can match the agreement of their teacher model with aggregated human ratings. Beyond benchmark validation, we demonstrate that distilled privacy evaluators can serve as scalable automatic metrics for quantifying privacy reduction in text de-identification systems[1]. Our contributions are threefold:

1. We curate a large corpus of 200,000 texts, automatically annotated for privacy sensitivity using a state-of-the-art open LLM.
2. We distill these LLM-generated privacy judgments into lightweight encoder models that achieve strong agreement with human anno-

[1]Models, code, and data are available at https://github.com/gabrielloiseau/privacy-distillation

Proceedings of LEGAL2026 and CALD-pseudo 2026 @ LREC 2026, pages 53–61
12 May 2026. ©ELRA Language Resources Association (ELRA), 2026

Rating	Name	Description	Dataset (%)
1	Harmless	Completely free of any private or sensitive information, including direct or indirect identifiers.	46.01
2	Mostly not private	May contain some indirect identifiers, but is largely free of sensitive or personal information.	16.58
3	Somewhat private	Contains some direct or indirect identifiers and can be considered moderately personal.	16.81
4	Very private	Contains several direct or indirect identifiers and clearly includes personal information.	14.21
5	Extremely private	Contains highly sensitive personal information or direct identifiers.	6.38

Table 1: Privacy rating annotation scheme and resulting dataset distribution

tations ($\alpha = 0.737$), surpassing the teacher model's own human alignment ($\alpha = 0.716$), while enabling efficient and fully local inference.

3. We demonstrate that distilled privacy evaluators function as scalable automatic metrics for assessing privacy reduction in text de-identification systems, and outline how compact privacy models open new research directions for privacy-aware NLP evaluation and system design.

2. Related Work

Privacy Evaluation in NLP. In privacy-preserving NLP, evaluation commonly relies on proxy metrics such as re-identification success rates, simulated attacks, plausible deniability, or semantic similarity measures (Shahriar et al., 2025). While these metrics capture specific threat models, they do not directly reflect how humans perceive the sensitivity of a text. Complementary research on automated privacy policy analysis (Wilson et al., 2016) and anonymization benchmarks (Lison et al., 2021; Pilán et al., 2022; Loiseau et al., 2025) provides structured evaluation frameworks, primarily focusing on entity-level redaction quality. However, these approaches do not measure text-level privacy sensitivity or its alignment with human judgment across domains. Recent work has proposed *LLM-as-a-Judge* as a scalable alternative to human evaluation for many NLP tasks (Li et al., 2024; Chiang and Lee, 2023; Bavaresco et al., 2025; Li et al., 2025), with potential for modeling perceived privacy risk (Meisenbacher et al., 2025).

LLM Distillation. Knowledge distillation (Hinton et al., 2015) transfers capabilities from large teacher models to smaller student models. In NLP, it has produced efficient transformer variants such as DistilBERT (Sanh et al., 2019) and compressed generative LLMs into lightweight classifiers. Distillation can also rely solely on predicted labels, enabling black-box knowledge transfer when logits are unavailable (Chen et al., 2024).

3. Methodology

3.1. Privacy Annotation Framework

We adopt the five-point Likert privacy sensitivity scale introduced by Meisenbacher et al. (2025) and detailed in Table 1, ranging from 1 (*Harmless*) to 5 (*Extremely private*). The scale operationalizes text-level privacy sensitivity by considering both direct identifiers (e.g., names, contact details) and broader contextual signals, including topical sensitivity (e.g., health conditions, legal situations), self-disclosure of personal experiences, and indirect identifiers that could enable re-identification in context. Sensitivity under this scale is therefore not limited to the presence of named entities or demographic attributes, but also encompasses the overall nature and intimacy of the disclosed content. The scale was previously validated through a large-scale human annotation study, and the resulting survey data were publicly released, providing a human-aligned target for supervision.

3.2. LLM Annotation and Distillation Pipeline

Data. We construct a corpus from the 10 publicly available datasets of user-written text from the original study spanning diverse domains: Blog Authorship Corpus (BAC), Enron Emails (EE), Medical Questions (MQ), Reddit Confessions (RC), Reddit Legal Advice (RLA), Mental Health Blog (MHB), Reddit Mental Health Posts (RMHP), Trustpilot Reviews (TR), Twitter (TW), and Yelp Reviews (YR). We sample 20,000 texts per dataset, excluding those used in the original human benchmark, resulting in approximately 200,000 texts. All data is in English; extending the approach to other languages remains future work. Additional details about each dataset are reported in Appendix A.

Teacher Model Annotation. We use the open-weight Mistral Large 3 (Mistral AI, 2025) as a teacher model to assign privacy sensitivity scores. We employ the structured prompting strategy of Meisenbacher et al. (2025), which provides explicit scale definitions and enforces discrete ratings. The

Dataset	Avg tokens	$\bar{S}$	% Priv.
MHB	272	3.31	77.4
RMHP	225	2.92	64.7
RC	306	2.89	60.6
RLA	283	2.84	60.3
MQ	136	2.34	44.6
EE	332	2.16	33.5
BAC	264	1.78	23.5
TR	69	1.28	5.2
YR	129	1.21	2.6
TW	50	1.11	1.7
All	**207**	**2.18**	**37.4**

Table 2: Per-dataset statistics, sorted by mean privacy score. Avg tokens: mean BPE token count. $\bar{S}$: mean teacher-assigned privacy score (1–5). % Priv.: percentage of texts rated ≥ 3 (Somewhat private or above).

full annotation prompt is provided in the Appendix B. This yields a large, automatically labeled dataset reflecting LLM-based privacy judgments.

Table 1 shows the target rating distribution of the resulting dataset. The class distribution is notably imbalanced: nearly half of the texts (46%) are rated as harmless, while the most sensitive category accounts for only about 6% of samples, reflecting the natural scarcity of highly private content in everyday online communication.

Table 2 provides a per-dataset breakdown, revealing variation in both text length and privacy sensitivity across domains. Health and confession-oriented domains (MHB, RC, RMHP, RLA) contain the highest proportions of private content, driven by self-disclosure of personal experiences, medical conditions, and sensitive life events. In contrast, review and microblog platforms (TR, TW, YR) are overwhelmingly rated as harmless (less than 6% rated somewhat private or above), consistent with their public-facing, non-personal communication norms. Intermediate domains such as emails (EE) and blog posts (BAC) reflect a mixture, where privacy signals arise from incidental identifiers (names, contact details) rather than topical sensitivity. This diversity is essential for training a privacy evaluator that generalizes across the contextual factors that shape perceived sensitivity. Table 3 provides examples illustrating each rating level across different domains.

Student Models. We distill these annotations into encoder-based classifiers trained for 5-class classification. We evaluate four models: Ettin-150M, Ettin-17M (Weller et al., 2025), BERT-base (Devlin et al., 2019), and ModernBERT-base (Warner et al., 2024). All models are fine-tuned using the same training recipe: learning rate 2×10^{-5} with 10% linear warmup, batch size 16, and 3 epochs. Due to its large size, the dataset is split into 90% training, 5% validation, and 5% test sets. We select the best checkpoint by validation macro F1.

4. Experiments

We evaluate whether distilled encoder models can (1) learn the LLM-defined privacy task, and (2) preserve alignment with human privacy judgments. Our evaluation therefore combines standard classification metrics on our held-out test set with agreement-based analysis on the publicly released human benchmark from Meisenbacher et al. (2025). We quantify agreement using Krippendorff's α (Krippendorff, 2011), an inter-rater reliability coefficient defined as $\alpha = 1 - \frac{D_o}{D_e}$, where D_o denotes the observed disagreement and D_e the disagreement expected by chance; $\alpha = 1$ indicates perfect agreement and $\alpha = 0$ corresponds to chance-level agreement. We report two complementary agreement scores: agreement with the average human rating per text, and the average pairwise agreement with all individual annotations, for which we also report the standard deviation across annotators.

4.1. Learning the Distilled Task

We first assess how well models learn the 5-class privacy classification task on our held-out test set. Table 4 reports accuracy, macro F1, mean absolute error[2], and per-class F1 scores.

The Ettin-150M model achieves 74.9% accuracy and 68.1 macro F1, substantially outperforming majority (45.9%) and random (20.0%) baselines. Performance is strong across the full privacy spectrum, including the most sensitive class (C5), where F1 reaches 68.6. Results are comparable to ModernBERT-base, while clearly surpassing BERT-base and the smaller Ettin-17M variant. F1 for the intermediate classes (C2–C4, ranging from 58 to 64) is lower than for the extreme classes (C1 at 91.5, C5 at 68.6). This is expected for an ordinal scale where adjacent categories are difficult to distinguish: texts at the boundary of "Mostly not private" and "Somewhat private" are inherently ambiguous. Importantly, classification errors for these middle classes are predominantly adjacent-class confusions (e.g., predicting C2 instead of C3), as reflected in the low MAE. Overall, these results confirm that privacy sensitivity, as defined

[2]Mean Absolute Error (MAE) captures the average absolute distance between predicted and true privacy levels, treating the task as ordinal. Unlike accuracy or F1, MAE penalizes predictions proportionally to how far they deviate from the correct class (e.g., predicting 5 instead of 1 is penalized more than predicting 2 instead of 1), making it particularly suitable for ordered label spaces such as Likert-scale privacy ratings.

Domain	Rating	Example
TW	1	"Happy First Day of Spring!"
TR	2	"I have not received my item, even though I had an email from Royal Mail stating it had been delivered on 27/12/24"
EE	3	"Rick, Attached are drafts of the letters notifiying Williams, Apache and Lone Star that ENA will be acting as agent for Tenaska IV. Please review the letters and get back to me"
MQ	4	"Hi, Im asking for a friend that went to the hospital today by ambulance. she is 61 years old and has been sick for over a month."
MHB	5	"I have suffered severe anxiety and depression for a very long time (first panic attack eight years old). I've seen plenty of psychologists ect but nothing seems to work."

Table 3: Examples illustrating the range of privacy ratings across domains.

Model	Params	Acc. ↑	Macro F1 ↑	MAE ↓	Per-class F1 ↑				
					C1	C2	C3	C4	C5
Ettin-150M	149M	74.9	68.1	0.28	91.5	58.3	58.5	63.6	68.6
ModernBERT-base	149M	73.7	67.2	0.28	91.9	57.4	57.7	62.8	68.2
BERT-base	110M	73.3	65.9	0.29	91.3	55.4	55.7	60.6	66.4
Ettin-17M	17M	71.1	62.5	0.34	90.1	51.2	52.8	57.5	60.7

Table 4: Classification performance on the held-out test set. Accuracy and F1 scores are in %. MAE is the mean absolute error on the 1–5 ordinal scale (range: 0–4). All encoder models are trained with identical hyperparameters. For reference, a Majority Class baseline achieves 45.9% accuracy (12.5 macro F1), while a Random baseline achieves 20.0% accuracy (18.0 macro F1).

Comparison	Krippendorff's α
Ettin-150M vs. Human avg	0.737
Ettin-150M pairwise w/ humans	0.514 ($\pm$0.265)
Ettin-17M vs. Human avg	0.708
Ettin-17M pairwise w/ humans	0.498 ($\pm$0.259)
Mistral Large 3 (675b) vs. Human avg	0.716
Mistral Large 3 (675b) pairwise w/ humans	0.502 ($\pm$0.264)
Mistral-7b vs. Human avg	0.563
Mistral-7b pairwise w/ humans	0.409 ($\pm$0.250)
Inter-Human (overall)	0.39
Inter-Human (pairwise avg)	0.54

Table 5: Agreement metrics on the 250-text benchmark from Meisenbacher et al. (2025). "Human avg" denote agreement with the average rating of all human annotators.

by the teacher model, can be learned reliably by lightweight encoders without architecture-specific tuning.

4.2. Alignment with Human Privacy Judgments

We next evaluate agreement with the 677 human annotations from the original benchmark, which covers 250 texts in total (25 from each dataset). Table 5 presents the central findings. The distilled Ettin-150M model achieves $\alpha = 0.737$ agreement with the *average human rating*. Notably, this exceeds the agreement of its teacher model, Mistral Large 3 ($\alpha = 0.716$). For completeness, we also report results for Mistral-7b (Jiang et al., 2023), which achieves substantially lower agreement compared to both Mistral Large 3 and our distilled encoder models.

When compared pairwise with individual human annotators, the model achieves $\alpha = 0.514$ (± 0.265), closely matching the inter-human pairwise average ($\alpha = 0.54$). This suggests that disagreements between the model and individual humans are of the same magnitude as disagreements among humans themselves. Our models align well on the general perception of privacy, whereas they cannot capture the unique perspectives and experiences of all represented annotators.

4.3. De-Identification Evaluation

To demonstrate a practical application, we evaluate our model's ability to assess anonymization quality using the Text Anonymization Benchmark (TAB) (Pilán et al., 2022). TAB comprises English-language court cases from the European Court of Human Rights, with expert annotations of entity mentions categorized as DIRECT identifiers (e.g., names, passport numbers), QUASI-identifiers (e.g., age, nationality, profession), or NO_MASK. Using the 555-document test split, we create four versions of each document: original, DIRECT-masked (1,612 entities replaced with `[REDACTED]`), QUASI-masked (19,197 entities), and fully masked (both types).

Table 6 reveals three key patterns. First, masking DIRECT identifiers ($\Delta = 0.34$) has a larger per-entity effect than masking QUASI-identifiers ($\Delta = 0.23$), despite far fewer entities (1,612 vs. 19,197), yielding higher privacy impact per entity for DIRECT identifiers. This aligns with established personally identifiable information (PII) categorizations: names and other direct identifiers are inherently more individualizing than demographic attributes.

Second, comprehensive masking ($\Delta = 1.86$) pro-

Condition	$\bar{S}$	Δ	% Class 1
Original	3.25	–	25.2
Mask DIRECT	2.91	0.34	30.5
Mask QUASI	3.02	0.23	28.5
Mask ALL	1.39	1.86	84.1
Mask 30% random	3.56	-0.31	17.3

Table 6: Privacy scores on TAB test set. $\bar{S}$: mean score (1–5). Δ: reduction from original. % Class 1: proportion rated "Harmless."

duces a larger reduction than the sum of individual effects ($0.34 + 0.23 = 0.57$), revealing a strong interaction between identifier types. When both direct and quasi-identifiers are present, direct identifiers enable identification of the person's identity while quasi-identifiers provide additional sensitive information, thereby increasing the overall privacy risk of the text.

Third, after full masking, 84.1% of documents are rated "Harmless" (class 1), compared to only 25.2% in the original. This demonstrates that TAB's expert-defined masking scheme effectively reduces model-perceived privacy sensitivity. These results validate that our classifier captures privacy-relevant information consistent with expert annotations.

As a sanity check, we also randomly replace 30% of words with `[REDACTED]` tokens. Rather than reducing privacy, random masking *increases* the mean privacy score ($\bar{S} = 3.56$, $\Delta = -0.31$). This occurs because uninformed redaction disrupts coherence while preserving identifying content. This confirms that the classifier is sensitive to *what* is masked, not merely to the presence of masking tokens.

5. Discussion & Future Work

Performance Measurement. A notable outcome is that the distilled Ettin-150M slightly exceeds the teacher model in agreement with the *average* human rating. This does *not* imply that the student is intrinsically "more correct" than the teacher using our approach; rather, distillation can act as a *denoising* process. Training on a large volume of teacher-labeled examples can smooth prompt-level stochasticity and compress the teacher's reasoning into a deterministic decision boundary that generalizes better on a small human benchmark. Future work should explicitly test this hypothesis by studying more teacher behaviors or varying the amount of distillation data.

Use cases. Beyond benchmarking, an on-device privacy sensitivity classifier unlocks workflows that are difficult or undesirable with API-based LLM judges: (i) *Dataset curation*: assigning sensitivity scores to large corpora to route high-risk examples for manual review, filtering, or access control before model training; (ii) *Privacy-aware evaluation for rewriting/anonymization*: using the score as an automatic metric to compare de-identification or privatization systems across datasets and parameter settings, complementing attack-based proxies; (iii) *User-facing privacy assistance*: real-time warnings in writing assistants (e.g., "this message contains likely identifying details") and suggestions for minimal edits. This is particularly valuable given evidence that users routinely leak PII when interacting with external LLMs (Mireshghallah et al., 2024).

Future Work. Compact evaluators enable research that is otherwise compute- or policy-constrained. First, they make it feasible to study *privacy signals at scale* via attribution and counterfactual edits, helping disentangle identifiability cues (names, locations, unique events) from topic sensitivity (health, legal issues, mental health). Second, the model can serve as a *training signal*: combined with a utility measure (e.g., semantic similarity), it can define privacy–utility trade-offs and support search or learning procedures that find minimal changes that reduce privacy sensitivity. Third, future work should move beyond a single "global" notion of privacy by incorporating *context* (audience, purpose, setting) and exploring *personalization* with small amounts of user-provided preference data. Finally, robustness remains open: calibrating scores, dealing with out-of-domain inputs, and auditing domain- and demographic-dependent failure modes are essential before deploying the model as part of automated pipelines.

6. Conclusion

We presented a knowledge distillation approach for creating efficient privacy sensitivity classifiers from LLM judgments. Responding to calls for lightweight privacy evaluation models, we distilled Mistral Large's privacy assessments into a 150M-parameter Ettin encoder that achieves strong agreement with human annotations while enabling private and faster inference. Our evaluation on the Text Anonymization Benchmark demonstrates that the classifier captures meaningful differences between direct and quasi-identifiers in expert-annotated documents, validating its utility for de-identification assessment. We release our code, models and dataset to support reproducible privacy evaluation in NLP.

Limitations

Our models inherit the privacy notion and potential biases of the teacher LLM: privacy is compressed

into a single 1–5 sensitivity score, which may conflate multiple dimensions such as identifiability and topic sensitivity. The training data is English-only; multilingual transfer remains untested. Privacy is contextual (Nissenbaum, 2004), yet the classifier evaluates texts largely in isolation without explicit information about audience, purpose, or setting.

Teacher labeling can be stochastic due to the non-deterministic nature of large language models (Song et al., 2025); using multiple teachers, or a small amount of human-labeled calibration data could reduce noise and improve robustness. We also did not systematically study alternative teacher models or distillation strategies. Finally, the score should not be interpreted as a formal privacy guarantee or as a proxy for adversarial re-identification risk: it captures *perceived* sensitivity under the adopted scale.

Ethical Considerations

This work processes potentially sensitive user-generated content. All source datasets are publicly available and have been previously used in research. Our classifier is intended for evaluating privacy-preserving methods and supporting privacy research, not for making decisions about individuals or for surveillance purposes. We caution against using the model as a hard gate without human oversight: the privacy scale is a subjective construct, and model scores should inform rather than replace human judgment.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback. We are also grateful to the authors of Meisenbacher et al. (2025), whose work on LLM-based privacy evaluation provided the foundation for this study.

Bibliographical References

Jane Bambauer, Alan Mislove, et al. 2022. What do we mean when we talk about privacy? A survey of privacy definitions and approaches. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1774–1784.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

Sravani Boinepelli, Tathagata Raha, Harika Abburi, Pulkit Parikh, Niyati Chhaya, and Vasudeva Varma. 2022. Leveraging mental health forums for user-level depression detection on social media. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5418–5427, Marseille, France. European Language Resources Association.

Hongzhan Chen, Ruijun Chen, Yuqi Yi, Xiaojun Quan, Chenliang Li, Ming Yan, and Ji Zhang. 2024. Knowledge distillation of black-box large language models.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Cynthia Dwork. 2006. Differential privacy. In *Automata, Languages and Programming*, pages 1–12, Berlin, Heidelberg. Springer Berlin Heidelberg.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Dirk Hovy, Anders Johannsen, and Anders Søgaard. 2015. User review sites as a resource for large-scale sociolinguistic studies. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 452–461, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. *Departmental Papers (ASC)*.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.

Jonathan Li, Rohan Bhambhoria, and Xiaodan Zhu. 2022. Parameter-efficient legal domain adaptation. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 119–129, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yizhong Li et al. 2024. LLMs-as-judges: A comprehensive survey on LLM-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. 2021. Anonymisation models for text data: State of the art, challenges and future directions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4188–4203, Online. Association for Computational Linguistics.

Gabriel Loiseau, Damien Sileo, Damien Riquet, Maxime Meyer, and Marc Tommasi. 2025. Tau-eval: A unified evaluation framework for useful and private text anonymization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 216–227, Suzhou, China. Association for Computational Linguistics.

Stephen Meisenbacher, Alexandra Klymenko, and Florian Matthes. 2025. LLM-as-a-Judge for privacy evaluation? Exploring the alignment of human and LLM perceptions of privacy in textual data. In *Proceedings of the 2025 Workshop on Human-Centered AI Privacy and Security (HAIPS '25)*. ACM.

Niloofar Mireshghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. 2024. Trust no bot: Discovering personal disclosures in human-LLM conversations in the wild. In *First Conference on Language Modeling*.

Mistral AI. 2025. Mistral large 3. *Mistral AI Blog*.

Helen Nissenbaum. 2004. Privacy as contextual integrity. *Washington Law Review*, 79(1):119–158.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The text anonymization benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(4):1053–1101.

Yaxuan Ren, Krithika Ramesh, Yaxing Yao, and Anjalie Field. 2025. How do we measure privacy in text? a survey of text anonymization metrics. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1532–1544, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Hossein Shahriar et al. 2025. A survey on privacy-preserving techniques in natural language processing. *Information Fusion*, 104:104358.

Yifan Song, Guoyin Wang, Sujian Li, and Bill Yuchen Lin. 2025. The good, the bad, and the greedy: Evaluation of LLMs should not ignore non-determinism. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4195–4206, Albuquerque, New Mexico. Association for Computational Linguistics.

Welderufael B. Tesfay et al. 2016. Challenges in automated privacy policy analysis. *IEEE Security & Privacy*.

Benjamin Warner, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, et al. 2024. ModernBERT: A modern approach to encoder-only transformers. *arXiv preprint arXiv:2412.13663*.

Orion Weller, Kathryn Ricci, Marc Marone, Antoine Chaffin, Dawn Lawrie, and Benjamin

Van Durme. 2025. Seq vs seq: An open suite of paired encoders and decoders. *arXiv preprint arXiv:2507.11412*.

Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, et al. 2016. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1340.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*.

A. Dataset Sources Description

To construct a diverse and coherent valid privacy corpus, we aggregate user-written texts from ten publicly available datasets spanning blogs, emails, health-related forums, legal advice, reviews, microblogs, and online discussions. Together, these datasets cover varying genres, platform norms, and disclosure styles, enabling us to capture multiple dimensions of privacy risk: explicit identifiers, contextual clues, health and legal sensitivity, and stylometric linkability.

Blog Authorship Corpus (BAC). BAC consists of long-form blog posts, capturing personal storytelling and opinionated writing with strong stylistic signatures—a setting commonly associated with authorship and user profiling tasks. Such stylometric consistency makes it a natural domain for studying linkability-based privacy risks, beyond overt identifiers. We use the HuggingFace version of the corpus (https://hf.co/datasets/tasksource/blog_authorship_corpus).

Enron Emails (EE). EE comprises semi-private organizational communication, where texts often include operational details, social relations, and direct contact information. This domain contributes a complementary privacy profile: not necessarily self-disclosure, but high likelihood of concrete identifiers and workplace context. We use the HuggingFace release (https://hf.co/datasets/sujan-maharjan/enron_email_dataset), totaling 362,180 emails from the sent items folder.

Medical Questions (MQ). The Medical Question Answering dataset contains over one million user-written medical question-answer pairs; we keep only the questions. These texts are representative of information-seeking and triage-like tasks in health NLP, and they frequently contain sensitive attributes (symptoms, medications, ages, conditions), making MQ a core pillar of the corpus for health-related privacy. We use the processed HuggingFace version (https://hf.co/datasets/Malikeh1375/medical-question-answering-datasets), comprising 246,678 questions.

Mental Health Blog (MHB). MHB contains posts from a mental health forum (2011–2020), reflecting long-form self-disclosure and peer support—a genre often used for mental-health-related classification and risk detection. Privacy signals here are typically driven by sensitive topics and personal experiences, even when explicit identifiers are absent (Boinepelli et al., 2022). (This dataset is not hosted on HuggingFace.)

Reddit Confessions (RC). This corpus (https://hf.co/datasets/SocialGrep/one-million-reddit-confessions) consists of one million posts from four confession-

oriented subreddits, emphasizing anonymous self-disclosure and narrative accounts. As a domain, it broadens our corpus toward "voluntary disclosure" settings, where privacy sensitivity is often rooted in intimate content rather than formal identifiers.

Reddit Legal Advice (RLA). RLA contains nearly 100k informal legal help-seeking posts from */r/legaladvice* (Li et al., 2022). We use the HuggingFace version (https://hf.co/datasets/jonathanli/legal-advice-reddit), which provides 79,136 posts. This domain introduces a distinct mix of sensitive situational details (disputes, employment issues, alleged wrongdoing) and contextual clues (locations, timelines, involved parties).

Reddit Mental Health Posts (RMHP). RMHP provides over 150k posts from five mental-health-related subreddits. Compared to MHB, RMHP adds a different platform style (shorter, conversational, often more fragmented) while covering similar sensitive themes. This contrast helps ensure our privacy corpus captures both topic sensitivity *and* platform-/genre-specific expressions of disclosure. We use the HuggingFace version (https://hf.co/datasets/solomonk/reddit_mental_health_posts), comprising 151,288 posts.

Trustpilot Reviews (TR). Trustpilot reviews are consumer-facing evaluations of services and products, commonly used in sentiment and sociolinguistic analyses (Hovy et al., 2015). While reviews may appear "less private" on average, they regularly include incidental identifiers (names, locations, service interactions) and thus provide valuable coverage of *everyday* privacy leakage in ostensibly public feedback. We use the HuggingFace dataset (https://hf.co/datasets/Kerassy/trustpilot-reviews-123k), containing 123,181 *en-US* reviews.

Twitter (TW). TW represents short, public microblog posts, often used for topic, sentiment, stance, or event-oriented analyses. Including Twitter ensures the corpus captures privacy signals in highly compressed text, where identifying cues can arise from usernames, mentions, location hints, or contextual references rather than explicit PII. We sample from the HuggingFace version of the 100M tweets corpus (https://hf.co/datasets/enryu43/twitter100m_tweets), containing 88,084,332 tweets.

Yelp Reviews (YR). YR contains user reviews of local businesses—a domain frequently associated with sentiment and recommendation tasks, grounded in real-world places and experiences. This setting contributes privacy risks tied to location context, routines, and named venues (even without explicit identifiers). We use the HuggingFace release (https://hf.co/datasets/yashraizad/yelp-open-dataset-reviews), which contains 6,990,280 reviews.

