Balancing Privacy and Utility in Personal LLM Writing Tasks: An Automated Pipeline for Evaluating Anonymizations

Stefan Pasch[1], Min Chul Cha[2]
[1]Division of Social Science & AI, Hankuk University of Foreign Studies, Seoul, Republic of Korea
[2]Division of Media Communication, Hankuk University of Foreign Studies, Seoul, Republic of Korea
(Corresponding Author)
stefan.pasch@outlook.com, minchulcha@hufs.ac.kr

Abstract

Large language models (LLMs) are widely used for personalized tasks involving sensitive information, raising privacy concerns. While anonymization techniques exist, their impact on response quality remains underexplored. This paper introduces a fully automated evaluation framework to assess anonymization strategies in LLM-generated responses. We generate synthetic prompts for three personal tasks—personal introductions, cover letters, and email writing—and apply anonymization techniques that preserve fluency while enabling entity backmapping. We test three anonymization strategies: simple masking, adding context to masked entities, and pseudonymization. Results show minimal response quality loss (roughly 1 point on a 10-point scale) while achieving 97%-99% entity masking. Responses generated with Llama 3.3:70b perform best with simple entity masking, while GPT-4o benefits from contextual cues. This study provides a framework and empirical insights into balancing privacy protection and response quality in LLM applications.

1 Introduction

The intersection of AI governance and data protection has garnered significant attention from academia (Yermilov et al., 2023; Staab et al. 2023), industry, (AWS, 2023; Azure, 2024) and regulatory bodies (European Data Protection Supervisor, 2025). As large language models (LLMs) become widely adopted, concerns regarding privacy risks in user interactions have increased. Particularly, the substantial costs of hosting LLMs, along with restricted access to certain proprietary models, pose significant challenges for individuals and small enterprises seeking to deploy LLMs locally. As a result, many rely on external LLM services, increasing privacy risks (Mao et al., 2024). Moreover, LLMs are frequently used in tasks that involve sensitive personal or corporate information, such as their names, company information, or location information. This raises critical questions about how anonymization strategies impact both privacy protection and response quality in these real-world use cases.

Existing research has primarily focused on privacy protection from adversarial attacks, such as attribute inference and re-identification risks (Staab et al., 2023; Chen et al., 2023). Approaches like differential privacy (Igamberdiev and Habernal, 2023) and prompt obfuscation (Sun et al, 2024) have been explored to mitigate these risks. However, these methods often concentrate on preventing external inference attacks rather than evaluating the direct trade-offs between anonymization and response quality in personal tasks.

While some studies have examined the utility of anonymized text, they often primarily focus on traditional NLP benchmarks like text classification or summarization (Yermilov et al., 2023; Riabi et al., 2024). However, the impact of anonymization on personalized, user-driven tasks, where coherence and contextual relevance are crucial, remains underexplored. Moreover, existing anonymization methods can degrade response quality, limiting real-world usability. Many privacy-enhancing techniques also rewrite entire user inputs, making it harder to retain original context and provide users with responses that align with their initial prompts.

In practice, however, many users engage large language models (LLMs) for tasks that involve

Proceedings of the Sixth Workshop on Privacy in Natural Language Processing, pages 32–41

sensitive personal or corporate information, such as drafting personal introductions, job applications, or emails. This raises concerns about how anonymization techniques affect the quality of LLM-generated responses in these personalized contexts, as there may be a trade-off between AI governance practices and response quality (Pasch, 2025).

In this paper, we analyze the effect of different anonymization techniques on personalized tasks and their impact on response quality. We introduce an automated end-to-end workflow to evaluate LLM-generated responses, encompassing the following steps:

1. **Creation of Synthetic Personal Prompts**: We generate prompts using LLMs for three writing tasks involving personal information: personal introductions, cover letters, and emails.
2. **Entity Identification**: Utilizing a BERT-based Named Entity Recognition (NER) model, we identify entities within these generated prompts.
3. **Anonymization Strategies**: We employ various anonymization techniques, enriching the initial entities using a local guardrail model to either provide context or substitute them with comparable pseudonyms.
4. **LLM Response Generation**: The anonymized prompts are input into LLMs to generate responses, simulating behavior in an unprotected environment.
5. **De-Anonymization**: We replace the masked entities in the responses with their original values.
6. **Evaluation**: We assess response quality using the LLM-as-a-Judge method and evaluate privacy by examining entity matches and LLM inference capabilities.

Our findings indicate that anonymization only slightly impacts response quality, with most settings showing a decrease of less than one point on a ten-point scale after de-anonymization. Notably, 97% to 99% of entities are effectively anonymized, demonstrating significant privacy enhancements. For responses generated by the Llama 3.3:70b model, a straightforward anonymization and de-anonymization approach outperforms more complex methods involving contextualization or pseudonymization. Conversely, for GPT-4o-generated responses, adding context further improves response quality.

