CAPID: Context-Aware PII Detection for Question-Answering Systems

Mariia Ponomarenko[1], Sepideh Abedini[1,2], Masoumeh Shafieinejad[2], D. B. Emerson[2], Shubhankar Mohapatra[1], Xi He[1,2]
[1]University of Waterloo, [2]Vector Institute
Correspondence: m2ponoma@uwaterloo.ca

Abstract

Detecting personally identifiable information (PII) in user queries is critical for ensuring privacy in question-answering systems. Current approaches typically redact all PII, disregarding the possibility that some may be contextually relevant to the user's question, thereby degrading response quality. Large language models (LLMs) may help determine which PII is relevant; however, due to their closed-source nature and lack of privacy guarantees, they are unsuitable for processing sensitive data. To achieve privacy-preserving PII detection, we propose CAPID, a practical approach that fine-tunes a locally owned small language model (SLM) that filters sensitive information before it is passed to LLMs for QA. However, existing datasets do not capture the context-dependent relevance of PII needed to train such a model effectively. To address this gap, we propose a synthetic data generation pipeline that leverages LLMs to produce a diverse, domain-rich dataset spanning multiple PII types and levels of relevance. Using this dataset, we fine-tune an SLM to detect PII spans, classify their types, and estimate contextual relevance. Our experiments show that relevance-aware PII detection with a fine-tuned SLM substantially outperforms existing baselines in span, relevance and type accuracy while preserving higher downstream utility under anonymization.

1 Introduction

In today's digital era, individuals frequently disclose personal information while interacting with online platforms such as conversational assistants and chatbots, particularly when seeking advice or posing questions (Saffarizadeh et al., 2018). These disclosures often involve personally identifiable information (PII), raising significant privacy concerns. Regulatory frameworks, such as GDPR (Tikkinen-Piri et al., 2018), have been established to protect personal data and ensure responsible handling of sensitive information.

Example 1.
Original query: *I'm a warehouse supervisor with chronic back pain from lifting heavy boxes. I live in Springfield and have two children. How can I reduce fatigue after long shifts?*

Generic PII redaction: *I'm a [OCCUPATION] with [HEALTH] from lifting heavy boxes. I live in [LOCATION] and have [FAMILY]. How can I reduce fatigue after long shifts?*

Context-aware redaction: *I'm a warehouse supervisor with chronic back pain from lifting heavy boxes. I live in [LOCATION] and have [FAMILY].* ***How can I reduce fatigue after long shifts?***

The example illustrates how context-aware redaction preserves personal information relevant to reasoning about the user's question. For the question "How can I reduce fatigue after long shifts?", information about the user's occupation (warehouse supervisor) and condition (back pain) is valuable to interpreting the cause of fatigue, whereas location and family details are unrelated and thus safely masked.

To protect user privacy, numerous privacy tools have been developed to detect and redact PII (Allal et al., 2023; Pilán et al., 2022). However, most of these tools (Microsoft, 2021; Amazon, 2025) do not account for the contextual relevance of the information they flag. As a result, they can obscure information essential for accurate and contextually appropriate response. In certain settings, retaining specific sensitive information is justified (Nissenbaum, 2004), as some private details are directly relevant to a user's goal. Although most existing approaches focus on general PII detection, some recent studies have begun to explore context-sensitive methods (Shen et al., 2025; Dou et al., 2024; Ngong et al., 2025). At the same time, LLMs have demonstrated remarkable performance across a range of tasks (Brown et al., 2020). Yet their widespread use through third-party APIs (e.g., OpenAI, Anthropic) raises privacy concerns, as user queries containing sensitive data may be transmitted to external

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 4: Student Research Workshop, pages 320–331

servers. To mitigate these concerns, fine-tuning local models for specific privacy-preserving tasks becomes essential. Nevertheless, to the best of our knowledge, no datasets or models have been publicly released for context-sensitive PII detection, and there is a lack of evaluation of how such context-sensitive redaction affects downstream application performance, such as question answering with LLMs. To this end, we present the following contributions.

1. Introducing CAPID, a synthetic dataset for context-aware PII detection. CAPID focuses on the relevance of PII spans with respect to a given question across diverse topics. The dataset is designed to support fine-tuning and evaluation of context-aware models that must reason not only about the presence of PII, but also about whether such information should be retained or masked in the question-answering tasks.

2. Showing the effectiveness of CAPID by training and evaluating several SLMs, including Llama-3.1-8B and Llama-3.2-3B, for context-aware PII detection, achieving an accuracy score improvement from 0.68 to 0.79 in classifying PII relevance compared to GPT-4.1-mini.

3. Exhibiting that relevance-aware anonymization preserves significantly more downstream answer utility than existing anonymization baselines using an LLM-as-a-judge approach. This is demonstrated by collecting and annotating real user queries from Reddit and evaluating LLM-generated answers under different masking strategies.

We open-source the code with the dataset[1] and the model[2].

