Semantically-Aware LLM Agent to Enhance Privacy in Conversational AI Services

Jayden Serenari
Department of Computer Science
University of Pittsburgh
Pittsburgh, Pennsylvania, USA

Stephen Lee
Department of Computer Science
University of Pittsburgh
Pittsburgh, Pennsylvania, USA

***Abstract*—With the increasing use of conversational AI systems, there is growing concern over privacy leaks, especially when users share sensitive personal data in interactions with Large Language Models (LLMs). Conversations shared with these models may contain Personally Identifiable Information (PII), which, if exposed, could lead to security breaches or identity theft. To address this challenge, we present the Local Optimizations for Pseudonymization with Semantic Integrity Directed Entity Detection (LOPSIDED) framework, a semantically-aware privacy agent designed to safeguard sensitive PII data when using remote LLMs. Unlike prior work that often degrade response quality, our approach dynamically replaces sensitive PII entities in user prompts with semantically consistent pseudonyms, preserving the contextual integrity of conversations. Once the model generates its response, the pseudonyms are automatically depseudonymized, ensuring the user receives an accurate, privacy-preserving output. We evaluate our approach using real-world conversations sourced from ShareGPT, which we further augment and annotate to assess whether named entities are contextually relevant to the model's response. Our results show that LOPSIDED reduces semantic utility errors by a factor of 5 compared to baseline techniques, all while enhancing privacy.**

***Index Terms*—privacy, LLM, Personally Identifiable Information (PII)**

I. Introduction

Conversational AI systems are rapidly being adopted across domains ranging from customer service and healthcare to creative writing and programming assistance. These systems rely on Large Language Models (LLMs) hosted on remote servers, requiring user inputs that may include potentially sensitive data, to be transmitted for inference. When users unknowingly share Personally Identifiable Information (PII), such as names and addresses, they expose themselves to privacy risks, regulatory liabilities, and possible data breaches [1].

Pseudonymization is a common approach [2] used to protect users' privacy by replacing PII with entities of the same class. For example, a city name like *Chicago* might be substituted with *Los Angeles* to obscure the original data while maintaining the overall structure of the input. However, if a system relies on specific details for accuracy, altering key information may lead to misleading or incorrect results. For instance, if a user asks, "What is the population of Chicago?" and the system modifies it to "What is the population of Los Angeles?", the semantic integrity of the query is compromised. This highlights a key challenge in privacy-preserving techniques — ensuring that user data remains protected without distorting the intended *semantic meaning* of their input.

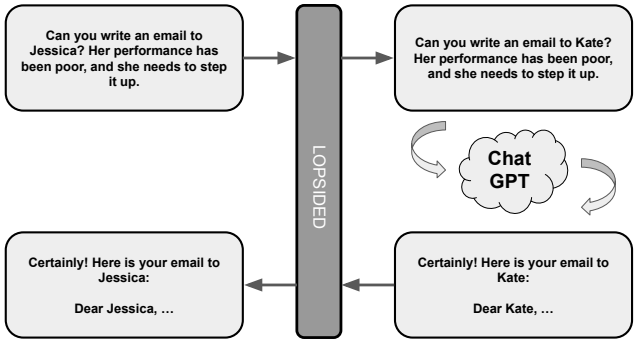

Fig. 1: The LOPSIDED privacy agent system design.

More recently, LLMs have been explored for PII removal in conversational AI services [3]. For example, Hide and Seek (HAS) [4] anonymizes any PII within a prompt before it is transmitted to a cloud-based language model and then de-anonymizes the LLM's response. This approach is subject to the same pitfall as stated previously. This gap motivates a key research question: *How can we preserve user privacy through pseudonymization while ensuring that LLM responses remain semantically faithful to the original query?*

Currently, all techniques rely on an *all-or-nothing approach*, where all private entities are removed or none are. This can render the system unusable for users, as they may not be able to effectively use the tool if essential information is removed. Our key insight is to develop a more nuanced approach that selectively pseudonymizes PII data and generates a semantically similar response. For example, if a user asks about the weather in Palo Alto, it could be replaced with San Jose, maintaining privacy while preserving the accuracy of the response.