B. Teacher Prompt

We use a structured prompt that enforces the discrete 1–5 privacy scale:

> Provide a **privacy rating** scoring on how sensitive or private the user_text, for example including personal information or indirectly identifiable information.
>
> Give your answer on a scale of 1 to 5, where 1 means that the user_text is not sensitive or private at all, and 5 means that text contains highly sensitive or private information.
>
> Here is the scale you should use to build your answer:
> 1: The user_text is harmless: completely free of any private or sensitive information, either direct or indirect identifiers.
> 2: The user_text is mostly not private: may contain some indirect identifiers, but is mostly free of sensitive information.
> 3: The user_text is somewhat private: contains some direct or indirect identifiers, and can be deemed somewhat personal information.
> 4: The user_text is very private: contains several direct or indirect identifiers, and it is clear that this text contains personal information.
> 5: The user_text is extremely private: contains highly sensitive information, such as direct personal identifiers, and the text is clearly something that should not be made public.

Birds of a Feather: Do Embedding Representations of Personal Information Flock Together?

Maria Irena Szawerna[1], Simon Dobnik[2]
[1]Språkbanken Text, SFS, University of Gothenburg, Sweden
[2]CLASP, FLoV, University of Gothenburg, Sweden
{maria.szawerna,simon.dobnik}@gu.se

Abstract

Personally identifiable information (PII or PI) can appear in a wide variety of linguistic data, posing both ethical and legal challenges for conducting research and developing applications involving such texts. In this paper, we investigate the alignment between automatic clustering of FastText and Transformer embedding representations of personal information spans sourced from essays written by adult learners of Swedish as a second language and the general and detailed personal information labels assigned to these spans by expert annotators. Our goals are to assess the extent of overlap between the semantic categories and evaluate the semantic coherence of the human-assigned classes, which may have implications for de-identification procedures. We observe that while contextual embeddings, especially ones from a specialized word-in-context model, produce relatively good clustering results, they only partly map to the human understanding of how to classify personal information.

Keywords: personal information, PII, de-identification, pseudonymization, anonymization, clustering

1. Introduction and Prior Research

The presence of personally identifiable information (PII, PI)[1] in language data poses undeniable ethical and legal challenges. There is a need for the development of tools aimed at automatizing the time-consuming task of personal information detection, followed by redaction or labeling and replacement. PI detection and labeling is a ubiquitous step in Named Entity Recognition-like approaches to de-identification of such data (Lison et al., 2021; Volodina et al., 2025). Many such approaches rely on (contextual) embedding representations of the tokens in the text to carry out the classification (cf. Grancharova and Dalianis, 2021 or Pilán et al., 2022), as rule-based methods can only capture some PI types that show less diversity in terms of form (Volodina et al., 2020). However, it has been shown that such systems can be sensitive to how internally consistent personal information classes are (Sierro et al., 2024; Szawerna et al., 2025).

In this exploratory paper, we set out to address the question to what degree do embedding representations of PI words and phrases capture the semantic knowledge of humans, where that semantic knowledge is approximated by a PI taxonomy developed by human annotators (i.e., division of spans identified as PI into classes salient for humans). Our goal is to improve the semantic understanding of PI annotation labels and to assess how representations used by models align with a human-devised taxonomy, which could help improve both PI detection and labeling methods and the taxonomies themselves. From the PI annotation perspective, such findings could be relevant for identifying the categories or their parts that are particularly confusing for the models and may benefit from a reinterpretation of the human-assigned labels or by identifying new categories salient for the detection and labeling model. Determining which embedding models permit a good level of distinguishing between different human-assigned labels yields insights into what models are worth investigating for automatic PI detection and labeling in practice. Investigating the semantic alignment between humans and language models in this specific domain may also hint at some more general patterns. In that sense, our work is reminiscent of language games pioneered by Steels and Belpaeme (2005), where they evaluate the similarity between natural language categories and categories in a an automatically-induced language emergent from situational grounding of two artificial agents.

We address our research question through clustering embedding representations of personal information. Clustering embeddings to understand the distributional properties of language data has previously been employed by e.g. Hertzberg et al. (2022) in the domain of political dogwhistles; while there are several differences in our approaches, the main one is our comparison being conducted against ground truth labels, whereas theirs tries to determine whether the successfulness of clustering correlates with inter-annotator agreement.

Given that per their definition, some of the PI categories are rather internally diverse, we expect to only see a partial overlap between automatic clustering and the pre-existing annotation. We also expect the embedding representations sourced from models with better understanding of the role that

[1]Henceforth often simply 'personal information.'

Proceedings of LEGAL2026 and CALD-pseudo 2026 @ LREC 2026, pages 62–72
12 May 2026. ©ELRA Language Resources Association (ELRA), 2026

context plays for the meaning of a word or a phrase to yield more distinct clusters that better map to at least some of the human-assigned categories.

2. Materials

In our experiments, we use 947 texts (totaling 301095 tokens) from SweLL-pilot and SweLL-gold (Volodina, 2024; Språkbanken Text, 2025b).[2] These two SweLL corpora are collections of essays written by learners of Swedish as a second language (L2). Many of these texts contain various kinds of PI, which are pseudonymized in the released versions of the corpora; however, we use the essays with the original PI intact.

The PI spans in the SweLL data are annotated with PI categories (see Megyesi et al. (2021) and Volodina et al. (2020)). This taxonomy is hierarchical, with 7 overarching general categories and 37 possible detailed PI categories. For instance, in the fictitious example of *mitt namn är Sonja och jag är 29* 'my name is Sonja and I am 29', *Sonja* would be labeled by an expert annotator as the detailed class `firstname_female` (which belongs to the general category `personal_name` together with surnames, masculine names, etc.), and *29* would be marked as `age_digits` (belonging to the general category `age`). In our data only 32 of those detailed categories are present.[3] Both singular tokens and multi-word expressions may be annotated as PI; 3348 tokens constituting 3076 spans are annotated as PI in our data. Table 2 in Appendix A shows the detailed counts of the annotated tokens and phrases alongside information as to which detailed categories correspond to which general ones.[4] As that table shows, some classes are much more frequent in the data than others, which is likely to negatively affect the discriminability of the infrequent classes.

As we are working with Swedish data, we chose three models trained for this language to obtain embedding representations of the PI spans:

1. `Kubord-fasttext - Dagens Nyheter 2010-2024 - token` (Språkbanken Text, 2025a): one of the FastText embedding models for Swedish (see Bojanowski et al. (2016)). Embedding size: 300. Henceforth FastText;

2. `KB/bert-base-swedish-cased` (Malmsten et al., 2020): the Swedish version of the original BERT model (Devlin et al., 2019). Embedding size: 768. Henceforth KB-BERT;

3. `pierluigic/xl-lexeme` (Cassotti et al., 2023): a specialized multilingual word-in-context (WiC) model based on XLM-RoBERTa-large (Conneau et al., 2020). Embedding size: 1024. Henceforth XL-LEXEME.

The FastText embeddings serve as a non-contextual baseline. KB-BERT has previously been used in many token classification tasks for Swedish, including PI detection and labeling applications (e.g., by Grancharova and Dalianis (2021), Vakili et al. (2022), or Szawerna et al. (2024)). XL-LEXEME belongs to a similar language model category as KB-BERT, but as it is specialized for word-in-context tasks, it may capture more of the nuances of personal information. Embeddings for each token or subword token in a PI span were obtained from each model. Maximum input size of 100 KB-BERT subword tokens was used for the masked language models to ensure that a comparable context was provided for the phrase in question. For KB-BERT, the last-layer representations were obtained, as those are more sensitive to semantics and context (Jawahar et al., 2019). In the cases of multi-token spans, a mean of the embeddings was obtained for FastText and BERT to preserve dimensionality; it was possible to directly obtain an embedding for the whole span from XL-LEXEME. These three sets of embeddings will henceforth be referred to as embedding types.

3. Methods

In order to evaluate the alignment between the embeddings of different PI spans and the human-assigned labels, we perform automatic clustering on the embeddings. We first reduce the embedding size using Uniform Manifold Approximation and Projection (UMAP, McInnes et al., 2020). This step already helps capture the underlying patterns and speeds up computation. We then perform a parameter search for four clustering algorithms: Hierarchical Density-Based Clustering (Campello et al., 2013), Affinity Propagation (Frey and Dueck, 2007), Mean Shift (Fukunaga and Hostetler, 1975), and Agglomerative Clustering,[5] in their scikit-learn implementations (Pedregosa et al., 2011). These four algorithms all permit varied cluster size, which is essential given the uneven distribution of human-annotated PI classes in our data. We evaluate the intrinsic quality of the clustering using the silhouette score (Rousseeuw, 1987), which is a measure of

[2] SweLL access can be requested at `https://sunet.artologik.net/gu/swell`

[3] `initials, area, url, personid_nr, account_nr, license_nr` are absent.

[4] The latter is also explained better in Megyesi et al. (2021); Volodina et al. (2020); Szawerna et al. (2025).

[5] To the best of our knowledge, there is no single citation for hierarchical agglomerative clustering, and only various linkage methods have standard references, see Müllner (2011).

how well data points fit their clusters and how well-bounded those clusters are on a scale between -1 and 1, and select the best clustering algorithm and parameters for each embedding type.

We calculate extrinsic measures for the selected results, comparing the emergent clusters to the human annotation. We focus on completeness (data points from one ground truth class being grouped in one cluster), homogeneity (the internal purity of clusters relative to the ground truth), and the combined v-score (Rosenberg and Hirschberg, 2007) as interpretable measures of specific properties of the clustering relative to the human annotation at both the detailed and general label level. We consider homogeneity to be more important than the other two in understanding how the machine clustering relates to the human-identified classes; it is clear that completeness will be much lower in clustering outcomes which result in hundreds of clusters, but as long as those are internally homogeneous, one can conclude that the clustering simply splits a human-assigned category into even more granular ones. Additionally, we calculate entropy (Shannon, 1948) per cluster and normalize it[6] to further inspect how pure the specific clusters are, analogous to how Dobnik and Kelleher (2013) or Dobnik and Kelleher (2014) use this measure to assess the purity of semantic categories against a set of labels.

4. Results and Discussion

The best silhouette scores were obtained for all embedding types by the HDBSCAN algorithm when the outlier category that it predicts was excluded from the calculation (0.83 for FastText, 0.67 for KB-BERT, and 0.72 for XL-LEXEME).[7]

Given that the silhouette score ranges from -1 (very bad) to 1 (perfect), these scores are good, and it is unsurprising to see clustering algorithms that eliminate outliers perform well on the intrinsic metric. However, as shown in Table 1, between 18 and 42% of the data was excluded as outliers, indicating that a large part of the data is hard to cluster cleanly. Across all embedding types, nearly all detailed PI categories are represented among the outliers. When inspecting the items identified as outliers, some trends can be noted, such as classes that are generally infrequent being more likely to have a large proportion of outliers, personal names and dates being hard to cluster with FastText embeddings, or KB-BERT struggling with geographic and transportation classes.[8] Results for XL-LEXEME stand out here, with the lowest number of embeddings that are treated as outliers and a high homogeneity score. While the KB-BERT embeddings result in a large number of outliers, the number of detected clusters is the closest to the number of detailed labels in the human taxonomy and the lowest out of the three outcomes. Finally, FastText embeddings result in noticeably worse results than the contextual embeddings.

An interesting, but not entirely unexpected, observation can be made by comparing the scores relative to the detailed and general human-assigned labels. Overall, completeness and v-score are lower when the comparison is made to the general classes, as the number of clusters is always much larger than the 7 general classes, meaning that multiple clusters will consist of examples from one such class. Homogeneity improves noticeably for contextual embeddings when the comparison is made to general human-assigned labels instead of detailed. This indicates that even though not all clusters are pure, elements that belong to different detailed classes but the same overarching general classes are still grouped together. This does not hold for the FastText embeddings, implying that the clusters there have more random impurities.

This is further corroborated by the per-cluster entropy scores pictured in the histograms in Figure 1.[9] A cluster being perfectly homogenous relative to ground truth means it has an entropy of 0, whereas an entropy of 1 means a maximally random assortment of human-assigned labels in the cluster. For FastText embeddings, there are relatively minor differences between the entropy scores for detailed and general labels. What can be noticed is that comparing to general labels leads to a small increase in the lower entropy scores, whereas comparing to detailed labels is what is responsible for entropy scores above 0.5. A similar pattern occurs in the case of the contextual embeddings, but the differences are more pronounced, especially in the case of XL-LEXEME, where the percentage of near-zero entropy clusters skyrockets when the comparison is made to general labels. This indicates that while the XL-LEXEME embeddings appear to be the best for clustering PI (with a relatively small number of outliers and high homogeneity), they do not permit the same fine-grained distinctions as the human annotation and instead group the information differently at that level of detailedness, though within the same general classes. When it comes to grouping the samples according to the detailed classes, KB-BERT appears to perform best, with propor-

[6] Entropy of a cluster X here is defined as $H(X) = -\sum_{x \in X} p(x) log_2 p(x)$ and it is normalized by the maximum possible entropy for the cluster, i.e. $-log_2(|X|)$.

[7] Parameter details can be found in Appendix A.

[8] The detailed counts can be found in Appendix A.

[9] Visualized using pandas (pandas development team, 2020), matplotlib (Hunter, 2007), and seaborn (Waskom, 2021)

Embedding	Clustering	Homogeneity	Completeness	V-score	N clusters	N outliers
FastText	HDBSCAN	0.69 (0.67)	0.44 (0.24)	0.54 (0.35)	66	1132
KB-BERT	HDBSCAN	0.72 (0.84)	0.53 (0.35)	0.61 (0.50)	41	1304
XL-LEXEME	HDBSCAN	0.77 (0.86)	0.50 (0.31)	0.60 (0.46)	70	561

Table 1: Metrics per embedding type for the best clustering results. Scores in black are compared against human-assigned detailed labels, whereas the (gray scores) are relative to the general labels.

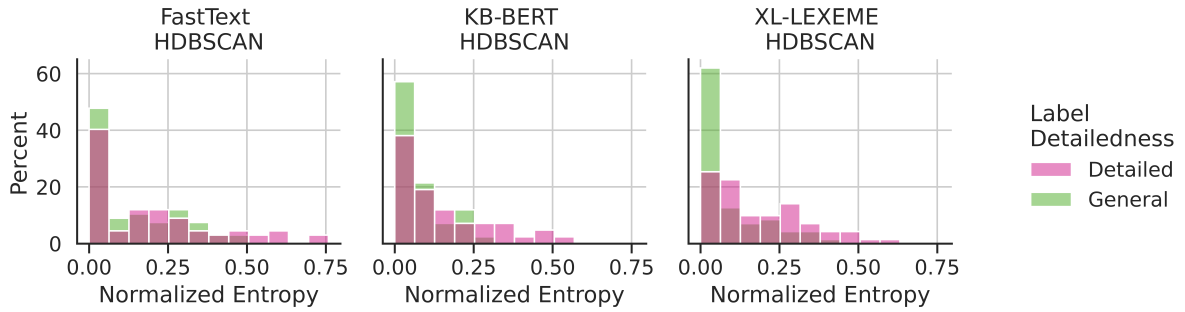


Figure 1: Histograms of entropy distribution, in percent for comparability across embedding types.

tionally only slightly fewer 0-entropy clusters than FastText, but with a tendency for overall lower entropy scores. Given that KB-BERT's general-level entropy is only slightly lower than XL-LEXEME's, this model's embeddings could be interpreted as more versatile when it comes to PI labeling.

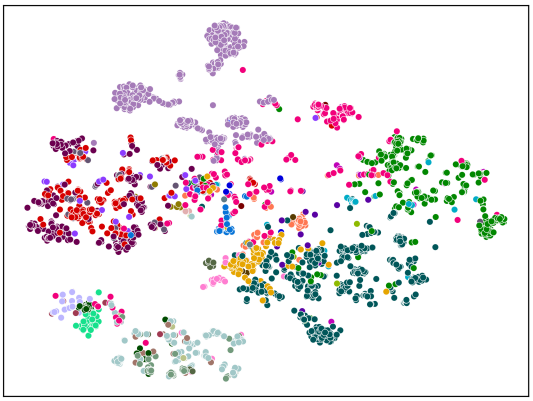
This can be visualized by reducing the dimensionality of the embeddings to 2 using UMAP and plotting the datapoints. Figure 2 shows this data for XL-LEXEME embeddings, colored according to detailed and general human annotation (Figure 2a, Figure 2b) and with the clusters assigned by HDBSCAN (Figure 2c). While the correspondence between colors and actual labels is not provided due to the number of labels, the differences between detailed human labels and the HDBSCAN clusters are quite apparent.[10]

It can be observed that for the contextual embeddings, some human-assigned labels correspond rather uniformly to a given area in the embedding space; for certain other categories, this distinction is much more blurred. For example, in Figure 2a, the pastel purple[11] points constituting the two large, distinct clusters near the top of the figure belong to the `fam` detailed category, which includes words describing relatives; this category is quite distinct and shares only a small overlap with the `sensitive` detailed category, marked in bright pink to the right and below the `fam` clusters. The `sensitive` category, in turn, does not appear to cluster well and is dispersed rather broadly, overlapping with several other categories. While Figure 2b does show that `personal_name` (red, left side of the figure), `other` (yellow, top of the figure), and `geographic` (purple, right side of the figure) seem to generally be confined to their own areas of the embedding space, there is a relatively distinct area at the bottom left, which is made up of `date`, `age`, and `other`; these are predominantly examples of PI that has to do with numbers (both in digit and string format), and that this characteristic was very salient for the model. Finally, Figure 2c shows — same as the entropy analysis did — that the automatically identified clusters subdivide the general labels rather finely, but not the same way that human annotators divided them. For instance, the aforementioned distinct `fam` groupings get clustered much more finely.

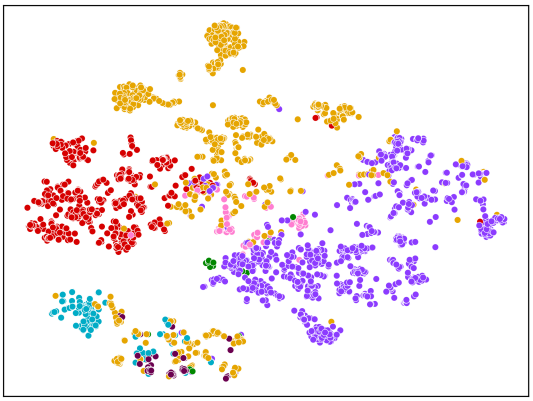
Inspecting the samples contained in the automatically detected clusters that make up `fam` shows the different types of family relations captured by XL-LEXEME. One cluster is made up of words like *bror* 'brother,' *farfar* 'paternal grandfather,' or *pappa* 'dad.' Separate clusters emerge for *moster* 'maternal aunt,' *faster* 'paternal aunt' together with *styvmamma* 'stepmom,' for *mamma* 'mom' and *mormor* 'maternal grandmother,' as well as for various forms of the word *syster* 'sister.' Another cluster contains words relating to children (*barn* 'child,' *son* 'son,' *syskon* 'sibling'). Yet another cluster within `fam` groups together more distant or loose relations and gender-agnostic ones, with *kusin* 'cousin,' *pojkvän* 'boyfriend' and *brorsfru* 'sister-in-law' clustered to-

[10] See Appendix A for plots for FastText and KB-BERT.

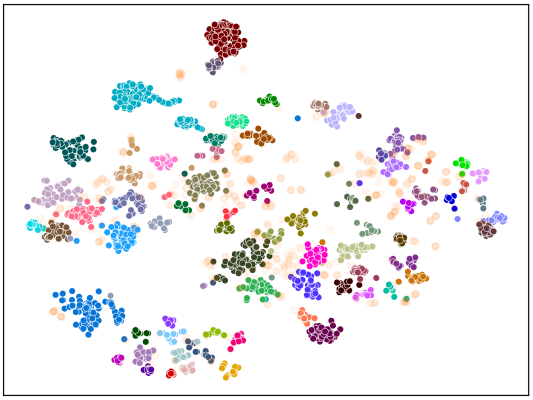
[11] We acknowledge the difficulty of relating this description to the figure in grayscale or for people with color vision deficiency, which is why we additionally try to provide the location of the discussed points in the figure. Due to the number of labels, using shapes rather than color to distinguish between categories was decided against as it made the plots largely unreadable.



(a) Detailed human labels



(b) General human labels



(c) HDBSCAN labels

Figure 2: XL-LEXEME scatterplots. Due to the number of classes the legend is not provided. For HDBSCAN, the outliers are marked with translucent orange points.

gether, and one more consists of various forms of the word *familj* 'family.'

Investigating the clusters emerging for the general category of *geographic* illustrates the opposite situation, where various detailed categories (e.g. `city`, `place`, `geo`) are grouped together based on a similarity in spelling or location. This shows that the most salient characteristics of the embeddings are not what is the most salient about a given entity to a human annotator in the context of annotating personal information. This issue seems to appear more frequently with proper nouns.

These results indicate that while some semantic similarity is captured by clustering, the original classes become fragmented. If the goal is for the embedding representations to mirror the human annotation, this could perhaps be mitigated both by tweaking how the embeddings are obtained (which layer, what context window) and by what the ground truth reference is. On the other hand, an analysis of the contents of the emergent clusters may help refine the taxonomy used to annotate PI: in the case of the broad `fam` detailed category, it is clear that it could be subdivided into e.g. female relatives, male relatives, and children at the very least.

5. Conclusions and Future Work

In this paper, we explored the effectiveness of automatic clustering of authentic PI spans from Swedish texts, represented using three different types of embeddings, in order to increase our understanding of the semantics of PI labels and their alignment with computational representations. We observed that in both non-contextual and contextual embeddings, a certain number of PI instances are hard to cluster, but a specialized word-in-context model struggled less with this issue. Clustering algorithms tend to identify more clusters than the human-assigned detailed PI classes. Their boundaries sometimes align with the human annotation, depending on the embedding and annotation type. Impurities in clusters identified for the contextual embeddings tend to stem from semantically similar concepts being grouped together (e.g., different types of geographical information) and natural subdivisions form within certain clusters.

In the future, it could be interesting to use this approach to try to identify which models' representations (and from which layers) are sensitive to the differences between PI types and non-personal information with the goal of establishing how and what to train for PI detection and labeling, and whether the performance in experiments such as ours correlates with that and what effect model fine-tuning has on these representations. As shown, investigating what sets the separate clusters containing the same human-annotated class apart could be an interesting way to potentially help refine the taxonomy used for annotating PI. Comparing which human-assigned labels have the lowest inter-annotator agreement and which kinds of personal information are the hardest to cluster could bring more nuance to an analysis of this type. Finally, semantic relatedness between various PI clusters could perhaps be exploited in studies on semantic coherence of pseudonyms used to replace personal information.

Limitations

A natural limitation of this research is that it is conducted only on one genre of texts. However, to the best of our knowledge, there exists no other PI-annotated corpus in Swedish or another corpus annotated with the same categories as SweLL that is possible for us to access, which would be a very valuable comparison allowing us to generlize our observations about the nature of personal information. Similarly, the use of authentic PI data severely limits the reproducibility of this study; however, it shows our compliance with legal and ethical standards. We believe that this methodology can be successfully applied to any other PI-annotated dataset, making the research replicable, if not reproducible.

Another limitation is that this comparison only includes three models from which embeddings are obtained. While still fewer than for some other languages, there are many models that can, to some extent, handle Swedish text and that could be included in a larger-scale comparison.

Our experiment does not clearly assess the usefulness of the embedding representations for PI detection (i.e., telling it apart from the surrounding non-personal context), but only for its subsequent classification.

Since a part of the representations stand for multiword expressions, the way in which they are calculated (a mean of the constituent embeddings for FastText and KB-BERT) could make them harder to cluster and result in them being rejected as outliers.

Any more qualitative analysis of the purity of the identified clusters was hindered by the sheer number of clusters and the fact that we were comparing three such results.

Ethical Considerations

Research about PI and de-identification is, in large part, fueled by ethical considerations and legal requirements when it comes to processing language data. Continued exploration of such questions can contribute to a better understanding of the effectiveness and the consequences of de-identification, as well as help improve the methods employed; in the case of this paper, it could inform the choice of model for PI detection tasks and perhaps assist with the development of refined PI taxonomies.

As we are using authentic PI, which is not available in the current release of the corpus that we are working with, we cannot go into a too detailed analysis of clusters (we cannot provide specific, authentic PI span examples), nor can we share the data or the embeddings used in the analysis. We assess the risks of information leakage from the results that we provide to be low, as they are only shown aggregated and without any references back to the texts that the phrases are extracted from, and all the experiments were conducted locally.

Acknowledgments

This work was possible thanks to the funding from the Swedish Research Council. The work was conducted within the research environment project *Grandma Karl is 27 years old: Automatic pseudonymization of research data* with the funding number 2022-02311 for the years 2023-2029. It was also supported by *Språkbanken*, which is jointly funded by its 10 partner institutions and the Swedish Research Council (2025–2028; project id 2023-00161). The second author is also supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the *Centre for Linguistic Theory and Studies in Probability (CLASP)* at the University of Gothenburg.

We would also like to thank our colleagues Ricardo Muñoz Sánchez and Felix Morger for their advice and support.

6. Bibliographical References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606*.

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-Based Clustering Based on Hierarchical Density Estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Simon Dobnik and John D. Kelleher. 2013. Towards an automatic identification of functional and geometric spatial prepositions. In *Proceedings of PRE-CogSsci 2013: Production of referring expressions - bridging the gap between cognitive and computational approaches to reference*.

Simon Dobnik and John D. Kelleher. 2014. Exploration of functional semantics of prepositions from corpora of descriptions of visual scenes. In *Proceedings of the Third Workshop on Vision and Language*, pages 33–37, Dublin, Ireland. Dublin City University and the Association for Computational Linguistics.

Brendan J. Frey and Delbert Dueck. 2007. Clustering by Passing Messages Between Data Points. *Science*, 315(5814):972–976.

Keinosuke Fukunaga and Larry Hostetler. 1975. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40.

Mila Grancharova and Hercules Dalianis. 2021. Applying and Sharing pre-trained BERT-models for Named Entity Recognition and Classification in Swedish Electronic Patient Records. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 231–239, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Niclas Hertzberg, Robin Cooper, Elina Lindgren, Björn Rönnerstrand, Gregor Rettenegger, Ellen Breitholtz, and Asad Sayeed. 2022. Distributional properties of political dogwhistle representations in Swedish BERT. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 170–175, Seattle, Washington (Hybrid). Association for Computational Linguistics.

John D. Hunter. 2007. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. 2021. Anonymisation Models for Text Data: State of the art, Challenges and Future Directions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4188–4203, Online. Association for Computational Linguistics.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction.

Beáta Megyesi, Lisa Rudebeck, and Elena Volodina. 2021. Swell pseudonymization guidelines. *GU-ISS Forskningsrapporter från Institutionen för svenska, flerspråkighet och språkteknologi*, GU-ISS 2021-02.