2 Related Work

Most existing PII detection systems are built on transformer-based NER models that identify a small, fixed set of entity types such as names, locations, and organizations (Microsoft, 2021; Amazon, 2025). Subsequent research has pursued finer-grained, domain-specific detection using synthetic data (Jangra et al., 2025), knowledge-graph supervision (Papadopoulou et al., 2022a), federated learning (Hathurusinghe et al., 2021), or LLM-based generation (Ngong et al., 2025) to expand coverage of PII and self-disclosure. Other works fine-tune large encoder models for span-level self-disclosure detection (Dou et al., 2024) or use LLMs to infer a wide range of personal attributes from text (Staab et al., 2024). Despite these advances, current methods fail to capture which PII are contextually relevant, often leading to excessive redaction and loss of information essential for accurate response generation (Pal et al., 2024; Larbi et al., 2022; Lukas et al., 2023).

Some recent efforts attempt to address this limitation by fine-tuning models for contextual PII detection. However, relevance is primarily defined through the distinction between public and private information (Xiao et al., 2024). In contrast, we fine-tune SLMs to achieve a more nuanced understanding of PII relevance within the context of a user's question. Ngong et al. (2025) estimated contextual relevance of PII using pretrained SLMs. However, relying solely on pretrained models can lead to lower accuracy than task-specific fine-tuning. Furthermore, existing corpora to fine-tune such models are limited in scope and quality. The Text Anonymization Benchmark focuses primarily on legal text and a narrow range of identifiers (Pilán et al., 2022). The pii-masking-300k dataset (AI4Privacy, 2022) provides broad topical coverage but limits annotations to direct identifiers, such as names or contact information. Other attributes in the text, such as occupation, education, or health, while not uniquely identifying an individual, can still disclose personal details and may therefore warrant masking. In our work, we treat these self-disclosed attributes as part of the privacy surface that should be protected. Dou et al. (2024) similarly annotate such attributes, labeling 4.8K spans with importance scores across 2.4K Reddit posts. Unfortunately, the released data omitted these scores and defined importance only at the message level. Similarly, Shen et al. (2025) developed an evaluation dataset for query-related PII detection; however, it is restricted to the job domain, and relevant PII is trivially linked to queries via explicit references, limiting generalization to more natural interactions. Consequently, no existing dataset adequately supports modeling the contextual relevance of PII across diverse scenarios.

[1] https://github.com/MariaPonomarenko38/CAPID
[2] https://huggingface.co/ponoma16/capid-llama8b-lora

3 Problem Statement

We consider a question-answering system backed by an externally hosted LLM, treated as untrusted (Wang et al., 2025). A user query consists of a *context*, C, and *question*, Q. Context refers to the textual background accompanying a question and provides essential information for accurately interpreting the question and deriving the correct answer. A context may contain sensitive text spans, defined as contiguous sequences of tokens that disclose personal information about the user. These spans are referred to as personally identifiable information or **PII** and denoted

$$P = \{p_1, p_2, \ldots, p_n\}, \quad p_i \subseteq C.$$

Each p_i is a contiguous subsequence of tokens within the context C and is associated with a type

$$t(p_i) \in \mathcal{T},$$

where $\mathcal{T}$ denotes the set of possible PII types, such as nationality, occupation, or medical condition.

To preserve user privacy, their query is commonly anonymized either by eliminating the PII or replacing them with abstracted forms (Dou et al., 2024). However, certain PII are essential for generating accurate responses and are intentionally shared as part of the user's goal, yet in many real-world cases, users include irrelevant PII in the context C that are unnecessary for answering the question Q. Therefore, to balance privacy and utility, we selectively retain only those PII that align with user intentions (see Example 1).

The goal is to assign each p_i a binary relevance label indicating whether it should be retained or masked prior to answering the question. This is formalized by a relevance function

$$r : P \times Q \to \{0, 1\},$$

which maps each $p_i \in P$ to a relevance score conditioned on the question Q, where 1 denotes high relevance and 0 denotes low relevance.

The problem is therefore defined as follows. Given a context C and a question Q containing a set of PII spans

$$P = \{p_1, p_2, \ldots, p_n\},$$

the task is to predict, for each $p_i \in P$, both a type label $t(p_i)$ and a binary relevance label $r(p_i, Q) \in \{0, 1\}$. The type label enables the model to distinguish among categories of sensitive information, thereby improving interpretability and supporting category-specific redaction strategies. The relevance label, in turn, captures the contextual importance of each PII span with respect to the question.

4 CAPID

To address limitations in the existing literature and facilitate training of local models for the mappings defined in the earlier section, we propose a principled LLM-based pipeline for generating context-aware PII detection datasets. As illustrated in Figure 1, the pipeline consists of a three-stage generation process followed by a rigorous manual evaluation. Synthetic samples are produced using GPT-4.1-mini and GPT-5. We provide comprehensive generation configurations and prompt templates in Appendix A and B, respectively.

4.1 Topics Generation

One limitation of previous work is the narrow domain coverage of the generated samples. If a dataset is dominated by a small set of PII types or topics, models risk overfitting and may not generalize to other settings. To encourage diverse representation across PII types, contexts, and questions, the LLMs are conditioned on carefully designed prefixes. Specifically, we begin with a broad list of PII types based on the taxonomy introduced in (Papadopoulou et al., 2022b; Dou et al., 2024), and reorganize them into more fine-grained categories: occupation, health, demographic, finance, age, education, location, organization, relationship, sexual orientation, belief, name, code (e.g., structured identifiers), datetime, and appearance. Thereafter, all possible unordered pairs among these types (except name and code) are enumerated, resulting in 78 distinct combinations. Each pair is then used as a prefix to generate 10 topics in which both types of PII are contextually relevant. For each topic, 20 subtopics are generated to expand thematic variety, producing 15,600 topic–subtopic pairs. After de-duplication, 11,663 unique triplets of the form (PIIType1, PIIType2, subtopic) are retained and serve as the basis for subsequent sample generation.

