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.

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.

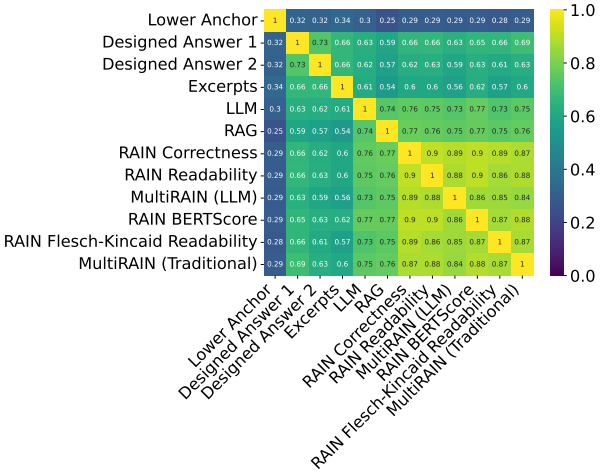
Percentiles: We computed two percentile-based thresholds. For interval metrics, we used the 10th and 90th percentiles as outer bounds and the 25th and 75th percentiles as the interquartile range. For other metrics, we used either the 10th or 25th percentile as a lower bound. The 10th percentile excludes 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

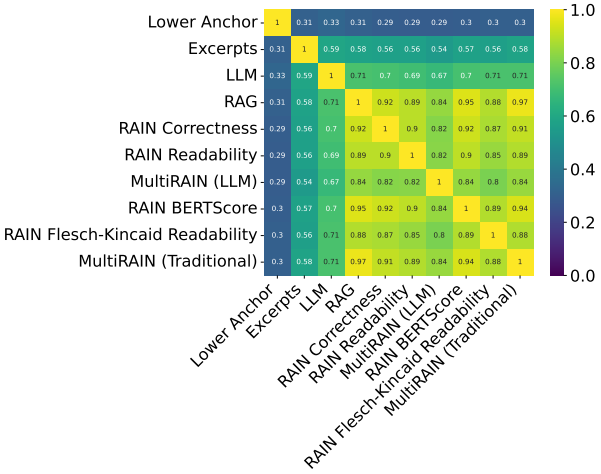
While expert-generated answers serve as an upper bound, a lower anchor is provided by random-word answers to all questions, with words drawn from the vocabulary of system-generated answers. We used an average sentence length of 18 words for the lower-anchor answers, matching the average sentence length 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-

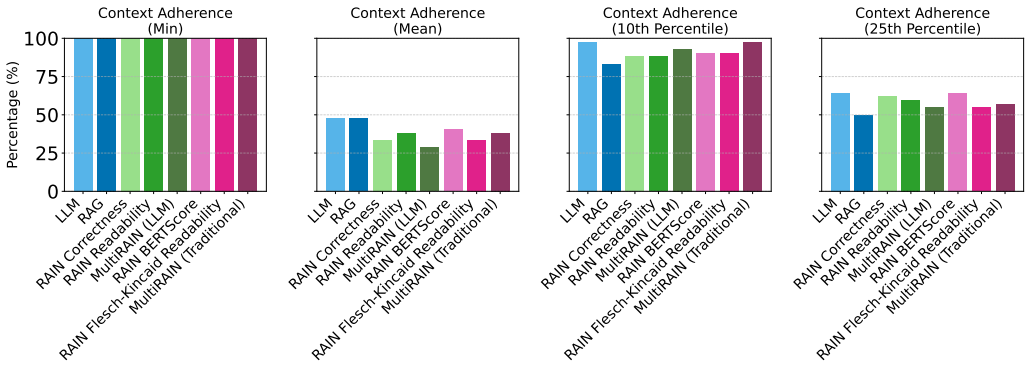


Figure 2: Comparison of "at least as good" as human expert answers under different thresholds in the expert-generated dataset.

