Towards Privacy Preservation in AI Summarization: Balancing Privacy and Completeness

Anonymous ACL submission

Abstract

With the rapid integration of AI in virtual meeting platforms, automatic summarization has become essential for productivity across sectors. While text summarization has seen significant progress, dialogue-based summarization remains underexplored, with efforts largely focusing on improving quality and addressing domain adaptation. Privacy concerns, however, are often neglected, exposing sensitive information, particularly in critical settings like healthcare, finance, and legal interactions. This paper introduces a privacy-sensitive taxonomy addressing diverse scenarios and explores strategies to safeguard privacy in AI-generated summaries. Our hybrid approach combines rule-based and learning-based techniques to address direct and indirect privacy threats while maintaining content accuracy. Using a specialized dataset curated around our taxonomy, we fine-tuned large language models and evaluated them with human and automated metrics, including Privacy and Completeness Scores. The results demonstrate the effectiveness of these models in mitigating privacy risks, offering a strong foundation for advancing privacy-preserving AI technologies while balancing privacy and completeness.

1 Introduction

With the integration of AI technologies in virtual meeting platforms like Google Meet, Zoom, and Microsoft Teams (Google, 2024; Zoom, 2023; Microsoft, 2024b, 2023), the automatic generation of summaries in remote collaboration environments —be it for meetings, codes, documents, or entire repositories — has become a powerful tool to enhance productivity and manage information flow. However, this advancement brings significant privacy concerns to the forefront. As these platforms process vast amounts of sensitive data, ensuring privacy is critical to prevent unauthorized access, data breaches, and compliance violations. Regulations like GDPR, CCPA, and HIPAA impose strict privacy requirements, yet breaches persist, highlighting the need for better data management. By prioritizing privacy, these summaries help enforce compliance and minimize data exposure across various fields, enabling seamless collaboration while upholding the privacy and compliance essential to digital ecosystems. Figure 4 compares the summary generated by the current baselines for a given conversation with an ideal target summary. The application of Privacy-preserving summaries have further been discussed in Appendix A. (General Data Protection Regulation (GDPR), 2021; Security Metrics, 2024; U.S. Department of Health and Human Services, 2021)

Currently, a lot of work has already been done in the field of text summarization as can be seen from the works of Yadav et al. (2022), Goyal et al. (2023) Hariri (2024), Shakil et al. (2024), and Zhang et al. (2023). A point to note is that although Dialogue-based summarization has become increasingly important across domains, yet the task remains largely unexplored at hand with even less focus on associated Privacy concerns. Some of the earlier works exploring Dialogue-based tasks like those by Wang et al. (2022), Gao et al. (2023) and Zhu et al. (2023) using smaller neural summarization models, and the more recent ones using LLMs like the works of Li et al. (2024b), Ramprasad et al. (2024), Tang et al. (2024) and Tian et al. (2024) , are all mainly focused for maintaining the overall quality of the summary generated, working on factors like Factual Consistency, Hallucinations and Domain Adaptation using curated datasets and trained models, with not much discussions done on Privacy. The work done by Dou et al. (2024) does address privacy in the form of *self-disclosures* by developing a taxonomy and fine-tuning models for better results, but we came across a few limitations including a more pronounced focus on a user-identifiable level and reduced scope of overall extensibility under different settings, elaborated in the next section.

Gumusel et al. (2024) identified significant privacy concerns in AI-powered chatbots like ChatGPT, including monitoring, data aggregation, and unauthorized sharing—risks that highlight potential privacy breaches in AI-driven summarization tools for virtual meetings if not properly managed. Moreover, Ruane et al. (2019) discussed the broader ethical implications of deploying *Conversational Agents* across various sectors, emphasizing the importance of handling data sensitively to avoid privacy breaches and prevent biases or misrepresentation in generated summaries.

Current privacy-preserving strategies can be broadly categorized into three main approaches:

- **Prompt-based masking** use task-specific prompts to guide models in masking personally identifiable information (PII) automatically but struggles with edge cases (Wang et al., 2023; Sivarajkumar et al., 2024)
- **Rule-based checklists** rely on predefined rules like direct matching using regex patterns or checklists trained using Named Entity Recognition (NER) models to mask PII consistently but struggle in contexts when dealing with new types of sensitive data or indirect references like partial private key or a masked credit card number in a non-standard format (Soomro et al., 2017; Sivarajkumar et al., 2024)
- **Learning-based approaches** leverage models trained on large datasets containing labeled PII to autonomously identify and mask sensitive information (Zheng et al., 2024; Sanh et al., 2022)

2 Our Contributions

The study understands privacy as protecting sensitive data against unauthorized access and is among the first to address this problem in depth. We propose a novel taxonomy for identifying sensitive data, drawing from literature and real-world conversational scenarios across twelve settings (Figure 2). Each setting is structured into categories, subcategories, and elements, prioritized by sensitivity (High, Medium, Low). Inspired by Zhang et al. (2024) and Fu et al. (2024), our approach integrates rule-based precision with learning-based adaptability, training LLMs on a dataset capturing diverse privacy breaches to generate aligned, privacy-preserving responses. We then proceeded

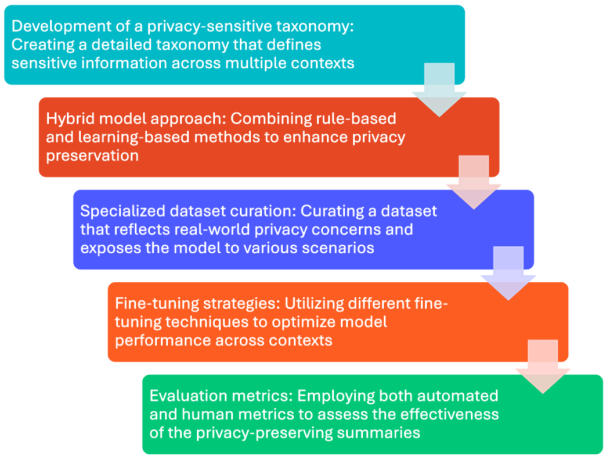


Figure 1: An overview of the systematic approach used to generate and verify privacy-preserving summaries in our research

to the dataset curation process which addressed gaps in existing datasets by generating synthetic conversations using GPT-4o across settings, having for each data point a dialog, metadata mapped to privacy categories, summaries, violation labels, and revised summaries. To ensure realistic scenarios, 200 data points from four public benchmarks (DialogSum, SAMSum, etc.) were picked based on their use cases and integrated with our dataset to mimic real-world settings. We then tested not only on these datapoints but also on the ai-masking-400k dataset (Figure 8), achieving high accuracies of detection to show that our approach masks sensitive information in actual real-world settings as well. Finally, the experiments involved fine-tuning seven LoRA-based models on Phi3.5-mini, testing diverse techniques among overfitting analysis, early stopping, and preference optimization methods. Evaluations used GPT-4-based Privacy and Completeness scores (1-5 scale), NLP metrics (ROUGE, BERTScore, MoverScore), and Human evaluation (Consistency, Relevance, Coherence, Privacy) with Kappa scores to measure inter-rater agreement. This was then followed by the interpretations based on the results thus obtained. Figure 1 gives an overview of the systematic approach used to generate and verify privacy-preserving summaries in our research.

3 Relevant Works

Differential Privacy The introduction of differential privacy into language models provides foundational insights into privacy preservation. Li et al. (2024a) introduce a comprehensive evaluation

framework for language models, assessing privacy vulnerabilities through simulated attacks. However, its focus on cryptographic and DP metrics means it may not fully account for the subtleties of natural language like semantic nuances and contextual implications, risking disclosure of personally identifiable information (PII) or sensitive personal opinions, resulting in privacy breaches. Mu et al. (2024) use differential diversity prompting to adapt to the context of the task, making them more versatile and effective in handling diverse reasoning challenges. The study enhances reasoning capabilities but lacks mechanisms to assess and manage sensitive information, posing risks in regulated fields like healthcare or finance. This oversight may lead to increased privacy violations, potentially compromising compliance with various regulatory bodies.