4.2 PII, Context and Question Generation

As illustrated in the right-hand section of Figure 1, context is generated using sample-wise decomposition (Long et al., 2024), a step-by-step strat-

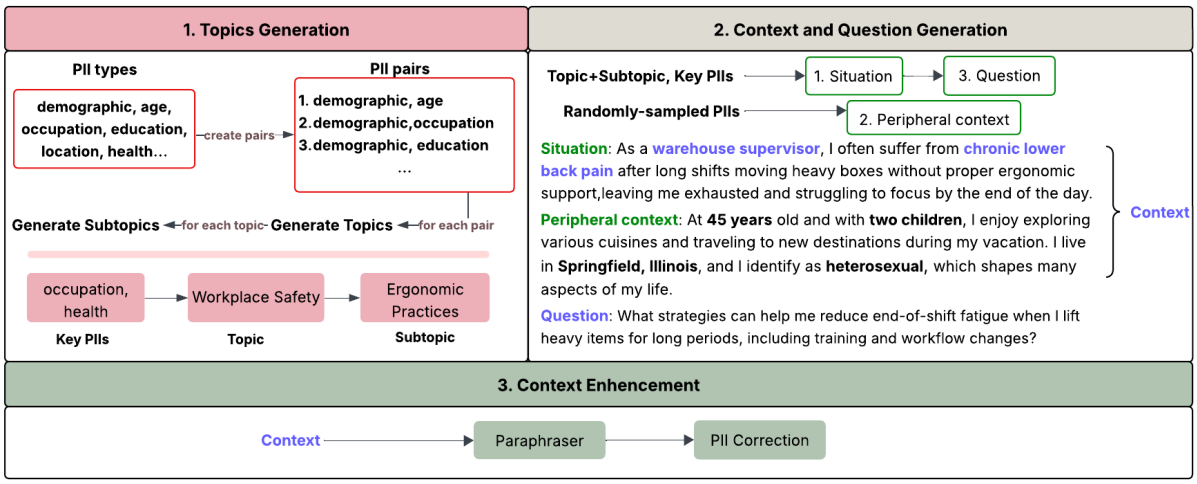

Figure 1: The three-stage sequential pipeline for generating the dataset. Stage 1: Topics Generation, which conditions the LLM for subsequent sampling. Stage 2: PII, Context and Questionv Generation, involving sample-wise decomposition to create a context containing both relevant and irrelevant PII, followed by situational question formulation. Stage 3: Optimization for Relevance and Coherence, where various techniques are applied to augment the contextual data.

egy in which each context is constructed from two components: the situation and the peripheral context. This structure ensures that each sample includes both relevant and irrelevant PII. The situation describes the central scenario that naturally motivates the subsequent question and contains all relevant PII associated with the topic–subtopic pair. The peripheral context provides unrelated information surrounding the event, enabling control over contextual relevance. For the generation of PII values, the model is conditioned on a prefix, either the partially generated context or the topic-subtopic phrase, to promote contextual variation and logical consistency. Question generation follows a two-step process. First, the LLM generates a question that may reflect certain key PII, for example, "What are effective ways to manage tiredness caused by **ongoing treatment**?", which hints at a medical condition. Second, a refinement prompt abstracts these cues, producing a neutral variant such as "What are effective ways to reduce tiredness?" while preserving the question's intent. This approach ensures that the relevance of the PII is implicit, thereby more closely simulating real-world conversational patterns in which individuals often pose abstract or broad questions.

4.3 Context Enhancement

The context is then paraphrased using a prompted LLM, which rephrases and restructures the text to improve fluency and diversity while ensuring that high-relevance PII do not consistently appear at the beginning of the passage. Finally, a consistency check is performed to ensure that all original PII values remain unchanged after these modifications. The system iterates through all PII; if an exact string match is not found in the modified context, an LLM is prompted to identify the closest textual span, and the corresponding span label is updated accordingly to maintain data integrity throughout the augmentation process.

4.4 Data Validation

To ensure the quality and reliability of the generated dataset, we manually verify and correct the annotations produced by the LLM. This post-processing step is essential because, for any given pair of key PIIs, it is challenging to automatically generate a question that strictly adheres to our core criteria: a high-relevance PII must be strongly and indispensably linked to the question such that answering correctly is impossible or exceedingly difficult without it. Five annotators are trained to identify and reclassify PII that initially appears to be of low relevance but is, in fact, highly relevant for answering the question, and vice versa. They also ensure that the questions do not contain abstracted forms of relevant PII. In cases where this is impossible, certain linguistic cues, such as the types of the high-relevant PII, are permitted.