This study contributes to the literature on LLM privacy in two major ways:

- Providing an end-to-end framework to evaluate anonymization strategies for personal writing tasks with LLMs.
- Assessing the effectiveness of various anonymization techniques in both privacy and response quality in personal writing tasks.

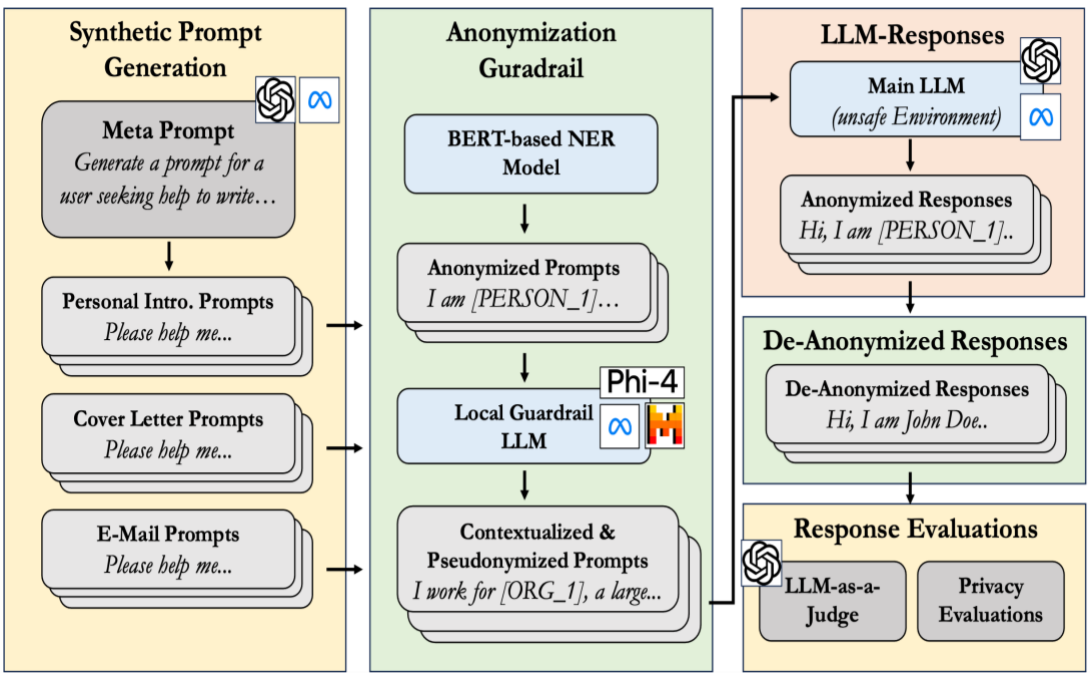

Figure 1. Overview of end-to-end anonymization and de-anonymization workflow

2 Methodology

Our approach presents a fully automated end-to-end workflow for evaluating anonymization strategies in LLM-based interactions, as depicted in Figure 1. Moreover, Figure 2 illustrates the different anonymization strategies. The pipeline spans synthetic prompt generation, guardrail-based anonymization, response generation, de-anonymization, and evaluation, ensuring a systematic assessment of privacy protection and response quality. To achieve this, we leverage two main categories of models:

Main LLM Models (Response Generation): These models are responsible for generating responses to user prompts. They reflect how proprietary AI systems process user inputs in real-world applications. We experiment with two state-of-the-art LLMs to evaluate the effects of anonymization on response quality: (i) ChatGPT 4o, and (ii) Llama 3.3:70b. While Llama can be locally deployed, we use it primarily to mimic proprietary AI systems, given its state-of-the-art performance, ensuring a controlled yet representative evaluation of anonymization effects.

Guardrail models: These models anonymize the text input. This first includes a NER model for entity masking and different LLMs to provide context or pseudonymize the entities. We specifically select open-source models for the guardrail tasks to enable deployment in controlled environments. The models used for these tasks are: Llamac3.3:70b, Llama 3.1:8b-instruct, Phi4:14b, and Mistral:7b.

2.1 Synthetic Prompt Generation

The first step in our workflow involved creating a dataset of prompts designed to assess response quality for personal tasks. Existing datasets in LLM anonymization research primarily focus on inferring personal information from text data or prompts (Yukhymenko et al., 2024). However, to the best of our knowledge, no dataset exists where user prompts explicitly request assistance for personal tasks that necessitate the inclusion of personal details such as names, locations, and affiliated organizations. We focus on three distinct personal tasks—personal introductions, cover letters, and business emails—as they represent common real-world scenarios in which users seek AI-generated text assistance while involving sensitive personal information.

Personal Introduction: Personal introductions are frequently used in professional and social settings, including networking events, biographies, and job-seeking platforms (Xu et al., 2023). These introductions typically contain personally identifiable information (PII) such as names, current and past employers, and locations.