Privacy Frameworks Dou et al. (2024) addressed privacy risks in online self-disclosures by developing language models trained on Reddit data to detect and abstract sensitive information using a predefined taxonomy. The study demonstrated promising results in minimizing privacy breaches. However, the major focus on personal identifiers along with the static taxonomy limits the flexibility to adapt to new contexts of sensitive information, while reliance on Reddit posts reduces the models' effectiveness in diverse linguistic and cultural contexts as well. This work might benefit from a dynamic taxonomy and a more inclusive dataset spanning various platforms and scenarios.

Fideslang Ethyca (2023a,b) is a technology company specializes in privacy engineering, focusing on helping organizations to streamline privacy compliance with global regulations like GDPR. In this pursuit, Ethyca developed Fideslang, an open-source privacy taxonomy that categorizes data types, uses, and subjects, enabling developers to embed privacy directly into the software development lifecycle. While effective in this regard, its rule-based structure is limited to software systems and lacks adaptability to unstructured interactions where its generic categorizations might not fully capture the subtleties of different contexts. To address this primary issue, a new privacy taxonomy overcoming the predefined limitations of the existing taxonomy is needed, enabling dynamic adaptation and consistent privacy protection across diverse scenarios through context-aware, sensitivity-based classifications .

Current Baselines In enhancing the safety and reliability of interactions involving LLMs, both the ShieldGemma project (Zeng et al., 2024) and Llama Guard (Inan et al., 2023) have made significant strides with ShieldGemma focusing on advanced content moderation models to detect harmful content such as hate speech and harassment, while Llama Guard classifying safety risks associated with user prompts and AI responses through a structured safety risk taxonomy. However, both initiatives lack an adaptive framework for managing sensitive information across contexts and have datasets, though effective for detecting harmful content, lack coverage of complex privacy scenarios, limiting their real-world applicability. Our research addresses these gaps by incorporating diverse real-world cases, proposing an adaptable taxonomy, and training robust models to balance privacy and utility. With strong results, our work sets a new benchmark for privacy-preserving AI, ensuring both safety and contextual sensitivity in AI interactions.

4 Privacy Taxonomy

The question of what constitutes privacy and what information is considered sensitive is central to ongoing debates and studies like those conducted by Li et al. (2023), where the authors emphasize that privacy can be understood as the safeguarding of sensitive and personal information that individuals or institutions hold, against any kind of unauthorized access, and by Veritas Technologies (2023), where privacy is defined as the individual's control over their personal and sensitive data, protecting such data from unauthorized access and breaches. The multifaceted nature of privacy leads to the definition of a dynamic entity that changes with the context and setting of a conversation. Within each setting, elements are considered sensitive on varying levels and require masking to prevent accidental leakage (Figures 2 and 5) . To address these complexities, based on existing literature, datasets and most common scenarios we came across, we have proposed a taxonomy encompassing 12 settings - Family and Relationships, Healthcare settings, Employment, Finances, Social Media, Legal Proceedings, Political Activities, Religious Contexts, Sexual Orientation and Gender Identity, Travel and Location, and Education, along with a Generic setting, covering any information that

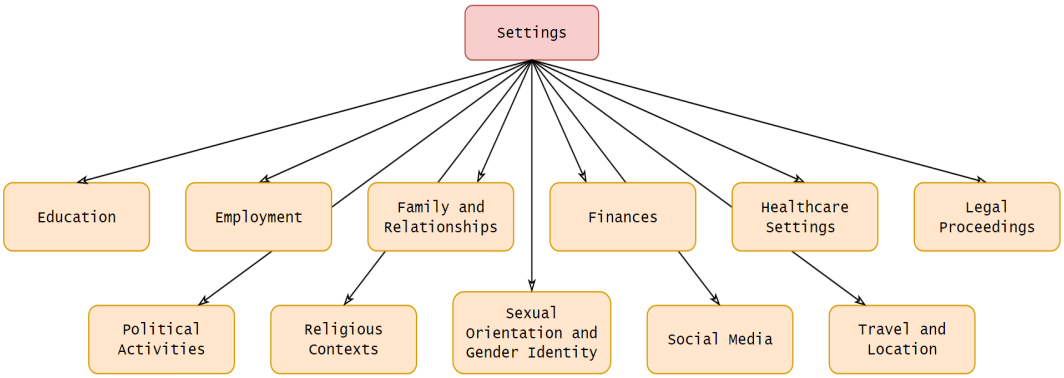


Figure 2: An overview of the Taxonomy showing the different settings considered

comes under PII. The settings were chosen to cover most of the sensitive information that typically arise in regular conversations in our day-to-day lives and is at risk of being exposed. We delve deeper into each setting, identifying all the possible different sensitive categories, sub-categories, and elements, organized according to the different levels of priority or sensitivity—High, Medium, or Low. We follow the Fideslang notation given by Ethyca (2023b), representing any element as <setting>.<sensitivity_level>.<category>.<subcategory (if any)> , with each of the levels mentioned in snake_case. For example, Work History from Figure 5 (b) would be represented as employment.high_sensitivity.work_history. While a strict demarcation is impractical, our approach aligns with general privacy concepts and perceptions of sensitivity, organizing privacy-sensitive information into hierarchies and clusters, and enabling a holistic view of potential risks. Our goal is not to achieve perfect privacy masking but to balance it with completeness, ensuring that all the necessary information is delivered without significant leakage of sensitive data, adhering to accepted privacy standards.

Moreover, this taxonomy can also can be incorporated into a dynamic pipeline where it would serve as a foundational "safety layer", ensuring alignment with existing privacy regulations, and LLMs can then be used to propose new categories, either deeper into an existing setting or a new one entirely, with human experts validating these additions to ensure legal and ethical alignment. This would ensure that the taxonomy remains both comprehensive and adaptable to new scenarios.

5 Dataset Curation

Existing datasets often focus on narrow aspects like hate speech or explicit identifiers, missing indirect privacy risks such as inferences or metadata, which are critical in domains like healthcare, law, and finance. Our dataset addresses this by covering both explicit and subtle privacy violations. We generated 1,100 synthetic data points using GPT-4o, each containing six key columns: setting, dialog, metadata mapped to privacy categories, privacy-preserving summary, evaluation labels for violations, and corrected summaries addressing identified privacy risks. The process followed an approach consisting of five key steps:

- **Step 1: Dialog Generation** We generated conversations between participants based on our taxonomy, covering different privacy-sensitive situations. For each setting we generated around 100 conversations, infusing a few minor settings and their related sensitive elements. We passed the major setting and the minor settings in the prompt, along with our taxonomy to help generate the required conversations.
- **Step 2: Metadata Extraction** Next, we extracted all relevant metadata from the conversation, mapping it to the appropriate privacy categories in the taxonomy. Here we provided the conversation generated in the previous step along with the taxonomy as input in the prompt.
- **Step 3: Summary Generation** In the third step, a privacy-preserving summary was generated from the conversation. For the inputs, we provided the Conversation and the Taxonomy. Guided by the taxonomy, this summary aimed to remove sensitive information while retaining key elements to provide an overall idea of the conversation.
- **Step 4: Summary Quality** After the initial summary, we identified privacy violations by

providing the Summary and the Metadata generated above with the prompt and asked GPT-4o to check if anything sensitive in the metadata is leaked into the summary. In case of minor, low sensitivity or no violations, the summary was labeled as "GOOD", otherwise "BAD" along with the violations in the same manner as in the Taxonomy.

- **Step 5: Summary Correction** If a summary was labeled as "BAD," a corrective step was taken where We provided in the input prompt the Summary generated along with the Violations identified in the previous step . We then obtained a revised summary generated by addressing the violations found in the earlier summary.

To ensure data quality, we first manually verified 30 initial datapoints and used them for In-Context Learning (ICL) with GPT-4o to generate better datapoints. These were **not** partial checks but a structured seed dataset for guiding ICL. Since privacy can be subjective, each generated conversation was then **manually reviewed** to ensure it seemed natural and aligned with realistic possibilities. While synthetic data formed the majority, we incorporated 200 real-world examples from benchmark datasets—DialogSum, SAMSum, ConvoSumm, and TweetSum—selecting 50 examples from each to improve real-world connectivity and mimic realistic scenarios. Edge cases were deliberately included for completeness to ensure broad coverage of privacy-sensitive situations. After curating the entire dataset, **full manual verification** was done on all the datapoints to ensure alignment with real-world privacy needs. The final dataset had 1,300 data points, with 1,065 for training and 235 for testing, ensuring a comprehensive privacy coverage. The importance of this dataset is further elaborated in Appendix B.1.