To standardize this nuanced judgment, we provide detailed annotation guidelines, a custom-built Streamlit tool to efficiently edit PII types, questions, context, and relevance scores, a comprehen-

sive video tutorial, and a set of example annotations illustrating various cases (see annotated examples in Appendix F). Given constraints on time and resources, the refinement process yields a final dataset of 2,307 samples, partitioned into a training set of 2,107 entries and a test set of 200 entries, which we consider sufficient for subsequent fine-tuning of an SLM. This work is essential for creating a coherent, diverse, and consistently annotated evaluation benchmark.

As shown in Table 1, the relevance distribution varies considerably across PII types. Categories such as occupation, health, demographic information, and location show a roughly balanced split between high and low relevance, indicating that they often play a meaningful role in answering the associated questions. In contrast, attributes such as relationship, education, age, and organization are relatively unimportant, meaning they tend to appear as peripheral details rather than as information required to derive the answer. Finally, name and code are almost always irrelevant, indicating that explicit identifiers are rarely necessary to resolve the question.

PII Type	Total Count	High Prop	Low Prop
occupation	1202	0.52	0.48
health	1226	0.56	0.44
demographic	1214	0.48	0.52
finance	1103	0.38	0.62
age	1085	0.26	0.74
education	975	0.24	0.76
location	917	0.48	0.52
organization	986	0.26	0.74
relationship	950	0.19	0.81
sexual orientation	932	0.21	0.79
belief	684	0.29	0.71
name	464	0.01	0.99
code	526	0.00	1.00
datetime	665	0.29	0.71
appearance	640	0.28	0.72

Table 1: Total PII counts and proportions of high and low relevance in the CAPID dataset.

5 Evaluation

5.1 Model Training Performance

We fine-tune Llama-3.2-3B and Llama-3.1-8B using the Unsloth framework (Daniel Han and team, 2023) with 4-bit quantization and LoRA adaptation (Hu et al., 2021) to perform span extraction, PII type prediction, and contextual relevance estimation. Training follows a standard causal language modeling formulation in which only the JSON-formatted label section of each formatted prompt contributes to the loss. Each input sample contains an Alpaca-style instruction (Appendix D), a context C, a question Q, and the expected structured PII annotations. Additional training details appear in Appendix C.

To benchmark against existing approaches, we also evaluate the method proposed by Ngong et al. (2025), which analyzes user input, detects contextually unnecessary details, and reformulates prompts to preserve intent while minimizing disclosure. Although their approach is not designed as a PII detection tool, it identifies sensitive details in the user query and partitions the input into `related_context` and `not_related_context`. This separation provides a suitable basis for comparison, allowing us to contrast our relevance-based PII annotations with their categorization of information as contextually necessary or unnecessary. We include Microsoft Presidio as a representative rule-based PII detection system that identifies and anonymizes predefined PII types without modeling contextual relevance, thereby serving as a non-contextual baseline. In addition, we compare our fine-tuned models with GPT-4.1-mini (using the prompt provided in the Appendix D.2). Although it is not suitable for PII-sensitive deployment due to privacy constraints, we include it as a baseline illustrating the performance of a proprietary LLM.

Model performance is evaluated along three dimensions: (i) span, (ii) PII type, and (iii) relevance. Span metrics quantify the model's ability to precisely identify PII-containing spans within the context. PII-type metrics assess whether the predicted type (e.g., occupation, nationality, or location) is correct, given correct span detection, ensuring that type classification is evaluated only when the PII span is located correctly. Relevance metrics measure whether the model can accurately judge the contextual importance of each PII instance with respect to the question.

For span quality, we report both micro-averaged precision, recall, and F1, as well as coverage, computed using a hybrid token–character F1 score: single-token spans are matched via character-level alignment while multi-token spans are scored using token overlap between predicted and gold spans. For PII type and relevance prediction, we report accuracy computed only over correctly matched spans. Type accuracy measures whether the predicted PII category matches the gold label, while relevance accuracy measures whether the predicted

Model	Span				Type	Relevance		
	P	R	F1	Cov.	Acc.	Acc.	Low Acc.	High Acc.
GPT-4.1-mini	0.8724	0.9438	0.8986	0.8957	0.9008	0.8396	0.8772	0.7254
Microsoft Presidio	0.7020	0.4393	0.5070	0.7992	0.3138	–	–	–
Llama-3.1-8B	0.4080	0.7018	0.4813	0.5294	0.5285	0.5129	0.5991	0.3050
Llama-3.1-8B (FT)	**0.9650**	**0.9598**	**0.9603**	**0.9606**	**0.9674**	**0.9306**	**0.9413**	**0.8704**
Llama-3.2-3B (FT)	0.9650	0.9608	0.9608	0.9606	0.9674	0.9306	0.9413	0.8704
Llama-3.1-8B (Ngong et al., 2025)	0.5704	0.7896	0.6439	0.6973	–	0.7002	0.8635	0.3408
Llama-3.2-3B (Ngong et al., 2025)	0.5997	0.4323	0.4708	0.6427	–	0.5925	0.8033	0.0050

Table 2: PII detection performance on the CAPID test set (200 samples). Type and relevance metrics are conditioned on correct spans. GPT-4.1 mini is an untrusted proprietary baseline.