To address this challenge, we propose LOPSIDED, a lightweight framework that balances PII removal and semantic response preservation. Our work focuses on maximizing user privacy by locally pseudonymizing sensitive information before it is transmitted to remote LLMs. The privacy agent, shown in Figure 1, operates as an intermediary between the user and the cloud. It intercepts the request, pseudonymizes

the prompt by replacing PII, and de-anonymizes it before presenting it to the user. This ensures that privacy-sensitive data is not exposed, while maintaining the relevance of the system's response. We note that there are situations where a replacement could completely alter the meaning. In such cases, we prioritize maintaining the utility of the response. Since pseudonymization must occur locally, we explore the use of small models that can be deployed on the user's device. Our key contributions are as follows:

LOPSIDED Design: We formulate the problem of semantic-aware privacy for conversational AI, where the goal is to pseudonymize named entities while preserving the utility of the LLM response. To address this problem, we introduce LOPSIDED, a framework which ensures that named entities can be modified without disrupting the semantic integrity of the response, making it both privacy-preserving and semantically accurate.

Semantic-aware Privacy Dataset: We augment the ShareGPT dataset, which contains real-world ChatGPT conversation histories, by annotating named entities for relevance to the prompt. This results a novel 866-sample evaluation dataset, designed for testing semantic-aware privacy agents.

Evaluation and Analysis: We evaluate our technique using real-world prompts and compare it against several baseline methods. Our analysis shows that prior work often prioritizes privacy at the expense of utility, while our approach reduces utility errors by a factor of 5 while still enhancing privacy.

II. Background

Personally Identifiable Information refers to any information that can be used to identify an individual, either directly or indirectly. PII is a key concept in privacy and data protection, as its exposure can lead to identity theft, fraud, and other security risks [5], [6]. The definition of what constitutes PII can vary, but generally, it includes both direct and indirect identifiers [7]. PII can be used in many malicious attacks, including the recent uptick in "Pig Butchering" scams [8]. Attackers can use PII to lure unsuspecting victims into a false state of trust by using personal information against them, such as the identities of close family, workplace details, or contact information.

Prior studies have highlighted that identifying PII is a significant challenge [7], [9]. To address these challenges, most techniques rely on Named Entity Recognition (NER), a method used to identify and classify entities such as names, locations, and organizations within text. Existing methods, such as SpaCy [10] and NLTK [11], leverage language models to perform NER. These models identify entities such as names, locations, organizations, and other relevant categories, classifying them into predefined labels like `[PERSON]` or `[GPE]` (Geopolitical Entity). Prior work has adopted spaCy as part of their pipeline to identify these named entities [4].

III. Related Work

Prior work has extensively studied the leakage of PII and privacy breaches through traffic originating from mobile devices and websites [12]. These studies highlight how data transmitted over networks can expose sensitive information. Such leaks enable third parties — including advertisers, data brokers, and malicious actors — to track users across different services. Such information can be correlated to build detailed user profiles without explicit consent. Recent research has also highlighted similar privacy risks arising from LLMs [13]. Today's growing reliance on LLMs introduces new challenges, as these models can inadvertently reveal sensitive information learned during training. Many companies openly acknowledge using personal data to train their models [14], which increases potential privacy risks. For instance, there is research that demonstrates that models may expose sensitive details such as phone numbers, email addresses, full names, and location information [15]. As a result, there has been recent work focused on mitigating these privacy concerns in language models [4], [16]. These mitigation techniques often involve the use of differential privacy guarantees during the training pipeline. Additionally, efforts have been made to reduce PII leakage in language models specifically [17]. However, much of this work primarily focuses on privacy, often neglecting the preservation of utility and the semantic integrity of the generated outputs. In contrast, our approach offers a balanced solution that effectively guards against unintentional PII leakage while still enabling users to interact with GPT-like models

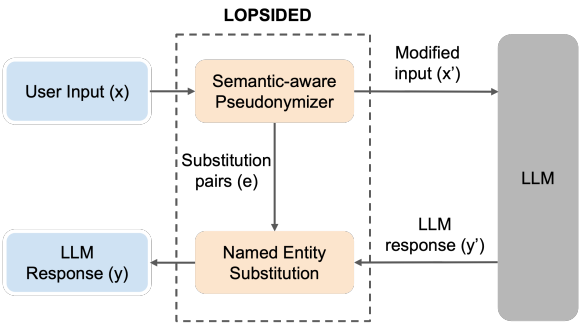

Fig. 2: Overall workflow of LOPSIDED framework.