Daniel Müllner. 2011. Modern hierarchical, agglomerative clustering algorithms.

The pandas development team. 2020. pandas-dev/pandas: Pandas.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization. *Computational Linguistics*, 48(4):1053–1101.

Andrew Rosenberg and Julia Hirschberg. 2007. V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, Prague, Czech Republic. Association for Computational Linguistics.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Claude E. Shannon. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423.

Maria Sierro, Begoña Altuna, and Itziar Gonzalez-Dios. 2024. Automatic Detection and Labelling of Personal Data in Case Reports from the ECHR in Spanish: Evaluation of Two Different Annotation Approaches. In *Proceedings of the Workshop on Computational Approaches to Language Data*

Pseudonymization (CALD-pseudo 2024), pages 18–24, St. Julian's, Malta. Association for Computational Linguistics.

Luc Steels and Tony Belpaeme. 2005. Coordinating perceptually grounded categories through language: a case study for colour. *Behavioral and Brain Sciences*, 28(4):469–489.

Maria Irena Szawerna, Simon Dobnik, Ricardo Muñoz Sánchez, Therese Lindström Tiedemann, and Elena Volodina. 2024. Detecting Personal Identifiable Information in Swedish Learner Essays. In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pages 54–63, St. Julian's, Malta. Association for Computational Linguistics.

Maria Irena Szawerna, Simon Dobnik, Ricardo Muñoz Sánchez, and Elena Volodina. 2025. The Devil's in the Details: the Detailedness of Classes Influences Personal Information Detection and Labeling. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 697–708, Tallinn, Estonia. University of Tartu Library.

Thomas Vakili, Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. 2022. Downstream Task Performance of BERT Models Pre-Trained Using Automatically De-Identified Clinical Data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4245–4252, Marseille, France. European Language Resources Association.

Elena Volodina, Yousuf Ali Mohammed, Sandra Derbring, Arild Matsson, and Beata Megyesi. 2020. Towards Privacy by Design in Learner Corpora Research: A Case of On-the-fly Pseudonymization of Swedish Learner Essays. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 357–369, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Elena Volodina, Simon Dobnik, Therese Lindström Tiedemann, Ricardo Muñoz Sánchez, Maria Irena Szawerna, and Lisa Södergård. 2025. Towards Shared Standards for Pseudonymization of Research Data. In *Proceedings of the 2nd Huminfra Conference*.

Michael L. Waskom. 2021. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021.

7. Language Resource References

Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada. Association for Computational Linguistics.

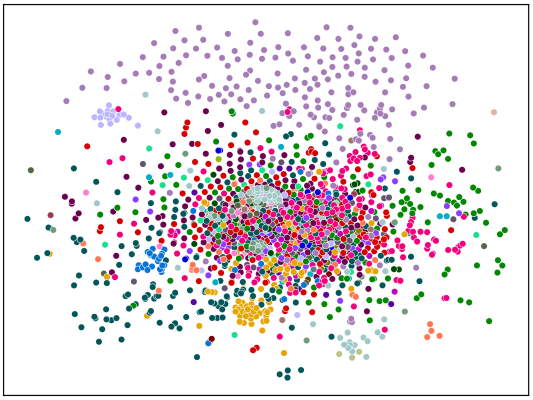
Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with Words at the National Library of Sweden – Making a Swedish BERT.

Språkbanken Text. 2025a. Kubord-fasttext - dagens nyheter 2010–2024 - token.

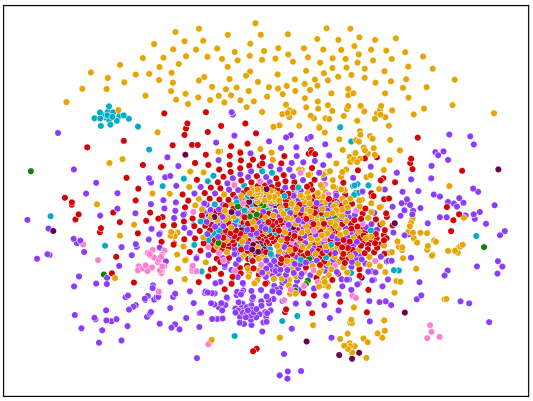
Språkbanken Text. 2025b. Swell.

Elena Volodina. 2024. On two swell learner corpora – swell-pilot and swell-gold. In *Proceedings of the Huminfra Conference (HiC 2024), 10-11 January, 2024, Gothenburg, Sweden*, pages 83–94, Linköping. Linköping University Press.

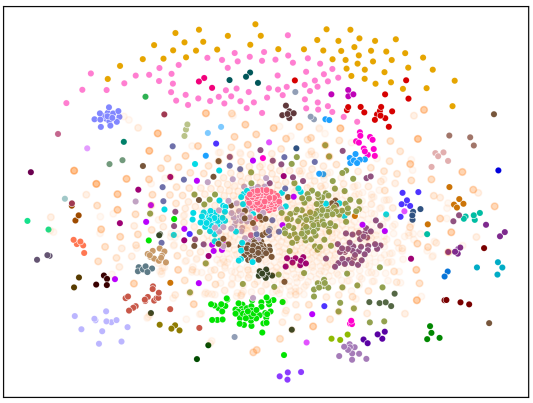
A. Appendix



(a) Detailed human labels

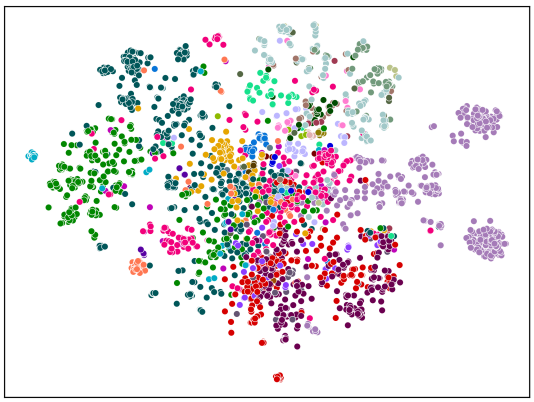


(b) General human labels

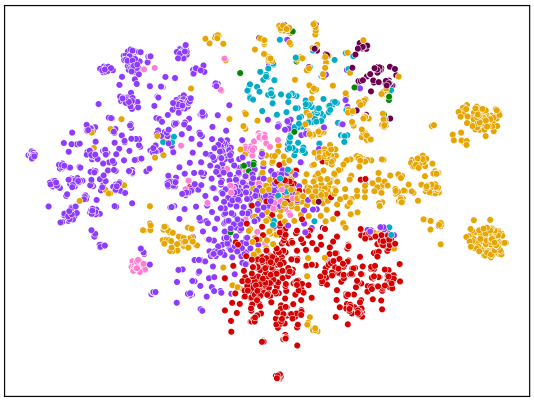


(c) HDBSCAN labels

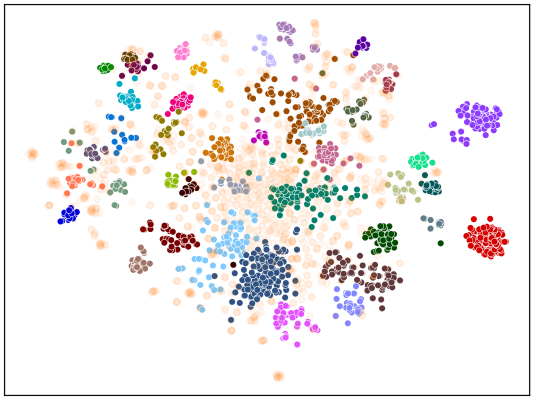
Figure 3: FastText scatterplots. Due to the number of classes the legend is not provided. For HDBSCAN, the outliers are marked with transluscent orange points.



(a) Detailed human labels



(b) General human labels



(c) HDBSCAN labels

Figure 4: KB-BERT scatterplots. Due to the number of classes the legend is not provided. For HDBSCAN, the outliers are marked with transluscent orange points.

CATEGORY	TOKENS	PHRASES
personal_name	**624**	**612**
firstname_male	234	228
firsname_female	289	287
firstname_unknown	49	47
middlename	1	1
surname	51	49
geographic	**1186**	**1135**
city	587	561
geo	17	17
country	401	399
place	112	94
region	39	36
street_nr	21	21
zip_code	9	7
institution	**160**	**111**
school	69	44
work	2	2
other_institution	89	65
transportation	**20**	**19**
transport_name	6	5
transport_nr	14	14
age	**94**	**94**
age_digits	82	82
age_string	12	12
date	**179**	**165**
day	27	27
month_digit	9	9
month_word	46	46
year	53	53
date_digits	44	30
other	**1085**	**940**
phone_nr	7	6
email	10	10
other_nr_seq	170	168
extra	40	32
prof	14	12
edu	7	5
fam	467	453
sensitive	370	254
TOTAL	**3348**	**3076**

Table 2: Token and phrase (MWE) counts for the PII spans in our data. General categories, given in **bold**, appear above the corresponding detailed labels.

Embedding	Clustering	Best parameters	Silhouette
FASTTEXT	HDBSCAN	cluster_selection_method='leaf', min_cluster_size=12	0.83
KB-BERT	HDBSCAN	cluster_selection_method='leaf', min_cluster_size=17	0.68
XL-LEXEME	HDBSCAN	cluster_selection_method='eom', min_cluster_size=14	0.72

Table 3: Best clustering algorithm and parameters per embedding type.

CATEGORY	FASTTEXT	KB-BERT	XL-LEXEME
personal_name	**356 (58.17%)**	**216 (35.29%)**	**87 (14.22%)**
firstname_male	135 (59.21%)	105 (46.05%)	39 (17.11%)
firsname_female	159 (55.40%)	75 (26.13%)	23 (8.01%)
firstname_unknown	28 (59.57%)	22 (46.81%)	11 (23.40%)
middlename	1 (100.00%)	-	-
surname	33 (67.35%)	14 (28.57%)	14 (28.57%)
geographic	**327 (32.78%)**	**594 (52.33%)**	**246 (21.67%)**
city	166 (29.59%)	273 (48.66%)	88 (15.69%)
geo	11 (64.71%)	12 (70.59%)	5 (29.41%)
country	123 (30.89%)	211 (52.88%)	118 (29.57%)
place	40 (42.55%)	51 (54.26%)	17 (18.09%)
region	19 (52.78%)	27 (75.00%)	10 (27.78%)
street_nr	10 (47.62%)	17 (80.95%)	6 (28.57%)
zip_code	3 (42.86%)	3 (42.86%)	2 (28.57%)
institution	**42 (37.84%)**	**49 (44.14%)**	**22 (19.82%)**
school	13 (29.55%)	23 (52.27%)	4 (9.09%)
work	2 (100.00%)	2 (100.00%)	1 (50.00%)
other_institution	27 (41.45%)	24 (36.92%)	17 (26.15%)
transportation	**9 (47.37%)**	**16 (84.21%)**	**9 (47.37%)**
transport_name	4 (80.00%)	5 (100.00%)	1 (20.00%)
transport_nr	5 (35.71%)	11 (78.57%)	8 (57.14%)
age	**21 (22.34%)**	**33 (35.11%)**	**3 (3.19%)**
age_digits	21 (25.61%)	26 (31.71%)	3 (3.66%)
age_string	-	7 (58.33%)	-
date	**81 (49.09%)**	**57 (34.55%)**	**41 (24.85%)**
day	13 (48.15%)	7 (25.93%)	7 (25.93%)
month_digit	4 (44.44%)	4 (44.44%)	3 (33.33%)
month_word	14 (30.43%)	23 (50.00%)	21 (45.65%)
year	31 (58.49%)	17 (32.08%)	9 (16.98%)
date_digits	19 (63.33%)	6 (20.00%)	1 (3.33%)
other	**251 (26.70%)**	**339 (36.06%)**	**153 (16.28%)**
phone_nr	2 (33.33%)	-	-
email	5 (50.00%)	-	1 (10.00%)
other_nr_seq	28 (16.67%)	92 (54.76%)	14 (8.33%)
extra	16 (50.00%)	17 (53.12%)	13 (40.62%)
prof	9 (75.00%)	3 (25.00%)	8 (66.67%)
edu	4 (80.00%)	5 (100.00%)	3 (60.00%)
fam	63 (13.91%)	83 (18.32%)	24 (5.30%)
sensitive	124 (48.82%)	139 (54.72%)	90 (35.43%)

Table 4: Outlier counts per embedding type (for its best associated clustering method) by original human-assigned class. The value in brackets is the % of all the phrases of this type that the outliers constitute. General classes are given in bold.

Modelling Legal Compliance in a Consent Wizard Application as Part of a Research-Centered and User-Oriented Data Infrastructure

Aliena Strathmann[†‡] **Marc-Levin Joppek**[§‡] **Maryam Mohammadi**[†‡] **Katja Politt**[†‡‖]
Paul T. Schrader[§‡] **Annett Jorschick**[†‡] **Hendrik Buschmeier**[†‡]
[†] Faculty of Linguistics and Literary Studies, [§] Faculty of Law, [‡] CRC 1646 'Linguistic Creativity in Communication', Bielefeld University, Bielefeld, Germany
[‖] Rostock University, Rostock, Germany
{firstname.lastname}@uni-bielefeld.de

Abstract

Recent research calls for data management infrastructures that explicitly operate within the bounds of ethical and legal constraints, and facilitate adherence to Open Science principles by integrating automated support for planning, collection, storage, use, reuse, and sharing of data within. Legal and ethical requirements of data processing have become increasingly complex, introducing administrative barriers to scientific research investigating data generated by human participants, which encompasses a vast majority of humanities research. In response to this, we present RUDI ("Research-centered User-oriented Data Infrastructure"), a modular framework grounded in an interdisciplinary approach informed by legal, computational and linguistic expertise. This paper introduces its first component: a configurable and dynamically adaptive consent form generator in the form of a wizard web application. We outline how legal aspects are modelled within, and highlight its concrete benefits for administrative aspects of research. Further, we discuss the contextualisation of data within the research domain by leveraging the use of standardised ontology within the framework.

Keywords: informed consent, data management, data life cycle, personal data, GDPR, linguistics, ontology

1. Introduction

Research involving data generated by human participants operates at the intersection of two normative commitments. On the one hand, Open Science practices seek to promote transparency and collaboration benefiting all of society, and 'FAIR' principles (Wilkinson et al., 2016) correspondingly dictate that research data should be Findable, Accessible, Interoperable, and Reusable. On the other hand, strict ethical and legal limitations on data collection, processing, management, use and re-use reinforced by the General Data Protection Regulation in the European Union (GDPR, 2016) and similar legislation adopted elsewhere (e.g., Brazil, Canada, Israel, California) aim to protect the privacy and autonomy of the individuals who are the 'data subjects', i.e., the original sources of said data. Both objectives are of considerable importance, but often in conflict with each other. Reconciling them presents a substantial challenge for researchers who collect, process, and store human-generated data.

In response, recent work has argued for data management infrastructures that embed legal and ethical requirements directly into user-friendly and comprehensive technical architecture that supports research workflows, integrates and promotes awareness of legal and ethical compliance aspects, and facilitates Open Science practices across research projects and throughout the entire data life cycle (e.g., Siegert et al., 2020; Kamocki and Witt, 2024; Jorschick et al., 2024).

In this context, we present 'RUDI' (Research-centered User-oriented Data Infrastructure), developed within the 'INF' project of the Collaborative Research Center CRC 1646 *Linguistic Creativity in Communication* at Bielefeld University, Germany, where heterogeneous study designs, data types, and participant populations across projects are the norm rather than the exception.

RUDI is an interdisciplinary infrastructure framework developed in close collaboration between legal, linguistic, and technical experts. Conceptually, it specifies how legal norms, ethical constraints, and research-specific requirements can be represented in a structured and machine-actionable way. Its central, practical goal is providing a comprehensive data management platform that implements these specifications and supports researchers in planning, data collection, storage, controlled access, use and re-use of human-generated data.

Core features of this data management platform should enable researchers to

(i) inform and automate the process of creating meaningful (i.e., GDPR-compliant and ethically sound) informed consent forms and related materials for participants,

(ii) collect, store, and access individual instances of participant consent pertaining to specific collected data points,

(iii) index and contextualise available data within a research domain in order to facilitate sharing

Proceedings of LEGAL2026 and CALD-pseudo 2026 @ LREC 2026, pages 73–80
12 May 2026. ©ELRA Language Resources Association (ELRA), 2026

and reuse in accordance with Open Science principles, and conversely

(iv) locate, retrieve and reuse available data within a research domain restricted to the boundaries of participant consent.

While originating within the field of linguistics and spanning the specific research domains of the CRC 1646, the platform is designed to be adaptable to the requirements of any field of research involving human participants.

In this paper, we present the first component of RUDI and the data management platform: the web-based consent wizard. The technical implementation of the principles outlined above adheres to established principles of interface design, in particular Nielsen's ten usability heuristics (Nielsen, 1994), and to ensure an efficient and user-friendly workflow. Additionally, we employ an iterative development process that incorporates continuous user testing and systematic integration of user feedback (Matera et al., 2006).

2. Consent Wizard Web Application

The first stage of the platform's implementation, mapping to its core feature (i), is the 'consent wizard', a dynamic and configurable web application that allows researchers to easily generate consent forms and other information material for participants that are tailored to the specific context and requirements of a study. They are both for participants to ensure informed consent as well as for the researchers to be informed of legal specifications pertaining to their research from the participants' perspective.

The consent wizard, pending user feedback integration as part of iterative development, is functionally implemented as a web application and (for now) stand-alone component of the data management platform.[1] It maps a set of legal properties to a study based on the researcher's inputs on a dynamically generated, questionnaire-style form, and automatically generates corresponding output documents for potential participants to review and sign.

The current implementation still requires the researcher's manual involvement both in having participants sign and then managing the resulting consent forms. At this point, researchers are the only immediate users of the wizard. Future versions will incorporate participants as users of the platform by providing them the possibility to fully or partially consent[2] from within the platform. This consent data is stored in a database, eliminating the need for manually distributing and managing (signed) legal documents.

The wizard application is where the vast majority of legal aspects of the framework are situated and modelled in. The following section details how this is implemented in practice.

3. Modelling Legal Compliance

With modularity as a core design principle of the platform infrastructure, the wizard's software architecture heavily relies on the use of configuration templates. This facilitates legal and ethical compliance by leveraging knowledge from law and ethics experts for configuration, as well as allowing for swift and comfortable adaptation of the wizard to future changes in legislation.

Dynamic form generation. The *core configuration template* of the wizard defines a set of legally and ethically relevant "properties" that a study may assume, with the choices informed by legal expert counsel. Examples of these properties include:

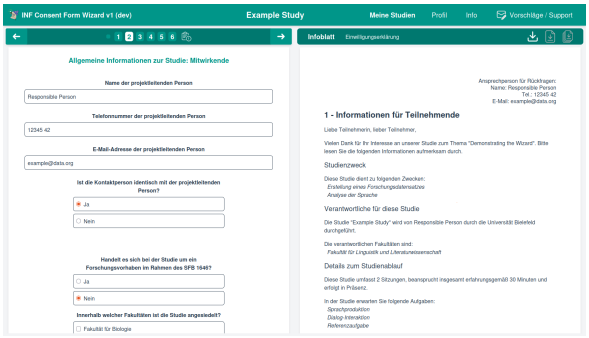
- processing of personal data according to the GDPR (see Section 3.2); represented as a binary value,
- permitted age groups of the participants; represented as a list of standardised string keys,
- participants under 14 years of age; represented as a binary value which may be programmatically inferred from the list of permitted age groups.
- whether legal guardian consent may be required for the participants; represented as a binary value.

A corresponding *form steps template* contains instructions for the program to dynamically assemble a questionnaire that maps the researcher's answers to this set of properties as concrete values, with the ability to conditionally render pages and questions (or other components) based on specific property values. Figure 1 exemplifies this: if the user specifies that the contact person of the project differs from the person responsible for the project, additional questions are displayed to collect the contact person's information.

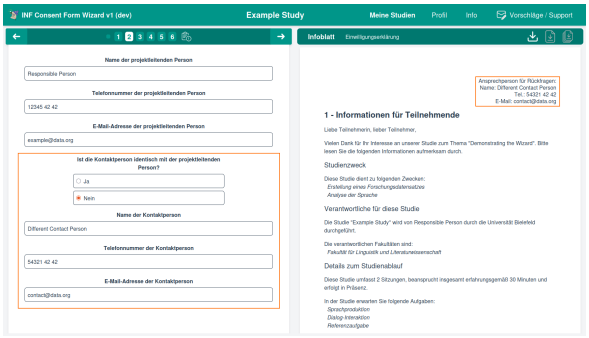
The wizard then uses these values to adapt text parts within the output documents intended for the participant, showing changes to the researcher in a live preview. The output forms, comprised of participant information and consent forms, are generated from configurable XML files based on document templates provided by a legal expert.

[1] https://purl.org/crc1646/RUDI-wizard

[2] Partially consenting means opting out of consent to specific data processing steps, e.g., publication or third-party sharing.



(a) The user specifies that the contact person is identical with the individual responsible for the project.



(b) The user specifies that the contact person differs from the individual responsible for the project.

Figure 1: Example demonstrating the consent wizard's conditional rendering of input form components. Additional input components are rendered in (b), and the output is adapted accordingly.

With every input, the wizard evaluates completeness and validity of the researcher's answers. If the form is determined to be sufficiently completed, the researcher is able to download the finalised consent and information forms in PDF format. In future versions it will be possible to make the output forms accessible to participants within the platform for review and electronic signing.

3.1. Modelling Legal Constraints

A third configurable template defines *autofill rules* based on legal constraints informed by expert counsel, and instructs the wizard to conditionally set and 'lock' certain properties based on present sets of property values per study instance.

Example of legal guardian consent. Study participation may require guardian consent under certain circumstances, e.g., for children or people with mental disabilities affecting their cognition (Schrader and Joppek, 2025). This is due to the fact that the GDPR requires the data subject's capacity to consent in order for consent to be effective. If the person giving consent is not capable of doing so, consent must be given by their legal representative. The GDPR does not define in detail when exactly the data subject is capable of giving consent and how this is determined. Essentially, it is important that the person giving consent is able to sufficiently understand the data processing that concerns them.

Usually, the researcher conducting the study must independently assess whether personal data are involved. Criteria such as the purpose, type, and scope of data processing as well as mental maturity can be included in the assessment. Nevertheless, it may not be possible to make an unequivocal judgment in individual cases. In cases of doubt, both the consent of the person concerned and that of their representative should therefore be obtained as a precautionary measure.

Fixed age limits may be considered as a possible solution. In the case of consent by minors, Art. 8 GDPR provides partial age thresholds for processing operations in certain contexts. However, due to the narrow scope of the provision, these requirements cannot be easily transferred to consent into research contexts and instead provide, at best, rough guidance. For other vulnerable groups, such as people with mental disabilities, such age limits do not apply. It therefore remains that the determination of capacity to consent is fundamentally case-specific.

Guardian consent within the wizard. The wizard explicitly asks whether the researcher presumes that legal guardian consent may be required, with an expandable information display summarising the legal situation outlined above. In our present *autofill rules* configuration template, the requirement of legal guardian consent is automatically set for participants under 14 years of age, or for participants under 18 years of age if the researcher has selected any special risks to the participants' mental or physical well-being associated with study-participation.

Based on this configuration setting, the autofill evaluation routine determines for each study instance whether either of the above conditions applies,[3] and locks the binary choice component to affirmative input, as shown in Figure 2. Furthermore, a corresponding explanation is displayed inside a yellow box and a clickable information display additionally outlines relevant legal situations where guardian consent may be required as described above. The resulting output documents are adapted accordingly, with additional forms for parents or legal guardians to sign.

[3]This mechanism is presently simplified, as we only have access to information pertaining to the study, not yet to information pertaining to the individual participant. Future versions that integrate participants as direct users of the platform will adapt this.

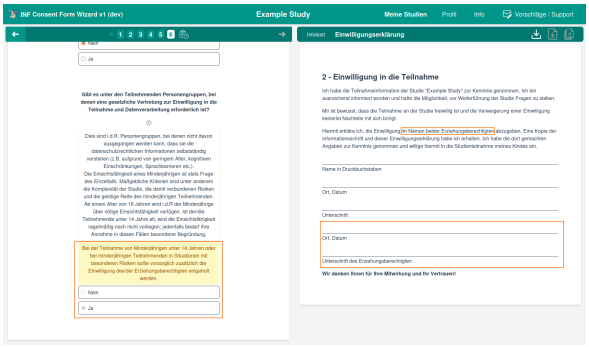


Figure 2: Example demonstrating a locked input component as a result of the autofill evaluation routine. The binary choice input is locked based on the researcher's prior input, with a corresponding explanation shown in a yellow box.

This same process can be applied to any number of explicit legal constraints based on an arbitrary set of determining parameters. Values that are conditionally (and automatically) set may correspond to (and override) parameter values that the user may otherwise set directly, as demonstrated in the example. In principle, they may also map to latent parameters without being directly bound to any specific input component; allowing full administrative control over setting conditions that determine the flow of information via the input form steps and study parameter templates.

3.2. Considerations for Personal Data

An important aspect of data management guidance for research projects in general, but especially for the consent wizard, is to clearly distinguish between *personal* and *non-personal* data in a way that is understandable to researchers. Understanding what is considered personal data is of central importance: the GDPR (2016) applies only when personal data are actually processed in accordance with Art. 2 para. 1 GDPR. This means, among other things, that the processing must be carried out in accordance with the principles of Art. 5 para. 1 GDPR, be based on a legal basis in accordance with Art. 6 para. 1 GDPR and the data subjects must be adequately informed about the data processing in accordance with Art. 13, 14 GDPR.

The term 'personal data' is defined in Art. 4 (1) GDPR. According to this, personal data are any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is someone who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier, or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person. Recital 26 GDPR specifies this as follows: to determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. Ultimately, whether an information is relatable to an individual must be determined from the controller's perspective. As a consequence, pseudonymised data cannot necessarily be regarded as personal data in every case (C-413/23 P, para. 86; of Justice of the European Union, 2025). Whether or not personal data is processed in a specific situation, often depends on the circumstances of the individual case. For reasons of transparency, obtaining consent under data protection law purely as a precautionary measure and providing data protection information in accordance with Art. 13, 14 of the GDPR without first reliably establishing whether personal data are in fact being processed should be avoided. Nevertheless, even in cases of exclusively anonymous data processing, consent to study participation is advisable for ethical reasons (e.g., Gauthier et al., 2010).

Data protection training courses or fact sheets are particularly suitable for communicating to researchers who are not legally experienced which data is considered personal data in individual cases. These measures can initially create a sound basic understanding. If there are still uncertainties in a specific case, these can be resolved through low-threshold counselling if necessary. In the long term, however, the aim is to minimise external input as much as possible: The categorisation of whether or not personal data is being processed should be computationally supported within the tool.