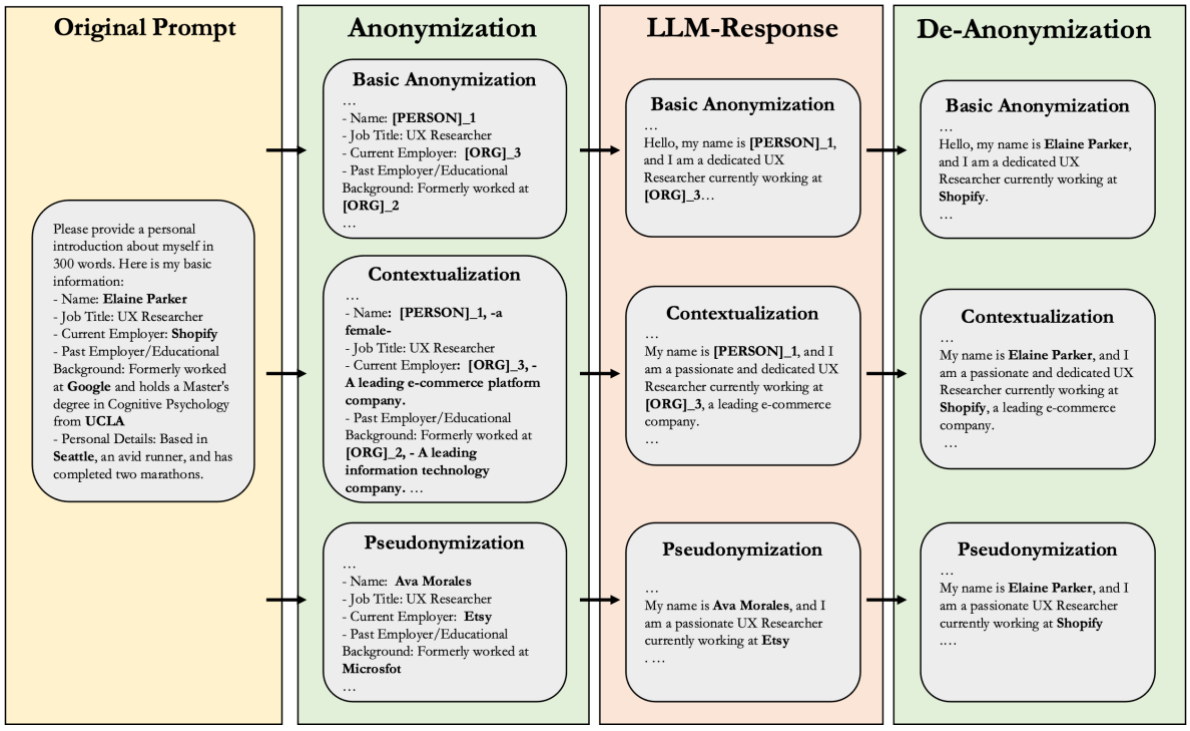

Figure 2. Overview of Anonymizations and Pseudonymizations

Cover Letter: Cover letters are a critical component of job applications and have been increasingly generated or refined using AI-powered writing assistants (Zinjad et al., 2024). Since cover letters include personal details such as work history, employer names, and sometimes personal aspirations, they provide a rich context for studying anonymization strategies in structured yet personalized texts.

Business Email: Email communication is a widely studied domain in NLP, particularly in business and professional settings (Jovic and Mnasri, 2024). Emails often contain sensitive information about organizations, job roles, and ongoing projects, making them a relevant task for evaluating anonymization methods while preserving coherence and intent.

By selecting these tasks, we aim to explore how anonymization affects the quality of LLM-generated outputs in contexts where personal information is integral to the content.

To create this dataset, we employed a meta-prompting approach, where an LLM was prompted to generate a single synthetic prompt for a given task. This process was repeated 50 times per task, resulting in a total of 150 prompts per LLM model. Importantly, all experimental steps were conducted twice, using two different LLMs—Llama3.3:70B and ChatGPT-4o—to generate independent prompt datasets.

Each meta-prompt included:

- Explicit task instructions (e.g., generating a personal introduction, cover letter, or email).
- A requirement to include realistic names, locations, and organizations that actually exist.
- A directive to ensure prompts were formulated from the perspective of a user seeking quick assistance, rather than overly refined or context-heavy instructions. This was done because initial trials revealed that the generated prompts were often too polished and provided a lot of context, resembling pre-written templates rather than spontaneous user queries.

2.2 Anonymize Prompts

In this study, we employ a BERT-based transformer model for Named Entity Recognition (NER) to anonymize prompts. Specifically, we utilize the *XLM-RoBERTa-large-finetuned-conll03-english* model (Conneau et al., 2020). Our choice of a BERT-based NER model is motivated by two primary factors: First, BERT-based models have achieved state-of-the-art results in various NER benchmarks (Conneau et al., 2020). Second, BERT-based models are increasingly being integrated into guardrail solutions to ensure safety and compliance in AI applications (Zheng et al. 2024).