IV. LOPSIDED Design

Figure 2 illustrates the overall workflow of the LOPSIDED framework, which consists of two main components: *semantic-aware pseudonymization* and the *named entity substitution* module. Unlike prior methods, the semantic-aware pseudonymization module is designed to generate semantically appropriate replacement entities, referred to as *pseudonyms*, for sensitive information while preserving the meaning of both the input prompt and the response derived from it. Specifically, when the user provides an input prompt, the semantic-aware pseudonymization module identifies and replaces private named entities, ensuring that this does not impact the overall response. The sensitive named entities are then stored locally for later use. The sanitized prompt is subsequently sent to the remote cloud provider, where a response is generated. Once returned, the named entity substitution module utilizes

```
{
  "changed_entities": [
    {
      "explanation": "The name 'Raven' can be
        any dog name...",
      "original_entity": "Raven",
      "new_entity": "Shadow"
    },
      ...
  ],
  "modified_prompt": "Write a short story
      about Shadow..."
}
```


Fig. 3: GPT-4o pseudonymization output.

both the locally stored private named entity information and the generated response to produce the final output. As a result, the user receives a response that protects privacy while maintaining the utility of the original prompt.

A. Semantic-aware Pseudonymization

This component substitutes named entities within a user prompt with pseudonyms. Formally, let x represent the original input prompt. We train a pseudonymization model $\mathcal{P}$ to generate an output consisting of a modified prompt x' and a set of entity pairs $e = \{(e_{\text{orig}}, e_{\text{pseudo}})\}$, where e_{orig} is a named entity identified in the original prompt x and e_{pseudo} is the corresponding pseudonym or replacement entity used in x'.

1) Data collection for Pseudonymization: To train the model, we first collect a dataset of sentences containing sensitive named entities. Since manually modifying each sentence is both time-consuming and costly, we leverage GPT-4o to automate this process. We prompt GPT-4o to generate semantically appropriate replacement entities for the sensitive information in the sentences, resulting in a modified version of the input prompt. Figure 3 illustrates a sample response from GPT-4o. This approach allows us to create a supervised dataset, which we then use to distill the knowledge from GPT-4o into our model.

2) Model Training.: The pseudonymization model $\mathcal{P}$ is trained on the curated dataset $\mathcal{D}_{\text{pseudo}}$ to learn to generate contextually appropriate pseudonyms while preserving the semantic meaning of the prompt. Each training example consists of the original input x, the modified prompt x' with pseudonymized entities, and the corresponding entity mapping e.

Formally, the model $\mathcal{P}$ is trained to maximize the likelihood of producing both the pseudonymized prompt and the entity mapping given the original input:

$$\max_{\mathcal{P}} \mathbb{E}_{(x,e,x')\sim\mathcal{D}_{\text{pseudo}}} \log p_{\mathcal{P}}(e, x'|x) \tag{1}$$

where $e = \{(e_{\text{orig}}, e_{\text{pseudo}}\})$ is a set of named entity substitution pairs, and x' is the modified prompt containing the pseudonymized entities.

B. Named Entity Substitution

The key goal of the named entity substitution model S, also referred to as the replacement model for simplicity, is to reconstruct the original response y as transparently as possible. Specifically, the model ensures that the final output is semantically faithful to the response that the remote LLM would have generated from the unmodified input x, while still protecting user privacy. To achieve this, note that after pseudonymization, the remote LLM produces a response y' based on the modified input x'. Since y' may contain pseudonyms rather than the original entities, the substitution model S uses the stored reverse mapping $e' = \{(e_{\text{pseudo}}, e_{\text{orig}})\}$ to restore the original entities, ensuring that the output remains semantically consistent with what the LLM would have produced for x.

For example, if the modified query *"What is the weather in San Jose?"* yields y' = *"The weather in San Jose is sunny."*, the substitution model applies e to produce y = *"The weather in Palo Alto is sunny."*. In this way, the user receives an accurate answer, while the sensitive entity never leaves the local environment.

Formally, let x' represent the modified input and y' denote the response generated by the remote LLM using x'. We train a substitution model S to reconstruct the response y, which would have been generated by remote LLM from the original input x. The model S takes the modified remote LLM response y', and uses a set of named entity mapping $e' = \{(e_{\text{pseudo}}, e_{\text{orig}})\}$ to restore the original entities within a response y.