Personal data within the wizard. We currently determine whether personal data are processed by letting the researcher make this distinction via a binary selection. In cases of uncertainty, we refer the user to a third-party tool ('iVA'; Herklotz and Oberländer, 2022), which guides them through a four-step decision process. In future versions, this decision process will be integrated into the wizard's own questionnaire.

If personal data are determined to be processed, information forms are required to reflect

- which particular personal data is collected for the purpose of the research goal, and
- the period for which the data will be stored,

as well as further information concerning data processing, such as sharing and publication. When this is the case, the questionnaire steps of the wizard are expanded accordingly, and a separate information form on the processing of personal data is

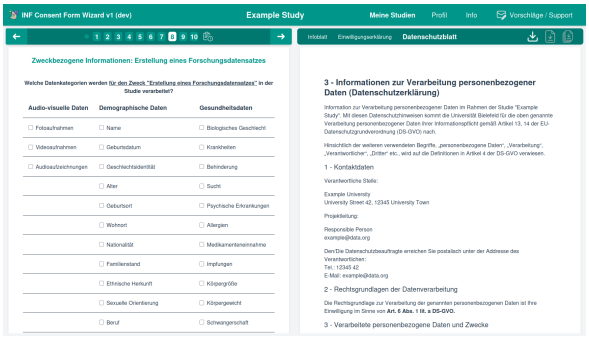


Figure 3: Example showing one of the extrapolated questionnaire steps if personal data is collected as part of the study, inquiring about specific personal data types collected and the data retention period per data processing goal. An additional information form with GDPR-specific information for the participant is populated with the additional input, and added to the set of output documents.

added to the output documents (see Figure 3) to comply with Art. 13, 14 of the GDPR.

3.3. Availing Legal Expertise to Users

Besides *researchers* who design studies and collect, use, and manage data and represent the main user group of the current stage of platform development, *users* also encompass study *participants* who provide data ('data subjects' in GDPR) in future development stages. For both user groups, we generally assume limited legal expertise.

For researchers, breaking down legal concepts into modular sets of questionnaire components and omitting any aspects that are irrelevant to their specific situation substantially reduces the burden of navigating legal considerations as laypeople. However, beyond just alleviating administrative workload by automating the consent and data management process, our goal is also to enhance awareness of legal aspects by embedding educational support seamlessly into the interaction process within the platform. To this end, the wizard features the ability for legal experts to embed extendable information displays into each questionnaire component via the form steps template: Figure 4 shows an example component that asks for (optional) information pertaining to the disclosure of the study's source of funding. When clicked, the display informs the researcher that the source of funding should be disclosed especially if there is reason to assume that this knowledge might influence the participant's decision. Another example of an information display is shown in Figure 2.

Additionally, the wizard's form steps template allows for full configurability in displaying conditional warnings and reminders: for example, a visually highlighted warning is shown that any *published*

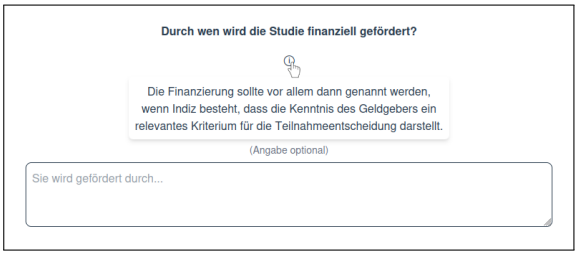


Figure 4: Example showing an information display for a text input component of the wizard questionnaire, pertaining to when disclosure of the study's source of funding is advisable. Clicking the info button displays the additional text box.

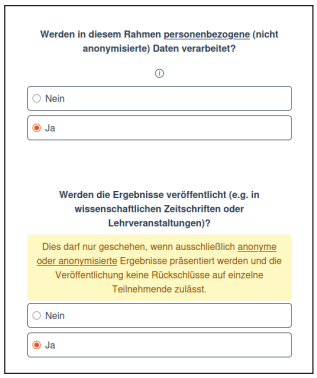


Figure 5: Example of a visual highlight warning, which is only displayed if both binary choice components shown in the image are answered affirmatively by the user.

results must always be anonymous or anonymised, and may never allow for drawing conclusions about the identity of individual participants (see Figure 5). This is only shown if the researcher selects that (a) personal data is being processed for the study, as well as (b) there is an intent to publish the results, e.g., in scientific journals, or use them within lectures or seminars. This is achieved by outfitting input components with warning text and corresponding 'warnIf' conditions within the form steps template, which are dynamically evaluated at runtime based on the current study parameter values.

This dynamic evaluation of study parameter conditions also allows for displaying situational suggestions for best practices and measures that may simplify the downstream data management process. Examples of this are yet to be implemented in the current version of the wizard; we are presently attempting to determine specific sets of personal data collection conditions that allow for anonymisation or pseudonymisation, which would both protect the data subject's privacy and avoid special consideration requirements under GDPR. Those conditions could then be embedded into the template and algorithmically flagged, and enable the application to inform the researcher of the possibility when given.

Another related feature that promotes researcher awareness of legal ramifications is the automated, conditional setting of values and locking of inputs based on a separate constraints template, as shown in the example already mentioned in Section 3.1: if a specific set of values set earlier dictates a future value, the researcher is informed by clear visual feedback and explanation (see Figure 2).

Finally, to benefit both user groups, great care is taken to translate legal concepts into easily understood (albeit legally accurate) language, tailored to their respective perspectives; this reflects the position of the European Data Protection Board (EDPB) that consent-related information must be formulated in clear and plain language that is understandable to the average person (Board, 2020, p. 18). The use of configurable templates where possible ensures a quick integration and addressing of user feedback from both groups, e.g., by amending phrasing that is perceived as confusing by users in either the study questionnaire or output forms.

4. Integrating Linguistic Perspective

Modular configurability of the platform facilitates the iterative development process and accommodates various legal and ethical aspects – but beyond that, it also provides crucial infrastructure for integrating aspects and requirements that are relevant to the specific research domain. Mapping studies to a modular set of data collection-related properties enables multifaceted downstream processing, including contextualisation of the collected data within a larger research ecosystem. Combined with integrated recording and storage of participants' individual consent choices, it allows for automated decisions regarding which operations (e.g., processing, sharing, publication, or controlled access) are permissible for specific datasets or even individual 'data points' (Jorschick et al., 2024).

This integrated approach sets our tool apart from prior approaches that implement standardised guidance, such as the 'DARIAH Consent Form Wizard' (Hannesschläger et al., 2020), which also supports template-based generation of consent forms, or 'Ethiktool' (Bendixen et al., 2025, 2026), which provides software-guided collection of information relevant for ethics review while generating participant information and privacy-related documents. Both tools treat the generated documents as the final output, rather than as one step in the data life cycle.

4.1. Leveraging Ontologies

Following the FAIR principles, we investigate an ontology-based implementation to enhance the findability of collected data. The use of inconsistent terminology risks creating broken links between related data and may result in data being lost in search processes. Principles of data visibility and reusability require that newly collected data are semantically linked to existing resources. This is particularly important here, given that RUDI comprises multiple modules which must work together consistently and integrate seamlessly with external resources. Ontological resources are thus shared across all modules. Mohammadi et al. (2026) identify the relevant data types and associated (meta)data within the linguistics domain. Although such information could be coded directly into the platform, we use ontologies, taxonomies and controlled vocabularies, to facilitate semantic alignment and connect our data to the broader semantic Web.

Given the growing need to address personal data considerations in different domains, numerous authors have proposed corresponding legal taxonomies (e.g., Pandit et al., 2019). Since linguistic demographic data substantially overlap with categories of personal data, we adopt and adapt established ontologies and standards wherever possible. For widely used classifications, we rely on ISO standards, including ISO 639-3 for language codes (ISO, 2023) and ISO 3166-1 for country codes (ISO/IEC, 2020). We also use domain-specific vocabularies such as BioPortal and the WHO International Classification of Diseases for medical information. To model personal data, we use the GDPR-aligned Data Privacy Vocabulary (DPV; Pandit et al., 2025; Esteves et al., 2025).

Within the context of the CRC 1646 research domains, we are developing the eXperimental Linguistics taxonomy (XLing), which defines a minimal set of field-specific concepts. Crucially, XLing entries are linked to CLARIN vocabularies to enable future integration and reuse. We employ established semantic web standards, including Resource Description Framework Schema (RDFs), Dublin Core Terms (dcTerms), and the Simple Knowledge Organization System (SKOS). Notably, these schemata have been implemented dynamically within the wizard application and will be extended to downstream components of the platform. This ensures that additions or changes in values can be integrated immediately at any stage, while practical usage of the platform can, in turn, inform further development of the ontologies.

5. Summary and Future Perspectives

In this work, we have introduced the present implementation of our consent wizard application, which is being developed as the first component of a more comprehensive, modular and configurable research data management platform. As a central, practical goal, this web-based platform is situated

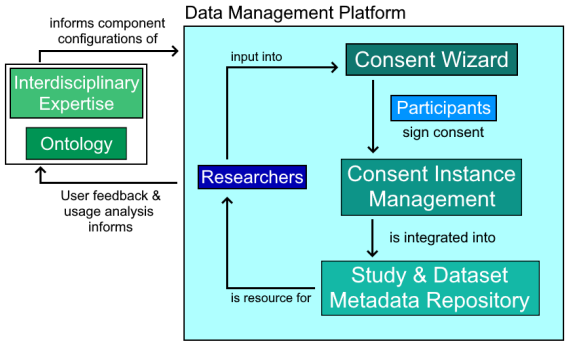


Figure 6: A visualisation of the planned data management platform, its components and its users situated within the Research-centered User-oriented Data Infrastructure (RUDI).

within our interdisciplinary infrastructure framework RUDI, which aims to support user-friendly linguistic research data management throughout the entire data lifecycle, by intuitively integrating the legal and ethical requirements of data management and facilitating Open Science practices and the FAIR principles. Figure 6 provides a simplified, general visual overview of how the wizard and the data management platform integrate into RUDI.

We have detailed how legal and ethical aspects of data processing are modelled and presented within the existing wizard application, and how they interface with linguistic research domain contextualisation aspects. Further, we have discussed the role of standardised ontology within RUDI, presented XLing as a taxonomy that is tailored to the specific research context of the CRC 1646, and how it integrates into the wizard component of the data management platform.

Further platform development. Pending server-side deployment, the consent wizard is to enter its first iterative development cycle with feedback from researcher users. The next concrete development milestone of the surrounding data management platform is the integration of an online database. Persistent, server-side storage enables implementing user management, which in turn enables (i) integrating participants as direct users, allowing for direct signing and storage of consent instances within the application and thus introducing the second core component of the platform, and (ii) sharing of study design templates between researchers. It also allows for collection and evaluation of usage (meta)data of the wizard and platform, laying the groundwork for further iterative platform development and meta-analysis within the surrounding research domain.

Following this, we plan to integrate the consent management capabilities into an open repository of metadata about empirical datasets; expanding the platform to act as a hub that facilitates discovery of legally reusable data and collaboration between the projects within its domain, thus fully encompassing the operationalisation of both legal compliance and Open Science practices.

General future goals of the project. While the platform is intended to be a stand-alone application, a general objective of the project is to establish compatibility with existing infrastructure supporting research workflows specific to Bielefeld University and, in the long term, with data infrastructures across Europe. An immediate goal is localisation of the wizard, the platform, and its configurations into English, which poses the challenge of creating legally accurate translations of, e.g., the wizard's input forms and output documents.

Pending completion of development, we intend to release the source code of the data management platform as a fully configurable and research domain-agnostic open-source project.

6. Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): SFB 1646/1 2024 – 512393437, project INF.

7. Bibliographical References

Alexandra Bendixen, Eva-Maria Berens, Thomas G. G. Wegner, Wolfgang Einhäuser, and Katarina Blask. 2025. Data (re-)use in human-participant research: Guided composition of informed consent forms. In *Abstracts of the 2nd Conference on Research Data Infrastructure (CoRDI)*. Zenodo.

Alexandra Bendixen, Thomas G. G. Wegner, and Wolfgang Einhäuser. 2026. Facilitating ethics application and review for interdisciplinary human-participant research via software-based guidance and standardization. In Bertolt Meyer, Ulrike Thomas, and Olfa Kanoun, editors, *Hybrid Societies: Humans Interacting with Embodied Technologies*, volume 1, pages 311–317. Springer, Cham, Switzerland.

European Data Protection Board. 2020. Guidelines 05/2020 on consent under Regulation 2016/679. Technical report, European Data Protection Board. Adopted on 4 May 2020.

Beatriz Esteves, Delaram Golpayegani, Georg P. Krog, Harshvardhan J. Pandit, Julian Flake, and Paul Ryan. 2025. Data Privacy Vocabulary (DPV), version 2.2. Final community group report, World Wide Web Consortium, Wakefield, MA, USA.

Janel Gauthier, Jean Pettifor, and Andrea Ferrero. 2010. The universal declaration of ethical principles for psychologists: A culture-sensitive model for creating and reviewing a code of ethics. *Ethics and Behavior*, 20:179–196.

GDPR. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council. *Official Journal of the European Union*, L 119:1–88.

Vanessa Hannesschläger, Walter Scholger, and Koraljka Kuzman. 2020. The DARIAH ELDAH consent form wizard. In *DARIAH Annual Event 2020: Scholarly Primitives*. Zenodo.

Markus Herklotz and Lars Oberländer. 2022. iVA: Ein interaktiver Virtueller Assistent von BERD@BW zur Aufbereitung von Rechtsfragen im Bereich Open Science. In Vincent Heuveline and Nina Bisheh, editors, *E-Science-Tage 2021: Share Your Research Data*, page 306–313. heiBOOKS.

ISO. 2023. ISO 639:2023 – Code for individual languages and language groups. International Standard 639:2023, International Organization for Standardization (ISO), Geneva, Switzerland.

ISO/IEC. 2020. ISO/IEC 3166-1:2020 – Codes for the representation of names of countries and their subdivisions – Part 1: Country code. Standard 3166-1:2020, International Organization for Standardization (ISO), Geneva, Switzerland.

Annett Jorschick, Paul T. Schrader, and Hendrik Buschmeier. 2024. What can I do with this data point? Towards modeling legal and ethical aspects of linguistic data collection and (re-)use. In *Proceedings of the Workshop on Legal and Ethical Issues in Human Language Technologies @ LREC-COLING 2024*, pages 47–51, Torino, Italy. ELRA and ICCL.

Pawel Kamocki and Andreas Witt. 2024. Ethical issues in language resources and language technology – New challenges, new perspectives. In *Proceedings of the Workshop on Legal and Ethical Issues in Human Language Technologies @ LREC-COLING 2024*, pages 19–23, Torino, Italia. ELRA and ICCL.

Maristella Matera, Francesca Rizzo, and Giovanni Toffetti Carughi. 2006. Web usability: Principles and evaluation methods. In Emilia Mendes and Nile Mosley, editors, *Web Engineering*, pages 143–180. Springer, Berlin, Germany.

Maryam Mohammadi, Katja Politt, and Annett Jorschick. 2026. Assessing data management and compliance in large research collaborations via knowledge bases: A semi-structured interview approach. *F1000Research*, 15:37. [v1; peer reviews: 2 'approved with reservations'].

Jakob Nielsen. 1994. Heuristic evaluation. In Jakob Nielsen and Robert L. Mack, editors, *Usability Inspection Methods*, pages 25–62. Wiley, New York, NY, USA.

Court of Justice of the European Union. 2025. Judgment of 4 September 2025, Case C-413/23 P. ECLI:EU:C:2025:645, para. 86.

Harshvardhan J. Pandit, Beatriz Esteves, Georg P. Krog, Delaram Golpayegani, and Julian Flake. 2025. Data Privacy Vocabulary (DPV) – version 2.0. In *The Semantic Web – ISWC 2024*, pages 171–193, Cham, Switzerland. Springer.

Harshvardhan J. Pandit, Axel Polleres, Bert Bos, Rob Brennan, Bud Bruegger, Fajar J. Ekaputra, Javier D. Fernández, Roghaiyeh Gachpaz Hamed, Elmar Kiesling, Mark Lizar, Eva Schlehahn, Simon Steyskal, and Rigo Wenning. 2019. Creating a vocabulary for data privacy: The first-year report of data privacy vocabularies and controls community group (DPVCG). In *Proceedings of the 18th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE 2019)*, pages 714–730, Rhodes, Greece. Springer.

Paul T. Schrader and Marc-Levin Joppek. 2025. Datenbasierte forschung und einwilligungsunfähige. zulässigkeit der verarbeitung von daten minderjähriger und geistig eingeschränkter personen. *Zeitschrift für Datenschutzrecht*, 2025:613–618.

Ingo Siegert, Vered Silber-Varod, Nehoray Carmi, and Pawel Kamocki. 2020. Personal data protection and academia: GDPR issues and multimodal data-collections "in the wild". *Online Journal of Applied Knowledge Management*, 8:16–31.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018.

Balancing FAIR and GDPR: a governance framework for oral archives

Elvira Mercatanti°, Monica Monachini°, Giovanni Abete', Silvia Calamai", Sergio Canazza°°, Alessandro Casellato*, Virginia Niri", Cesarina Vecchia', Giulia Zitelli Conti*, Giada Zuccolo°°

CNR-ILC, Pisa°, Università degli Studi di Napoli Federico II', Università degli Studi di Siena"
Università degli Studi di Padova°°, Università Ca' Foscari Venezia*
elviramercatanti@cnr.it, monica.monachini@cnr.it, giovanni.abete@unina.it,silvia.calamai@unisi.it,
canazza@unipd.it, casellat@unive.it, virginia.niri@unisi.it, cesarina.vecchia@unina.it,
giulia.zitelliconti@unive.it, giada.zuccolo@studenti.unipd.it

Abstract

This paper presents a governance framework developed within the research project ROADS to support the sustainable management of oral archives, which constitute essential linguistic resources for interdisciplinary research and cultural heritage preservation. Oral archives raise complex ethical and legal challenges due to the hybrid nature of voice data, which function simultaneously as historical documents, scientific sources and biometric identifiers, thereby creating tensions between open science principles and data protection regulations. The proposed framework integrates FAIR principles (Findable, Accessible, Interoperable, Reusable) with Privacy by Design and the GDPR accountability principle through a multilayered approach. It introduces an access model that distinguishes between publicly available metadata and controlled access to identifiable audio materials, following trusted repository standards. The framework also incorporates consent management procedures and safeguards for legacy collections, enabling responsible data sharing while preserving scientific usability. More broadly, ROADS provides a transferable model to guide the transition from project-based archives to FAIR, sustainable and reusable research resources, ensuring compliance with data protection requirements and respect for the sensitivity of the documented contexts.

Keywords: GDPR, speech data, ethics, oral archives, FAIR principles, language resources

1. Introduction

Access to speech data is essential for linguistic research, language technologies and cultural heritage preservation (De Dominicis, 2002; Sornicola et al., 2019). A central challenge in oral archives lies in the hybrid nature of voice data: audio recordings simultaneously function as historical documents, scientific sources and biometric identifiers, creating tensions between open science principles and data protection regulations.

Managing oral archives requires an interdisciplinary approach integrating humanistic, archival, technological and legal expertise. Researchers must reconcile data openness and reuse with the protection of data-subject rights while ensuring long-term sustainability and integration into Digital Humanities ecosystems. In Italy, the digitization and reuse of oral archives remain limited due to heterogeneous practices, local preservation constraints and regulatory uncertainty. Recent regulatory frameworks, particularly the General Data Protection Regulation (GDPR), have reshaped how sensitive linguistic resources can be collected, processed and shared. Ethical principles such as fairness, transparency and trust are no longer abstract ideals but concrete design constraints for research infrastructures.

Within this context, the ROADS project has developed a governance framework to overcome the fragmentation of Italy's oral heritage through standardized scientific and regulatory practices. To guarantee the long-term sustainability of the model, ROADS relied on legal experts embedded within the participating institutions-professionals capable of mediating between the philosophy of Open Science and the stringent constraints imposed by the GDPR. Their contribution has been crucial in ensuring the legal integrity of the resources produced, making them fully legitimate and transparent scientific sources for future generations (Abete et al., 2026).

The paper is structured as follows: Section 2 describes ROADS and its pilot archive, together with the challenges encountered; Section 3 presents the proposed solutions for FAIR- and GDPR-compliant reuse; Section 4 outlines the conclusions.

2. The ROADS project as a model for oral data

ROADS[1] is a national Italian Project of Relevant National Interest designed to coordinate and sustain Italy's oral heritage by developing models and tools for the recovery, preservation, description and scholarly reuse of oral archives, adopting a FAIR-by-design approach integrating technical, ethical

[1] https://csc.dei.unipd.it/roads-project/index.html

Proceedings of LEGAL2026 and CALD-pseudo 2026 @ LREC 2026, pages 81–85
12 May 2026. ©ELRA Language Resources Association (ELRA), 2026

and legal considerations from the beginning of the data lifecycle (Abete et al., 2025b). Core project activities include a national survey of oral archives in Italian public universities, development of a management and access infrastructure, deposit of a pilot archive and training initiatives to build capacity for sustainable stewardship of oral sources (Abete et al., 2025a).

ROADS addresses two types of sources: (i) pre-existing historical oral archives, collected before the current regulatory framework; (ii) new oral interviews collected within the project, targeting a representative selection of researchers to gather information on the genesis of collections, their size and composition, archival criteria, and conservation status. This dual perspective enables differentiated methodological and legal solutions tailored to distinct data-production contexts, helping make FAIR principles operational (Calamai and Frontini, 2018; Wilkinson et al., 2018) even when normative and ethical constraints are heterogeneous.

A key element of the governance framework is the Data Management Plan (DMP), which specifies how data are collected, documented, preserved and shared. The DMP applies FAIR principles across all stages of the data lifecycle and links them to legal, ethical and technical requirements. It defines procedures for consent, anonymisation, licensing and access control, ensuring GDPR compliance and safeguarding sensitive historical and biometric information. This ensures that resources are usable for research while respecting the rights and privacy of data subjects.

2.1. The pilot archive

The selected pilot archive is the research collection of historian Gabriella Gribaudi[2], based on fieldwork conducted since 1974 on the social history of Southern Italy during World War II (Gribaudi, 1980, 1990, 2005, 2016, 2023). The collection comprises 148 audio carriers (audiocassettes, minicassettes, and digital audio tapes) and approximately 189 interviewees (born 1909-1945), including both direct witnesses and individuals who reported family narratives. Beyond its historical relevance, the archive is particularly suitable for linguistic analyses, as it contains rich, naturally produced speech with speaker and contextual diversity supporting phonetic, sociolinguistic, and discourse-oriented studies. This case provides a realistic benchmark for implementing differentiated access and reuse policies for historically sensitive, inherently identifiable oral data.

[2]Professor at the University of Naples Federico II and founder and first president of the Italian Oral History Association.

2.2. Challenges requiring legal and governance measures

Oral archives in the Italian context are often preserved locally, described with heterogeneous metadata practices, and shared under unclear conditions. This hinders discoverability, long-term sustainability, and reuse. A major source of complexity lies in the intrinsic nature of oral data: audio recordings contain not only biographical information and opinions, but also biometric traits (voice) and dialectal or cultural cues.

Within the project, a further limitation arises from the coexistence of legacy collections and newly collected data. Legacy recordings, such as the Gribaudi archive, were produced before the GDPR and often lack documentation that would now be expected (e.g., explicit consent forms, clear information on intended dissemination, and standardized provenance). In many cases, data subjects cannot be contacted due to decease or irretrievability, which makes it impossible to update consent and requires careful legal framing and proportionate safeguards. By contrast, newly collected interviews can be designed to meet GDPR transparency and accountability requirements from the outset, but they still contain inherently identifiable voice data and potentially sensitive contextual information. These two regimes create a practical governance challenge: a single infrastructure must support FAIRness and scientific usability while applying differentiated access and reuse rules aligned with the origin, documentation, and sensitivity of each dataset.

3. Solutions for FAIRness, legal and ethical compliance

Managing personal data within ROADS required fulfilling regulatory obligations to ensure legitimate research activity. These steps constitute the backbone of a protection system that safeguards individuals. In line with the accountability principle, ROADS formalized roles among partners, identified legal bases, defined secure preservation protocols, and implemented transparent information flows to data subjects.

3.1. Legal framing and governance

The project follows two parallel methodological tracks: the ethical valorization of historical oral archives and the collection of new testimonies, both aimed at building a sustainable and secure research infrastructure compliant with FAIR principles. Given the multicentric and interdisciplinary nature of the project, governance has been formalized through a joint controllership agreement

(Art. 26 GDPR) among ROADS partners. This instrument clarifies each partner's responsibilities, appoints a single contact point for exercising data-subject rights (Arts. 15–22 GDPR), and identifies the national research infrastructure CLARIN-IT as the technological custodian supporting long-term preservation.

One of the main challenges is that informed consent cannot be obtained for part of the legacy materials because many interviewees are deceased or no longer traceable. ROADS addresses this issue through a formal "Diligent Search" protocol, based on public notices on institutional websites that inform data subjects (or their heirs) about digitization and reuse while preserving the right to object on legitimate grounds. Although the GDPR does not apply to deceased people, this approach is aligned with Art. 2-terdecies of the Italian Privacy Code, which enables heirs to exercise the data subject's rights post mortem. The framework further balances the public interest in protecting cultural heritage with privacy safeguards by minimizing or removing identifying metadata and restricting access to full audio through controlled procedures (e.g., excerpts or partial access), thereby reducing the risk of unlawful or inappropriate reuse.

3.2. A multilevel system of legal bases

ROADS relies on an integrated set of legal bases, calibrated to the nature of the processing and the institutional mandate of the partners:

- Scientific research (Art. 6(1)(e) GDPR): core research activities are grounded in the public interest. This legal basis derives from a combined reading of the GDPR and Art. 2-ter of the Italian Privacy Code (D.Lgs. 196/2003), which recognizes scientific and historical research as a primary institutional function of Universities and Public Research Bodies. This framework ensures that processing is not contingent upon individual withdrawal where the data serves a broader collective scientific purpose, provided that the principle of data minimization is strictly observed.

- Archiving and Historical Research (Art. 9(2)(j) GDPR; Art. 2-sexies Italian Privacy Code): this legal basis is particularly relevant for legacy collections where consent cannot be obtained due to the decease of the data subject or their untraceability. In such cases, the framework incorporates Art. 2-terdecies of the Italian Privacy Code, which provides that heirs may exercise the relevant data protection rights on behalf of the deceased. Appropriate safeguards are implemented through a formal "Diligent Search" protocol (e.g., public notices and the right to object), combined with robust technical and organizational measures, including controlled access via ILC4CLARIN. In particular, the protocol requires the anonymization of descriptive metadata in legacy archives, ensuring that identifying references are protected and not publicly accessible, while restricting full audio access to authenticated researchers in order to mitigate risks such as unauthorized voice harvesting.