Once the entities are identified, we anonymize the prompt text by systematically replacing each detected entity with a structured placeholder that preserves its semantic role. Specifically, named entities are substituted with generic category-based markers to maintain coherence and allow for later de-anonymization. Each entity type is assigned a unique identifier that follows a consistent pattern across all prompts. For instance, a detected organization (e.g., *Google*) is replaced with ORG_1, a location (e.g., *New York*) is replaced with LOCATION_1, and a person's name (e.g., *John Doe*) is substituted with PERSON_1. If multiple entities of the same category appear in a prompt, they are enumerated sequentially.

This structured anonymization approach ensures that the prompts retain their original syntactic and semantic integrity while eliminating personally identifiable information (PII). The placeholders allow for the preservation of relationships between entities.

2.3 Contextualization of Entities

Anonymization of entities often results in loss of contextual information, which can affect the quality and coherence of generated responses. For example, both *Google* and *Stanford University* would be anonymized as ORG_X, obscuring the distinction between a large software company and a university. To mitigate this issue, we implement a contextualization step where guardrail LLMs provide enriched descriptions of the masked entities. This approach ensures that the semantic role of entities remains intact, allowing the main LLM models to generate more coherent and informative responses despite anonymization.

Each anonymized entity is passed to the guardrail LLM, which is prompted to generate a concise description of the entity without revealing its name. For instance:

- *Google* → "a large software company"

- *Stanford University* → "a private research university"

For personal names, the contextualization is limited to gender classification, where the guardrail model predicts whether the name is typically male or female. This step helps in preserving pronoun consistency in text generation while avoiding re-identification of individuals.

2.4 Pseudonymization of Entities

In an alternative anonymization setup, instead of contextualizing the masked entities, we apply pseudonymization, where each entity is replaced with a comparable but non-identical alternative. This approach retains the structural integrity of the text while obfuscating specific details.

To achieve this, we prompt our guardrail LLM models to generate substitutes for entities identified by the NER model. The replacements are chosen to be semantically similar but distinct from the original entity. For example:

- *John Doe → Frank Miller*
- *Google → Microsoft*
- *New York → Chicago*

The goal of this approach is to preserve the context of the text while preventing direct entity recognition. Unlike contextualization, where descriptions replace entity names, pseudonymization maintains the original sentence structure, allowing the text to remain fluent and natural without explicit entity masking.

2.5 LLM Response Generation

After setting up the different prompts with various anonymization techniques, we input these prompts into the main LLM models to generate responses. In the system prompt, we inform the model that the input contains entity markers (with contextual information where applicable) or pseudonyms. Additionally, we instruct the model not to modify the format of these entity markers to ensure that they can be accurately mapped back in later stages. Overall, responses are generated for four different anonymization setups: (i) The original prompts (no anonymization), (ii) the anonymized prompts with simple masking, (iii) the anonymized prompts with contextualized information, and (iv) the pseudonymized prompts.

2.6 De-Anonymization

For prompts that underwent entity masking, each anonymized entity (e.g., ORG_1, LOCATION_1, PERSON_1) is replaced in the LLM responses with its original name based on the entity mapping from the anonymization step. Similarly, in the pseudonymized setup, each substituted entity (e.g., *Microsoft* in place of *Google*) is reverted to its original counterpart.

This step ensures that we can evaluate the quality of the generated text in its original form while analyzing whether anonymization strategies introduced any distortions or inconsistencies in the output.

2.7 Evaluating the Response Quality

To assess the quality of the generated responses, we use an automated evaluation approach based on the LLM-as-a-Judge method (Zheng et al., 2023), a widely used technique for evaluating LLM-generated text.

For the primary evaluation, we adopt the single answer grading approach, where the LLM is presented with a single prompt-response pair and asked to rate the response on a scale from 1 to 10. To ensure consistency, we use the official single answer grading prompt from Zheng et al. (2023).

While LLM-as-a-Judge typically provides an overall quality score, anonymization techniques may affect different aspects of response quality in varying ways. Therefore, in addition to a single score, we follow Zhong et al. (2022) and evaluate responses across four key dimensions:

- Coherence – Logical structure and connectedness of ideas.
- Consistency – Internal consistency and factual alignment with the prompt.
- Fluency – Grammatical correctness and naturalness of the language.
- Relevance – Appropriateness and relevance of the response to the given prompt.

We compute an average of these four scores to provide a secondary measure of overall quality. Based on recent findings, we use GPT-4o as the evaluation model, as it has been shown to exhibit high alignment with human preferences in LLM-as-a-Judge comparisons (Raju et al., 2024).