1) Data collection for substitution model: We augment the dataset collected during the pseudonymization step by incorporating responses from GPT-4o. For each original input x and its corresponding modified version x', we query GPT-4o to generate both the original response y (for x) and the modified response y' (for x').

The process results in a dataset consisting of tuples of the form (original input, modified input, original response, modified response, named entity substitution pairs). We then use this dataset to train the substitution model S to reconstruct the original response y from the modified response y' and named entity substitution pairs e'.

2) Model training: We train the substitution model S on the dataset $\mathcal{D}_{\text{sub}}$ to maximize the likelihood of reconstructing y from y' given e'. The objective is:

$$\max_{\mathcal{S}} \mathbb{E}_{(y,e',y')\sim\mathcal{D}_{\text{sub}}} \log p_{\mathcal{S}}(y|y', e') \tag{2}$$

where $e' = \{(e_{\text{pseudo}}, e_{\text{orig}})\}$ is a set of pairs that provides the pseudonym and its corresponding named entity. This setup ensures that S learns to reintegrate entities in context, rather than performing naive string replacement.

V. Dataset

A. Data Collection

We use the ShareGPT dataset, a publicly available dataset, consisting of 70K ChatGPT conversation history of users [18].

Metric	Test Set	Training Set
# of Prompts	866	2595
# of Entities	1195	3696
Entities per Prompt	1.38	1.42
Avg # of Word Tokens	49.38	49.03
Avg Entity Length	6.80	7.24
Max Ents in a Prompt	8	31
# Prompts Req. Review	30	N/A
Rejections	20	N/A

TABLE I: Data statistics and validation summary.

Type	**Relevant**	**Irrelevant**
Person	228	363
Organization	160	54
Facility	5	1
City/Country	267	23
Landmark	26	4
Demographic	62	2
Total	748	447

TABLE II: Statistics of our human annotated dataset.

This dataset includes a wide variety of real user-generated conversations with AI systems, and is freely available.

We focus only on the first turn of conversation in the dataset, as expanding the context to include multiple turns would significantly increase the resources required for training the models. However, our approach is extendable to multi-turn conversations, and future work could explore how to efficiently handle longer context windows while maintaining the same level of privacy and semantic integrity.

As the majority of the samples in this corpus do not contain PII, we utilize Amazon Comprehend's PII Detection Service to filter for prompts that do. After running Comprehend on the dataset, the service flagged 3461 samples as containing PII. However, upon manually inspecting these samples, we found that the service is not always accurate and sometimes flags sentences that do not actually contain any PII. Nevertheless, we decided to retain these samples in the dataset. Later, when we use these prompts with GPT to identify named entities, the GPT responses for these flagged prompts contain no named entities, confirming that they do not actually contain PII. Table I summarizes the key statistics of our dataset.

We begin by tagging the named entities using spaCy, which also categorizes each entity (e.g., location, name). We then annotate whether the entities are relevant or irrelevant. We consider an entity relevant if substituting it would alter the meaning of the prompt or significantly impact the quality of the response. Irrelevant named entities can be safely replaced without affecting the response from an LLM.

We developed a custom interface designed to streamline this process, which enables annotators to easily label entities. A local LLM runs in the background, allowing users to test how our privacy agent can impact the model's responses. On the one side of the interface, annotators can view the model's original output without any privacy intervention. On the other side, they can observe the response generated after replacing the entity with a randomly generated one of the same type. This comparison is provided as guidance to assist annotators in making informed decisions. We recruited three graduate students from our lab, who volunteered to assist with the annotation process. Each was trained on how to use the interface and was instructed to label the entities in the dataset as relevant or irrelevant for PII. Given the substantial time and effort required for manual annotation, we limited the scope of the data annotation to the test dataset. This allowed us to evaluate how our approach performed in preserving privacy and semantic integrity.

B. Data Validation

We validated our data using a majority voting approach. Specifically, if two out of the three annotators agreed on an entity being relevant, then it was classified as relevant; if two out of the three annotators agreed that it was irrelevant, it was considered irrelevant. Annotator agreement was around 75%, with a free-marginal Kappa value of 0.64. Any number above 0 indicates better-than-chance agreement, while 1 indicates complete agreement. Our value shows that our annotators had substantial agreement [19].

