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.

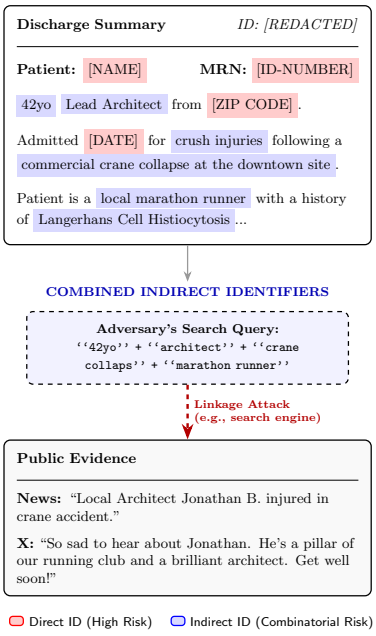

Figure 1: Illustration of re-identification risk through indirect identifier linkage in an artificial MIMIC-style document with already redacted PHI.

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.

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

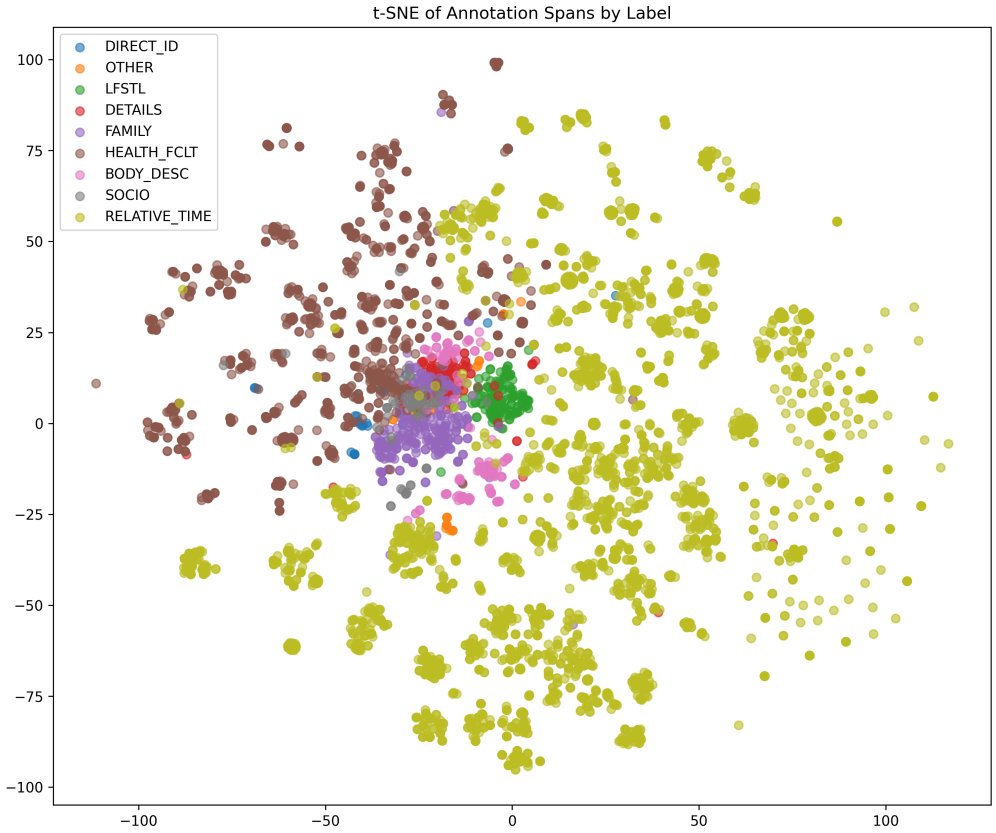

Figure 2: t-SNE projection of cosine similarity between annotated spans using BERT-based sentence embeddings.

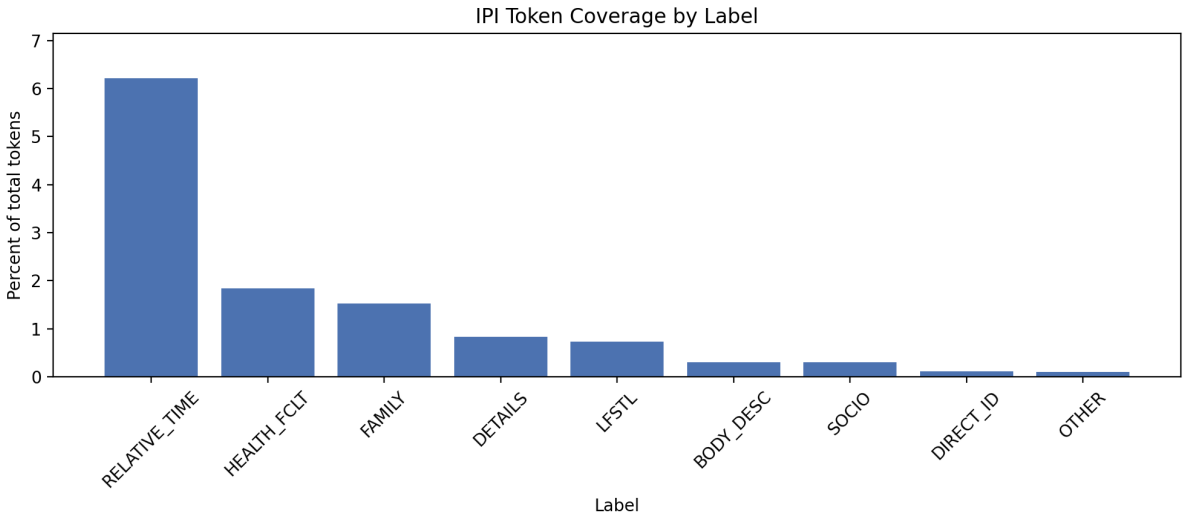

Figure 3: Token-level coverage of indirect personal identifier spans in the given dataset.

	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.