- Informed Consent and releases (Art. 6(1)(a) GDPR): in coordination with applicable copyright provisions, this legal basis governs optional and ancillary processing activities. Adopting a granular approach, ROADS distinguishes between essential research operations and specific authorizations—such as video dissemination, third-party reuse, or potential commercial exploitation—which remain fully revocable by the data subject at any time.

To illustrate this model, consider the ILC4CLARIN repository workflow for video interviews. The system adopts a differentiated access model: a curated "Partial Version" is made available for public dissemination, while the complete recording is distributed under a "Restricted" license. Access to the full version requires institutional Single Sign-On (e.g., IDEM/Edugain), thereby preventing unauthorized voice harvesting and ensuring traceability in line with GDPR accountability requirements. For legacy data lacking valid consent, the "Diligent Search" protocol is systematically applied.

3.3. Operational implementation: acknowledgement, stratified consent and transparency

The main complexity lies in the intrinsic nature of oral data: audio recordings may contain not only biographical information and opinions, but also biometric traits (voice) and dialectal or cultural cues. This requires a granular governance approach that distinguishes what is necessary for science from what pertains to dissemination. To translate this complexity into practice, the project implements a stratified consent architecture (Modules 01, 02, 03) that supports informed control:

MOD 01 - participation and originality: formalizes participation and a declaration that provided content does not infringe third-party rights.

MOD 02-A - mandatory acknowledgement of the information notice: supports scientific transparency and accountability, including awareness that identifying metadata may be publicly available for scholarly attribution and long-term findability.

MOD 02-B and MOD 03 - optional modules: separate choices on re-contact, dissemination-oriented

video use, and third-party reuse for external scientific/didactic purposes.

3.4. Security, preservation and scientific authorship

In accordance with what is stated in the information notice, and following the participant's acknowledgment of the FAIR protocols, the security of data processing is ensured through a multi-level, accredited access system designed to balance individual protection with the needs of the scientific community:

- Attribution and traceability: since these are methodological contributions provided by scholars, the identifying metadata (name and affiliation) are made public. This step is essential not only to meet FAIR requirements, but also to ensure the proper scientific authorship of the collected testimony and its traceability over time. The partial version of the archives, comprising descriptive metadata and curated audio excerpts, is available in Open Access to the general public. This version is indexed by general search engines to ensure maximum findability, but it is carefully processed to reduce the risk of re-identification.
- Multi-level and accredited access: while attribution is public, the full audio/video files are deposited in a restricted-access digital archive. Access is protected by authentication protocols and reserved exclusively for the international scholarly community, preventing indiscriminate dissemination of the content and of biometric data beyond the research context. Access to these integral versions is strictly shielded: they are not indexed by search engines and are accessible only to verified scholars through strong authentication protocols (e.g., IDEM/Edugain/Shibboleth) via the ILC4CLARIN certified repository. This prevents indiscriminate dissemination and automated voice-harvesting, limiting use to verified research contexts.
- Preservation: the data are deposited in the certified repository, which ensures protection against unauthorized access and the safeguarding of the information assets beyond the duration of the project, in accordance with Article 99 of the Privacy Code for purposes of public interest. Specifically, data processed for historical or scientific purposes are preserved indefinitely as a public information asset, provided they are not used for decisions affecting the specific data subject.

4. Conclusions

ROADS demonstrates that managing oral archives is a complex interdisciplinary challenge, combining legal compliance, ethical accountability and robust infrastructure. Challenges addressed include the hybrid nature of voice data, which simultaneously function as historical testimony, scientific sources and sensitive biometric identifiers, and the temporal stratification of archives, requiring harmonization of legacy pre-GDPR collections with current European standards.

The approach preserves the integrity of recordings in the transition from analog to digital formats and defines specific protocols to protect the privacy of witnesses, often deceased or unreachable. Structured governance, multilevel access, and rigorous transparency protocols make ROADS a reference model for Digital Humanities, ensuring historical memory is preserved within a fully legal and ethically robust ecosystem.

From a research-infrastructure perspective, the project operationalizes FAIR-by-design for inherently identifiable speech: public metadata enable discovery and attribution, while controlled access preserves scientific usability and data-subject rights. Components such as role allocation, joint controllership, calibrated legal bases, access tiers, and documentation practices can be replicated in other projects dealing with sensitive speech or oral history collections. Currently, the application and replicability of the defined model are being tested on project archives to support their transition into FAIR, sustainable and GDPR-compliant resources.

5. Acknowledgments

This work is supported by the PRIN PNRR 2022 project ROADS (MUR P20229S48H), by CLARIN-IT, the Italian node of the CLARIN ERIC research infrastructure, and by the H2IOSC Project-Humanities and Cultural Heritage Italian Open Science Cloud, funded by the European Union-NextGenerationEU (NRRP M4C2, project code IR0000029;see the project website at https://www.h2iosc.cnr.it/). We gratefully acknowledge Rosaria Deluca, Responsible for the Privacy Service of the CNR Research Area in Pisa, for her expert legal guidance and support of the ROADS project. Her advice and assistance were indispensable to our work.

6. Bibliographical References

Giovanni Abete, Silvia Calamai, Sergio Canazza, Alessandro Casellato, Elvira Mercatanti, and Monica Monachini. 2026. La filiera legale di ROADS. Una proposta FAIR per archivi orali analogici.

Giovanni Abete, Silvia Calamai, Sergio Canazza, Alessandro Casellato, Elvira Mercatanti, Monica Monachini, Virginia Niri, Cesarina Vecchia, Giulia Zitelli Conti, and Giada Zuccolo. 2025a. Fare didattica con gli archivi orali: la fonte orale al crocevia di discipline e di saperi.

Giovanni Abete, Cesarina Vecchia, Silvia Calamai, Alessandro Casellato, Sergio Canazza, Elvira Mercatanti, Monica Monachini, Roberta Ottaviani, Giulia Zitelli Conti, and Giada Zuccolo. 2025b. On the lifecycle of Italian oral archives: the ROADS project. In *La voce della grammatica. Nuove prospettive sull'interazione tra fonetica e morfologia, sintassi, lessico*, Università degli Studi di Urbino Carlo Bo. Associazione Italiana di Scienze della voce.

Silvia Calamai and Francesca Frontini. 2018. FAIR data principles and their application to speech and oral archives. *Journal of New Music Research*, (47):339–354.

Amedeo De Dominicis, editor. 2002. *La voce come bene culturale*. Carocci, Roma.

Gabriella Gribaudi. 1980. *Mediatori: antropologia del potere democristiano nel Mezzogiorno*. Rosenberg & Sellier.

Gabriella Gribaudi. 1990. *A Eboli. Il mondo meridionale in cent'anni di trasformazione*. Marsilio, Venezia.

Gabriella Gribaudi. 2005. *Guerra totale. Tra bombe alleate e violenze naziste. Napoli e il fronte meridionale 1940-1944.* Bollati Boringhieri, Torino.

Gabriella Gribaudi. 2016. *Combattenti, sbandati, prigionieri. Esperienze e memorie di reduci della seconda guerra mondiale*. Donzelli, Roma.

Gabriella Gribaudi, editor. 2023. *Terra bruciata. Le stragi naziste sul fronte meridionale*. Guida, Napoli.

Rosanna Sornicola, Giovanni Abete, Elisa D'Argenio, and Cesarina Vecchia. 2019. Raccontare un archivio di fonti orali: il progetto Voci, parole e testi della Campania. In Duccio Piccardi, Fabio Ardolino, and Silvia Calamai, editors, *Gli archivi sonori al crocevia tra scienze fonetiche, informatica umanistica e patrimonio digitale*, volume 6 of *Studi AISV*, pages 75–93. Officinaventuno, IT.

Mark D. Wilkinson, Susanna-Assunta Sansone, Erik Schultes, Peter Doorn, Luiz Bonino da Silva Santos, and Michel Dumontier. 2018. A design framework and exemplar metrics for FAIRness. *Scientific Data*, 5:180118.

Legal considerations in the use of synthetic data for AI development and finetuning: The case of LLMs4EU

Kossay Talmoudi, Khalid Choukri, Amélie Gourgeot, Florine Astruc
ELDA, ALT-EDIC
9 Rue des Cordelières 75013 Paris, 1 Pl. Aristide Briand 02600 Villers-Cotterêts
{Kossay, Khalid, Amelie}@elda.org, Florine.astruc@alt-edic.eu

Abstract

This paper examines the legal implications of using synthetic data to develop and fine-tune general-purpose AI models in the European Union, using the LLMs4EU project as a case study. It situates synthetic data within the Union's broader data policy and highlights it as a candidate tool for reconciling data availability with regulatory constraints. From a data protection perspective, it analyses whether and when synthetic data should be classified as "personal data" under the GDPR. From a copyright and contractual standpoint, the paper assesses the risks that synthetic datasets may embed infringing content or derive from other models, in light of the GEMA v. OpenAI ruling on memorised works and emerging analyses of liability for AI-generated outputs, and considers the constraints imposed by model licensing and acceptable-use policies on using models to generate training data for other models. The paper concludes that synthetic data can play a valuable role in mitigating legal risks and enabling compliant AI development in LLMs4EU, but only if its generation and use are embedded in robust governance frameworks that address data protection, copyright and contractual obligations across the entire data value chain.

Keywords: Synthetic data, finetuning, training data

1. Introduction

LLMs4EU is an EU-funded project that aims to fine-tune general-purpose AI models capable of addressing concrete, domain-specific cases, with a particular consideration given to linguistic diversity. This objective presupposes access to diverse datasets, which in turn raises recurrent legal and practical challenges concerning both personal and non-personal data. In this context, the project considers synthetic data as an important component of the training mix, complementing human-generated data and potentially reducing the dependency on scarce or legally constrained datasets.

Recent technical work has stressed that, under current trajectories, the stock of publicly available human-generated text will be insufficient to sustain large-scale LLM training, with projections of exhaustion between now and 2032 if present trends continue (Villalobos et al., 2024). This scarcity is a focal point for the Union's data policy, in which the 2020 European Strategy for Data and the subsequent Data Union Strategy of 2025 urge for a systematic increase in data availability for innovation and competition. Such efforts to limit data scarcity are also reflected in the adoption of data space initiatives such as the Language Data Space.

In the framework of the legal bundle referred to as the EU data laws, the Data Act adopts a broad and technology-neutral understanding of "data" in its article 2(1), agnostic of form, source and structure, and encompasses personal and non-personal data, thereby opening the door to the application of multiple legal regimes depending on provenance and use. Synthetic data are not therefore specifically defined in this legal instrument, but such agnostic definition englobes it.

Synthetic data is increasingly presented as a means to reconcile the objective of quality data and the respect of legal constraints. In LLMs4EU, synthetic data thus appears as a technical response to data scarcity and as a potential regulatory tool to mitigate legal risks mainly related to data protection and intellectual property rights.

The turn to synthetic data also reflects the sensitivity around copyright in the context of generative AI training although emphasis on synthetic data as a possible way to reduce reliance on protected material is confronted to the fact that such data presents its set of legal challenges. LLMs4EU must therefore navigate this regulatory landscape and identify under which conditions synthetic data can lawfully support model development in the Union.

2. Synthetic data as an opportunity for AI developers

2.1 Definitions of synthetic data

Synthetic data can be defined in functional terms as data generated by an algorithm that statistically resembles real-world datasets but does not directly reproduce any specific record. It serves multiple functions such as filling gaps where data is missing or cannot be accessed, or to exhibit particular distributions or properties that are hard to obtain in practice. Synthetic data has been for example used operationally for debiasing through techniques such as the Synthetic Minority Over-sampling Technique (SMOTE), which creates synthetic examples of under-represented classes in supervised learning.

Proceedings of LEGAL2026 and CALD-pseudo 2026 @ LREC 2026, pages 86–90
12 May 2026. ©ELRA Language Resources Association (ELRA), 2026

Contemporary synthetic data for AI training is predominantly generated using deep-learning models. In the context of LLMs, synthetic text can be used to augment existing corpora, to create specialised or domain-specific datasets, and to explore hypothetical scenarios that would be difficult to document otherwise.

Recently, so-called data distillation approaches have been proposed to compress large training corpora into smaller, high-utility synthetic sets. In such approaches, models are trained or fine-tuned on structured synthetic datasets generated by a base model with the aim of preserving performance while reducing dataset size. Such synthetic data can be generated in an environment in which a large language model is prompted to produce outputs that are then re-used as training inputs, either for itself (self-training) or for other models. Such iterative loops promise cost-effective scaling of training data. It is to be noted that empirical work in the LLM field suggests that carefully curated synthetic data can, under certain conditions, support further training or fine-tuning of models, although repeated training on model-generated data can also introduce biases and degradation (Shumailov et al., 2024)

2.2 Synthetic data as a mitigator of legal risk in AI training

2.2.1 Synthetic data in support of data protection compliance

Advanced analytical techniques have shown that even heavily processed and cleaned datasets used to train AI may contain sufficient information to re-identify individuals or to infer sensitive attributes, particularly when combined with other data. This undermines the effectiveness of traditional anonymisation methods and supports the view of supervisory authorities and the European Data Protection Board (EDPB) that anonymisation must be robust and context-specific to prevent re-identification, since in cases where there is a mere prevention of attribution, the data cannot be considered as anonymous (European Data Protection Board, 2025). The question therefore remains whether synthetic data constitute “personal data” within the meaning of Article 4(1) General Data Protection Regulation (GDPR) where they allow information relating to an identifiable person to be inferred. Synthetic data is often proposed as a potential solution, as it can, in principle, either be fully anonymous in fully synthetic data, or at least cut the link between records and identifiable persons while preserving utility in the case of partially synthetic data (Zhang et al., 2022).

Some researchers (Gal and Lynskey, 2023) have argued that synthetic data challenges the conceptual foundations of data protection law because, even where a synthetic dataset no longer contains any record corresponding to an actual individual, it can still be used to make decisions and inferences that affect real persons. Their analysis suggests that many synthetic datasets should be treated as personal data whenever they allow for individual-level impacts, even if the data points themselves are artificially generated.

In all cases, synthetic data produced via other models are dependant in its compliance on whether the underlying model is trained lawfully or not. The EDPB’s opinion on AI models emphasises that developers must ensure that models are not trained on unlawfully processed personal data and that data protection principles apply across the lifecycle of AI models, including training, validation and deployment (European Data Protection Board, 2024). The Joint Research Centre’s work on synthetic data in digital finance indicates that, with appropriate rules for generation, synthetic datasets can yield analytical results closely aligned with those obtained from original data while ensuring compliance with confidentiality requirements (European Commission’s Joint Research Centre, 2024).

Synthetic data, depending on how its generation is carried out, must not be regarded as a way to circumvent data protection constraints, but rather as a privacy-enhancing instrument that aims to minimise privacy risks, provided that its use remains grounded in a compliant governance framework, as clear risks remain to be mitigated, as analysed in section 3.3.

2.2.2 Mitigating copyright risks through synthetic data

Synthetic data may mitigate copyright risk in LLM fine-tuning because it can reduce direct dependence on protected material. Where a fine-tuning corpus is built from human-authored works, the legal exposure arises principally from the possibility that the training set contains protected expression, whether in full works, substantial parts, or fragments that remain recognisable in downstream outputs. This is sensitive in generative AI, where the model may reproduce or closely imitate protected sequences, and where the line between lawful learning and infringing reproduction is often contested (Tyagi, 2025).

Synthetic data offers a partial answer to the copyright question because, when it is generated from a model rather than from direct reuse of protected expression, it is less likely to replicate copyrighted works verbatim or to preserve expressive choices that are protected by copyright. In other words, the legal risk is not eliminated, but it is shifted: the compliance question moves from the downstream corpus to the upstream generation process. If the synthetic corpus is generated by model, thus without directly copying protected text, and if it is sufficiently transformed so that it no longer

contains substantial parts of any pre-existing work, it is generally less problematic from a copyright perspective than a corpus assembled by scraping or reusing protected works directly.

That said, synthetic data is not a categorical panacea. If the synthetic generator itself is trained on copyrighted material without legal scrutiny, or if it emits outputs that reproduce protected expression with sufficient similarity, the resulting corpus may still carry copyright risk. This is why LLMs4EU should treat synthetic data as a risk-reduction technique rather than as a substitute for copyright clearance. The relevant operational question whether the outputs are sufficiently detached from protected source works is analysed in section 3.3.

3. Legal hurdles to the adoption of synthetic data in AI development

Synthetic data used in LLM training lies at the intersection of several legal regimes that may apply cumulatively. From a data protection perspective, the qualification of synthetic datasets as personal or non-personal data determines the applicability of the GDPR. From a copyright perspective, synthetic data derived from models trained on protected works may constitute adaptations or reproductions, and their generation may require appropriate legal scrutiny. In addition, contractual provisions in model licences and data-sharing agreements can impose further restrictions on the generation and reuse of synthetic data, especially in relation to acceptable uses and redistribution.

3.1 Model licensing and acceptable uses

Many foundation models are distributed under licences that include acceptable use policies prohibiting certain applications, such as generating abusive content, violating privacy or using the model to develop competing models. Such provisions may explicitly or implicitly restrict the generation of large synthetic corpora for downstream model training or may condition such uses on obtaining additional permissions. For LLMs4EU, which contemplates using existing models to generate synthetic data, careful scrutiny of such acceptable use clauses is necessary to ensure that the planned uses do not constitute misuse and thus do not result in a breach of contractual obligations.

The emergence of “open-source” or “open-weight” models has raised debates about what openness entails in the AI context, including whether there are restrictions on commercial use, re-distribution of weights, or model-as-a-service deployment. Even where model weights are openly accessible, associated licences may limit training on model outputs or prohibit using the model to generate data that is then used to train another model that competes with the original. In LLMs4EU, relying on such models to create synthetic training data requires an analysis of licence compatibility, particularly when the aim is to finetune AI models that may themselves be released under open policies.

In addition to model licences, and depending on the mechanism of data synthesis, licences governing the data used to generate synthetic data may impose downstream obligations. Where synthetic data is generated from licensed datasets, the question arises whether the synthetic corpus is a derivative work or whether it falls outside the scope of the licence. Given that some rights-holders and collecting societies take the view that output generation constitute reproductions and adaptations, contractual terms may attempt to extend protection to synthetic derivatives as well. In LLMs4EU, both the licensing of input data and the allocation of rights and responsibilities over synthetic outputs must therefore be considered.

3.2 Synthetic data constituted of copyright infringing outputs

AI-generated outputs that closely resemble pre-existing works used in training can give rise to copyright infringement, and liability may attach both to the user who inputs the prompts and to the developer or provider who made the model available (Rosati, 2025). For LLMs4EU, this entails that synthetic data used for further training must be assessed in terms of the concrete risk that it contains infringing sequences that could be reproduced or amplified in downstream models. This entails that the models used to generate the synthetic data need to be assessed on whether they are resilient enough to possibilities of them “leaking” the original data it was trained on, which was feasible with early generative AI models that were not subjected to sufficient alignment procedures (Carlini et al., 2021).

The GEMA v. OpenAI decision is an important case in this regard as it clarified that AI models can embody protected works in a way that triggers copyright liability. The Munich Regional Court held that specific song lyrics were “physically fixed” in the model, that they could be indirectly perceived through prompts, and that this memorisation constituted reproduction within the meaning of Article 2 of the InfoSoc Directive. The court further rejected defences based on quotation, parody or other limitations and ordered OpenAI to provide information and pay damages, finding that the company acted at least negligently despite legal uncertainty.

Synthetic datasets used for training may consist wholly or partly of outputs generated by models that have been trained on unlicensed or infringing

data. In such cases, even if the synthetic data does not contain verbatim reproductions, it may still be tainted by the initial unlawfulness of the training process or may in practice reproduce protected expression when prompted in specific ways. In this sense, research on memorisation and extraction in large language models shows that generative systems can regurgitate training material (Ahmed et al., 2026). The possible persistence of such behaviour makes it necessary for LLMs4EU test synthetic corpora for near-duplication, long-span overlap, and other forms of recoverable expression before relying on them for fine-tuning. Indeed, a synthetic dataset derived from other models cannot automatically be regarded as free from legal constraints.

The legal analysis of synthetic data in LLMs4EU focuses on the specific layer of the value chain at which synthetic outputs are generated and reused. Even when the consortium does not directly process the original human-generated data when creating synthetic datasets, it may still incur responsibility for using synthetic data that embed protected expression or personal information in a way that is functionally equivalent to using the original data.

3.3 Residual data protection risks in synthetic data

When synthetic data is generated by models trained on datasets containing personal data, traces of that data can be found in the output. Synthetic data can therefore be deemed "not inherently anonymous" (Achterberg et al., 2025). This is true as synthetic data can still encode patterns that are traceable to specific individuals.

Privacy evaluations assessed on synthetic data shows that the reidentification risk is residual, at least compared to real world datasets: A study conducted on synthetic data in the field of child and adolescent mental health performed three attack-based privacy evaluations on the synthetic data used. The evaluation shows that while "the overall risk remains quite low", "there is some potential for linking data to individuals" (Haizoune et al., 2026). From a legal point of view, such risk is sufficient for the GDPR to apply. Such research is also in line with previous findings that distinguish between partially synthesised data that remains “vulnerable to membership inference” and fully synthesised data that remains quite resilient when met with adversarial attacks (Zhang et al., 2022).

For LLMs4EU, this implies that synthetic datasets must be subjected to rigorous privacy assessment and testing to demonstrate whether there are risks in reidentifying the legal persons in synthetic datasets. In practice, this assessment should rather include structured tests of whether individuals can still be singled out. A first step is to evaluate the generation process itself: the consortium should verify whether the synthesis method preserves rare combinations, exact values, or patterns that could disclose information about identifiable persons. A second step is to conduct re-identification testing, for instance by attempting membership inference, attribute inference, and reconstruction attacks on the synthetic dataset, in order to measure whether an attacker could recover data about specific individuals. A third step is to document all mitigation measures when these are carried out. In this sense, compliance should be demonstrated through clear methodologies and not only a declaration of the synthesised nature of the data.

4. Conclusion

Synthetic data offers significant opportunities for LLMs4EU in terms of mitigating legal risks, challenging data scarcity and enhancing representativeness, particularly for under-served languages and domains. It can reduce direct reliance on personal data and copyrighted works in training and facilitate wider sharing of data. Nonetheless, synthetic data cannot be assumed to be legally neutral; its generation and use remain embedded in the broader framework of data protection, copyright and contractual law.

Given the persistent risks of re-identification, infringing outputs and contractual non-compliance, LLMs4EU should endorse synthetic data as part of its training strategy only under robust safeguards. These include careful selection of models used for generation, ensuring that acceptable use policies and licences permit synthetic data creation for model training and that their own training processes are sufficiently transparent and reliable not to introduce additional compliance risks; rigorous privacy and copyright audits of synthetic corpora; and documentation of generation processes as part of accountability under the GDPR and the AI Act. In particular, using models to generate data that will then train other models should be explicitly permitted and appropriately governed. It needs to be highlighted that when synthetic datasets are treated as legally “easier” to use and share, there is a risk that they will be used in contexts where their limitations are not sufficiently understood, which could lead to potential liability.

The legal status of synthetic data will remain dynamic as courts and regulators confront new cases, including further decisions on AI training and output liability and evolving interpretations of personal data and copyrightability. In LLMs4EU, synthetic data will be thus treated as a component of a broader legal and technical governance framework that is periodically reassessed in light of new case law, regulatory guidance and technical evidence on the behaviour of models trained on synthetic data.

5. Acknowledgements

This work is supported by the LLMs4EU project, funded by the European Union through the Digital Europe Programme (DIGITAL) under the grant agreement 10119847.

6. Bibliographical references

Achterberg, J., van Dijk, B., Waseem, H.M., Gallos, P., Epiphaniou, G., Maple, C., Haas, M. Spruit, M. (2025). The Data Sharing Paradox of Synthetic Data in Healthcare. *arXiv preprint arXiv:2503.20847*.

Ahmed, A., Cooper, A. F., Koyejo, S., & Liang, P. (2026). Extracting books from production language models. *arXiv preprint arXiv:2601.02671.*

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., Raffel, C. (2021). Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21),* pages 2633-2650.

European Commission, Joint Research Centre, (2024) Synthetic data in the Data Hub of the Digital Finance Platform, *Publications Office of the European Union, Luxembourg*

European Data Protection Board, (2025) Guidelines 01/2025 on Pseudonymisation

European Data Protection Board, (2024) Opinion 28/2024 on certain data protection aspects of AI models

Gal, M. Lynskey, O. (2023), Synthetic Data: Legal Implications of the Data-Generation Revolution, *109 Iowa Law Review*

Goodfellow, J. Pouget-Abadie, J. Mirza, M. Xu, B. Warde-Farley, D. Ozair, S. Courville, A. Bengio, Y. (2014) Generative adversarial networks, In *Proceedings of the 27th International Conference on Neural Information Processing Systems, Volume 2,* pages 2672-2680

Haizoune, M., Leventhal, B. L., Pant, D., Nytrø, Ø., Koochakpour, K., Koposov, R.A., Øhlckers, L.R., Skokauskas, N. (2026). Balancing Privacy and Utility in Child and Adolescent Mental Health Services Research: Retrospective Cohort Study on Synthetic Data Generation. *JMIR Medical Informatics, 14*, e71819.

Rosati, E. (2025) Infringing AI: Liability for AI-generated outputs under international, EU, and UK copyright law, *European Journal of Risk Regulation 16(2)* p. 611

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., Gal, Y. (2024) AI models collapse when trained on recursively generated data. *Nature 631*, pages 755–759.

Tyagi, K. (2025) Synthetic Data, Data Protection and Copyright in an era of Generative AI, *16 JIPITEC* p.176

Villalobos, P. Ho, A. Sevilla, J. Besiroglu, T. Heim, L. Hobbhahn, M. (2024) Will we run out of data? Limits of LLM scaling based on human-generated data, *arXiv preprint arXiv:2211.04325*

Zhang, Z., Yan, C., & Malin, B. A. (2022). Membership inference attacks against synthetic health data. Journal of biomedical informatics, *125*, 103977.