C. Data Analysis

Table II highlights the key characteristics of the human-annotated dataset. We observe that, for each named entity category, the number of relevant and irrelevant samples varies. In general, relevant tags occur 1.6 times more frequently than irrelevant tags (Table II), which aligns with the intuition that users typically include information that is relevant to the task. However, 37% of the samples were deemed irrelevant, indicating that these named entities can be safely replaced without affecting LLM's response.

VI. Evaluation

A. Baseline Methods

Microsoft Presidio [20]. This data protection tool focuses on accurately detecting private information in text for anonymization or removal. It prioritizes privacy but does not consider the utility of the entities it removes. *Presidio w/ Replacement* modifies the anonymizer by assigning identifiers to entities it replaces. This is stored as a mapping to the original, and is restored by the Presidio Deanonymizer.

Hide-and-Seek (HaS) [4]. This privacy framework uses an LLM to anonymize and deanonymize prompts to LLM. The model focuses on privacy but does not consider semantic meaning. As a result, it may lead to responses that are semantically inconsistent to the user's intended query. *HaS (fine-tuned)* is trained on our dataset to improve its performance in chat scenarios.

B. Training

We use a Gemma 2 2b model, and fine-tune it on our dataset. The model is fine-tuned using LoRa and 4 bit quantization on an A6000 GPU.

C. Metrics and Definitions

1) ROUGE and BLEU: BLEU score measures how many words in the sample appear in the reference, which shows the precision. ROUGE score measures the degree of between the sample and reference, which shows recall. These methods are commonly used [21], [22] in machine translation and summarization to evaluate the response quality.

2) LLM-as-a-Judge: LLM-as-a-Judge [23] uses LLMs to rate a response on a scale, with guidelines on response quality, relativity, and fulfillment. We use GPT-4o-mini as our judge, due to its low cost. This allows us to achieve higher score accuracy than if using a smaller local model. As a reference, we have ran both Llama3-1b and GPT-4o-mini on our dataset without anonymization, and they were rated 8.17 and 9.32 respectively. We consider these the upper and lower bound for our experiments.

3) Syntheticity: We define a response to be synthetic if there are any perceived perturbation of said response on the user side of the pipeline. For example, any hallucinations caused by erroneously replaced entities would be evidence of a synthetic response.

4) Privacy and Utility Errors: We evaluate the model using two complementary metrics: privacy and utility errors. **Privacy errors** occur when entities are *not* replaced when they should have been. **Utility errors** are defined as the amount of relevant entities that were incorrectly replaced by the model.

VII. Results

A. Baseline Performance

We begin by comparing our approach to baseline techniques. In our experiment, we modify the prompt and compare the output generated by the privacy agent to the output produced by GPT-4 alone, without any privacy interventions. To evaluate the performance of each approach, we use standard metrics such as ROUGE and BLEU scores, which assess the quality and similarity of the generated responses [22].

Table III compares the performance of various privacy agents. As shown, LOPSIDED outperforms other techniques in terms of ROUGE/BLEU scores. In general, models that were not fine-tuned exhibit lower performance. Notably, LOPSIDED achieves higher scores, indicating that its responses are closer to the ground truth (i.e., the original, unmodified response) compared to other baseline techniques. Additionally, models that have been fine-tuned tend to have lower scores, further highlighting the effectiveness of LOPSIDED in preserving semantic integrity of the LLM response. Also included in this table are the results from our LLM-as-a-Judge evaluation. This evaluation gives a a reasonably accurate estimate to what a human evaluation of our framework's output would be. LOPSIDED not only has the highest score of all the frameworks, but it also scores only .13 points lower than our upper bound.

B. Privacy and Utility Evaluation

Next, we evaluate the performance of our substitution model in balancing privacy and utility. For this evaluation, we use the human-annotated test set, which contains labels indicating whether each entity in a prompt is relevant or irrelevant. By comparing the output of the substitution model with these labels, we can measure how well it maintains the utility of the response by ensuring that relevant entities remain intact, while effectively substituting irrelevant or private entities.