Model	Span				Type	Relevance		
	P	R	F1	Cov.	Acc.	Acc.	Low Acc.	High Acc.
GPT-4.1-mini	0.7586	**0.9128**	0.8107	0.9098	**0.8928**	0.6896	0.5924	0.6923
Microsoft Presidio	0.7162	0.5005	0.5625	0.8360	0.6711	–	–	–
Llama-3.1-8B	0.3493	0.5669	0.3968	0.4894	0.2895	0.4277	0.5283	0.1678
Llama-3.1-8B (FT)	**0.8618**	0.8135	**0.8159**	**0.9135**	0.8606	**0.7994**	**0.6823**	**0.8004**
Llama-3.2-3B (FT)	0.8251	0.8089	0.7973	0.8872	0.8366	0.7195	0.6530	0.6679
Llama-3.1-8B (Ngong et al., 2025)	0.4902	0.7100	0.5572	0.5816	–	0.6072	0.54633	0.5161
Llama-3.2-3B (Ngong et al., 2025)	0.6258	0.4570	0.4835	0.6031	–	0.5078	0.5728	0.1456

Table 3: PII detection on the Reddit set (150 samples). Type and relevance metrics are conditioned on correct spans. GPT-4.1 mini is an untrusted proprietary baseline.

binary relevance label is correct. To provide finer-grained insight into relevance performance, we additionally report accuracy separately for low-relevance and high-relevance PII spans.

Across all metrics in Table 2, fine-tuned models substantially outperform alternative baselines. Llama-3.1-8B (FT) raises span F1 from 0.48 to 0.96 and relevance accuracy from 0.51 to 0.93 compared to the pre-trained only model. This demonstrates that relevance estimation and span-aware PII detection strongly benefit from task-specific supervision. Although GPT-4.1-mini achieves a comparable span recall of 0.94, it does not match the performance of the fine-tuned models. We observe lower performance for Microsoft Presidio and Ngong et al. (2025) for the following reasons. Presidio is limited to a narrower set of PII categories than those considered in our evaluation, which reduces its recall and type accuracy when the PII types are broader. In contrast, the method of Ngong et al. (2025) is not designed for precise PII detection at the span level; it frequently identifies context fragments that are not truly sensitive, resulting in a high number of false-positive spans and, consequently, weaker span and relevance scores.

5.2 Downstream Performance

To understand the behavior of our approach beyond controlled synthetic settings, we evaluate it using real Reddit data and measure the utility impact of different anonymization strategies.

5.2.1 Evaluation on Reddit Data

In addition to synthetic samples, we evaluate the model's performance on text authored by real users, in which linguistic structure, ambiguity, tone, and contextual cues exhibit substantially greater variability. We collect 150 Reddit excerpts that contain naturally occurring personal information. We source content from a diverse set of subreddits, including `r/movetojapan`, `r/movetoscotland`, `r/confessions`, and `r/jobs`, where users frequently disclose sensitive personal information when asking for advice or describing life circumstances. Manual annotation of relevance is challenging, as determining whether a PII attribute is required to answer a question often requires domain-specific expertise (e.g., immigration, employment regulations, or mental health counseling). To ensure consistent and accurate labeling we construct the questions for 100 samples ourselves, allowing unambiguous identification of relevant versus irrelevant PII. For the remaining 50 samples, we preserve the original user-written Reddit questions.

We compare our fine-tuned models with the same baselines as in the Table 2. As shown in Table 3, GPT-4.1-mini achieves strong span detection, but tends to over-predict PII spans, reflected in very

high recall relative to precision. Notably, our fine-tuned Llama-3.1-8B achieves substantially higher relevance accuracy than GPT-4.1-mini (0.7994 vs. 0.6896), with substantial gains on both low- and high-relevance PIIs, while being much smaller and trained exclusively on our dataset. Finally, the relevance predictions from Ngong et al. (2025) are substantially weaker, confirming that contextual relevance of PII remains a challenging modeling problem. We have also analyzed accuracy by PII type on the Reddit dataset, comparing Llama-3.1-8B fine-tuned on our data with GPT-4.1-mini. Table 4 shows that performance varies across types, with each model outperforming the other in different categories while both achieve perfect accuracy for some types (e.g., name and organization).

Type	Llama-3.1-8B (FT)	GPT-4.1-mini
health	0.9473	0.9500
location	0.8351	0.9414
sexual orientation	0.8000	1.0000
occupation	0.9400	0.8474
age	0.8870	0.9558
relationship	0.9464	0.9218
name	1.0000	1.0000
education	0.7741	0.8815
appearance	0.8000	0.8333
code	-	1.0000
organization	1.0000	1.0000
finance	0.9714	0.9000
datetime	0.7027	0.9000
demographic	0.8181	0.8163

Table 4: Type accuracy by PII type on the Reddit dataset. Metrics are reported only for correctly detected spans. A dash (-) denotes that no spans of the given PII type were detected.