2.8 Privacy Evaluation

In addition to assessing response quality, we evaluate whether the anonymized text effectively preserves privacy. To measure this, we use two complementary approaches:

1. **Entity Matching** – We conduct a simple entity match by comparing the originally identified entities with those present in the anonymized prompts and responses. This allows us to check if any masked entities leak into the anonymized versions.
2. **LLM-Based Inference Attacks** – Inspired by Staab et al. (2023), we test whether an LLM (ChatGPT-4o) can infer masked or pseudonymized entities. The model is prompted to guess the original entities based on the anonymized text, simulating a potential privacy risk where an AI system could re-identify anonymized information.

Since the entities in our dataset were originally generated by LLMs, they tend to be commonly known entities (e.g., *Harvard University* or *Google*). This likely overestimates the model's ability to predict masked entities, as real-world anonymization would often involve more unique or less widely known names. Nevertheless, this measure provides a useful benchmark for comparing the relative differences between anonymization setups, particularly in assessing whether adding contextual descriptions or pseudonyms increases the likelihood of entity re-identification.

For both privacy measures, we define privacy as the inverse of the number of identified entities, calculated as:

$$Privacy = 1 - \frac{Identified\ Entities}{Total\ Entities}$$

When comparing the effectiveness of different anonymization strategies, we measure privacy before de-anonymization since de-anonymization occurs outside the "unsafe environment" in our setup.

3 Results

3.1 Evaluation of Anonymization Strategies

Table 1 presents the results for utility and privacy across different anonymization strategies. As expected, responses to original (non-anonymized) prompts achieve the highest scores across utility metrics, with an LLM-as-a-Judge score of 9.95 (ChatGPT-4o) and 9.76 (Llama 3.3:70B). Privacy scores are naturally low, as all original entities remain intact.

Across all specifications, we observe that anonymized and pseudonymized responses (without de-masking) exhibit lower quality scores. For example, basic anonymization results in a drop in utility, with ChatGPT-4o scoring 3.09 and Llama 3.3:70B scoring 3.19 in overall LLM-as-a-Judge evaluations. This is unsurprising, as these transformations alter the structure of the original prompt, potentially reducing the coherence and contextual accuracy with the initial prompt.

However, once the initial entities are reinserted into the anonymized or pseudonymized responses (i.e., after de-anonymization), response quality significantly improves. The LLM-as-a-Judge score of de-anonymized responses reaches 9.37 (ChatGPT-4o) and 8.41 (Llama 3.3:70B), indicating that while anonymization impacts output quality, de-anonymization can effectively restore much of the lost information.

When comparing different anonymization techniques, we find that simple anonymization followed by de-anonymization performs surprisingly well. Notably, for Llama 3.3:70B-generated responses, this basic anonymization-de-anonymization approach outperforms all other anonymization strategies.

For GPT-4o-generated responses, however, the results vary depending on the guardrail model used. We find that for all guardrail models except Mistral 7B, contextualized anonymization slightly outperforms the simple masking technique. For instance, the contextualized de-anonymized responses using Phi-4 14B achieve an LLM-as-a-Judge score of 9.70 (ChatGPT-4o), slightly higher than 9.37 for basic de-anonymization.

Regarding privacy scores, we observe that Llama-generated contextualization perform comparable to simple anonymization-de-anonymization when assessed using entity matching. Specifically, Llama 3.3:70B contextualized anonymization retains a privacy score of 0.99 (entity match), similar to basic anonymization. However, for Phi-4 and Mistral-generated contexts, a slightly higher number of tagged entities appear in responses, suggesting an increased risk of entity leakage when adding contextual information. For instance, the privacy score (entity match) of Phi4:14b drops to 0.95. Similarly, we find high risk of revealing entities for Phi and Mistral generated pseudonymization.

Using the LLM inference method to assess privacy risks, we find an increased privacy risk for all contextualization methods. For example, ChatGPT-4o contextualization (Phi-4 14B) has an