Evaluating Encoder- and LLM-Based Approaches for Robust Indirect Personal Identifier Detection

Christoph Otto[1,4*], **Ibrahim Baroud**[1,3*], **Akiko Aizawa**[2], **Sebastian Möller**[1,3], **Roland Roller**[1], **Lisa Raithel**[1,3,5,6]
[1]German Research Center for Artificial Intelligence (DFKI), [2]National Institute of Informatics, Tokyo,
[3]Technische Universität Berlin, [4]University of Potsdam,
[5]BIFOLD – Berlin Institute for the Foundations of Learning and Data,
[6]Charité – Universitätsmedizin Berlin, Institute for Artificial Intelligence in Medicine
{christoph.otto, roland.roller}@dfki.de
{ibrahim.baroud, raithel, sebastian.moeller}@tu-berlin.de
aizawa@nii.ac.jp

Abstract

Removing explicit protected health information does not fully eliminate re-identification risk in clinical text. Contextual attributes such as socio-economic status, institutional affiliations or detailed life circumstances may still enable linkage attacks. These heterogeneous and often sparsely distributed elements are referred to as Indirect Personal Identifiers, i.e., textual elements that are not always identiyfing in isolation but may enable re-identification when combined with external knowledge. They extend de-identification beyond fixed identifier lists and pose new modeling challenges. Therefore, we present a systematic comparison of encoder-only models, prompt-based LLMs and hybrid pipelines for span-level IPI detection in English discharge summaries. A fine-tuned RoBERTa-large model improves on an existing baseline and substantially outperforms ChatGPT-5.2, achieving 0.906 micro-F1 and 0.724 macro-F1, compared to 0.509 micro-F1 and 0.487 macro-F1. Our findings indicate that IPI detection constitutes a distinct modeling regime characterized by class imbalance and high intra-class variability, where scaling model capacity alone does not guarantee macro-level robustness. We show that supervised encoder models currently provide the most reliable foundation for extending anonymization guarantees and future research.

Keywords: anonymization, privacy, de-identification, indirect personal identifiers

1. Introduction

Clinical natural language processing (NLP) depends on access to large collections of medical documents such as discharge summaries. These texts contain rich diagnostic and procedural detail, which makes them invaluable for research, but they also include information that may enable patient re-identification. Reliable privacy protection is therefore a prerequisite for responsible data sharing and reproducible clinical NLP.

Traditionally, privacy in clinical text has been achieved through de-identification, i.e., the detection and removal of explicitly defined personal health information (PHI) such as names, addresses, and dates, following regulatory frameworks like HIPAA[1]. However, the absence of explicit identifiers does not always eliminate re-identification risk. Research on indirect identifiers showed that combinations of seemingly benign demographic attributes can uniquely identify large portions of the population (Sweeney, 2002). Similar concerns arise in clinical free text. Even after removal of direct PHI, residual contextual traits such as occupation, family structure, or living situation may narrow the set of possible individuals (Feder et al., 2020). These markers are rarely identifying in isolation, but may become identifying when combined with external knowledge or contextual inference.

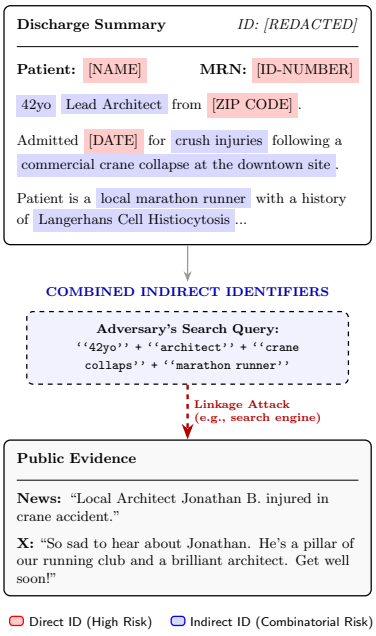


Figure 1: Illustration of re-identification risk through indirect identifier linkage in an artificial MIMIC-style document with already redacted PHI.

This broader, risk-oriented understanding of pri-

*These authors contributed equally.

[1]https://www.hhs.gov/hipaa/

Proceedings of LEGAL2026 and CALD-pseudo 2026 @ LREC 2026, pages 91–101
12 May 2026. ©ELRA Language Resources Association (ELRA), 2026

vacy is framed in regulations such as the GDPR, which emphasizes "acceptable re-identification risk" rather than fixed identifier lists [2]. Building on this perspective, Baroud et al. (2025) introduced annotation guidelines for Indirect Personal Identifiers (IPIs), which are textual spans that may contribute to re-identification despite not being explicit PHI. Compared to PHI, IPIs are heterogeneous, often infrequent and exhibit substantial lexical and semantic variability. Examples include detailed body descriptions, socio-economic circumstances or references to specific institutions (Figure 1).

Extending de-identification to include IPIs offers therefore stronger anonymization guarantees, yet also introduces new modeling challenges. While LLMs have demonstrated strong performance on several clinical information extraction tasks (Erez et al., 2025; Hu et al., 2026), prior comparative studies report that fine-tuned encoder models are competitive in supervised span-level extraction and clinical de-identification, often outperforming zero-shot LLM approaches (Kocaman et al., 2023; Diaz Ochoa et al., 2025). It is therefore an open question how these performance trends generalize to IPI detection. Hence, in this work, we systematically evaluate encoder-based models, LLM-based approaches and hybrid pipelines for span-level IPI detection in English clinical discharge summaries. We answer the following question:

RQ1: Which modeling paradigm yields robust and high-recall IPI detection suitable for privacy-preserving anonymization pipelines?

Our contributions are threefold: First, we provide a systematic comparison of fine-tuned encoders and frontier LLM-based approaches for span-level IPI detection. Second, we establish a strong encoder baseline, improving upon the best results reported in Baroud et al. (2025). Third, we show that IPI detection constitutes a distinct modeling regime characterized by severe class imbalance and high intra-class variability, where scaling model capacity alone does not ensure macro-level robustness.

2. Related work

Prior work relevant to this study spans between privacy-preserving text processing and comparative analyses of encoder-based and LLM-based model approaches in clinical NLP.

De-identification and Privacy of Clinical Text Large-scale clinical corpora such as MIMIC-III (Johnson et al., 2016) and MIMIC-IV (Johnson et al., 2024) provide central datasets for public de-identified electronic health records. Building on these resources, annotation efforts such as i2b2 (Stubbs and Uzuner, 2015) enabled supervised approaches for PHI detection in clinical text. In addition, prior work has addressed de-identification from a data governance and publication perspective, proposing minimum standards for preparing clinical datasets prior to sharing or journal publication (Hrynaszkiewicz et al., 2010).

Privacy research has advanced risk-based anonymization frameworks that account for adversarial re-identification through auxiliary information (El Emam and Arbuckle, 2013). Earlier work on indirect identifiers, mainly in structured data, demonstrated that combinations of seemingly benign attributes can enable re-identification, even when explicit identifiers are removed (Sweeney, 2002). Subsequent work focused on formalizing such indirect information beyond well-defined PHI categories that may carry re-identification risk in clinical text data (Feder et al., 2020; Baroud et al., 2025).

Encoder-only vs. LLM comparison in clinical text Recent studies compare encoder-based models and LLMs for clinical information extraction (IE). Instruction-tuned LLMs have shown strong performance in structured extraction tasks, sometimes outperforming fine-tuned BERT models (Erez et al., 2025; Hu et al., 2026). However, results appear to be task-dependent, at higher computational cost, and fine-tuned encoder models remain competitive in specific NER settings (Kocaman et al., 2023; Diaz Ochoa et al., 2025). IPI are semantically heavily diverse, i.e., in discharge summaries, they may range from lifestyle behaviors (e.g, *long-term smoker*), family context (e.g., *widowed*) to hospital references. It therefore remains an open empirical question whether performance trends observed for clinical IE tasks generalize to IPI detection.

3. Data and Methods

We use the annotations introduced by Baroud et al. (2025), which define IPIs as span-level textual elements that may contribute to re-identification risk despite not being explicit identifiers. The annotation schema consists of nine categories with different types of potentially sensitive information: *BODY_DESC*, *SOCIO*, *DETAILS*, *DIRECT_ID*, *FAMILY*, *HEALTH_FCLT*, *RELATIVE_TIME*, *LFSTL* and *OTHER*. These labels cover a wide range of attributes, including physical appearance, socio-econonomic and demographic characteristics, institutional references, temporal expressions and lifestyle factors that may reveal identifying information when combined. For example, in Figure 1, individual mentions such as a patient's occupation (e.g., *lead architect*), details about a specific event (e.g., *commercial crane collapse*), or lifestyle traits

[2] https://tinyurl.com/eu-lex-32016R0679

(e.g., *local marathon runner*) can form a distinctive combination that enables re-identification when linked with external sources, while direct identifiers are redacted.

The dataset consists of 100 de-identified discharge summaries from MIMIC-III (Johnson et al., 2016). Annotations are performed at span-level to preserve clinically relevant information, while isolating potentially identifying information and avoiding the removal of entire sentences. The corpus contains 6199 annotated spans with an inter-annotator agreement of 0.87. The label distribution is imbalanced: the majority of annotations represent information about relative time or health facilities and personnel, while other information, such as events or socio-economic and criminal history occur rarely.

Methodically, we evaluate three classes of approaches for IPI detection: encoder-based models, LLM-based methods and hybrid pipelines combining both. Even comparatively structured labels such as *RELATIVE_TIME* exhibited substantial variation in expression. We therefore focus on learning-based approaches that better capture contextual variability. Performance is measured using relaxed span-level precision, recall and F1-score following the evaluation protocol from Baroud et al. (2025). Further, to ensure comparability across models, all input documents are processed to chunks of up to 512 tokens to address the context window limitations of encoders.

Encoder-based detection As an encoder-based baseline, we fine-tune transformer models for span-level IPI classification. After preliminary testing, we found that RoBERTa-large (Liu et al., 2019) achieved the strongest and most stable fine-tuning performance. In particular, domain-specific encoders such as BioBERT (Lee et al., 2019) and ClinicalBERT (Huang et al., 2019), as well as a more recent ModernBERT (Warner et al., 2024), did not provide gains in macro-level performance against other models for IPI categories. We hypothesize that this reflects the domain agnostic complexity of IPI, i.e., semantically diverse classes rather than strict clinical jargon, which limits the benefit of domain-specific pretraining. We therefore adopt RoBERTa-large as the encoder in all encoder-only and hybrid experiments.

LLM-based detection For LLM-based IPI detection, we use both open- and closed-source state-of-the-art models, DeepSeek-V3.2 and ChatGPT-5.2 (OpenAI, 2025; DeepSeek AI, 2025) (via Microsoft Azure). We evaluate two configurations: (i) a single-stage few-shot prompting setup (LLM-Fewshot) that directly extracts and labels IPI spans, and (ii) a two-stage LLM pipeline in which the model first proposes candidate spans and then assigns IPI labels in a separate step. The two-stage design aims to test whether decoupling span detection and label assignment may improve reliability for minority classes. In particular, separating candidate generation from classification allows the model to first leverage high-recall span extraction before label assignments. We additionally conducted exploratory parameter-efficient fine-tuning experiments (QLoRA) with Qwen-14B. However, these did not result in consistent performance improvements over few-shot prompting frontier LLMs.

Label	Span Recall	Covered / Total
BODY_DESC	0.931	27 / 29
SOCIO	0.857	12 / 14
DETAILS	0.633	19 / 30
DIRECT_ID	0.500	2 / 4
FAMILY	0.764	55 / 72
HEALTH_FCLT	0.712	257 / 361
RELATIVE_TIME	0.636	638 / 1003
LFSTL	0.800	28 / 35
OTHER	0.857	6 / 7
Overall Micro	**0.671**	–
Overall Macro	**0.743**	–

Table 1: Recall of the LLM-based filtering stage via ChatGPT-5.2.

LLM-based filtering We additionally evaluate an LLM-based filtering stage that identifies candidate spans prior to downstream classification. Filtering is done at sentence level for all IPI categories.

Hybrid pipeline Inspired by recent work on decomposing NER pipelines (Chen et al., 2024), in the hybrid setup, candidate spans proposed by the LLM-based filtering stage are passed to a fine-tuned RoBERTa-large encoder for final classification. This pipeline combines the contextual knowledge of LLM-based candidate generation with the efficiency of encoder-based classification.

4. Results and Discussion

Table 2 summarizes performance across modeling paradigms and addresses RQ1. The encoder-only model achieves the strongest overall results (micro-F1 0.90; macro-F1 0.72), clearly outperforming both LLM-based and hybrid configurations. This extends the findings of Baroud et al. (2025), who report a BERT-base baseline (micro-F1 0.78; macro-F1 0.50) and similarly observe weaker performance for LLM-based approaches. While their study provides initial results of LLMs for IPI detection, our results show that even more recent and larger models continue to exhibit lower precision and recall on this task. All LLM-based approaches in our

	LLM-Few (DS)			LLM-Few (GPT)			LLM-Pipeline (GPT)			Encoder-only			Hybrid (GPT/BERT)		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Overall Performance															
Micro Avg.	0.509	0.302	0.379	0.459	0.572	0.509	0.443	0.601	0.510	0.859	0.958	**0.906**	0.885	0.531	0.664
Macro Avg.	0.464	0.463	0.380	0.485	0.616	0.487	0.446	0.684	0.484	0.660	0.824	**0.724**	0.791	0.523	0.606
Per-label Performance															
BODY_DESC	0.254	0.552	0.348	0.253	0.828	0.387	0.343	0.828	0.485	0.759	0.759	**0.759**	0.824	0.483	0.609
SOCIO	0.647	0.786	0.710	0.524	0.786	0.629	0.520	0.929	0.667	0.813	0.929	**0.867**	0.909	0.714	0.800
DETAILS	0.188	0.433	0.263	0.382	0.433	**0.406**	0.302	0.533	0.386	0.291	0.533	0.377	0.750	0.100	0.177
DIRECT_ID	0.010	0.500	0.019	0.001	0.250	0.003	0.002	0.500	0.005	0.300	0.750	0.429	0.400	0.500	**0.444**
FAMILY	0.862	0.347	0.495	0.913	0.583	0.712	0.814	0.667	0.733	0.793	0.958	**0.868**	0.860	0.681	0.760
H_FCLT	0.541	0.493	0.516	0.582	0.629	0.605	0.541	0.601	0.570	0.854	0.953	**0.901**	0.858	0.587	0.697
REL_TIME	0.807	0.146	0.247	0.899	0.437	0.588	0.832	0.470	0.600	0.897	0.967	**0.931**	0.896	0.497	0.639
LFSTL	0.412	0.200	0.269	0.473	0.743	0.578	0.458	0.771	0.575	0.682	0.857	**0.760**	0.625	0.571	0.597
OTHER	0.455	0.714	0.556	0.333	0.857	0.480	0.207	0.857	0.333	0.556	0.714	0.625	1.000	0.571	**0.727**

Table 2: Comparison of LLM-based and encoder-based approaches for IPI detection. Best F1 per row is shown in bold, worst F1 is underlined. Hybrid combines ChatGPT-based filtering with BERT classification. We report Precision (P), Recall (R) and F1-scores.

experiments show substantially lower macro performance, indicating instability across IPI categories. While few-shot prompting achieves competitive recall in several cases, precision remains consistently low. The hybrid pipeline improves precision relative to LLM-only setups, but remains constrained by the recall bottleneck of the LLM filtering stage (Table 1), preventing it from surpassing the encoder baseline.

LLMs as unreliable detectors Across LLM-only configurations, we observe a recurring high-recall/low-precision pattern, particularly for rare IPI categories such as *DETAILS* or *DIRECT_ID*. Prompted LLMs frequently overgenerate candidate spans when a text span weakly suggests personal relevance, leading to false positives. This behavior negatively impacts macro-level robustness, given the label imbalance of the data. Notably, the two-stage LLM-Pipeline setup improves precision relative to few-shot prompting, suggesting that decomposing detection into candidate proposal and relabeling reduces false positives. Nevertheless, performance variability across minority categories persists and overall macro-F1 remains below the encoder-only model.

Encoder robustness under label imbalance In contrast, the fine-tuned RoBERTa-large model demonstrates more stable performance across both frequent and minority categories. High F1-scores for *RELATIVE_TIME* and *HEALTH_FCLT*, combined with comparatively consistent behavior on less frequent labels, suggest that supervised fine-tuning enables the encoder to learn annotation-aligned decision boundaries even under skewed class distributions. Rather than relying on broad semantic coverage, the encoder appears to benefit from task-specific boundary learning grounded in the annotation guidelines.

Hybrid pipelines: complementary but limited gains The hybrid configuration occupies an intermediate position. While LLM-based prefiltering improves precision compared to few-shot prompting, it does not outperform the encoder-only baseline. Gains are most visible for categories such as *FAMILY* and *LFSTL*, where LLM candidate generation appears beneficial. However, the recall bottleneck of the filtering (Table 1) stage limits improvements for rare categories such as *DETAILS* and *DIRECT_ID*, reducing the overall macro-level.

Error analysis Qualitative inspection of model errors aligns with these quantitative trends. LLM-based approaches frequently overgenerate spans when textual cues imply personal relevance, resulting in false positives. For example, descriptive statements about treatment circumstances or generic life events are often labeled as IPI despite lacking meaningful identification risk. Encoder-only errors, in contrast, more often show confusion between semantically adjacent categories rather than missed detections. Mentions of healthcare organizations (e.g., “All Care VNA of Greater [Location]”) are occasionally misclassified as *DIRECT_ID* instead of *HEALTH_FCLT*, indicating boundary ambiguity between institutional and direct identifiers. From a privacy perspective, such category confusions are less critical than false negatives, since the information is still identified and can be addressed during downstream generalization.

Implications for IPI modeling Taken together, these findings suggest that the performance advantages of LLMs reported for other clinical extraction tasks do not directly transfer to IPI detection. IPI

categories are heterogeneous and often sparsely represented, making stable calibration under limited supervision crucial. In this setting, increased model capacity alone does not guarantee improved robustness or better macro-F1 performance.

Importantly, indirect personal identifiers should not remain identifiable in published text, but simple deletion is often not an appropriate solution. In the given dataset, annotated IPI spans account for 11.85% of all tokens across 100 discharge summaries. Removing all detected IPI would therefore substantially degrade document informativeness. Instead, IPI handling is better framed as controlled generalization rather than deletion, where sensitive details are rewritten into broader (sub)group-level descriptions while preserving semantic utility.

Given that IPI detection performance is already robust for major categories, future work should focus on developing reliable generalization strategies on top of these models. Promising approaches include category-specific rule-based generalization, prompt-based LLM rewriting approaches and differentially private rewriting methods that balance privacy guarantees and semantic accuracy (Meisenbacher and Matthes, 2024). Furthermore, the consistent weaknesses observed for minority IPI categories across all models highlight the need for structured and reliable synthetic data generation to improve the coverage for underrepresented classes (Vakili et al., 2025; Shimizu et al., 2025; Kweon et al., 2024).

5. Conclusion

In summary, our results reinforce the assumption that IPI constitute a distinct and challenging task for mitigating privacy risks. While LLMs offer strong general-reasoning capabilities, our experiments show that fine-tuned encoder-based models remain more reliable in the IPI task setting. These findings highlight the importance of careful model calibration and motivate future work that moves beyond detection toward principled rewriting and synthetic data generation strategies for personal indirect identifiers.

Limitations

Our experiments are conducted on a single dataset derived from discharge summaries, following the annotation scheme introduced by Baroud et al. (2025). While this dataset provides a realistic and challenging benchmark for IPI detection, the findings may not fully generalize to other clinical document types.

Additionally, our evaluation focuses on structured span-level detection, requiring models to return exact substrings in a predefined JSON format. Large language models may be disadvantaged in this setting, as their strengths lie in generative reasoning rather than precise boundary extraction and structured output compliance. It is therefore possible that LLMs would perform more competitively in alternative formulations of the task, such as direct privacy-preserving rewriting or controlled generalization of IPI content.

Ethics Statement

The dataset used in this work is available after conducting an appropriate training and already de-identified. We do not attempt to re-identify individuals and solely focus on identifying residual information that may contribute to re-identification risk.

Additionally, methods for detecting IPIs could potentially be misused to facilitate re-identification. However, our work is explicitly designed for risk mitigation and improving privacy-preserving data sharing. We do not release tools or resources intended for adversarial use.

Acknowledgments

This work was partially conducted during an internship at the National Institute of Informatics (NII) in Tokyo, at the Aizawa Laboratory. We gratefully acknowledge funding from the German Federal Ministry of Research, Technology and Space (BMFTR) through the project VERANDA (16KIS2046K) and through the grant BIFOLD26B.

6. Bibliographical References

Ibrahim Baroud, Lisa Raithel, Sebastian Möller, and Roland Roller. 2025. Beyond de-identification: A structured approach for defining and detecting indirect identifiers in medical texts. In *Proceedings of the Sixth Workshop on Privacy in Natural Language Processing*, pages 75–85, Albuquerque, New Mexico. Association for Computational Linguistics.

Wei Chen, Lili Zhao, Zhi Zheng, Tong Xu, Yang Wang, and Enhong Chen. 2024. Double-checker: Large language model as a checker for few-shot named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3172–3181, Miami, Florida, USA. Association for Computational Linguistics.

DeepSeek AI. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *arXiv:2512.02556*.

Juan G. Diaz Ochoa, Natalie Layer, Jonas Mahr, Faizan E Mustafa, Christian U. Menzel, Martina Müller, Tobias Schilling, Gerald Illerhaus, Markus Knott, and Alexander Krohn. 2025. Optimized bert-based nlp outperforms zero-shot methods for automated symptom detection in clinical practice. *Frontiers in Digital Health*, Volume 7 - 2025.

Khaled El Emam and Luk Arbuckle. 2013. *Anonymizing Health Data: Case Studies and Methods to Get You Started*, 1st edition. O'Reilly Media, Inc.

Ely Erez, Sedem Dankwa, McKenzie Tuttle, Afsheen Nasir, Prashanth Vallabhajosyula, Eric B. Schneider, Roland Assi, and Chin Siang Ong. 2025. Instruction-tuned large language models for clinical data extraction: Creating an aortic measurement database from ct radiology reports. *Journal of Healthcare Informatics Research*, 9(4):587–605.

Amir Feder, Danny Vainstein, Roni Rosenfeld, Tzvika Hartman, Avinatan Hassidim, and Yossi Matias. 2020. Active deep learning to detect demographic traits in free-form clinical notes. *Journal of Biomedical Informatics*, 107:103436.

Iain Hrynaszkiewicz, Melissa L Norton, Andrew J Vickers, and Douglas G Altman. 2010. Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. *BMJ*, 340.

Yan Hu, Xu Zuo, Yujia Zhou, Xueqing Peng, Jimin Huang, Vipina K Keloth, Vincent J Zhang, Ruey-Ling Weng, Cathy Shyr, Qingyu Chen, Xiaoqian Jiang, Kirk E Roberts, and Hua Xu. 2026. Information extraction from clinical notes: are we ready to switch to large language models? *Journal of the American Medical Informatics Association*, page ocaf213.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv:1904.05342*.

V. Kocaman, D. Talby, and H. Ul Hak. 2023. RWD143 beyond accuracy: Automated de-identification of large real-world clinical text datasets. *Value in Health*, 26(12):S532.

Sunjun Kweon, Junu Kim, Jiyoun Kim, Sujeong Im, Eunbyeol Cho, Seongsu Bae, Jungwoo Oh, Gyubok Lee, Jong Hak Moon, Seng Chan You, Seungjin Baek, Chang Hoon Han, Yoon Bin Jung, Yohan Jo, and Edward Choi. 2024. Publicly shareable clinical large language model built on synthetic clinical notes. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5148–5168, Bangkok, Thailand. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Stephen Meisenbacher and Florian Matthes. 2024. Just rewrite it again: A post-processing method for enhanced semantic similarity and privacy preservation of differentially private rewritten text. In *Proceedings of the 19th International Conference on Availability, Reliability and Security*, ARES '24, New York, NY, USA. Association for Computing Machinery.

OpenAI. 2025. Openai gpt-5 system card. *arXiv:2601.03267*.

Seiji Shimizu, Ibrahim Baroud, Lisa Raithel, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2025. RecordTwin: Towards creating safe synthetic clinical corpora. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14714–14726, Vienna, Austria. Association for Computational Linguistics.

Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05):557–570.

Thomas Vakili, Aron Henriksson, and Hercules Dalianis. 2025. Data-constrained synthesis of training data for de-identification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27414–27427, Vienna, Austria. Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference.

7. Language Resource References

Ibrahim Baroud, Lisa Raithel, Sebastian Möller, and Roland Roller. 2025. Beyond de-identification: A structured approach for defining and detecting indirect identifiers in medical texts. In *Proceedings of the Sixth Workshop on Privacy in Natural Language Processing*, pages 75–85, Albuquerque, New Mexico. Association for Computational Linguistics.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. 2024. MIMIC-IV. *PhysioNet*. Version 3.1.

Alistair Johnson, Tom Pollard, and Roger Mark. 2016. MIMIC-III Clinical Database. *PhysioNet*. Version 1.4.

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus. volume 58, pages S20–S29. Supplement: Proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data.

References

A. Data Statistics and Visualization

To investigate whether IPI categories possess distinct semantic patterns, we project their BERT-based sentence embeddings[3] into a lower-dimensional space (see Figure 2). Our analysis reveals that IPI categories do not form well-defined, linearly separable clusters but exhibit significant semantic overlap. Even frequent labels show high intra-class variance, while rare categories are often subsumed within broader semantic regions. This suggests that IPIs are not characterized by static lexical patterns but are defined through contextual nuances. The combination of strong class imbalance and highly diverse surface realizations reflects the complex narrative structure of discharge summaries. Consequently, IPI detection serves as a rigorous test for evaluating model performance in realistic IE settings, as it requires distinguishing semantically diverse spans under limited supervision.

Statistic	Value
Documents	100
Total tokens	144529
Covered tokens (IPI)	17124
Coverage (%)	11.85

Table 3: Corpus statistics and token-level coverage of indirect personal identifier (IPI) spans.

B. Prompt Templates and Additional Modelling Details

We note model results on the LLM pipeline and Hybrid setup with DeepSeek-V3.2 as the LLM component (Table 4), including the recall of LLM-based filtering stage (Table 5).

For reproducibility, decoding parameters are fixed with temperature set to 0 and top_p (nucleus sampling) set to 1.0. Here, top_p restricts token selection to the smallest set of tokens whose cumulative probability exceeds a threshold. All models in our work are evaluated on the same train/development/test split (60/15/25) as introduced by Baroud et al. (2025).