5.2.2 Utility Analysis

To assess the practical effect of relevance-aware anonymization, we evaluate utility trade-offs on the Reddit benchmark using the same samples as above. For each context-question pair, we have GPT-4 generate answers to the provided questions under two anonymization settings: (1) full masking, where all detected PII is anonymized, and (2) low-relevance masking, where only PII marked as low relevance is anonymized while high-relevance PII is preserved. To quantify utility, we use the same LLM as a judge, providing it with the original, unmasked context and question, and asking it to decide which answer is more accurate and useful. Additionally, the experiment is replicated using a different judge LLM, Claude Sonnet 4.5, decoupling answer generation and evaluation.

The utility scores in Table 5 report the proportion of cases in which the answer derived from the low-relevance masked context setting is preferred over the fully masked answer. As shown in Table 5, relevance-sensitive anonymization with our fine-tuned model consistently yields higher response utility than Ngong et al. (2025), improving utility by 22% on Reddit and 28% on the CAPID test set. The prompts for evaluating downstream performance are provided in Appendix E.

Method	Dataset	GPT-4	Claude
Llama-3.1-8B (FT)	Reddit	0.80	0.79
	CAPID test set	0.79	0.73
Llama-3.1-8B (Ngong et al., 2025)	Reddit	0.58	0.48
	CAPID test set	0.51	0.43

Table 5: Utility preservation scores showing the proportion of cases in which the *low-relevance masked* context leads to a better answer compared to the *fully masked* context.

6 Conclusion

This work introduces a context-aware approach to PII detection and anonymization, addressing a core limitation of existing systems that treat all personal information as equally sensitive. By modeling not only which spans constitute PII but also whether each attribute is essential for downstream task performance, our method enables selective preservation of high-relevance information while masking only what is truly unnecessary. Across both synthetic and naturally occurring Reddit data, our fine-tuned Llama model substantially outperforms strong baselines, including GPT-4-mini and Microsoft Presidio, in span detection, type assignment, and relevance classification. Moreover, relevance-aware masking yields consistently higher answer utility than fully masked anonymization, demonstrating that preserving contextually important PII can materially improve model performance in settings where retention of PII required to accurately perform a downstream task is justified.

Limitations

Our approach highlights several opportunities for refinement and future work. First, current models struggle with very long sequences, and the quality of relevance estimation degrades as contexts become large and information-dense. Although this can theoretically be mitigated by chunking or

summarization, improving long-context reasoning remains an important direction. Second, the quality of relevance predictions is noticeably higher when the associated question contains linguistic cues that indirectly signal informational needs (e.g., terms such as "local", "near me", "in my area" for location–critical questions). In fully neutral formulations where no hints are present, the relevance distinction becomes more ambiguous, making prediction harder even for humans. Third, our framework operates in a domain-agnostic manner and assumes that relevance is reasonably assessable by a non-expert annotator. However, domain-specific settings such as immigration, legal advice, or medical diagnosis have their own rules and contextual dependencies that can determine the contextual relevance of a PII element with respect to the question. Developing customizable or domain-adaptive relevance policies, potentially informed by expert knowledge, would make the method more broadly usable in specialized applications. Although CAPID reduces the number of PIIs revealed to LLMS, it currently uses binary scoring for sensitivity allocation. Hence, all detected PII, including the revealed highly relevant ones, are considered highly sensitive. Future research is needed to extend CAPID to continuous sensitivity scores, and adjust accordingly to limit privacy leakage even further.

References

AI4Privacy. 2022. Pii masking 300k dataset. https://huggingface.co/datasets/ai4privacy/pii-masking-300k. Accessed: 2025-10-03.

Loubna Ben Allal, Raymond Li, Denis Kocetkov, Chenghao Mou, Christopher Akiki, Carlos Munoz Ferrandis, Niklas Muennighoff, Mayank Mishra, Alex Gu, Manan Dey, Logesh Kumar Umapathi, Carolyn Jane Anderson, Yangtian Zi, Joel Lamy Poirier, Hailey Schoelkopf, Sergey Troshin, Dmitry Abulkhanov, Manuel Romero, Michael Lappert, and 22 others. 2023. Santacoder: don't reach for the stars! *Preprint*, arXiv:2301.03988.

Amazon. 2025. Amazon comprehend. https://aws.amazon.com/comprehend/. Accessed: 2025-10-03.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, and Wei Xu. 2024. Reducing privacy risks in online self-disclosures with language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13732–13754, Bangkok, Thailand. 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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Shalini Jangra, Suparna De, Nishanth Sastry, and Saeed Fadaei. 2025. Protecting vulnerable voices: Synthetic dataset generation for self-disclosure detection. *Preprint*, arXiv:2507.22930.

Iyadh Ben Cheikh Larbi, Aljoscha Burchardt, and Roland Roller. 2022. Which anonymization technique is best for which nlp task? – it depends. a systematic study on clinical text processing. *Preprint*, arXiv:2209.00262.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082, Bangkok, Thailand. Association for Computational Linguistics.

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Beguelin. 2023. Analyzing Leakage of Personally Identifiable Information in Language Models . In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 346–363, Los Alamitos, CA, USA. IEEE Computer Society.

Microsoft. 2021. Presidio. https://microsoft.github.io/presidio/. Accessed: 2025-10-03.

Ivoline C. Ngong, Swanand Ravindra Kadhe, Hao Wang, Keerthiram Murugesan, Justin D. Weisz, Amit Dhurandhar, and Karthikeyan Natesan Ramamurthy. 2025. Protecting users from themselves: Safeguarding contextual privacy in interactions with conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26196–26220, Vienna, Austria. Association for Computational Linguistics.