Table 1: Utility and Privacy Scores by Anonymization Strategy

LLM Response Model	ChatGPT				Llama			
Dimension	**Utility**		**Privacy**		**Utility**		**Privacy**	
Metric	**Avg 4D**	**Score**	**Entity Match**	**LLM Inf.**	**Avg 4D**	**Score**	**Entity Match**	**LLM Inf.**
Baseline: Original Response	9.97	9.95	0.00	0.10	9.89	9.76	0.00	0.11
Basic Anonymization								
Anonymized Response	6.72	3.09	0.97	0.83	6.35	3.19	0.99	0.86
De-Anonymized Response	9.75	9.37	0.97	0.83	9.41	8.41	0.99	0.86
Contextualization (Anonymized)								
Contextualization: Phi4 14b	7.18	3.18	0.88	0.46	6.50	3.49	0.95	0.42
Contextualization: Llama3.3 70b	7.10	3.20	0.97	0.62	6.48	3.47	0.99	0.61
Contextualization: Llama3.1 8b	7.07	3.17	0.97	0.59	6.49	3.37	0.99	0.58
Contextualization: Mistral 7b	7.18	3.19	0.94	0.54	6.13	3.24	0.97	0.53
Contextualization (De-Anonymized)								
Contextualization: Phi4 14b	9.86	9.70	0.88	0.46	9.17	8.03	0.95	0.42
Contextualization: Llama3.3 70b	9.83	9.53	0.97	0.62	9.11	7.51	0.99	0.61
Contextualization: Llama3.1 8b	9.82	9.60	0.97	0.59	9.16	7.72	0.99	0.58
Contextualization: Mistral 7b	9.72	9.46	0.94	0.54	8.79	7.14	0.97	0.53
Pseudonymization (Pseudonyms)								
Pseudonymization: Phi4 14b	3.86	1.58	0.78	0.85	3.64	1.32	0.82	0.87
Pseudonymization: Llama3.3 70b	3.81	1.48	0.97	0.98	3.64	1.23	0.98	0.99
Pseudonymization: Llama3.1 8b	3.85	1.60	0.95	0.97	3.61	1.23	0.98	0.98
Pseudonymization: Mistral 7b	3.93	1.79	0.77	0.87	3.60	1.23	0.78	0.82
Pseudonymization (De-Anonymized))								
Pseudonymization: Phi4 14b	7.77	6.04	0.78	0.85	7.27	5.06	0.82	0.87
Pseudonymization: Llama3.3 70b	9.29	8.57	0.97	0.98	9.01	7.43	0.98	0.99
Pseudonymization: Llama3.1 8b	9.37	8.69	0.95	0.97	8.75	7.03	0.98	0.98
Pseudonymization: Mistral 7b	6.65	4.61	0.77	0.87	5.38	3.21	0.78	0.82

Utility reflects response ratings using the LLM-as-a-Judge method. Avg. 4D corresponds to the average score of 4 dimensions of response quality: Coherence, consistency, fluency, and relevance. Score reflects a single overall score for the output.

LLM inference score of 0.46, while basic anonymization is at 0.83, suggesting that adding descriptions makes it easier for an LLM to reconstruct the original entities. In contrast, pseudonymization decreases this risk, with Llama3.3:70b pseudonymization reaching privacy scores of 0.98 (ChatGPT-4o) and 0.99 (Llama 3.3:70B), indicating that substituting entities with comparable alternatives can be an effective method to obscure true entities.

3.2 Evaluation by Task Type

We also analyzed differences in response quality and privacy scores across task types. Figure 3 presents response quality (measured as the average score across four dimensions) and privacy (measured using LLM inference) for selected anonymization strategies.

Overall, we found that results remained consistent across task types. However, for both GPT-4o and Llama-generated responses, the drop in response quality of anonymized prompts was most pronounced in cover letter. This is unsurprising, as cover letters require personalized and highly structured writing, making anonymization more disruptive to specific entity information. Consistent with this, we observe that contextualization had a strong positive effect on cover letters, particularly for GPT-4o, where it outperformed simple masking. For Llama-generated responses, contextualization also had a