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,

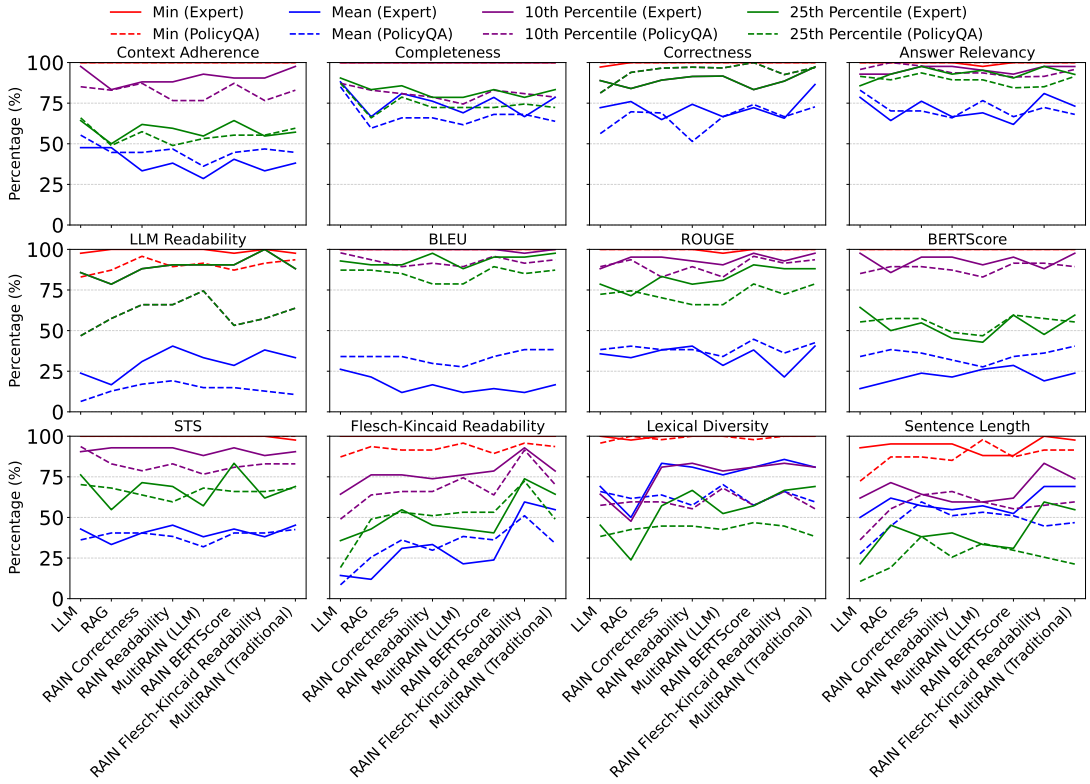


Figure 3: Performance trends for both evaluation datasets across systems for each evaluation metric. The mean is chosen as a threshold for exemplary purposes.

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.

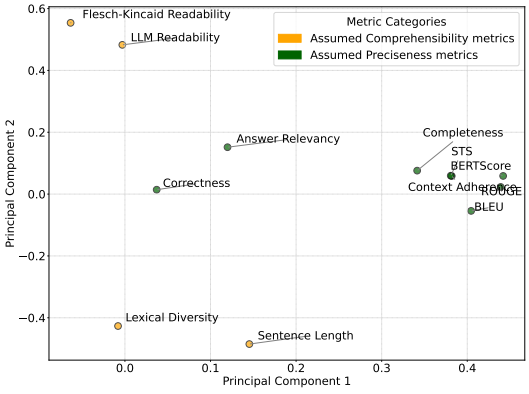


Figure 4: The 2D PCA projection shows relationships between text evaluation metrics. Metrics are colored based on their assumed relationship, i.e., precision (dark green) and comprehensibility (yellow).