[3] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

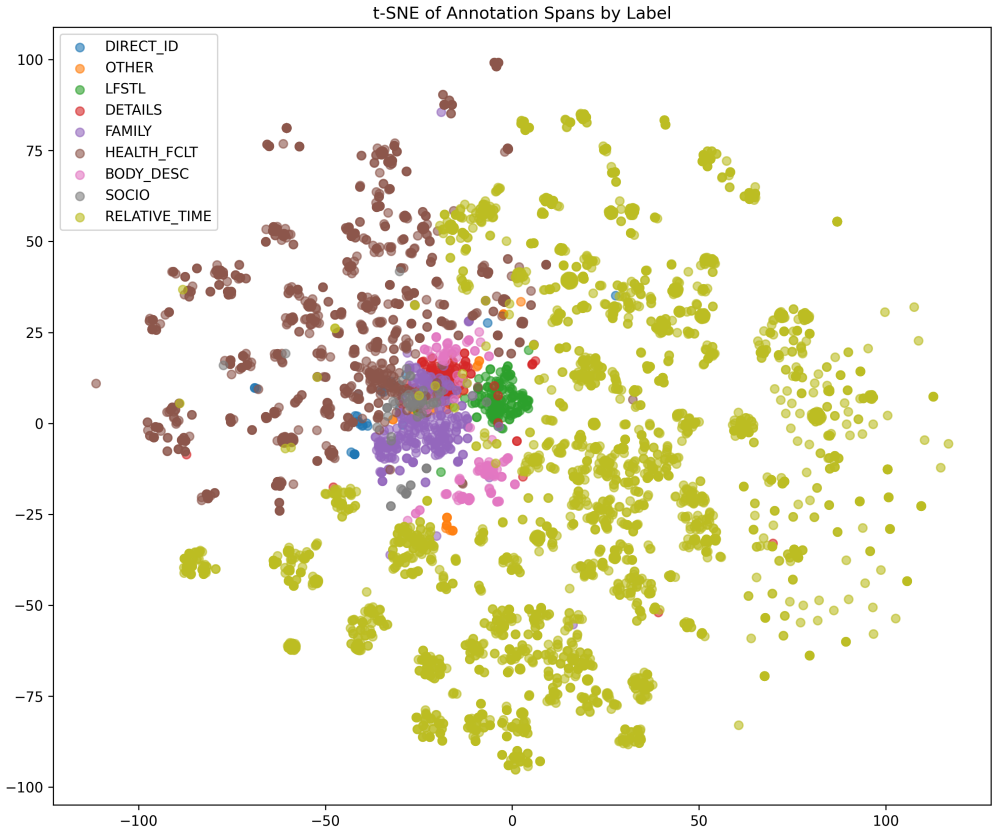


Figure 2: t-SNE projection of cosine similarity between annotated spans using BERT-based sentence embeddings.

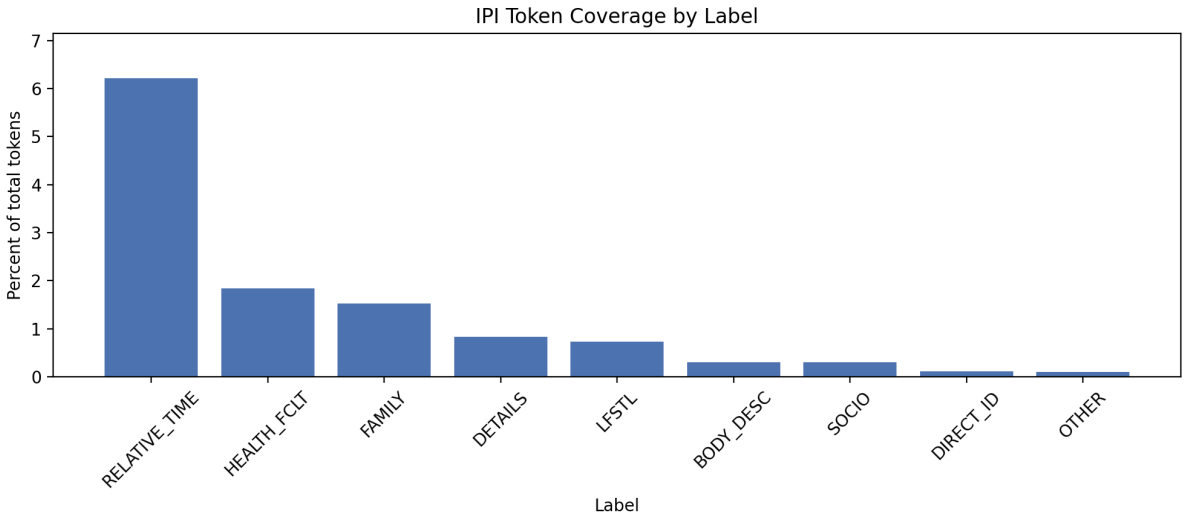


Figure 3: Token-level coverage of indirect personal identifier spans in the given dataset.

	LLM-Pipeline (DS)			Hybrid (DS/BERT)		
	P	R	F1	P	R	F1
Overall Performance						
Micro Avg.	0.384	0.251	0.304	0.827	0.350	**0.492**
Macro Avg.	0.437	0.490	0.402	0.738	0.428	**0.516**
Per-label Performance						
BODY_DESC	0.339	0.690	**0.455**	0.625	0.345	0.444
SOCIO	0.579	0.786	0.667	1.000	0.786	**0.880**
DETAILS	0.254	0.567	**0.351**	0.500	0.200	0.286
DIRECT_ID	0.000	0.000	0.000	0.250	0.500	**0.333**
FAMILY	0.739	0.472	0.576	0.808	0.583	**0.677**
H_FCLT	0.392	0.266	0.317	0.755	0.393	**0.517**
REL_TIME	0.770	0.144	0.242	0.854	0.302	**0.446**
LFSTL	0.531	0.486	**0.508**	0.846	0.314	0.458
OTHER	0.333	1.000	0.500	1.000	0.429	**0.600**

Table 4: Comparison of LLM-based approaches for IPI detection with DeepSeek-V3.2. Best F1 per row is shown in bold. Hybrid combines DeepSeek-V3.2.-based filtering with BERT classification. We report Precision (P), Recall (R) and F1-scores.

Label	Span Recall	Covered / Total
BODY_DESC	0.724	21 / 29
SOCIO	0.857	12 / 14
DETAILS	0.633	19 / 30
DIRECT_ID	0.750	3 / 4
FAMILY	0.708	51 / 72
HEALTH_FCLT	0.601	217 / 361
RELATIVE_TIME	0.446	447 / 1003
LFSTL	0.771	27 / 35
OTHER	1.000	7 / 7
Overall Micro	**0.517**	–
Overall Macro	**0.721**	–

Table 5: Recall of the LLM-based filtering stage via DeepSeek-V3.2.

Few-Shot LLM Detection Prompt Template

System Message

You are an expert annotator of indirect personal identifiers (IPI) in clinical text. Your task is to detect span-level IPI instances for the following labels: `<IPI_LABELS>`.

Return *only* valid JSON in the following format:

```
{
  "spans": [
    {"text": "exact substring from input", "label": "LABEL"}
  ]
}
```

Rules:

- Each span must be an exact substring copied verbatim from TEXT.
- The label must be one of `<IPI_LABELS>`.
- Prefer minimal spans covering the identifying content.
- If no IPI is present, return: `"spans": []`.

Annotation guidelines: *(full label definitions from (Baroud et al., 2025) included verbatim in the prompt).*

User Message (Few-Shot + Inference)

```
Fewshot Examples
TEXT:
<<<
<example text>
>>>
Return JSON only.


### NOW ANNOTATE THIS TEXT
TEXT:
<<<
<input chunk (max 512 tokens)>
>>>
Return JSON only.
```

Figure 4: Prompt template used for the LLM-Fewshot configuration. Five randomly sampled IPI fewshot examples from the validation set are included, and full annotation guidelines are embedded in the system message.

LLM-Filtering Prompt Template

System message.

You are a high-recall, permissive filter for indirect personal identifiers in clinical text.

Task: given TEXT, return snippets that could plausibly contain any of these labels: `<IPI_LABELS>`.

Annotation guidelines: *(full label definitions from (Baroud et al., 2025) included verbatim in the prompt).*

Rules:

- Return *only* valid JSON: `{"snippets": ["..."]}`.
- Each snippet must be an exact substring copied verbatim from TEXT.
- Be maximally inclusive: include a snippet if it might contain an IPI, even if unsure.
- If none, return `{"snippets": []}`.

User message.

```
TEXT:
<<<
<input chunk>
>>>
Return JSON only.
```

Figure 5: Prompt template used for the LLM-based filtering stage, designed to maximize recall by returning candidate snippets for downstream classification.

VEIL: A Benchmark for Value-Preserving Entity Identification Limitation

Darina Gold[1], **Shadi Rastegar**[1], **Alina Liebel**[1], **Alessandra Zarcone**[1,2]
[1]Fraunhofer IIS, [2]Technische Hochschule Augsburg
{firstname.secondname}@iis.fraunhofer.de

Abstract

Large Language Models (LLMs) are linked to several issues regarding Personally Identifiable Information (PII). PII can occur in the training data and can thus be accidentally leaked or extracted with malicious intent, or it can be inputted in LLM-based technologies by users through their prompts. A viable strategy to limit the LLMs' exposure to PII is to filter input and output data by de-identifying PII, including personal names. This however poses a challenge: a name could refer to a private person in a context containing sensitive information (e.g., *Michelangelo is an atheist*), or it could refer to a famous artist in another context (e.g., *Michelangelo's Sistine Chapel*), and masking the latter may hinder the LLMs' capabilities in general-knowledge tasks. We tackle the problem of personal name de-identification and focus on the decision of which personal names need to be removed (and which should be kept), based on context. We present VEIL, a challenging benchmark for Value-preserving Entity Identification Limitation, for context-aware de-identification decisions on LLM training data, and compare the performance of different state-of-the-art systems on the task.

Keywords: de-identification benchmark, data privacy, context-sensitive de-identification

1. Introduction

Large Language Models (LLMs) are typically trained on large amounts of training data, built from publicly available datasets, which may contain personally identifiable information (PII). This makes them vulnerable to prompt-based attacks, which may successfully extract personal data (Carlini et al., 2021; Mireshghallah et al., 2024). Training LLMs on data containing PII is not only potentially harmful, but can also conflict with a fundamental human right, that is the Right to Privacy[1]. Data protection regulations (e.g., the General Data Protection Regulation or GDPR in the European Union) require providers to uphold the principle of data minimization, that is the amount of personal data processed should be proportionate to pursue the legitimate interest at stake[2]. Lawful data processing for LLM training would thus require the removal of any unnecessary PII (e.g., passwords, email addresses, names) from the training data[3].

While removing all elements potentially containing PII from training data may be the most privacy-compliant strategy, such strategy may also negatively impact a range of downstream tasks. Removing the names of people currently holding an office, those of historically-significant figures, or those of artists, authors, and other cultural icons, would arguably remove widely-recognized general knowledge. This could in turn potentially degrade performance in knowledge-intensive question answering (e.g., TriviaQA, Joshi et al., 2017), reasoning tasks (e.g., CommonsenseQA, Talmor et al., 2019, HellaSwag, Zellers et al., 2019), information extraction (e.g., TACRED, Zhang et al., 2017, FewRel, Han et al., 2018), as well as slot filling and entity linking (e.g., TAC-KBP, Getman et al., 2018)—as suggested by first results comparing several de-identification strategies (masking, removal, pseudonymization)[4] (Berg et al., 2020; Lothritz et al., 2023).

This raises the question of how to determine which personal names should be removed, anonymized or at least de-identified from LLM training data and which can or should be preserved, in order to strike an ideal balance between privacy compliance and performance on downstream tasks. A name could refer to a private person in a context containing sensitive information (e.g., *Michelangelo is an atheist*), or it could refer to a famous artist in a nonsensitive context (e.g., *Michelangelo's Sistine Chapel*). Masking the latter may hinder the LLM's capabilities in general-knowledge tasks, while keeping the former may reveal sensitive data[5]. A simi-

[1]Article 12 of the United Nation's Universal Declaration of Human Rights, Article 8 of the European Convention on Human Rights.

[2]See the following opinion from the European Data Protection Board on the topic in the context of AI models: https://www.edpb.europa.eu/system/files/2024-12/edpb_opinion_202428_ai-models_en.pdf

[3]See some mitigation strategies here (Page 69): https://www.edpb.europa.eu/system/files/2025-04/ai-privacy-risks-and-mitigations-in-llms.pdf

[4]We use *de-identification* to refer to masking techniques which may not exclude the possibility for re-identification and we use *anonymization* to refer to a process which effectively prevents future re-identification.

[5]Religious beliefs fall under a special category of data

Proceedings of LEGAL2026 and CALD-pseudo 2026 @ LREC 2026, pages 102–115
12 May 2026. ©ELRA Language Resources Association (ELRA), 2026

lar balancing act is typically done by news organizations, as they have to decide what information should remain private and what should be disclosed in order to uphold another fundamental right, that is the Freedom of Expression Right[6]. For public figures, disclosure of certain information may be justified in order to safeguard the public's right to be informed, for example when reporting allegations of unlawful financial benefits received by a politician.

In this paper we focus on one type of personal data, that is personal names in context, and (1) draw inspiration from the Council of Europe's Guidelines on safeguarding privacy in the media (Council of Europe, 2019) to propose practical guidelines to drive decision-making when including or excluding personal names from LLM training data based on context (the name itself as well as the context in which it appears). We then (2) curate and annotate the VEIL dataset, which we introduce as a benchmark for Value-Preserving Entity Identification Limitation. To our knowledge, it is the first benchmark to include different combinations of person categories and contexts as well as the decision on whether their names should be de-identified, and (3) evaluate state-of-the-art approaches on the task of deciding what personal names should be kept and what names should not. VEIL is available under CC-BY-4.0 license[7].

We argue that the decision to de-identify should be context-driven and show that existing approaches struggle to distinguish between cases where the personal names should be de-identified and cases where context justifies keeping the names. Developing systems which can do this effectively is valuable for both de-identifying LLM training data and for filtering input and output data, as avoiding potential leaks and can make LLM-based technologies more robust with regard to privacy. Systems which can effectively perform context-based de-identification (aimed at preserving data utility) are also needed to systematically evaluate the impact of context-based de-identification on downstream performance, as compared to a complete de-identification.

2. Related Work

When LLM-based technologies deal with data containing PII, the focus is on one or more of the following aspects: (1) PII identification, (2) comparison of different de-identification methods, e.g., pseudonymization or masking, (3) evaluation of how de-identification of PII affects downstream tasks.

The necessity for anonymization has always been clear in the clinical NLP domain, where Protected Health Information (PHI) in medical texts has been identified using either pattern matching (regular expressions, rules, gazetteers) or machine learning methods (Meystre et al., 2010), and several benchmarks and shared tasks have encouraged work in this area (Stubbs and Uzuner, 2015; Stubbs et al., 2017). Outside the clinical NLP community, benchmarks have been created containing personal emails or text messages in an already anonymized form (Medlock, 2006; Eder et al., 2020; Patel et al., 2013) or legal documents (de Gibert et al., 2022; Pilán et al., 2022). The assumption here is however that any PII should be anonymized, if possible.

Wikipedia as evaluation data Using data containing PII poses a number of ethical and legal problems and, outside the clinical domain, it is challenging to obtain access to data which has a high density of PII. For this reason, Wikipedia biographies are also used to evaluate de-identification methods (e.g., Chow et al., 2008; Sánchez and Batet, 2016; Lison et al., 2021; Hathurusinghe et al., 2021; Papadopoulou et al., 2022), the choice being motivated by practicality (the biographies are publicly available and contain a large amount of personal names) as well as by reduced ethical concerns (the names in Wikipedia are considered to be acceptable to share). However, as argued by Pilán et al. (2022), Wikipedia mentions may not be representative of what PII typically needs to be removed from a distributional point of view. We argue that Wikipedia is actually exemplary of information which should not be removed to preserve data utility and thus use it to harvest data which is acceptable to keep as is.

Impact on downstream performance It is reasonable to assume that PII de-identification has a negative impact on the representation of de-identified text and consequently also on downstream tasks (Obeid et al., 2019). Studies assessing the extent of this impact for different de-identification methods, however, are rare. Deleger et al. (2013) have compared the effect of PHI de-identification on medication name extraction, but found no differences in performance. Meystre et al. (2014) identified an effect on clinical information extraction, interestingly affecting eponyms (names derived from proper names of persons or locations, e.g., *Alzheimer's disease* or *Achilles tendon*). Obeid et al. (2019) found no significant difference on a mental status classification task using original vs. de-identified data. Berg et al. (2020) com-

protected under Article 9 of the GDPR.

[6] Article 19 of the United Nation's Universal Declaration of Human Rights, Article 10 of the European Convention on Human Rights.

[7] https://huggingface.co/datasets/IIS-NLP/VEIL

	Private Individuals	Public Figures	Historical / Fictional
Private Contexts	Yes	Yes	No
Nonsensitive Contexts	Yes	No	No

Table 1: The conditions in our dataset, with indications if the personal names should be de-identified or not.

pared different de-identification techniques, that is pseudonymization (replacement with a surrogate), replacement by PHI class (e.g., *Eva* → <*First Name*>), masking with XXXX, and complete removal of the affected sentences, and found that they affect performance on downstream NER tasks to different degrees, with pseudonymization yielding the best results. More recently, Vakili et al. (2024) pose the problem of distinguishing between patient names (to anonymize) and eponyms (to keep), in order to limit the loss of relevant medical information. Lothritz et al. (2023) focus on personal names in order to address a broader palette of tasks at different levels of difficulty outside the clinical domain, finding a negative effect of training data de-identification on downstream performance, with the best results coming from pseudonymization. Previous work thus points in the direction of pseudonymization as the best de-identification strategy to preserve data utility.

Like Meystre et al. (2014), we are interested in preserving data utility, and like Lothritz et al. (2023), we focus on personal names and are interested in general-domain data. Preserving data utility has also been a focus of differential privacy efforts (Rodriguez-Garcia et al., 2019; Domingo-Ferrer et al., 2021; Lison et al., 2021). However, to the best of our knowledge (with the exception on work on eponyms in the clinical domain), previous work has focused on de-identifying anything which could potentially constitute PII or on how to best mask it, rather than on making context-based decisions on information needs to be de-identified and what should be kept to limit performance loss in downstream tasks.

3. The Dataset

3.1. Relevant Categories and Guidelines

To create VEIL, a benchmark for Value-Preserving Entity Identification Limitation, we focus on six conditions, which result from a combination of two variables (2 x 3 design, see an overview in Table 1): the person mentioned in the text (whether they are a **private individual**, a **public figure**, or a **historical figure / fictional character**), and the context where the person is mentioned (a **private context** or a **nonsensitive context**). We ground our definition of these categories on the guidelines of the Council of Europe and the European Court of Human Rights concerning the protection of privacy of public figures and private individuals in the media (Council of Europe, 2019).

We define three categories of personal names:

Private individuals are people who have not entered the public domain and are generally considered to have stronger expectation of privacy. Names of private individuals should always be anonymized, regardless of context. The data for the *private individual* conditions in VEIL is always generated synthetically and does not pertain to real private individuals.

Public figures are people who are active in a field of public concern, e.g., in politics, the economy, the arts, the social sphere, or sports. These may include people who are less known but still have a role in public life, as well as celebrities who are widely known to the public, even if they do not have institutional roles. Their right to keep their private life private is protected when they engage in purely private activities (*private context* condition, e.g., if a famous skier spends time with their family in their private time), but it may be restricted if the reporting does contribute to a matter of public interest, in which case the freedom of expression may prevail. In our dataset we thus allow for cases where their names may not be anonymized (*nonsensitive context* condition, see below), that is contexts which match the field of public concern where they are active. The data for *public figures* in *private contexts* in VEIL is partially original and partially synthesized.

Historical figures and fictional characters refer to individuals who are not protected by privacy rights—either because they are fictional, or because they are historical figures who have been deceased for a substantial period (e.g., over a hundred years). For historical figures, privacy protection mostly applies to protecting their reputation and dignity after death (Rawindaran and Bentotahewa, 2024), but it is otherwise acknowledged very little by legislations of different countries (Schafer et al., 2023). Their names may be preserved even in more private contexts, for reasons of historical documentation / public interest. We group together historical figures and fictional characters in one category, as their names are not de-identified in any context.

Additionally, we define two possible context categories to drive the decision to de-identify or not, that is private and nonsensitive contexts. In order to ground our annotation even more in the guidelines

	Private Individuals	Public Figures	Historical / Fictional
Private	Earlier today, a local news outlet reported that *Silas Kline*, a freelance graphic designer from Tampa, was arrested on suspicion of driving under the influence early this morning.	*John Legend* walked down the steps of the Boston Community College, wondering what the building was called.	*Farinelli* was among thousands of boys castrated to preserve their high-pitched voices as they grew up.
Nonsensitive	*Rashad Barlow*, a community organizer in his hometown, advocated for a new generation of "safe, clean nuclear power plants" in the United States.	China celebrated another successful step forward in the slow but steady space program that President *Xi Jinping* has linked to his "dream" of national revival.	Over the centuries *Pluto*'s bitterness grew leading him to rebel several times against *Zeus*.

Table 2: Exemplary sentences for the six condition present in our dataset: Personal names of private individuals, a public figures, and historical figures / fictional characters, each in private and nonsensitive context.

of the Council of Europe (2019), we identified some relevant subdomains for each category from the law cases discussed in them for illustrative purposes and used them to extract relevant paragraphs to include in our benchmark.

Private contexts include *being a victim of sexual abuse, criminal activities, leisure activities, matters regarding children or other family members, one's home address or holiday destination, romantic relationships,* and *suffering from an illness*. Even for public figures, journalists have the obligation to respect their legitimate expectations to privacy, which is particularly relevant when they engage in purely private activities. Some exceptions are listed in Council of Europe (2019), where the reporting of private contexts contributes to a matter of public interest and therefore where the right to be informed prevails. However, as most of these cases require a deeper case-by-case consideration based on extended knowledge of each case, we do not consider exceptions where the right to be informed prevails over the right to privacy. In VEIL we synthesize all data for *private individuals* in *private contexts* and part of the data for *public figures* in *private contexts*.

Nonsensitive contexts are contexts which are arguably not private. Subdomains for this category include *improper use of public money, misuse of public office,* and *protection of national security or public safety*. When it comes to public figures, we also consider any context where they engage in a public role or in activities which match the field they are famous for or active in as nonsensitive, thus making this decision dependent on who the mentioned person is (e.g., a famous skier will not be de-identified in the context of a ski competition, but if a prime minister is skiing on their private holiday, their name will be). We synthesize the data only for *private individuals* in *nonsensitive contexts*.

Table 2 illustrates some examples for all six conditions in the dataset. We are aware that these categories are an oversimplification, as it is often left to the discretion of the journalists on a case by case decision. Yet, we argue that such distinctions can constitute helpful and grounded guidelines to navigate the decision of what information should be de-identified and what could be kept.

3.2. Dataset Creation

The dataset is composed of two types of data: original and synthetic. Since the original data in this work contained personal names considered private in this work, we generated synthetic data for the *private person* conditions. The original data for *public figure* in *nonsensitive context* and for *historical / fictional character* in *private* and *nonsensitive context* were deemed suitable to be included in the VEIL benchmark in their original form, as they corresponded to the three conditions where we would not de-identify personal names (Table 1). We also kept data for *public figure* in *private context*, but mixed it with synthesized data. The synthesis procedure, which builds upon the original data, is detailed in Section 3.3.

Overall, the dataset contains 1083 paragraphs, annotated with 2438 personal names. Approximately 70% of the data is original and 30% is synthetic. The dataset is exclusively in English.

Extraction We extracted the data for the VEIL benchmark from the DCLM corpus (DCLM-baseline-1.0, train split, Li et al., 2024), a corpus created by filtering the Common Crawl[8]. This allowed us to do without further filtering, while still using data which could realistically be used to train LLMs. We split documents into paragraphs using

[8] https://commoncrawl.org/

line breaks and kept those with 5 to 500 tokens. We then used BERT-base-NER[9] to identify and annotate personal names (PERSON / PER entities) in each paragraph and discarded paragraphs without any personal names. For each document, we extracted up to 3 paragraphs and stopped the extraction when we collected 30 000 paragraphs.

Filtering and Pre-Annotation After extracting the data, we automatically filtered and pre-annotated it. We checked if the personal names in each paragraph referred to people with corresponding Wikipedia pages and if they met further criteria: if they had a corresponding Wikipedia page and had died before 1925, they were pre-annotated as *historical figures*; if they were categorized in Wikipedia under *fictional females* or *fictional males*, they were labeled as *fictional characters*. We excluded paragraphs with at least one mention of someone *without* a corresponding Wikipedia page from further processing steps. We then extracted the occupation of people in *historical / fictional* and *public* from Wikipedia and computed the semantic similarity[10] between a person's occupation and two lists of keywords respectively for *public* and *nonsensitive contexts* extracted from examples in Council of Europe (2019). If the occupation of all people mentioned in the paragraph had a higher similarity to keywords for *private contexts* than to those for *nonsensitive context*, the paragraph was pre-annotated as *private context*, otherwise it was pre-annotated as *nonsensitive context*. At the end of this process, all our paragraphs were pre-annotated with one person category and one context category.

Annotation The pre-annotated contexts were manually annotated using INCEpTION[11] by one single expert annotator, who is one of the authors. She followed our guidelines (Section 3.1) to confirm or modify the pre-annotation (*private / public / historical or fictional* person, *private / nonsensitive* context) and annotated any co-reference of the same person with personal names (but she did not annotate pronouns or other co-references, as our focus is on personal names). The annotator kept annotating the data in batches to reach a number of datapoints per condition which was roughly comparable across different conditions.

Inter-Annotator Agreement In order to evaluate our expert annotation, a second annotator, who is also one of the authors, likewise annotated 100 randomly-extracted contexts from the pre-annotated data. Cohen's κ was used to measure inter-annotator agreement.

The name-level annotation yielded a strict match κ of .43 (both span and name label need to match) and a type match κ of .46 (the name label needs to match but a span overlap is enough), corresponding to *moderate agreement* according to Landis and Koch (1977). Sometimes one annotator missed a name, but if we excluded the names missed by one of the annotators the type match κ was as high as .63 (*substantial agreement*)[12]. Contexts were more challenging to annotate: the name-level annotation yielded a κ of .33, if we excluded the context for the entities missed by one of the annotators κ was as high as .52 - *moderate agreement*. At the paragraph level, we transformed the annotation into a binary decision (*de-identify* if a *private person* is present or if a *public figure* appears in a *nonsensitive context*). Comparing the two binary decisions for the two annotators yielded a paragraph-level κ of .40 (*fair agreement*)—likely driven by the more challenging context annotation.

	Private Individuals	**Public Figures**	**Historical / Fictional**
Private	synthesis from public and historical / fictional in private ctxt	original + synthesis from historical / fictional in private ctxt	original data
Nonsensitive	synthesis from public and historical / fictional in nonsensitive ctxt	original data	original data

Table 3: Original and synthesized data across contextual categories, including base categories used for synthesis

3.3. Data Synthetization

In order to obtain examples for the remaining categories (*private person* in *private* or *nonsensitive context* and *public figure* in *private context*), we first rephrased and then pseudonymized the data (as this is also the technique found to best preserve data utility by Berg et al., 2020 and Lothritz et al., 2023) using GPT-OSS-120B. Table 3 summarizes the synthesized data in VEIL and what sources were used for the synthetization, the prompts to generate synthesized versions of the data are provided in Appendix A.1.