6 Experiments

Model Prompting We had done an extensive analysis to check the adaptation of the models using prompting alone (zero-shot, one-shot, few-shot) but we observed that despite providing the complete taxonomy and incorporating few-shot examples for In-Context Learning (ICL), the generated summaries exhibited a lot of inconsistencies, with some successfully masking sensitive information while others inadvertently leaking private data, even though it had specifically been provided information about sensitive data including named entities (Appendix E.1 contains more details on this). Figure 9 shows a few cases where despite providing all information, prompting alone failed to adhere to some indirect as well as some very basic checks. These results were unreliable, as there was no consistent guarantee of privacy preservation across responses. Given the limitations of prompting-based approaches, we shifted our focus to fine-tuning models for the task.

Model Fine-tuning For testing our dataset in privacy-preserving summarization, we fine-tuned seven LoRA-based models using different techniques each with Phi-3.5-mini as the base (**Model 0**), chosen for its 128K token context length and strong dialogue-handling capabilities. **Model 1** analyzed overfitting by training on both correct and incorrect summaries, while **Model 2** applied early stopping to balance learning and generalization. **Model 3** was trained solely on correct summaries, serving as a benchmark for ideal conditions without dealing with potential privacy leaks explicitly. **Model 4** introduced corrected summaries post-privacy violations to teach correction mechanisms. **Model 5** used Direct Preference Optimization (DPO) to align with human preferences, while **Model 7** leveraged Odds Ratio Preference Optimization (ORPO) for the same with efficient handling of ambiguous privacy violations. **Model 6** generated both normal and privacy-preserving summaries simultaneously for training the model to balance completeness and privacy-preservation dynamically. These models were systematically designed to explore different optimization strategies, ensuring a complete evaluation of privacy protection in summarization. Section C elaborates further about the different techniques used to train the models along with the intuition behind them.

7 Evaluations

7.1 LLM-as-a-Judge

To evaluate the model responses, we employed Privacy and Completeness scores as metrics, using the LLM-as-a-judge evaluation technique. GPT-4 was used as the judge, scoring summaries on these two aspects based on a detailed scoring rubric with the original conversation, the generated summary, and the scoring criteria included. The Privacy score assesses the extent to which summaries preserve

Table 1: Model performance across privacy settings (Bold values compare models to the highest scores overall)

settings	Model 0	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	GPT-4o
Generic	0.3673	0.5714	0.3878	**0.9592**	0.6327	0.9796	**0.9388**	0.9796	0.7551
Education	0.2973	0.4865	0.5676	**0.9459**	0.6757	0.9730	**0.9459**	0.9730	0.7297
Employment	0.2083	0.6667	0.4375	**0.9583**	0.5000	0.9792	**0.9583**	0.9792	0.7500
Family and Relationships	0.2778	0.5370	0.3519	**0.9444**	0.5926	0.9815	**0.9444**	0.9630	0.7407
Finances	0.4524	0.6667	0.4286	**0.9524**	0.5238	0.9762	**0.9286**	0.9762	0.7619
Healthcare settings	0.4615	0.6346	0.3846	**0.9423**	0.5962	0.9808	**0.9231**	0.9808	0.8077
Legal Proceedings	0.2647	0.5294	0.4118	**0.9118**	0.6176	0.9706	**0.9118**	0.9706	0.7941
Political Activities	0.2778	0.5556	0.3056	**0.9167**	0.6944	0.9722	**0.9444**	0.9722	0.7778
Religious Contexts	0.2979	0.6170	0.5319	**0.9149**	0.5957	0.9787	**0.9149**	0.9787	0.8085
Sexual Orientation and Gender Identity	0.2564	0.5128	0.4872	**0.9487**	0.5385	0.9744	**0.9487**	0.9744	0.7436
Social Media	0.2286	0.6857	0.6000	**0.9143**	0.5429	0.9714	**0.9429**	0.9714	0.8000
Travel and Location	0.2121	0.6667	0.5455	**0.9394**	0.5455	0.9697	**0.9394**	0.9697	0.7273
Average	0.3063	0.5949	0.4466	**0.9387**	0.5870	**0.9763**	**0.9368**	**0.9743**	0.7668

Table 2: Comparison of Privacy and Completeness scores for Models across LLMs

Models	Phi-3.5			Phi-4			Qwen2.5		
	Model 0	Model 3	Model 6	Base Model	Model 3	Model 6	Base Model	Model 3	Model 6
Privacy Score	4.235	4.605	4.884	3.8205	4.6175	4.6562	3.6438	4.5857	4.3233
Completeness Score	4.270	4.051	4.047	4.4756	4.0424	4.0293	4.2439	3.9878	4.0537

sensitive information by effectively masking sensitive information while the Completeness score measures how well summaries retain the key information from the original conversation and convey the essential points. Scores were rated on a 5-point scale, ranging from 5 (perfect) to 1 (critical issues), with 4 indicating minor issues, 3 moderate gaps, and 2 significant shortcomings in privacy or completeness. It should be noted that GPT-4 here doesn't serve as an infallible judge, instead it serves as a preliminary evaluation tool in a multi-layered framework, using structured guidelines for Completeness and Privacy. Its assessments are validated against NLP metrics and human evaluations, ensuring a balanced, iterative approach that minimizes biases. Moreover, all conclusions drawn in the paper are then based on a combination of GPT-4's assessments, NLP metrics and human evaluations, ensuring a balanced approach that mitigates potential biases from any source. We also evaluated all models for various categories and subcategories across settings and obtained setting-wise and overall privacy accuracy scores. These, along with the LLM-as-a-Judge scores together, acted as a screening tool for us to determine which techniques effectively balanced Privacy and Completeness, informing the next steps of evaluation in our research.

Results Based on overall average scores across settings (Table 1), the percentage of acceptable summaries (Figure 7), and LLM metrics (Table 9), Models 3 and 6 effectively balanced privacy and completeness, achieving scores comparable to or surpassing the baseline GPT-4o and approaching the scores of the Ground Truth summaries. Consequently, we decided to focus on these two models for further analysis and experimentation. Model 3, trained solely on privacy-preserving summaries, achieved high scores in privacy (4.605) and completeness (4.051), while Model 6, generating both normal and privacy-preserving summaries simultaneously, also achieved similarly high scores in privacy(4.884) and completeness (4.047), reflecting their capability to manage the trade-offs effectively. Similar trends were observed with Phi-4 and Qwen2.5 (Table 2), confirming the robustness of Model 3 and Model 6 across LLM architectures with privacy scores around 4.5 and completeness above 4.0, faring much better than the respective base models overall.

7.2 NLP Metrics

We also used metrics that current models often rely on such as ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020), and MoverScore (Zhao et al., 2019), all meant to measure content quality but in different ways. While ROGUE focuses on text overlap of n-grams, BERTScore and MoverScore rely on semantic embeddings to evaluate the similarity between the generated and reference summaries. This semantic-based evaluation helps accommodate the different conversation and dia-

Table 3: Model Comparison across NLP Metrics with Phi-3.5 baseline

Models	ROUGE Scores				BERTScores			MoverScores		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	BERT-base	RoBERTa	DeBERTa	BERT-Tiny	BERT-Small	BERT-Medium
Model 3	0.4934	0.2062	0.3573	0.3572	0.7163	0.9156	0.7664	0.5716	0.5005	0.4530
Model 6	0.4998	0.2143	0.3680	0.3676	0.7236	0.9177	0.7715	0.5832	0.5112	0.4624
Base Model	0.4766	0.1952	0.3450	0.3471	0.7018	0.9111	0.7526	0.5591	0.4834	0.4323

Table 4: Model Comparison across NLP Metrics with Phi-4 baseline

Models	ROUGE Scores				BERTScores			MoverScores		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	BERT-base	RoBERTa	DeBERTa	BERT-Tiny	BERT-Small	BERT-Medium
Model 3	0.5112	0.2247	0.3729	0.3726	0.7192	0.9163	0.7682	0.5726	0.5028	0.4674
Model 6	0.4963	0.2052	0.3587	0.3688	0.7001	0.9107	0.7521	0.5623	0.4855	0.4239
Base Model	0.4317	0.1624	0.2883	0.2960	0.6425	0.8914	0.7009	0.4629	0.3976	0.3573

logue patterns encountered during testing, providing more flexibility in measuring summary quality. These metrics were computed using the Frugalscore Framework (Eddine et al., 2021) for efficient computation.