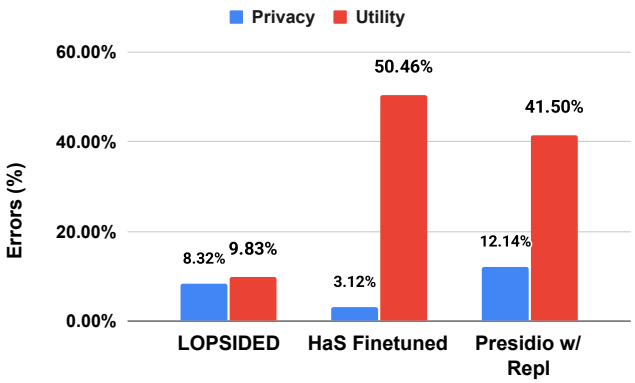

Fig. 4: Privacy and utility error rate comparisons.

Figure 4 shows that HaS achieves a low privacy error of 3% because it primarily focuses on substituting all named entities, regardless of their relevance. However, this comes at the cost of higher utility errors. LOPSIDED has only a slightly higher privacy error of 8%, but it achieves 5× fewer utility-related errors, demonstrating its ability to preserve relevant entities while still protecting private information. Compared to Presidio, LOPSIDED achieves lower privacy *and* utility errors.

C. Text Syntheticity Detection

Similar to Yermilov et al, we conduct a text syntheticity detection experiment. This analysis is necessary because pseudonymization can disrupt the relationships between named entities and their surrounding context, potentially leading to inconsistencies in downstream tasks. We combine responses from both pseudonymized and original texts and train a classification model using `bert-base-uncased` to distinguish between the two. A high classification accuracy indicates that pseudonymization introduces detectable artifacts, whereas a low accuracy suggests that modified texts closely resemble their original counterparts. Table IV presents the text syntheticity classification scores for different techniques. As shown, LOPSIDED achieves the lowest classification score, indicating that its pseudonymized texts are the most similar to the original ones.

VIII. Conclusion

LOPSIDED introduces a novel framework for pseudonymizing LLM API prompts while preserving the utility of the user's request. The methods we use introduce the key contribution of selectively changing private entities, alleviating the destructive effects of all-or-nothing solutions, while still enhancing the user's privacy protections. When comparing other state of the art privacy preserving methods, LOPSIDED protects the utility

	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4	LLM-as-a-Judge
LOPSIDED	**0.796**	**0.625**	**0.654**	**0.720**	**0.641**	**0.595**	**0.564**	**9.19**
HaS (finetuned)	0.461	0.226	0.284	0.149	0.108	0.096	0.090	8.17
HaS	0.149	0.102	0.125	0.139	0.129	0.124	0.121	N/A
Presidio	0.642	0.443	0.487	0.532	0.444	0.397	0.366	N/A
Presidio w/ Repl	0.655	0.454	0.497	0.541	0.453	0.405	0.374	7.29

TABLE III: Baseline performance comparisons. Bolded values are the highest scores.

Framework	Detectability (Avg)	Detectability (Final)
LOPSIDED	**46.59%**	**44.88%**
HaS Finetuned	71.84%	83.84%
HaS	85.65%	87.73%
Presidio	60.43%	62.19%
Presidio w/ Repl	51.36%	48.63%

TABLE IV: Syntheticity detection scores.

of named entities at a significantly greater rate. At the same time, we replace the irrelevant private entities at a similar, or better, rate. These privacy protections help users remain safer in a rapidly scaling world of AI systems. To support our claims, we present an evaluation dataset validated by human annotators to assess the effectiveness and utility of privacy agents. This dataset serves as a benchmark for evaluating similar privacy techniques and demonstrates the strengths of our approach. Our framework is extendable, with the ability for future researchers to extend the platform to ensure robustness in multi-turn scenarios, and more intricate replacement strategies. Supporting datasets and models can be found on our GitHub repository: *https://github.com/jmseren/LOPSIDED*.

Acknowledgments. This research was supported in part by the University of Pittsburgh Center for Research Computing and Data, RRID:SCR_022735, through the resources provided. Specifically, this work used the H2P cluster, which is supported by NSF award number OAC-2117681.