For *public person* in *private context*, original data cover $\sim 1/3$ of the data for this condition, the remaining $\sim 2/3$ was obtained through synthetization.

[9] https://huggingface.co/dslim/bert-base-NER

[10] Sentence Transformer model all-MiniLM-L6-v2, https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[11] https://inception-project.github.io/

[12] The annotators were not provided with guidelines on how to annotate the spans, only on guidelines for the type of context and type of name label.

	Private Individuals	Public Figures	Historical / Fictional
Private Contexts	369	190	227
Nonsensitive Contexts	273	752	627

Table 4: The number of personal names for each condition in VEIL.

To obtain the synthetized data, we used prompts aimed at replacing personal names as well as any other relevant information (e.g., number of children, locations) with realistic variations to create suitable *public person* in *private context* data points. In order to create synthetic data for *private person* in *private context* (from *public figure* and *historical / fictional character* in *private context*), and to create synthetic *private person* in *nonsensitive context* data (from *public figure* and *historical / fictional character* in *nonsensitive context*), we prompted the LLM to replace the personal names with a random first and last name while preserving gender and ethnicity. See Table 2 for some illustrative examples of the synthesized data.

Using an LLM-based data synthetization step was not without its challenges: we observed that at times the context still hinted at the person being a public figure, such as a famous athlete or author (while a synthetic private person was desired), even after the synthetization, and that some names often were repeatedly used, which we avoided by using additional prompts. One expert annotator (one of the authors) thus manually checked all synthesized instances to verify that the intended label (e.g., *private person* in *private context*) matched the synthesized version and that textual coherence within the paragraph was maintained. This resulted in the exclusion of some instances where these or other issues occurred.

3.4. Dataset Statistics

Overall Due to the difficulty to extract suitable data and the necessity to exclude data after the synthetization step, the final benchmark has some degree of class imbalance. Table 4 summarizes the number of instances included in VEIL for each condition. The *de-identify* paragraphs—that is those where either an instance of *private individual* or an instance of *public figure* in *private context* occurred—were 331 (36 paragraphs were original data and 295 synthetic data), whereas the other paragraphs (*do not de-identify*) were 804.

Lexical Diversity We calculated Type-Token Ratio (TTR) to assess lexical diversity in our small corpus. The *de-identify* paragraphs had an average TTR of .65 (original data) and .39 (synthetic data), the other paragraphs (*do not de-identify*) had a TTR of .41. As TTR is sensitive to corpus size, we also computed Measure of Textual Lexical Diversity (MTLD) scores (McCarthy and Jarvis, 2010), which for the *de-identify* paragraphs were on average 177.44 (original data) and 173.58 (synthetic data), and 147.63 for the *do not de-identify* paragraphs.

Genre	Original	Synthetic	Total
Academic Article	39	3	42
Biographical Text	174	97	244
Blog Post	31	24	55
Encyclopedia Entry	19	1	20
Fictional Narrative	185	94	279
Historical or Religious Narrative	142	4	146
News Article	125	35	160
Other	125	37	262
Total	840	295	1135

Table 5: Genre classification of paragraphs in VEIL with zero-shot classifier.

Genre We also conducted a brief genre assessment using an out-of-the box zero-shot classifier (deberta-v3-base-zeroshot-v1[13], a Transformer-based encoder model in the style of BERT), assigning texts to the following candidate labels: *academic article*, *fictional narrative*, *news article*, *blog post*, *biographical text*, *encyclopedia entry*, *historical or religious narrative*, and *other*. This approach provides an exploratory indication of genre distribution without requiring task-specific training data, though the results should be interpreted cautiously given the model's general-purpose nature. The results are shown in Table 5. The most frequent genres in both the original as well as the synthesized texts according to the classifier are *biographical texts* and *fictional narratives*.

4. Benchmarking on VEIL

We benchmark LLM-based approaches on a name de-identification task using VEIL. The first and second classification tasks are sequence-labelling tasks, where the models identifies each personal name and classified the name itself (person classification) and its context (context classification) according to the defined categories (*name-level decision*).

In a third classification task, if a paragraph contains at least one name that should be de-identified (a *private person* or a *public figure* in *private context*), the whole text is labeled as DE-IDENTIFY (*paragraph-level decision*).

[13] https://huggingface.co/MoritzLaurer/deberta-v3-base-zeroshot-v1

4.1. Prompt-based LLM classifiers

We use LLMs as classifiers by framing the task as a prompt and optionally providing few-shot examples to guide their predictions, which reflects a current state-of-the-art approach. We evaluate the following models: Qwen3-30B-A3B-GPTQ-Int4, Mistral-Small-3.1-24B, and Meta-Llama-3.3-70B. These models were selected because they can be run on-premise, which is important when working with private data. Using such models allows for secure, in-house processing, which is typically not the case for very large models hosted on external servers, which may not be suitable to process sensitive information. Our model choice aimed at providing diversity in geographic origin (Asia, Europe, North America), model size, and release period. Additionally, they differ from an architecture point of view: Qwen3-30B-A3B-GPTQ-Int4 uses a Mixture-of-Experts (MoE) approach, activating only a subset of parameters per toke, while the others are dense models, offering a comparison in efficiency and scalability. All models were tested using the same prompt while varying the number of in-context examples provided.[14]

Prompts Our prompts build directly on Rescriber, a prompt-based browser extension designed to identify and anonymize PII in user interactions with LLM-based conversational agents (Zhou et al., 2025). The original study evaluated the approach using Llama3-8B and GPT-4o.

We slightly modify the original prompt by explicitly instructing the model to consider not only the status of the person referred to, but also the context in which the name appears. The prompts are displayed in Appendix A.2.

The model returns the original paragraph, annotating person names across the two independent dimensions—person category and context, without explicitly linking them beyond the shared entity.

Zero-shot vs. few-shot We compare zero-shot, one-shot, and three-shot settings to assess whether few-shot (in-context) learning provides an advantage over zero-shot prompting. The different shot versions are also displayed in Appendix A.2.

4.2. Results

Tables 6, 7, and 8 summarize the benchmarking results. Across all settings, few-shot prompting outperforms zero-shot. While introducing a single example (1-shot) generally leads to a noticeable improvement, the performance difference between 1-shot and 3-shot settings is comparatively smaller.

[14]We used a temperature of 0.0, set max_tokens to 3072, and used nucleus (top_p) sampling of 0.95.

Model	Shot	P	R	F1
llama	0-shot	15.61	6.26	8.61
	1-shot	15.59	6.59	8.91
	3-shot	17.49	7.45	**9.96**
qwen	0-shot	11.10	0.77	1.43
	1-shot	12.67	3.76	5.69
	3-shot	17.56	5.74	8.62
mistral	0-shot	11.34	5.48	7.31
	1-shot	10.80	5.66	7.15
	3-shot	13.20	6.48	8.20

Table 6: Name-level results (person classification) for llama (Meta-Llama-3.3-70B), qwen (Qwen3-30B-A3B-GPTQ-Int4), and mistral (Mistral-Small-3.1-24B) as P (precision), R (recall) and F1 (F1-score) scaled 0–100.

Model	Shot	P	R	F1
llama	0-shot	9.23	3.79	5.00
	1-shot	11.25	5.04	6.85
	3-shot	11.09	5.01	6.79
qwen	0-shot	10.70	0.90	1.65
	1-shot	10.84	3.32	4.98
	3-shot	10.85	3.47	5.16
mistral	0-shot	8.59	4.51	5.65
	1-shot	10.38	6.16	**7.73**
	3-shot	10.30	6.07	7.64

Table 7: Name-level results (context classification) for llama (Meta-Llama-3.3-70B), qwen (Qwen3-30B-A3B-GPTQ-Int4), and mistral (Mistral-Small-3.1-24B) as P (precision), R (recall) and F1 (F1-score) scaled 0–100.

Meta-Llama-3.3-70B achieves the best results at the name-level person classification task (F_1 = 9.96) and at the paragraph-level decision (F_1 = 77.39), while Mistral-Small-3.1-24B performs best at the context classification task (F_1 = 7.73).

The name-level tasks are comparatively harder because, while identifying the entity was part of the task, we did not provide guidelines in the prompts to support the models' decisions on the span extension. Overall, the results indicate that the task is challenging also at the comparatively easier paragraph level decision, as these results must be interpreted in light of the strong majority baseline of 74.6%.

5. Conclusions and Further Work

We present VEIL, the first benchmark for value-preserving entity identification, introducing both the concept and the construction of a dataset that combines person and context categories to guide decisions on de-identifying personal names. Although VEIL is currently small, it can be easily extended with additional examples or categories, increasing

Model	Shot	P	R	F1
llama	0-shot	78.63	73.36	75.14
	1-shot	77.73	76.59	77.11
	3-shot	77.23	77.56	**77.39**
qwen	0-shot	52.34	50.93	47.97
	1-shot	74.31	63.06	64.27
	3-shot	73.25	68.10	69.58
mistral	0-shot	72.52	70.54	71.34
	1-shot	68.72	71.10	69.37
	3-shot	68.01	71.08	68.39

Table 8: Results of the paragraph-level decision for llama (META-LLAMA-3.3-70B), qwen (QWEN3-30B-A3B-GPTQ-INT4), and mistral (MISTRAL-SMALL-3.1-24B) as P (precision), R (recall) and F1 (F1-score) scaled 0–100.

robustness and generality. Our experiments with prompt-based LLM classifiers show that even relatively capable models struggle with the task, reflecting its inherent difficulty as evidenced also by modest inter-annotator agreement.

The models we evaluated were comparatively small and runnable fully on-premise; in future work, it could be interesting to explore the performance of larger models not deployable locally, to probe potential ceilings. Importantly, VEIL also enables investigation into whether distinguishing between types of individuals—specifically, whether their identities are of potential "public" interest—affects data utility for downstream applications. A classifier which scores well on VEIL and is able to perform context-based de-identification may reach an ideal trade-off between data utility and privacy preservation and may allow to evaluate the impact of a context-based de-identification (preserving data utility) on downstream tasks, as compared to a complete de-identification.

Overall, VEIL provides a systematic and practical foundation for context-sensitive, privacy-aware entity identification in NLP.

Limitations

Our 2x3 annotation simplified the problem of context-based decision on personal name de-identification, as in media it is often left to the discretion of the journalists to decide case by case what information to disclose. However, in order to implement scalable approaches, it is helpful to provide grounded guidelines to navigate this decision when processing text on a large scale.

Our work is limited to the English language, which is typically the most represented in LLM training data. As the task we propose is context-specific, language-specific benchmarks should be created to evaluate models on this task. However, due to several challenges in the creation of the dataset (finding suitable data representative of the task, manually annotating the data and synthetizing data), the resulting benchmark is small and has some degree of class imbalance. Our experience in creating the benchmark shows that is it not trivial to collect a large dataset for the task of context-based de-identification, as the current dataset was extracted from a first selection of 30,000 paragraphs. If a pre-filtered, large resource such as the DCLM corpus is not available for a given language, this would make the creation of a similar benchmark even more challenging. Furthermore, the human annotation yielded modest agreement values, mostly driven by the span decision (where no guideline was provided) but also by the context classification, where agreement was still rather low even when relaxing the span requirements. The lack of guidelines on the span extension in the prompts also affected the evaluation scores in the name-level labeling tasks.

Our work focuses on personal names, mostly ignoring other kinds of direct identifiers as well as quasi-identifiers, which in combination may enable re-identification if not removed or anonymized. We removed all original data annotated as *private person*, in order to minimize the risk that existing private persons would be re-identified and additionally prompted the LLM in the synthetization to alter all relevant details of *public figures* and their families in private contexts. We cannot exclude that the generated personal names were not actual names of private persons but they were not associated with further identifiers or quasi-identifiers.

Ethics Statement

This study analyzes publicly available data about identifiable public figures (e.g., researchers, politicians, artists) and additional non-public data. Public data were used only where relevant to public knowledge, while non-public data were handled with heightened care to protect privacy and minimize risk. All data use adheres to applicable ethical standards and principles of contextual privacy.

6. Acknowledgments

This work has been funded by the Free State of Bavaria in the DSgenAI project (Grant Nr.: RMF-SG20-3410-2-18-4) and the CHIASM project (Bayerische Staatsministerium für Wissenschaft und Kunst - StMWK). The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) under the NHR project ELMOD: Efficient language models for on-device deployment (Grant Nr.:

b239dc). NHR funding is provided by federal and Bavarian state authorities. NHR@FAU hardware is partially funded by the German Research Foundation (DFG) – 440719683. We would like to thank our anonymous reviewers and colleagues, Luzian Hahn, Viktor Hangya, Christian Kroos and Anna Leschanowsky, for the useful feedback.

7. References

Hanna Berg, Aron Henriksson, and Hercules Dalianis. 2020. The impact of de-identification on downstream named entity recognition in clinical text. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 1–11, Online. Association for Computational Linguistics.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650.

Richard Chow, Philippe Golle, and Jessica Staddon. 2008. Detecting privacy leaks using corpus-based association rules. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 893–901.

Council of Europe. 2019. Guidelines on safeguarding privacy in the media. Accessed October 2025.

Ona de Gibert, Aitor García-Pablos, Montse Cuadros, and Maite Melero. 2022. Spanish datasets for sensitive entity detection in the legal domain. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3751–3760, Marseille, France. European Language Resources Association.

Louise Deleger, Katalin Molnar, Guergana Savova, Fei Xia, Todd Lingren, Qi Li, Keith Marsolo, Anil Jegga, Megan Kaiser, Laura Stoutenborough, et al. 2013. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. *JAMIA-Journal of the American Medical Informatics Association*, 20(1):84–94.

Josep Domingo-Ferrer, David Sánchez, and Alberto Blanco-Justicia. 2021. The limits of differential privacy (and its misuse in data release and machine learning). *Communications of the ACM*, 64(7):33–35.

Elisabeth Eder, Ulrike Krieg-Holz, and Udo Hahn. 2020. CodE alltag 2.0 — a pseudonymized German-language email corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4466–4477, Marseille, France. European Language Resources Association.

Jeremy Getman, Joe Ellis, Stephanie Strassel, Zhiyi Song, and Jennifer Tracey. 2018. Laying the groundwork for knowledge base population: Nine years of linguistic resources for TAC KBP. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

Rajitha Hathurusinghe, Isar Nejadgholi, and Miodrag Bolic. 2021. A privacy-preserving approach to extraction of personal information through automatic annotation and federated learning. In *Proceedings of the Third Workshop on Privacy in Natural Language Processing*, pages 36–45, Online. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Guha, Sedrick Scott Keh, Kushal Arora, et al. 2024. DataComp-LM: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282.

Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. 2021. Anonymisation models for text data: State of the art, challenges and future directions. In *Proceedings of the 59th*

Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 4188–4203, Online. Association for Computational Linguistics.

Cedric Lothritz, Bertrand Lebichot, Kevin Allix, Saad Ezzini, Tegawendé Bissyandé, Jacques Klein, Andrey Boytsov, Clément Lefebvre, and Anne Goujon. 2023. Evaluating the impact of text de-identification on downstream NLP tasks. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 10–16, Tórshavn, Faroe Islands. University of Tartu Library.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Ben Medlock. 2006. An introduction to NLP-based textual anonymisation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Stéphane M Meystre, Oscar Ferrández, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. 2014. Text de-identification for privacy protection: a study of its impact on clinical text information content. *Journal of biomedical informatics*, 50:142–150.

Stephane M Meystre, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. 2010. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC medical research methodology*, 10(1):70.

Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. 2024. Can LLMs keep a secret? testing privacy implications of language models via contextual integrity theory. In *The 12th International Conference on Learning Representations*.

Jihad S Obeid, Paul M Heider, Erin R Weeda, Andrew J Matuskowitz, Christine M Carr, Kevin Gagnon, Tami Crawford, and Stephane M Meystre. 2019. Impact of de-identification on clinical text classification using traditional and deep learning classifiers. *Studies in health technology and informatics*, 264:283.

Anthi Papadopoulou, Pierre Lison, Lilja Øvrelid, and Ildikó Pilán. 2022. Bootstrapping text anonymization models with distant supervision. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4477–4487, Marseille, France. European Language Resources Association.

Namrata Patel, Pierre Accorsi, Diana Inkpen, Cédric Lopez, and Mathieu Roche. 2013. Approaches of anonymisation of an SMS corpus. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 77–88. Springer.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(4):1053–1101.

Nisha Rawindaran and Vibhushinie Bentotahewa. 2024. Death becomes data. In *Data Protection: The Wake of AI and Machine Learning*, pages 29–45. Springer.

Mercedes Rodriguez-Garcia, Montserrat Batet, and David Sánchez. 2019. Utility-preserving privacy protection of nominal data sets via semantic rank swapping. *Information Fusion*, 45:282–295.

David Sánchez and Montserrat Batet. 2016. C-sanitized: A privacy model for document redaction and sanitization. *Journal of the Association for Information Science and Technology*, 67(1):148–163.

Burkhard Schafer, Jo Briggs, Wendy Moncur, Emma Nicol, and Leif Azzopardi. 2023. What the dickens? post-mortem privacy and intergenerational trust. *Computer Law & Security Review*, 49.

Amber Stubbs, Michele Filannino, and Özlem Uzuner. 2017. De-identification of psychiatric intake records: Overview of 2016 cegs n-grid shared tasks track 1. *Journal of biomedical informatics*, 75:S4–S18.

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus. *Journal of biomedical informatics*, 58:S20–S29.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Vakili, Tyr Hullmann, Aron Henriksson, and Hercules Dalianis. 2024. When is a name sensitive? eponyms in clinical text and implications for de-identification. In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pages 76–80, St. Julian's, Malta. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 35–45.

Jijie Zhou, Eryue Xu, Yaoyao Wu, and Tianshi Li. 2025. Rescriber: Smaller-LLM-Powered User-Led Data Minimization for LLM-Based Chatbots. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–28.

A. Appendix

A.1. Prompts for data synthetization

A.1.1. Private person in nonsensitive context

Prompt 1

```
You are a strict data synthesization
assistant. Your task is to synthesize new
data by replacing every public figure's
name with a random first and last name
of a private person and also transforming
the scenario so it's plausible for
an ordinary person in a public context.
Strict rules:
- Replace the public figure's name with a
random first and last name. The new name
must match the original gender and ethnicity.
- Transform the context from a public
figure's scenario to a private individual's
scenario who is doing a public action.
Keep the theme of the action but adjust the
role and context so it makes sense for
an ordinary person.
- Do NOT explain your changes.
- Output ONLY the transformed text.
```

Prompt 2

```
You are a strict data synthesization
assistant. Your task is to synthesize new
data by replacing every name with a random
first and last name of a private person.
Strict rules:
- Replace ALL names with a random private
first and last name. The new name must
match the original gender and ethnicity.
Do NOT use the same first or last name
twice.
- Be consistent: the same original name must
map to the same pseudonym within a single text.
- Always generate NEW pseudonyms not used
before.
- Do NOT use the names Emily, Ethan, Harper,
Carter, Patel, Whitaker. Use more diverse
and uncommon names.
- Do NOT add explanations.
Return ONLY valid JSON in this format:
{
  "text": "...",
  "names_map": {
      "Original Name": "New Name"
  }
}
```

A.1.2. Private person in private context

Prompt 1 - from public person data

```
You are a strict data synthesization
assistant. Your task is to synthesize new
data by transforming the scenario so it's
plausible for an ordinary person.
Strict rules:
- Transform the given context to a private
individual's scenario. Keep the names and
theme of the action but adjust the role and
context so it makes sense for an ordinary
person.
- Do NOT change any of the names.
- Do NOT explain your changes. Do Not
output reasoning content.
- Do NOT add explanations.
Return ONLY valid JSON in this format:
{
  "original_text": "...",
  "synthesized_text": "..."
}
```

Prompt 1 - from historical/fictional figure data

```
You are a strict data synthesization
assistant. Your task is to synthesize new
data by transforming the scenario so it's
```

```
plausible for an ordinary person.
Strict rules:
- Transform the context from a historical
scenario to a private individual's scenario.
Keep the names and theme of the action but
adjust the role and context so it makes
sense for an ordinary person.
- Do NOT change any of the names.
- Do NOT explain your changes. Do Not
output reasoning content.
- Do NOT add explanations.
Return ONLY valid JSON in this format:
{
  "original_text": "...",
  "synthesized_text": "..."
}
```

Prompt 2

```
You are a strict data synthesization
assistant.
Your task is to synthesize new
data by replacing every person name
with a random first and last name of a
private person.
Strict rules:
- Replace ALL names with a random private
first and last name. The new name must
match the original gender and ethnicity.
Do NOT use the same first or last name
twice.
- Be consistent: the same original name
must map to the same pseudonym within a
single text.
- Always generate NEW pseudonyms not used
before.
- Do NOT use the names Emily, Ethan, Harper,
Carter, Patel, Whitaker. Use more diverse
and uncommon names.
- Do NOT change anything else.
- Do NOT add explanations.
Return ONLY valid JSON in this format:
{
  "text": "...",
  "names_map": {
      "Original Name": "New Name"
  }
}
```

A.1.3. Public person in private context

```
You are a strict data synthesization
assistant. Your task is to synthesize new
data by replacing every historical and
fictional name with a public person's name
and also transforming the scenario so it's
plausible for a non-historical context.
Strict rules:
- Replace the historical and fictional
names with a unique public name. The new name
must match the original gender and ethnicity.
- Names must be unique across the entire
dataset synthesization. If you used a name
once, do NOT use it again for other texts.
- Do not reuse any public person name that
has appeared in previous generations. Do not
use Emma Watson.
- Transform the context from a historical
scenario to a more modern and yet private
individual's scenario. Keep the theme of the
action but adjust the role and context so it
makes sense for the new public person.
- Do NOT explain your changes. Do Not output
reasoning content.
Return ONLY valid JSON in this format:
{
  "text": "...",
  "names_map": {
      "Original Name": "New Name"
  }
}
```

A.1.4. Public person in private context

```
You are a data synthesization
assistant.Your task is to synthesize new
data by replacing every historical and
fictional name with a public person's name
and also transforming the scenario so it's
plausible for a non-historical context.
Strict rules:
- Replace the historical and fictional names
with a unique public name. The new name must
match the original gender and ethnicity.
- Names must be unique across the entire
dataset synthesization. If you used a name
once, do NOT use it again for other texts.
- Do not reuse any public person name that
has appeared in previous generations. Do not
use Emma Watson.
- Replace each with a mononymous star name.
- Transform the context from a historical
scenario to a more modern and yet private
individual's scenario. Keep the theme of the
action but adjust the role and context so
it makes sense for the new public person.
- Do NOT explain your changes. Do Not
output reasoning content.
- Output ONLY the transformed text.
```

A.2. Prompts for classification

A.2.1. Prompt for detecting personal names

```
You are an expert in cybersecurity and
data privacy. Detect personal names of
public and non-public people.
```

```
Furthermore, treat historic characters
(died no later than 1925) and fictional
characters separately.

Return ONLY this JSON format:
{{
  "entity_type":
    "public_or_private_or_
     historic_or_fictional",
  "text": "NAME"
}}

{self.shot_examples[
  str(self.shots)
]["detect_pii"]}
```

A.2.2. Prompt for clustering personal names

```
Return only a JSON object. No explanation.
Cluster the a list of human personal names
from a text referring to one person.

Return ONLY this JSON format:
{{
    "NameVariant1":
    ["NameVariant1", "NameVariant2"]
}}

{self.shot_examples[
    str(self.shots)]
    ["cluster_pii"]}

        {prompt_body}
```

A.2.3. Prompt for classifying context

```
You are an expert in privacy and public
information classification.

Return ONLY this JSON:
{{
    "person": "{person_name}",
    "context": "PrivateContext"
                OR
                "PublicContext"
}}

{self.shot_examples[
    str(self.shots)]
    ["classify_context"]}

Text:
{text}

Person:
{person_name}
```

A.2.4. The different shot versions

```
{
"0": {
  "detect_pii": "",
  "cluster_pii": "",
  "classify_context": ""
},

"1": {
  "detect_pii": "\n        Example:\n
- \"Simone Biles won gold at the Olympics\"
  →\n
  {\"entity_type\":
    \"public\",
    \"text\": \"Simone Biles\"}\n",

  "cluster_pii": "\n        Example:\n
- \"Simone Biles won gold at the Olympics.
  Simone's the best gymnast.\"
  →\n
  {\"Simone Biles\":
 [\"Simone Biles\", \"Simone\"]}\n     ",

 "classify_context": "\n      Example:\n
  - \"Simone Biles\",
 \"Simone Biles won gold at the Olympics.\"
    →\n
    {\"person\": \"Simone Biles\",
 \"context\":\"PublicSetting\"}\n     "
},

"3": {
  "detect_pii":
  "Example:
- \"Simone Biles won gold at the Olympics.\"
  →
  {\"entity_type\":
    \"public\",
    \"text\": \"Simone Biles\"}
 - \"Marie Curie had two daughters,
 Irène and Ève.\"
 → {\"entity_type\":
    \"public\",
    \"text\": \"Marie Curie\"}
   {\"entity_type\":
    \"public\",
    \"text\": \"Irène\"}
  {\"entity_type\":
    \"public\",
    \"text\": \"Ève\"}
- \"John Doe's phone number is 555-1234.\"
 →
 {\"entity_type\":
    \"private\",
    \"text\": \"John Doe\"}",

  "cluster_pii":
  "Example:
```

```
  - \"Simone Biles won gold at the Olympics.
    Simone's the best gymnast.\"
    →
    {\"Simone Biles\":
    [\"Simone Biles\", \"Simone\"]}

    - \"Maria Salomea Skłodowska, or
   Madame Curie, earned two Nobel Prizes for
    her work in physics and chemistry, later
   becoming Prof. Curie and inspiring future
    scientists.\"
    →
    {\"Maria Salomea Skłodowska\":
    [\"Maria Salomea Skłodowska\",
    \"Madame Curie\", \"Prof. Curie\"]}

   - \"John Doe's phone number is 555-1234.
    She called him.\"
    →
    {\"John Doe\":
    [\"John Doe\"]}",

    "classify_context":
    "Example:
    - \"Simone Biles\",
  \"Simone Biles won gold at the Olympics.\"
    →
    {\"person\":\"Simone Biles\",
     \"context\": \"PublicSetting\"}

    - \"Marie Curie\", \"Marie Curie had two
      daughters, Irène and Ève.\"
      →
      {\"person\": \"Marie Curie\",
      \"context\": \"PrivateSetting\"}

   - \"John Doe\", \"John Doe's phone number
      is 555-1234.\"
      →
      {\"person\": \"John Doe\",
      \"context\": \"PrivateSetting\"}"
  }
}
```

Author Index