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;
```

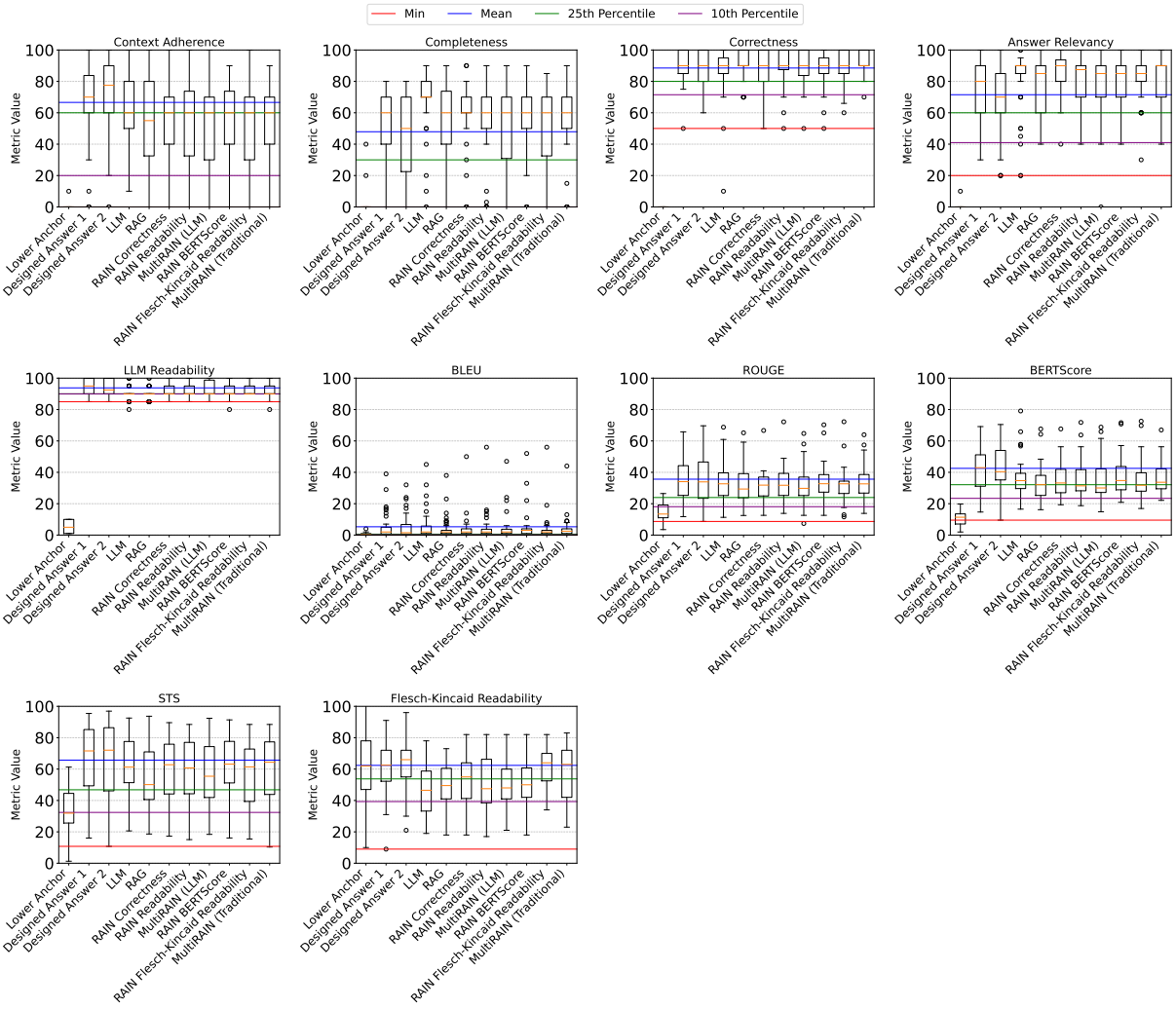


Figure 5: Raw metric scores and thresholds evaluated on the expert-generated dataset for all metrics where “bigger is better”. Threshold computation depends on the metric implementation (see Section 5.2).

12.2. Raw Metric Scores and Thresholds

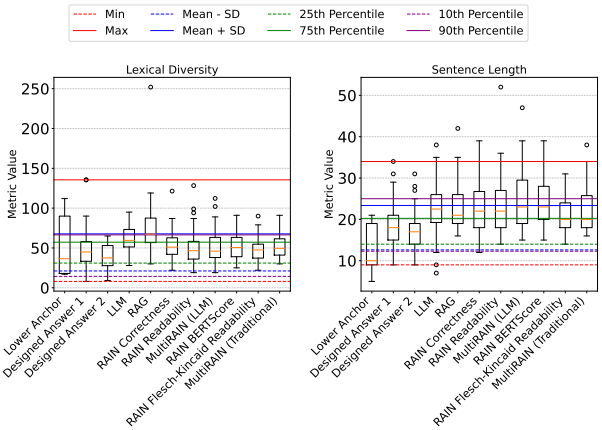


Figure 6: Raw metric scores and thresholds evaluated on the expert-generated dataset for all interval-based metrics (see Section 5.2 for information on threshold computation).

12.3. PCA

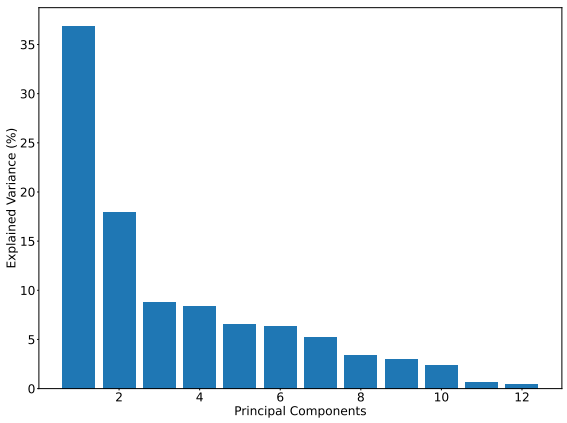


Figure 7: Explained variance by principal components. The first and second principal components are projected as x- and y-axes, respectively, in Figure 4.

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'.