Results Regarding the NLP metrics, we observed similar trends across LLMS Phi-3.5, Phi-4 and Qwen2.5 (Tables 3, 4 and 5), where Model 3 and Model 6 achieved the highest scores across all ROUGE metrics, suggesting better retained critical information while adhering to privacy constraints across different model architectures. They also delivered highest scores across all configurations of BERTScore, highlighting superior semantic understanding and alignment with ground-truth summaries. The models again emerged as the strongest across BERT-based student models in MoverScore, indicating ability to align summaries with input conversations while preserving semantic integrity. In all these cases they consistently outperformed the base model particularly for use cases requiring both context preservation and strong privacy safeguards, making them highly suitable for applications in sensitive domains.

7.3 Human Evaluation

While NLP metrics prioritize semantic similarity and coherence they often fail to assess privacy preservation, meaning a high score could still imply exposure of sensitive information. This paper thus advocates for a *Human evaluation* to ensure summaries meet privacy constraints while maintaining coherence, relevance, and factual accuracy to a standard generally considered acceptable. Our evaluation criteria, adapted from DialogSum (Chen et al., 2021) with an added Privacy parameter, include:

- **Consistency**: Measures whether the summary consistently reflects the original conversation.
- **Relevance**: Judges how well the summary retains essential information for completeness
- **Coherence**: Evaluates whether the summary logically flows and makes sense.
- **Privacy**: Assesses how well the summary effectively masks sensitive data

Evaluations used a binary scale (0 or 1) with inter-rater agreement measured via Cohen's and Fleiss' Kappa scores (McHugh, 2012). Initially, six evaluators, who had been given instructions on how to annotate using a clear evaluation criteria, assessed 10 conversations across seven fine-tuned models, the base model, GPT-4o, and ground truth summaries. After identifying top-performing models, a more focused or Distilled Evaluation followed on 20 additional conversations to validate findings, ensuring a rigorous and credible assessment of the models' effectiveness in generating privacy-preserving summaries. Further details about our choice of scale here have been discussed in Appendix E.3.

Results In the initial human evaluation, Model 3 showcased a strong performance, achieving high scores in Privacy (0.89) while also maintaining good results across other dimensions. Similarly, Model 6 demonstrated high performance, with a Privacy score of 0.88, reflecting its effectiveness in privacy-preserving summarization. Both models outperformed GPT-4o, which, despite strong overall performance, struggled with maintaining a decent score in Privacy (0.73). The distilled evaluations reinforced these findings, with Model 3 slightly improving its Privacy score while Model 6 showed further advancements, reaching 0.90 in Privacy. Both models continued to outperform GPT-4o, demonstrating their suitability for privacy-

Table 5: Model Comparison across NLP Metrics with Qwen2.5 baseline

Models	ROUGE Scores				BERTScores			MoverScores		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	BERT-base	RoBERTa	DeBERTa	BERT-Tiny	BERT-Small	BERT-Medium
Model 3	0.4365	0.1758	0.3649	0.3351	0.6830	0.9049	0.7374	0.5410	0.4685	0.4100
Model 6	0.4605	0.1984	0.3460	0.3557	0.6971	0.9099	0.7503	0.5636	0.4732	0.4390
Base Model	0.4192	0.1513	0.2738	0.3139	0.6258	0.8861	0.6849	0.4542	0.3885	0.3513

Table 6: Human evaluation of Models against GPT-4o and Groundtruth as baselines with Kappa scores

Models	Initial Evaluation				Distilled Evaluation			
	Consistency	Relevance	Coherence	Privacy	Consistency	Relevance	Coherence	Privacy
Model 3	0.88	0.83	0.84	0.87	0.93	0.85	0.86	0.88
Model 6	0.90	0.81	0.85	0.86	0.91	0.84	0.88	0.90
GPT-4o	0.91	0.80	0.82	0.75	0.92	0.82	0.84	0.72
Ground Truth	0.93	0.88	0.90	0.91	0.93	0.83	0.87	0.80
Cohen's Kappa (Avg)	0.798	0.716	0.817	0.744	0.813	0.729	0.832	0.761
Fleiss' Kappa	0.797	0.714	0.817	0.748	0.814	0.732	0.834	0.763

centric summarization tasks with minimal content quality compromise. The high Kappa scores, both Cohen's and Fleiss' scores closing 0.8 and above across all dimensions, validated these results with average scores increasing in the Distilled Evaluation, indicating strong agreement among evaluators and further validating that well-tuned models can deliver enhanced privacy protection without compromising summary quality.

8 Future Work

The study by Li et al. (2020) explores the impact of cultural differences on privacy and the need for dynamic categorization of sensitive elements according to contextual settings while the work of Liang (2019) talks about ways to implement user-level customization. Future work could aim to integrate our models into a dynamic pipeline, that allows individuals to provide relational information for tailored masking of sensitive data, adapting to context and user needs to address these concerns. The management of permission access to sensitive data could present challenges across organizations, which could be minimized by exploring ways for curation, training and inference locally within a secure environment. The model's performance after quantization could also be explored to enable edge computing on personal devices, enhancing privacy, reducing latency, and improving scalability.

9 Conclusion

This study addresses privacy-preserving summarization, balancing completeness with safeguarding privacy. We introduced a structured taxonomy identifying sensitive elements across diverse domains and curated a dataset integrating synthetic and real-world examples to ensure diversity and relevance. Seven models were fine-tuned on Phi 3.5, with **Model 3** and **Model 6** performing best. The robust evaluation framework, combining NLP metrics and human assessments, confirmed these models' capabilities in real-world settings. Model 3, trained on high-quality, privacy-aware examples, demonstrated an optimal balance by internalizing omission patterns without excessive false positives. Model 6 introduced a dual-output design, generating both standard and privacy-preserving summaries, further enabling dynamic adaptation to meet varying privacy requirements across scenarios. In contrast, models emphasizing strict redactions, such as Model 5 (DPO) and Model 7 (ORPO), prioritized privacy but compromised completeness at the cost of excessive content loss, highlighting the inherent trade-off between the two. While GPT-4o served as a benchmark, its lack of domain-specific privacy control emphasized the necessity of such a tailored, domain-focused training. Our findings demonstrate that no single paradigm universally resolves this trade-off but instead privacy-focused AI requires contextual awareness and adaptable architectures to optimize for the task. By establishing a structured taxonomy, dataset, and evaluation framework, this work provides insights for developing systems that mitigate privacy risks while maintaining content integrity, thus supporting compliance, collaboration, and innovation in digital ecosystems. The dynamic nature of privacy, however, demands ongoing refinement of these models to accommodate evolving scenarios and user-specific needs, which is something future works could focus on.

Limitations

Concerns regarding truly unbiased data hold for our use of GPT-4, GPT-4o, and human evaluators to assess the performance and utility of the models trained. One set of evaluations was done using LLMs, while another was done by human evaluators, making it important to acknowledge the possibility that the pre-trained models or evaluators may introduce their own biases when determining what constitutes sensitive information and what qualifies as a privacy violation. Further challenges associated with human evaluation include the increased time and effort required for evaluating the models and their subsequent re-training, if needed, as they are inherently more labor-intensive and slower compared to automated processes. While we attempt to mitigate the risk of subjective bias in human judgments by employing multiple evaluators and using standardized criteria, this approach does not fully eliminate the risk as some degree of bias may still persist since this is not a comprehensive solution.

Although our models have been tested on both synthetic and real-world datasets, they have not yet been deployed in real-world settings where their performance could be continuously monitored and we would be able to observe any violations when exposed to new settings and situations, not covered in the training phase. So, further testing in the real world across a broader range of datasets and varied scenarios is necessary to validate the model's general applicability as well.