Helen Nissenbaum. 2004. Privacy as contextual integrity. *Washington Law Review*, 79.

Anwesan Pal, Radhika Bhargava, Kyle Hinsz, Jacques Esterhuizen, and Sudipta Bhattacharya. 2024. The empirical impact of data sanitization on language models. *Preprint*, arXiv:2411.05978.

Anthi Papadopoulou, Pierre Lison, Lilja Øvrelid, and Ildikó Pilán. 2022a. 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.

Anthi Papadopoulou, Yunhao Yu, Pierre Lison, and Lilja Øvrelid. 2022b. Neural text sanitization with explicit measures of privacy risk. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 217–229, Online only. Association for Computational Linguistics.

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.

Kambiz Saffarizadeh, Maheshwar Boodraj, and Tawfiq Alashoor. 2018. Conversational assistants: Investigating privacy concerns, trust, and self-disclosure. In *ICIS 2017 Proceedings*, Proceedings of the International Conference on Information Systems. Association for Information Systems. AIS Electronic Library (AISeL). 38th International Conference on Information Systems: Transforming Society with Digital Innovation, ICIS 2017 : Transforming Society with Digital Innovation, ICIS 2017 ; Conference date: 10-12-2017 Through 13-12-2017.

Hao Shen, Zhouhong Gu, Haokai Hong, and Weili Han. 2025. Pii-bench: Evaluating query-aware privacy protection systems. *Preprint*, arXiv:2502.18545.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2024. Beyond memorization: Violating privacy via inference with large language models. *Preprint*, arXiv:2310.07298.

Christina Tikkinen-Piri, Anna Rohunen, and Jouni Markkula. 2018. EU General Data Protection Regulation: Changes and Implications for Personal Data Collecting Companies. *Computer Law & Security Review*, 34(1):134–153.

Shang Wang, Tianqing Zhu, Bo Liu, Ming Ding, Dayong Ye, Wanlei Zhou, and Philip Yu. 2025. Unique security and privacy threats of large language models: A comprehensive survey. *ACM Comput. Surv.*, 58(4).

Yijia Xiao, Yiqiao Jin, Yushi Bai, Yue Wu, Xianjun Yang, Xiao Luo, Wenchao Yu, Xujiang Zhao, Yanchi Liu, Quanquan Gu, Haifeng Chen, Wei Wang, and Wei Cheng. 2024. Large language models can be contextual privacy protection learners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14179–14201, Miami, Florida, USA. Association for Computational Linguistics.

A Generation Configuration

All model interactions are conducted using the OpenAI Responses API through a unified interface. We use `gpt-5-chat-latest` for synthetic data generation tasks and reasoning-based evaluations, and `gpt-4.1-mini` for question answering during utility evaluation. The generation parameters are set as follows: `temperature = 1.0`, `top_p` = 1.0. All models receive the prompt as a user message and return a single text output, ensuring consistent inference across all experiments.

B Prompt Templates

This section lists the prompt templates used in the various phases of dataset generation. Variables to be replaced with values to complete the prompts are typeset in bold and wrapped in braces.

B.1 Topic Generation

`Generate 20 topics that would require knowledge about` **`{PII_type_1}`** `and` **`{PII_type_2}`**.

`Topic should consist of 1-3 words. It should be something people might write about on forums.`

B.2 Subtopic Generation

`Generate 10 subtopics related to the topic` **`{topic}`**.

`Each subtopic should consist of 1-3 words.`

`Subtopic should be of a nature that when writing about it you could mention` **`{PII_type_1}`** `and` **`{PII_type_2}`**.

B.3 Situation Generation

`Topic:` **`{topic}`**

`Subtopic:` **`{subtopic}`**

`PII:` **`{pii_category}`** `-` **`{pii_category_value}`**
`{supporting_pii_category}` `-` **`{support_pii_category_val}`**

`Generate exactly one natural-sounding sentence that:`

1. Describes a realistic situation connected to the given topic and subtopic.
2. Includes PII in the text exactly in the format they are. Do not change them.
3. Describes a problem.
4. Uses "I" and makes it sound like a personal experience.
5. Is specific – includes at least one additional detail that makes the situation vivid (e.g., time, reason, feeling, or other context clues).
6. Keeps the sentence between 20–35 words.

B.4 Peripheral Context Generation

Generate some facts about the person.

They should be completely unrelated to the following text: **{topic}** – **{subtopic}**.

They should be about an unrelated subject.

The facts MUST include these private information (PIIs) exactly as written:

{low_relevance_piis}

These PIIs must appear in the text unchanged and in their original exact form.

Write one natural-sounding first-person sentence using "I", consisting of 20–25 words, in plain text.

B.5 Question Generation

You are given a short description of a situation. Your task is to generate a general question.

Analyze the topic of the issue in the situation and generate a general question on that topic.

The question should not contain the words **{pii_category}** and **{supporting_pii_category}**, as well as their rephrased forms.

Situation: **{situation}**

Make the provided question sound more personal by rewriting it with "I".
Question: **{intermediate_result}**

You are given the question.