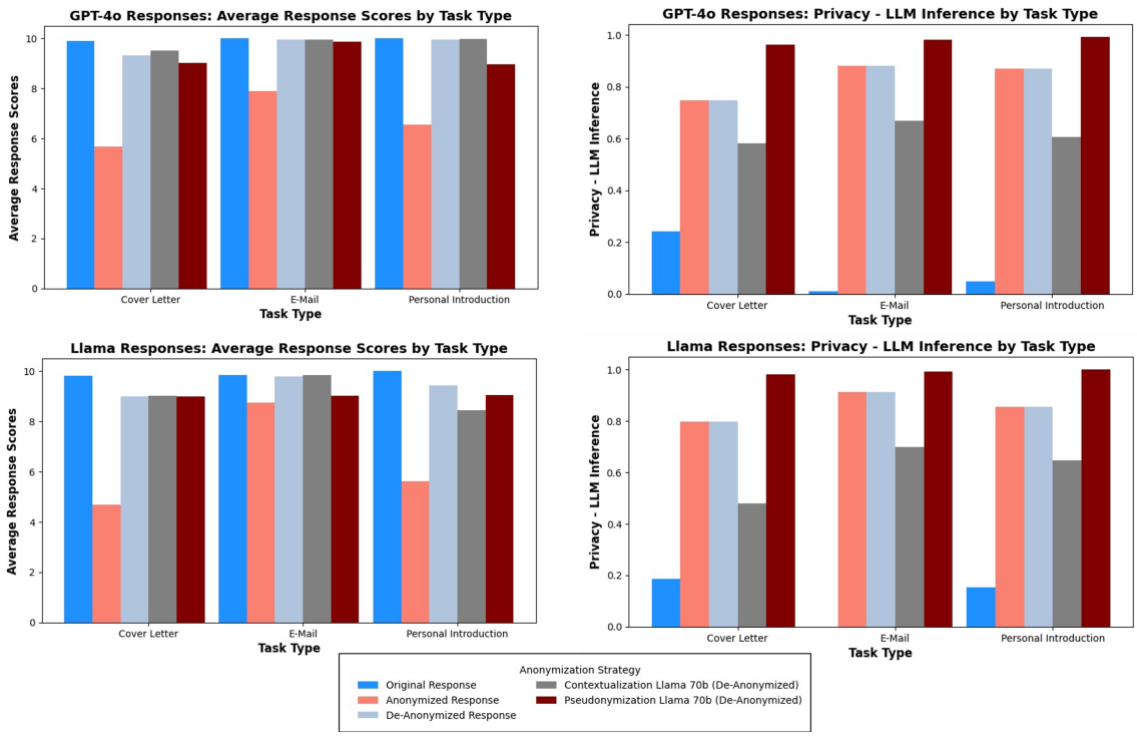

Figure 3. Evaluations by Task Type

moderate positive effect on cover letters, though its impact was smaller than for GPT-4o.

However, for Llama-generated responses, we found a notable drop in response quality for personal introductions when using contextualization. This suggests that while contextual descriptions help preserve coherence in structured tasks where tailoring responses for entities matters like cover letters, they may introduce unintended biases or distortions in more flexible, open-ended tasks like personal introductions.

Regarding privacy, we found that for both GPT-4o and Llama models, cover letters had the lowest privacy scores. This indicates that the contextual and job-specific details present in cover letter prompts may make it easier for LLMs to infer the original entities, reducing the effectiveness of simple anonymization strategies. Hence, for cover letters, we require strategies that better obscure entity identities, such as pseudonymization, which proved to be effective in preventing LLM inferences across all task types.

4 Discussion

Our results demonstrate that anonymization can effectively protect sensitive information while maintaining response quality in personalized LLM tasks. Across different anonymization strategies, we observe a minimal reduction in response quality (roughly 1 point on a 10-point scale), while achieving 97%-99% entity masking, indicating a strong privacy gain.

Interestingly, simple anonymization and de-anonymization methods (e.g., direct entity masking and backmapping) yield the best results for Llama-generated responses, suggesting that additional context can introduce unnecessary variability. Although prompts clarify that contextual information is provided solely as background information, we found that models often over-integrate these details into responses, such as mentioning that the user lives in an *East Coast city* in a cover letter. In contrast, GPT-4o benefits from contextualized anonymization, where entity replacements include descriptive labels. This indicates that some models may better leverage contextual cues to compensate for missing specific entity references.

Our findings highlight the importance of tailoring anonymization strategies to specific LLM architectures and task types, as different models interpret masked entities and contextual information differently. Additionally, we show that effective anonymization does not necessarily require complex transformations, as simpler

techniques achieve comparable privacy protection with minimal response degradation.

This study underscores the feasibility of deploying automated anonymization workflows for real-world, privacy-sensitive LLM applications. Future work could explore adaptive anonymization techniques, where models dynamically adjust anonymization levels based on task sensitivity and model behavior.

5 Limitations

While our study provides valuable insights into the effects of anonymization on LLM-generated responses, several limitations should be considered when interpreting our findings.

First, our analysis is limited to ChatGPT-4o and Llama models, meaning the results may not generalize to other large language models, such as Claude, Gemini, or Mistral, which may process anonymized prompts differently. Different LLM architectures may exhibit varying sensitivity to entity masking, contextualization, or pseudonymization, potentially leading to different response quality and privacy trade-offs. Future work could expand the analysis to a broader range of models to assess generalizability across LLM ecosystems.

Second, while we employ the LLM-as-a-Judge method to automate response quality evaluation, our study does not incorporate human raters. Although recent work suggests that ratings with GPT-4o align well with human preferences, LLM-based scoring may not fully capture nuances such as subtle coherence issues, tone, or factual correctness. Similarly, our evaluation does not explicitly assess truthfulness or detect hallucinations in de-anonymized responses. For example, a de-anonymized cover letter could introduce fabricated details not present in the original prompt. Future research could incorporate human evaluations and factual consistency checks to ensure that anonymization does not introduce unintended distortions or hallucinated content that may not be detected by AI-based scoring.

Third, our dataset consists of synthetically generated prompts rather than real user queries. While this allows for an automated workflow, real-world user prompts may introduce greater variation, ambiguity, or complexity that could affect both anonymization performance and response generation. In particular, one challenge is anonymizing lesser-known entities, such as small businesses or less prominent organizations, which LLM-based techniques may struggle to recognize. Since our synthetic prompts are LLM-generated, they may overrepresent well-known entities, whereas real-world inputs may include more unique or less widely recognized names that could be more challenging to identify and anonymize effectively. Future research could explore real-world anonymization cases to assess how different anonymization strategies perform in practical applications.