Ethics Statement

This study is conducted in accordance with the guidelines of the ACL Code of Ethics. We have rigorously filtered out any potentially offensive content and removed all identifiable information of the participants involved in the study to ensure confidentiality. The primary objective of this study is to develop a tool that mitigates privacy risks associated with dialogue-based summarizations, preventing both direct and indirect leakage of highly confidential and sensitive information. Our evaluations identified no potential risks that could adversely disadvantage any marginalized or otherwise vulnerable populations. We expect that this approach will lead to a net improvement addressing privacy concerns in existing and future models. The curated data is intended solely for research purposes only, and the views expressed in the data do not necessarily reflect the views of the research team or any of its members.

References

AI4Privacy. 2024. https://ai4privacy.com/.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. Dialogsum: A real-life scenario dialogue summarization dataset. *Preprint*, arXiv:2105.06762.

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. *Preprint*, arXiv:2311.09538.

Moussa Kamal Eddine, Guokan Shang, Antoine J. P. Tixier, and Michalis Vazirgiannis. 2021. Frugalscore: Learning cheaper, lighter and faster evaluation metricsfor automatic text generation. *Preprint*, arXiv:2110.08559.

Ethyca. 2023a. Data privacy compliance automation. https://ethyca.com/.

Ethyca. 2023b. Fideslang: A taxonomy for privacy engineering - data privacy management for developers.

Letian Fu, Gaurav Datta, Huang Huang, William Chung-Ho Panitch, Jaimyn Drake, Joseph Ortiz, Mustafa Mukadam, Mike Lambeta, Roberto Calandra, and Ken Goldberg. 2024. A touch, vision, and language dataset for multimodal alignment. *Preprint*, arXiv:2402.13232.

Mingqi Gao, Xiaojun Wan, Jia Su, Zhefeng Wang, and Baoxing Huai. 2023. Reference matters: Benchmarking factual error correction for dialogue summarization with fine-grained evaluation framework. *Preprint*, arXiv:2306.05119.

General Data Protection Regulation (GDPR). 2021. Fines / penalties - General Data Protection Regulation (GDPR).

Google. 2024. "Take notes for me" in Google Meet is now available.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2023. News summarization and evaluation in the era of gpt-3. *Preprint*, arXiv:2209.12356.

Ece Gumusel, Kyrie Zhixuan Zhou, and Madelyn Rose Sanfilippo. 2024. User Privacy Harms and Risks in Conversational AI: A Proposed Framework. *Preprint*, arXiv:2402.09716.

Walid Hariri. 2024. Unlocking the potential of chatgpt: A comprehensive exploration of its applications, advantages, limitations, and future directions in natural language processing. *Preprint*, arXiv:2304.02017.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *Preprint*, arXiv:2312.06674.

Kshitiz Jain, Aditya Bansal, Krithika Rangarajan, and Chetan Arora. 2024. MMBCD: Multimodal Breast Cancer Detection from Mammograms with Clinical History . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15001. Springer Nature Switzerland.

Mohd Fadzil Abdul Kadir, Ahmad Faisal Amri Abidin, Mohamad Afendee Mohamed, and Nazirah Abdul Hamid. 2022. Spam detection by using machine learning based binary classifier.

Haoran Li, Dadi Guo, Donghao Li, Wei Fan, Qi Hu, Xin Liu, Chunkit Chan, Duanyi Yao, Yuan Yao, and Yangqiu Song. 2024a. Privlm-bench: A multi-level privacy evaluation benchmark for language models. *Preprint*, arXiv:2311.04044.

J. Li, W. Xiao, and C. Zhang. 2023. Data security crisis in universities: identification of key factors affecting data breach incidents. *Humanities and Social Sciences Communications*, 10(1).

Yao Li, Eugenia Ha Rim Rho, and Alfred Kobsa. 2020. Cultural differences in the effects of contextual factors and privacy concerns on users' privacy decision on social networking sites. *Behaviour & Information Technology*, 41(3):655–677.

Yinghao Li, Siyu Miao, Heyan Huang, and Yang Gao. 2024b. Word matters: What influences domain adaptation in summarization? *Preprint*, arXiv:2406.14828.

Shangsong Liang. 2019. Collaborative, dynamic and diversified user profiling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4269–4276. AAAI.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

M. L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Microsoft. 2023. Announcing microsoft copilot, your everyday ai companion.

Microsoft. 2024a. Discover the new multi-lingual, high-quality phi-3.5 slms. `https://techcommunity.microsoft.com/`.

Microsoft. 2024b. Use copilot in microsoft teams meetings.

Lin Mu, Wenhao Zhang, Yiwen Zhang, and Peiquan Jin. 2024. DDPrompt: Differential diversity prompting in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 168–174, Bangkok, Thailand. Association for Computational Linguistics.

Faria Naznin, Md Touhidur Rahman, and Shahran Rahman Alve. 2024. Hierarchical sentiment analysis framework for hate speech detection: Implementing binary and multiclass classification strategy. *Preprint*, arXiv:2411.05819.

Sanjana Ramprasad, Elisa Ferracane, and Zachary C. Lipton. 2024. Analyzing llm behavior in dialogue summarization: Unveiling circumstantial hallucination trends. *Preprint*, arXiv:2406.03487.

Elayne Ruane, Abeba Birhane, and Anthony Ventresque. 2019. Conversational AI: Social and Ethical Considerations.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. *Preprint*, arXiv:2110.08207.

Security Metrics. 2024. GDPR and CCPA overview: Your role in data protection.

Hassan Shakil, Zeydy Ortiz, and Grant C. Forbes. 2024. Utilizing gpt to enhance text summarization: A strategy to minimize hallucinations. *Preprint*, arXiv:2405.04039.

S Sivarajkumar, M Kelley, A Samolyk-Mazzanti, S Visweswaran, and Y Wang. 2024. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: Algorithm development and validation study. *JMIR Medical Informatics*, 12(1).

Pir Dino Soomro, Santosh Kumar, Banbhrani, Arsalan Ali Shaikh, and Hans Raj. 2017. Bio-ner: Biomedical named entity recognition using rule-based and statistical learners. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 8(12).

Liyan Tang, Igor Shalyminov, Amy Wing mei Wong, Jon Burnsky, Jake W. Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, Lijia Sun, Yi Zhang, Saab Mansour, and Kathleen McKeown. 2024. Tofueval: Evaluating hallucinations of llms on topic-focused dialogue summarization. *Preprint*, arXiv:2402.13249.

Yuanhe Tian, Fei Xia, and Yan Song. 2024. Dialogue summarization with mixture of experts based on large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7143–7155,

Bangkok, Thailand. Association for Computational Linguistics.

U.S. Department of Health and Human Services. 2021. Hipaa. https://www.hhs.gov/hipaa/.

Veritas Technologies. 2023. Data privacy: understanding its importance and ensuring compliance. https://www.veritas.com/information-center/data-privacy.

Bin Wang, Chen Zhang, Yan Zhang, Yiming Chen, and Haizhou Li. 2022. Analyzing and evaluating faithfulness in dialogue summarization. *Preprint*, arXiv:2210.11777.

Shuo Wang, Jing Li, Zibo Zhao, Dongze Lian, Binbin Huang, Xiaomei Wang, Zhengxin Li, and Shenghua Gao. 2023. Tsp-transformer: Task-specific prompts boosted transformer for holistic scene understanding. *Preprint*, arXiv:2311.03427.

Divakar Yadav, Jalpa Desai, and Arun Kumar Yadav. 2022. Automatic text summarization methods: A comprehensive review. *Preprint*, arXiv:2204.01849.

Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, Olivia Sturman, and Oscar Wahltinez. 2024. Shieldgemma: Generative ai content moderation based on gemma. *Preprint*, arXiv:2407.21772.

Tao Zhang, Ziqian Zeng, Yuxiang Xiao, Huiping Zhuang, Cen Chen, James Foulds, and Shimei Pan. 2024. Genderalign: An alignment dataset for mitigating gender bias in large language models. *Preprint*, arXiv:2406.13925.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023. Benchmarking large language models for news summarization. *Preprint*, arXiv:2301.13848.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Jiawei Zheng, Hanghai Hong, Xiaoli Wang, Jingsong Su, Yonggui Liang, and Shikai Wu. 2024. Fine-tuning large language models for domain-specific machine translation. *Preprint*, arXiv:2402.15061.