REFERENCES

[1] T. Aura, T. A. Kuhn, and M. Roe, "Scanning electronic documents for personally identifiable information," in *Proceedings of the 5th ACM Workshop on Privacy in Electronic Society*, ser. WPES '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 41–50. [Online]. Available: https://doi.org/10.1145/1179601.1179608

[2] A. Stubbs, C. Kotfila, and Özlem Uzuner, "Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1," *Journal of Biomedical Informatics*, vol. 58, pp. S11–S19, 2015, supplement: Proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046415001173

[3] Y. Dou, I. Krsek, T. Naous, A. Kabra, S. Das, A. Ritter, and W. Xu, "Reducing privacy risks in online self-disclosures with language models," *arXiv preprint arXiv:2311.09538*, 2023.

[4] Y. Chen, T. Li, H. Liu, and Y. Yu, "Hide and seek (has): A lightweight framework for prompt privacy protection," 2023. [Online]. Available: https://arxiv.org/abs/2309.03057

[5] A. H. Seh, M. Zarour, M. Alenezi, A. K. Sarkar, A. Agrawal, R. Kumar, and R. Ahmad Khan, "Healthcare data breaches: Insights and implications," *Healthcare*, vol. 8, no. 2, 2020. [Online]. Available: https://www.mdpi.com/2227-9032/8/2/133

[6] B. Krishnamurthy and C. E. Wills, "On the leakage of personally identifiable information via online social networks," in *Proceedings of the 2nd ACM Workshop on Online Social Networks*, ser. WOSN '09. New York, NY, USA: Association for Computing Machinery, 2009, p. 7–12. [Online]. Available: https://doi.org/10.1145/1592665.1592668

[7] I. Pilán, P. Lison, L. Øvrelid, A. Papadopoulou, D. Sánchez, and M. Batet, "The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization," *Computational Linguistics*, vol. 48, no. 4, pp. 1053–1101, 2022.

[8] B. Han and M. B. and, "An anatomy of 'pig butchering scams': Chinese victims' and police officers' perspectives," *Deviant Behavior*, vol. 0, no. 0, pp. 1–19, 2025. [Online]. Available: https://doi.org/10.1080/01639625.2025.2453821

[9] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," in *Named Entities: Recognition, classification and use*. John Benjamins publishing company, 2009, pp. 3–28.

[10] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017, to appear.

[11] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[12] O. Starov, P. Gill, and N. Nikiforakis, "Are you sure you want to contact us? quantifying the leakage of pii via website contact forms," *Proceedings on Privacy Enhancing Technologies*, 2016.

[13] L. Rocher, J. M. Hendrickx, and Y.-A. De Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models," *Nature communications*, vol. 10, no. 1, pp. 1–9, 2019.

[14] S. Morrison, "The tricky truth about how generative AI uses your data," 2023, (Accessed Feb 2024). [Online]. Available: https://www.vox.com/technology/2023/7/27/23808499/ai-openai-google-meta-data-privacy-nope

[15] J. Lee, T. Le, J. Chen, and D. Lee, "Do language models plagiarize?" in *Proceedings of the ACM Web Conference 2023*, 2023, pp. 3637–3647.

[16] X. Li, F. Tramer, P. Liang, and T. Hashimoto, "Large language models can be strong differentially private learners," *arXiv preprint arXiv:2110.05679*, 2021.

[17] X. Zhao, L. Li, and Y.-X. Wang, "Provably confidential language modelling," *arXiv preprint arXiv:2205.01863*, 2022.

[18] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," March 2023. [Online]. Available: https://lmsys.org/blog/2023-03-30-vicuna/

[19] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977. [Online]. Available: http://www.jstor.org/stable/2529310

[20] O. Mendels, C. Peled, N. Vaisman Levy, S. Hart, T. Rosenthal, L. Lahiani *et al.*, "Microsoft Presidio: Context aware, pluggable and customizable pii anonymization service for text and images," Microsoft, 2018. [Online]. Available: https://microsoft.github.io/presidio

[21] Y. Graham, "Re-evaluating automatic summarization with bleu and 192 shades of rouge," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 128–137.

[22] K. Blagec, G. Dorffner, M. Moradi, S. Ott, and M. Samwald, "A global analysis of metrics used for measuring performance in natural language processing," 2022. [Online]. Available: https://arxiv.org/abs/2204.11574

[23] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging llm-as-a-judge with mt-bench and chatbot arena," 2023. [Online]. Available: https://arxiv.org/abs/2306.05685