Moreover, while our privacy evaluation effectively quantifies entity masking and assesses re-identification risks using LLM inference, it does not fully capture the severity of a single entity leakage. The current approach assumes that privacy loss is proportional to the number of entities disclosed, but in real-world applications, even a single leaked entity (such as a person's name) could constitute a significant privacy risk. This is particularly critical in tasks like cover letters and business emails, where context may allow an adversary to infer personal details even if only one entity is revealed.

Finally, our study employs a single anonymization approach, using a BERT-based NER model for entity recognition. While this approach is effective for structured anonymization, other anonymization techniques exist, including LLM-based NER. In addition, recent privacy-preserving prompt sanitization techniques, such as Casper (Chong et al., 2024), extend beyond NER by incorporating topic-based anonymization and rule-based filters. Future research could explore how different anonymization methods interact with various LLMs, assessing trade-offs between privacy effectiveness and response degradation.

Acknowledgments

This research was supported by the Culture, Sports, and Tourism R&D Program through a Korea Creative Content Agency grant funded by the Ministry of Culture, Sports, and Tourism in 2024 (Project Name: Development of multimodal UX evaluation platform technology for XR spatial responsive content optimization; Project Number: RS-2024-00361757).

References

AWS. (2023). Foundational data protection for enterprise LLM acceleration with Protopia AI Available at: https://aws.amazon.com/

Azure. (2024). Data, privacy, and security for Azure OpenAI Service Available at: https://learn.microsoft.com/en-us/legal/cognitive-services/openai/data-privacy?tabs=azure-portal

Chen, Y., Li, T., Liu, H., & Yu, Y. (2023). Hide and seek (has): A lightweight framework for prompt privacy protection. *arXiv preprint arXiv:2309.03057*.

Chong, C. J., Hou, C., Yao, Z., & Talebi, S. M. S. (2024). Casper: Prompt Sanitization for Protecting User Privacy in Web-Based Large Language Models. *arXiv preprint arXiv:2408.07004*.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2020). Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

European Data Protection Supervisor. (2025). Large Language Models (LLM). Available at: https://www.edps.europa.eu/data-protection/technology-monitoring/techsonar/large-language-models-llm_en

Igamberdiev, T., & Habernal, I. (2023). DP-BART for privatized text rewriting under local differential privacy. *arXiv preprint arXiv:2302.07636*.

Jovic, M., & Mnasri, S. (2024). Evaluating AI-Generated Emails: A Comparative Efficiency Analysis. *World Journal of English Language*, *14*(2).

Mao, Y., Liao, X., Liu, W., & Yang, A. (2024). A Practical and Privacy-Preserving Framework for Real-World Large Language Model Services. *arXiv preprint arXiv:2411.01471* .

Pasch, S. (2025). LLM Content Moderation and User Satisfaction: Evidence from Response Refusals in Chatbot Arena. *arXiv preprint arXiv:2501.03266*.

Raju, R., Jain, S., Li, B., Li, J., & Thakker, U. (2024). Constructing domain-specific evaluation sets for llm-as-a-judge. *arXiv preprint arXiv:2408.08808*.

Riabi, A., Mahamdi, M., Mouilleron, V., & Seddah, D. (2024). Cloaked classifiers: Pseudonymization strategies on sensitive classification tasks. *arXiv preprint arXiv:2406.17875*.

Staab, R., Vero, M., Balunović, M., & Vechev, M. (2023). Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298*.

Sun, X., Liu, G., He, Z., Li, H., & Li, X. (2024). DePrompt: Desensitization and Evaluation of Personal Identifiable Information in Large Language Model Prompts. *arXiv preprint arXiv:2408.08930*.

Yermilov, O., Raheja, V., & Chernodub, A. (2023). Privacy-and utility-preserving nlp with anonymized data: A case study of pseudonymization. *arXiv preprint arXiv:2306.05561*.

Yukhymenko, H., Staab, R., Vero, M., & Vechev, M. (2024). A Synthetic Dataset for Personal Attribute Inference. *arXiv preprint arXiv:2406.07217*.

Xu, C., Li, J., Li, P., & Yang, M. (2023). Topic-guided self-introduction generation for social media users. *arXiv preprint arXiv:2305.15138*.

Zheng, A., Rana, M., & Stolcke, A. (2024). Lightweight Safety Guardrails Using Fine-tuned BERT Embeddings. *arXiv preprint arXiv:2411.14398*.

Zheng, L., Chiang, W. L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., ... & Stoica, I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, *36*, 46595-46623.

Zhong, M., Liu, Y., Yin, D., Mao, Y., Jiao, Y., Liu, P., ... & Han, J. (2022). Towards a unified multi-dimensional evaluator for text generation. *arXiv preprint arXiv:2210.07197*.

Zinjad, S. B., Bhattacharjee, A., Bhilegaonkar, A., & Liu, H. (2024, July). Resumeflow: An llm-facilitated pipeline for personalized resume generation and refinement. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 2781-2785).