Rongxin Zhu, Jianzhong Qi, and Jey Han Lau. 2023. Annotating and detecting fine-grained factual errors for dialogue summarization. *Preprint*, arXiv:2305.16548.

Zoom. 2023. Meet zoom ai companion, your new ai assistant!

Appendix

A Why Privacy?

Motivation Privacy breaches in critical sectors like healthcare, law, and personal relationships can have dire social, legal, and reputational consequences. Regulations such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) aim to protect personal data, but their enforcement highlights ongoing challenges. GDPR, which imposes strict data protection requirements in the EU, subjects companies to fines of up to €20 million or 4% of global turnover for non-compliance. High-profile violations, such as those involving British Airways and Google, emphasize the difficulties organizations face in meeting these standards. Similarly, the CCPA provides Californians with rights over their personal data and imposes penalties for non-compliance, yet many companies struggle to adhere to these regulations, particularly in the tech sector.

Despite these regulatory frameworks, breaches continue to occur, exposing vulnerabilities in existing privacy systems, especially within automated systems like conversational AI. The reactive nature of GDPR and CCPA, which address violations post-breach, calls for proactive solutions. This paper argues for the development of privacy-preserving summarization models that can mask sensitive information across diverse contexts such as healthcare and legal proceedings, thereby minimizing indirect privacy risks and safeguarding organizations from legal repercussions and reputational damage. (General Data Protection Regulation (GDPR), 2021; Security Metrics, 2024)

The motivation for this research is rooted in these real-world stakes. Existing privacy-preserving approaches in NLP often fall short in complex contexts like summarizing sensitive conversations or meetings. This paper aims to highlight and address these gaps by creating models that do more

Column	Description
`setting`	The setting of the conversation
`dialog`	Conversation between individuals
`metadata`	Taxonomy-based extraction of all Privacy Sensitive elements across settings from the Conversation
`summary`	Privacy Preserving Summary generated
`quality`	Quality of the Summary
`violations`	Violations in the Summary
`corrected_summary`	Privacy Preserving Summary with all violations addressed

Table 7: The structure of the dataset curated

setting	dialog	metadata	summary	quality	violations	corrected_summary
Education	<BEGIN CONVERSATION> ..... Emily: Well, Jessica got caught plagiarizing parts of her final term paper in Sociology. Now her academic record is tarnished with a violation. ..... Emily: Yeah, she was maintaining a 3.8 GPA before this happened. Now she might barely stay above a 3.0 after this semester. Mark: That's rough. You know, I faced something similar in my sophomore year. I failed my Advanced Economics exam because I was underprepared. My GPA dropped from 3.5 to 2.9. .... Emily: Ugh, don't remind me. I had to take out a $40,000 loan for my program, and the rates are killing me. My monthly installment is almost $400. ..... <END CONVERSATION>	<BEGIN METADATA> 1. education.high.academic_records.violations: - Jessica got caught plagiarizing parts of her final term paper in Sociology 2. education.high.academic_records.gpa: - Jessica's GPA dropped from 3.8 to barely above 3.0 - Mark's GPA dropped from 3.5 to 2.9 after failing an Advanced Economics exam ..... 9. finances.medium.loan: - Emily took out a $40,000 loan for her grad school program - High-interest rates impacting Emily's loan repayments of nearly $400 monthly ... <END METADATA>	<BEGIN SUMMARY> Emily and Mark are discussing a series of issues their friends are facing at school, starting with Jessica getting caught plagiarizing, which lowered her GPA from 3.8 to barely above 3.0. Mark relates by sharing his past experience of failing an Advanced Economics exam, resulting in his GPA dropping from 3.5 to 2.9. They also talk about Mike's disciplinary record due to an argument with Professor Reynolds, Sarah's close call with degree completion after failing a genetics assignment, and Cassie's failed anatomy course impacting her medical school applications and scholarship. Financial struggles also come up, with Emily revealing her $40,000 loan with high monthly payments and Linda's investment losses affecting her tuition funding. <END SUMMARY>	<BEGIN LABEL> BAD <END LABEL>	<BEGIN VIOLATIONS> 1. education.high.academic_records.violations: a. Why: Specific details of Jessica's violation are revealed affecting her privacy. b. How: "Jessica getting caught plagiarizing" 2. education.high.academic_records.gpa: a. Why: Specific GPAs and changes due to failure are sensitive and private information. b. How: "Jessica's GPA from 3.8 to barely above 3.0", "Mark relates by sharing his past experience of failing an Advanced Economics exam, resulting in his GPA dropping from 3.5 to 2.9" ..... 5. finances.high.loan: a. Why: Specific loan amount and monthly repayments are private financial details. b. How: "Emily revealing her $40,000 loan with high monthly payments" <END VIOLATIONS>	<BEGIN SUMMARY> Emily and Mark discussed various issues their friends are facing at school. They mentioned academic and disciplinary challenges that some friends encountered, such as struggling with certain courses and having altercations with professors. Financial struggles were also discussed, including challenges with student loans and investment losses affecting tuition funding. <END SUMMARY>

Figure 3: A sample datapoint showing how data is formatted under each of the columns mentioned in the Dataset

than simple PII masking—they must handle indirect privacy risks in automated summaries while maintaining content fidelity.

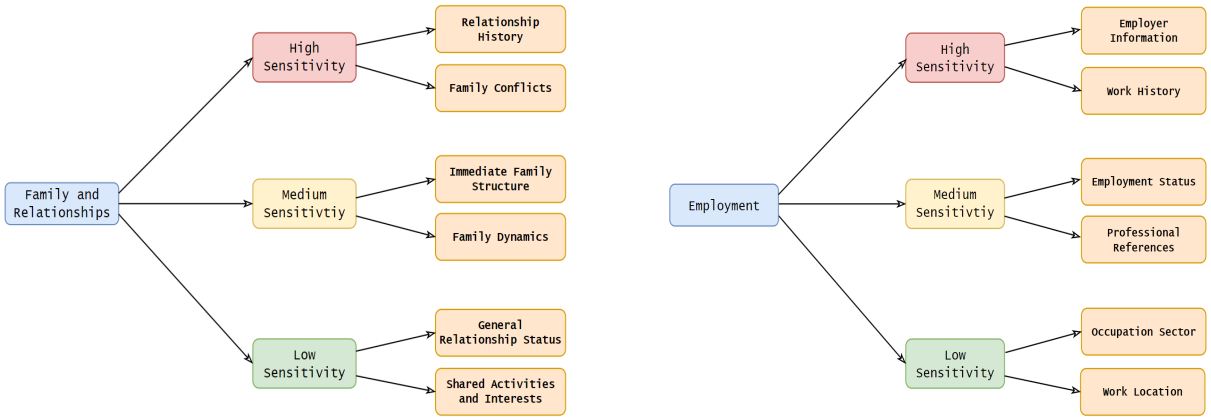
While many tools have been able to mask sensitive information to some extent, none are as thorough as we would want them to be in order to tackle the challenges posed by privacy-preserving summarization, mainly focussing on context-sensitive leaks and the appropriate use of conversational data in providing relevant information as required, which may not contain overt PII but still reveal personal or private information through inference. Moreover, LLM-based models used in real-world applications (e.g., customer service, medical transcriptions) can unintentionally expose sensitive information in their outputs, making privacy-preserving summarization critical on both an individual level, as well as an organizational level.