Remove all words that are related to **{relevant_pii_type_and_value_1}**.

Remove all words that are related to **{relevant_pii_type_and_value_2}**.

Question: **{question}**

Output only the modified question.

B.6 Paraphrased Context Generation

Rewrite this text so it sounds coherent.

The rewritten text should be in the first person.

Pay attention to how sentences start and how they are connected with each other.

Text: **{context}**

Do not change the spelling of these words: **{piis }**.

Output only the modified text.

B.7 Span Retrieval

Find a span in the text that is the most similar to **{pii}**.

Text: **{context}**

Output only the span in its original form.

B.8 PII Generation

Generate **{pii_category}** (private detail) that makes sense based on the context and existing private details about the person.

{pii_category_description}

Context: **{context}**.

Output only the generated **{pii_category}** (1–3 words).

The context is either the already generated part of the context or the phrase:

It should be the **{pii_category}** of the person that faces issues with **{topic}**–**{subtopic}**.

C Training parameters

The parameters for finetuning Llama-3.1-8B and Llama-3.2-3B using LoRA are presented in Table 6. The lightweight LoRA adaptation allows efficient training on limited computational resources while preserving the generative capabilities of the model.

Parameter	Value
LoRA rank (r)	32
LoRA alpha	32
Target modules	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj
Batch size / device	2
Grad. accumulation	4
Epochs	2
Warmup ratio	0.03
Optimizer	AdamW (8-bit) with weight decay of 0.01
Quantization	4-bit

Table 6: Training parameters for fine-tuning Llama-3.1-8B and Llama-3.2-3B.

D Fine-tuning and Pretrained Model Prompts

Below are the instruction templates used for the pretrained and fine-tuned Llama models. Both follow the same Alpaca-style prompt structure.

D.1 Shared Prompt Template

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction: {model-specific task description}

Input: Text: **{...}** Question: **{...}**

Response: (JSON output)

D.2 Pretrained Model Instruction

You are given the text and the question.

Find all PII (Personally Identifiable Information) in the text and output them separated by commas.

Classify each PII into one of the following types: health, location, sexual orientation, occupation, age, belief, relationship, name, education, appearance, code, organization, finance, datetime, demographic.

Classify their relevance to the question: 1 (high), 0 (low).

Example:

Text: "John Smith, a 22-year-old student from Canada, works for the University of Toronto."

Question: "What are the educational institutions mentioned in the text?"

Output: { "John Smith": {"type": "family", "relevance": "0"}, ... }

Do not include any explanations or extra text beyond this JSON structure.

D.3 Fine-tuned Model Instruction

You are given the text and the question.

Find all PII (Personally Identifiable Information) in the text and output them separated by commas.

Classify each PII into one of the following types: health, location, sexual orientation, occupation, age, belief, relationship, name, education, appearance, code, organization, finance, datetime, demographic.

Classify their relevance to the question: 1 (high), 0 (low).

When classifying the relevance, pay to attention to how each PII can be helpful for answering the question. When it is highly helpful, its a high (1) relevance.

Output the result in JSON format.

E Downstream Performance Prompts

E.1 Question Answering Instruction

Answer the question by taking into account the provided context.

Context: **{context}**.

Question: **{question}**.

Utilize the provided information as much as possible when formulating the answer.

The answer should be concise and approximately three sentences long.

Output only the answer.

E.2 Answer Evaluation Instruction

You are an expert evaluator.

You are given a context, a question, and two candidate answers.

Your task is to decide which answer is better supported by the context and more directly answers the question.

If two answers are equal then output equal.

Make sure to choose the most accurate answer based on the context available.

Context: **{context}**

Question: **{question}**

Answer A: **{answer_A}**

Answer B: **{answer_B}**

First, think step-by-step and discuss your reasoning. After, response with either "A" or "B" or "Equal" corresponding to your choice

F Annotation Examples

In the following examples we mark low-relevant PII in yellow and high-relevent PII in blue.

F.1 Example 1

Context: So here's my story: I'm 34 and spend my days as a preschool teacher, which I totally love! I live with a cognitive development disorder, but honestly, I make it work. My $36,500 annually keeps me living pretty comfortably, thank you very much. I snagged my Associate's Degree before diving into the world of tiny humans and finger paint, and oh yeah, I'm heterosexual.

Question: How can my issues affect my daily responsibilities?

Explanation: It is impossible to answer the question without knowing exactly what issues the person has and what the nature of their job is. However, their education, salary, and sexuality are irrelevant in terms of the question.

F.2 Example 2

Context: I want you to know that my journey has taken me from Canada to Brighton, England, where I've been thriving for the past three years. Being open about my bisexuality has truly transformed my life—it's allowed me to forge genuine, meaningful connections with others. At 22 years old, I'm navigating life with borderline personality disorder, and I'm proud to say I've created an incredible support system at Richardson Ltd, where the relationships I've built with my colleagues have become invaluable to me.

Question: I want to become a citizen, how easy that procedure will be for me in terms of legal docs?

Explanation: It is impossible to answer this question without knowing from where the person is and where they are residing. Age is also important information here. However, sexuality, health issues, and organization name "Richardson Ltd" are irrelevant PII.