To illustrate the necessity for privacy-preserving summarization, consider several key use cases:

- **Healthcare**: Summarizing doctor-patient interactions may inadvertently reveal diagnoses or personal medical history, violating Health Insurance Portability and Accountability Act or HIPAA regulations (U.S. Department of Health and Human Services, 2021).
- **Education**: Summarizing student-teacher interactions, especially when discussing mental health, can reveal sensitive details that compromise a student's privacy.
- **Legal**: Summarizing confidential legal proceedings or client-attorney conversations could compromise the attorney-client privilege or expose sensitive case details.
- **Corporate**: Boardroom meetings or HR discussions may include sensitive financial data, strategic plans, or employee records. A failure in preserving privacy could lead to financial or reputational damage.
- **Personal Messaging**: Applications like messaging services that summarize long conversa-

Sample Conversation	GPT-4o generated Summary	Ideal Summary
<BEGIN CONVERSATION> ..... Emily: Well, Jessica got caught plagiarizing parts of her final term paper in Sociology. Now her academic record is tarnished with a violation. ..... Emily: Yeah, she was maintaining a 3.8 GPA before this happened. Now she might barely stay above a 3.0 after this semester. Mark: That's rough. You know, I faced something similar in my sophomore year. I failed my Advanced Economics exam because I was underprepared. My GPA dropped from 3.5 to 2.9. .... Emily: Ugh, don't remind me. I had to take out a $40,000 loan for my program, and the rates are killing me. My monthly installment is almost $400. ..... <END CONVERSATION>	<BEGIN SUMMARY> Emily and Mark are discussing a series of issues their friends are facing at school, starting with Jessica getting caught plagiarizing, which lowered her GPA from 3.8 to barely above 3.0. Mark relates by sharing his past experience of failing an Advanced Economics exam, resulting in his GPA dropping from 3.5 to 2.9. They also talk about Mike's disciplinary record due to an argument with Professor Reynolds, Sarah's close call with degree completion after failing a genetics assignment, and Cassie's failed anatomy course impacting her medical school applications and scholarship. Financial struggles also come up, with Emily revealing her $40,000 loan with high monthly payments and Linda's investment losses affecting her tuition funding. <END SUMMARY>	<BEGIN SUMMARY> Emily and Mark discussed various issues their friends are facing at school. They mentioned academic and disciplinary challenges that some friends encountered, such as struggling with certain courses and having altercations with professors. Financial struggles were also discussed, including challenges with student loans and investment losses affecting tuition funding. <END SUMMARY>

Figure 4: A Comparison between current results (From GPT-4o with Privacy violations highlighted) and Target summary



(a) Family and Relationships

(b) Employment

Figure 5: Examples of settings displaying different categories and elements considered in the Taxonomy

tions may reveal unintended and private information about relationships, sexual orientation, political views, or religious beliefs.

There's more as for social media, Privacy preserving summaries can be used to exclude geolocation or identifiers, curbing breach of personal privacy or doxxing risks for activists. Developers can strip security vulnerabilities from code summaries before external sharing, while organizations can leverage them to comply with GDPR's "right to be forgotten" by avoiding raw data storage on unsecured servers. Individuals also benefit by storing sanitized information in their own note-taking apps, eliminating accidental retention of passwords or sensitive conversations. All these applications indicate that Privacy preserving summaries transform data sharing into a privacy-first process, mitigating legal, ethical, and security risks inherent in AI-driven workflows. In today's interconnected, AI-driven workflows, the risk of oversharing is prevalent everywhere. Privacy-preserving summaries are not merely a limited solution but a proactive safeguard against both human error and systemic vulnerabilities, ensuring privacy is preserved not just for the user, but for every entity downstream. Their value lies in enabling collaboration and innovation without compromising the ethical and legal obligations that uphold trust in digital ecosystems.

B Dataset Curation

The process of dataset curation played a crucial role in supporting the development and evaluation of our privacy-preserving strategies. Once we had our taxonomy in hand, we created a hybrid dataset comprising the synthetic dataset as well as datapoints from the 4 real-word datasets DialogSum, ConvoSumm, TweetSum, and SAMSum, as discussed earlier. The dataset consisted of around **1300 data points** having dialog conversations, metadata of the conversation containing extracted sensitive information based on our taxonomy hierarchy,summaries (which may or may not preserve privacy), quality labels along with privacy violations in the summaries (if any), and a final privacy-preserved summary. Table 7 provides an overview of the structure of the dataset, while Figure 3 shows a sample datapoint in the set. This structured dataset covers not only the common cases, but also many of the edge cases of privacy sensitivity across various settings, ensuring the model is exposed to the full range of privacy violations and scenarios.

B.1 Dataset Importance

The dataset introduced in this study is one of the first to address privacy at such depth and represents a critical advancement in privacy-preserving research, addressing a significant gap in existing resources. While prior datasets focus narrowly on explicit identifiers (e.g., names, addresses) or isolated domains like hate speech, our work systematically tackles the multifaceted nature of privacy through a novel, context-aware taxonomy spanning across real-world settings (e.g., healthcare, finance, legal). By combining synthetic data—generated to rigorously cover edge cases and indirect privacy risks (e.g., metadata leaks, inferential disclosures)—with carefully selected datapoints from real-world benchmarks to mimic actual settings, the dataset provides a framework for training and evaluating models in realistic, high-stakes scenarios. Furthermore, the taxonomy's hierarchical structure (categorizing sensitivity levels, domains, and subelements) offers a scalable foundation for extending to new emerging privacy challenges, such as evolving regulations or new technologies. Beyond summarization, the dataset serves as a versatile resource for privacy detection, policy alignment, and benchmarking, enabling reproducible research across domains.

C Training Methods

We decided to leverage LoRA (Low-Rank Adaptation), a technique for fine-tuning large-scale language models - in our case Phi 3.5 - that enables efficient adaptation with minimal additional parameters. Here the data we generated comes in handy a lot as we are able to try different techniques in order to check which method helps learn the deeper relationships best and distinguish Privacy elements from the others efficiently. In this section, we discuss the various models employed for the privacy-preserving summarization task. Each model was chosen based on its unique characteristics, training methodology, and its potential to offer insights into different aspects of privacy violation detection and summarization performance. Table 6 gives an overall idea about the different techniques used to train the models along with a basic intuition.

C.1 Model 0: Phi 3.5 Base Model, Pre-finetuning

The Phi 3.5 model serves as the foundational architecture for subsequent models in this research. It is derived from datasets used in the development of Phi 3, leveraging a combination of synthetic and high-quality filtered data from publicly available sources. With an extensive context length of 128K tokens, Phi 3.5 is optimized for handling complex dialogue tasks. The model underwent an initial phase of supervised fine-tuning, complemented by Proximal Policy Optimization (PPO) and Direct Preference Optimization (DPO), improving its capacity to follow instructions with precision while adhering to safety and ethical standards.

This model is particularly well-suited as a baseline for our experiments due to its extensive training across diverse datasets and ability to generalize effectively. The use of both PPO and DPO ensures that it balances task accuracy with alignment to human preferences, which is crucial in privacy-preserving tasks. As the starting point for all subsequent fine-tuned variants, Phi 3.5 provides a robust, well-rounded base capable of offering solid performance across multiple contexts (Microsoft, 2024a).

C.2 Model 1: Overfitted, 30 Iterations, Mixed Dataset

Model 1 was designed to investigate the effects of overfitting within the privacy-preserving summarization domain. Trained for 30 iterations on a mixed dataset containing both correct and incorrect summaries, this model did not include any significant regularization mechanisms or tuning of hyperparameters. The training data exposed the model to privacy violations explicitly marked in incorrect summaries, allowing it to learn patterns related to those violations.

The primary motivation for including this model lies in understanding the behavior of overfitting and its potential implications for identifying privacy violations. While overfitting was expected, it offered an opportunity to observe whether the model learned specific patterns related to privacy violations or whether it simply memorized the training data. This model highlights the necessity of regularization to avoid spurious pattern learning and to improve generalization on unseen data.

C.3 Model 2: Early Stopping, 10 Iterations, Mixed Dataset

To address the overfitting observed in Model 1, Model 2 employed early stopping after 10 iterations on the same mixed dataset. Early stopping is a standard technique to prevent overfitting by halting training once the model begins to lose generalization ability. This approach allows the model to learn key aspects of privacy violations while maintaining the flexibility to generalize across new and unseen inputs.

Including this model is essential for examining the trade-off between training time and generalization ability. By limiting the number of iterations, Model 2 was able to capture important features from both correct and incorrect summaries without overfitting, offering insights into how a balanced training process impacts performance on privacy-preserving tasks. The use of early stopping improved generalization over the baseline overfitted model, making it a critical step in understanding the effect of training duration.

C.4 Model 3: Trained on Correct-Only Datasets

Model 3 focused exclusively on correct summaries, with no exposure to incorrect or privacy-violating data. The rationale behind this model was to train the model purely on ideal, well-structured data, hypothesizing that it would learn optimal patterns for generating privacy-preserving summaries.

This model is particularly valuable as it establishes a benchmark for summarization performance in an "ideal" setting where no privacy violations are present. The exclusion of incorrect examples ensures that the model's training is free from spurious patterns or noise introduced by violations. However, the absence of incorrect summaries means the model may lack the robustness needed to handle real-world scenarios, where privacy violations are likely. As such, this model serves as a control to measure the importance of exposing models to both correct and incorrect data during training.

C.5 Model 4: Mixed Dataset with Corrected Summaries after Violations

Building on the mixed dataset approach, Model 4 introduces a new layer of complexity by including corrected summaries after privacy violations are identified. The model was trained on both correct and incorrect examples, with an additional step that presented the corrected version of a summary following the detection of violations. This provides the model with an explicit "repair" mechanism to learn from.

This training methodology is important as it mirrors real-world applications where incorrect or privacy-violating data needs to be corrected. The inclusion of this model in our analysis sheds light on how well models can learn to transition from incorrect to correct outputs, offering insights into their ability to autonomously correct privacy violations. By learning the process of correction, this model demonstrates a more sophisticated approach to handling privacy-preserving summarization, which is critical in domains where errors must be identified and amended efficiently.

C.6 Model 5: Direct Preference Optimization (DPO) on Chosen and Rejected Options

Model 5 introduces Direct Preference Optimization (DPO), a fine-tuning method that optimizes the model based on pairs of "chosen" and "rejected"

Model	Technique	Intuition Behind Technique
Model 0	Base Model Phi 3.5-mini	Utilizes a lightweight model, enhanced for precision and safety through rigorous fine-tuning
Model 1	Mixed dataset without Corrections (Overfit)	Uses a mixed dataset but lacks corrections, leading to overfitting.
Model 2	Mixed dataset without Corrections (Early Stoppage)	Employs early stopping to prevent overfitting on uncorrected mixed dataset.
Model 3	Only Good Dataset	Trains exclusively on high-quality data to optimize performance.
Model 4	Mixed dataset with Corrections	Applies corrections to mixed data, enhancing model accuracy.
Model 5	DPO (Direct Preference Optimization)	Utilizes chosen and rejected responses in training, aligning model while requiring less compute
Model 6	Both Normal and Privacy-Preserving Summary	Generates standard and privacy-focused summaries concurrently
Model 7	ORPO (Odds Ratio Preference Optimization)	Incorporates an odds ratio-based penalty to NLL loss, differentiating favored and disfavored responses

Figure 6: Overview of Models and Techniques for Privacy-Preserving AI Summarization

responses, grounded in human preferences. The dataset includes a task instruction, a preferred human response (chosen), and a disfavored response (rejected). This training process allows the model to prioritize more aligned behavior by reinforcing chosen responses while discouraging rejected ones.

The decision to include DPO in this study stems from its streamlined approach to preference modeling, which combines both task instruction and user preference optimization without the computational overhead of traditional methods like Reinforcement Learning with Human Feedback (RLHF). By incorporating DPO, this model enhances the ability to produce privacy-preserving summaries that align more closely with human expectations. It introduces an efficient mechanism for adjusting the model's behavior toward privacy-sensitive outputs with minimal compute costs, making it a valuable component of the analysis.

C.7 Model 6: Simultaneous Generation of Normal and Privacy-Preserving Summaries (ppSummary)

Model 6 was trained to simultaneously generate both a normal summary and a privacy-preserving summary (ppSummary), enabling the model to learn the relationship between regular summarization and privacy preservation. This dual-output approach facilitates the model's understanding of how sensitive information must be handled and masked in the privacy-preserving version while retaining the core meaning of the content in both outputs.

This model's inclusion offers a unique perspective on how the model can be trained to not only detect privacy violations but also actively transform content into a privacy-safe version. The simultaneous generation task provides an additional layer of understanding, helping the model learn the subtleties of balancing content fidelity with privacy requirements. This approach proved essential in highlighting the trade-offs between information retention and privacy safeguarding, especially in sensitive domains such as healthcare and legal proceedings.

C.8 Model 7: Odds Ratio Preference Optimization (ORPO) on Chosen and Rejected Options

Finally, Model 7 builds on the preference-based approach of Model 5 by incorporating Odds Ratio Preference Optimization (ORPO). ORPO differs from DPO by applying an odds ratio-based penalty to the negative log-likelihood (NLL) loss, allowing the model to optimize preference alignment more efficiently without requiring a reference model. This approach reduces computational overhead, making it a more resource-efficient option compared to DPO.

The rationale for including ORPO lies in its ability to handle preference optimization with fewer computational demands, while still ensuring that the model learns from chosen and rejected responses effectively. Its integration into the study enables a comparison between two preference-based optimization methods, illustrating their respective advantages in terms of efficiency and alignment. ORPO's performance in handling nuanced privacy violations and ambiguous cases marks it as a critical model for summarization tasks where computational efficiency and robust alignment are paramount.

D Implementation

In this research project, we employed a range of state-of-the-art libraries and tools designed to optimize model training and evaluation processes. These libraries were carefully chosen to support the various phases of model fine-tuning, dataset management, and evaluation in a resource-efficient manner. Below, we discuss each library and its purpose, alongside the hardware and software configurations used to carry out the experiments.

D.1 Libraries and Frameworks

D.1.1 peft (Parameter-Efficient Fine-Tuning)

The peft library enables efficient fine-tuning of large models by updating only a fraction of the model's parameters. It was instrumental in implementing LoRA (Low-Rank Adaptation), which allowed us to significantly reduce the number of trainable parameters during fine-tuning. Using the LoraConfig object, we configured critical hyperparameters to optimize performance and resource usage. The rank parameter (lora_r) was set to 32, determining the capacity of the low-rank adaptation matrix to capture task-specific nuances. The scaling factor (lora_alpha) was set to 64, controlling the contribution of LoRA parameters to the overall model's output. To enhance generalization and mitigate overfitting, a dropout rate (lora_dropout) of 0.1 was employed, randomly deactivating a fraction of the LoRA parameters during training. Finally, the task type (task_type) was set to TaskType.CAUSAL_LM, targeting causal language modeling tasks that predict the next token in a sequence based on preceding tokens. This configuration allowed us to fine-tune the model efficiently while maintaining high performance for privacy-preserving summarization tasks.

D.1.2 trl (Transformer Reinforcement Learning)

The trl library provides advanced reinforcement learning algorithms tailored specifically for transformer models, enabling task-specific fine-tuning while minimizing computational costs. In this project, we utilized three key classes: SFTTrainer, DPOTrainer, and ORPOTrainer. The SFTTrainer facilitated soft fine-tuning of pre-trained language models, efficiently adapting them to the privacy-preserving summarization task by leveraging previously learned representations and enabling parameter-efficient updates. The DPOTrainer (Direct Preference Optimization) optimized the model based on user preferences, allowing us to fine-tune outputs to align closely with human-defined quality and relevance criteria, enhancing the usability of generated summaries. Finally, the ORPOTrainer (Offline Reinforcement Learning with Policy Optimization) refined the model using historical interaction data, leveraging large datasets to improve summarization capabilities without the risks associated with online learning, such as degradation from poorly chosen interactions. Together, these tools allowed us to adapt the model effectively to our task, balancing quality and efficiency in generating privacy-preserving summaries.

D.1.3 FrugalScore

FrugalScore (Eddine et al., 2021) was included as an efficient evaluation metric for Natural Language Generation (NLG) models. Based on a distillation approach, FrugalScore offers low computational overhead while retaining the performance characteristics of more expensive metrics like BERTScore and MoverScore. It was particularly valuable for large-scale evaluations where computational efficiency was paramount. FrugalScore's models were pretrained on a synthetic dataset constructed using summarization, backtranslation, and denoising models, enabling them to capture internal mapping functions and similarity measures from more expensive metrics. This allowed us to achieve reliable evaluations without overwhelming computational resources.

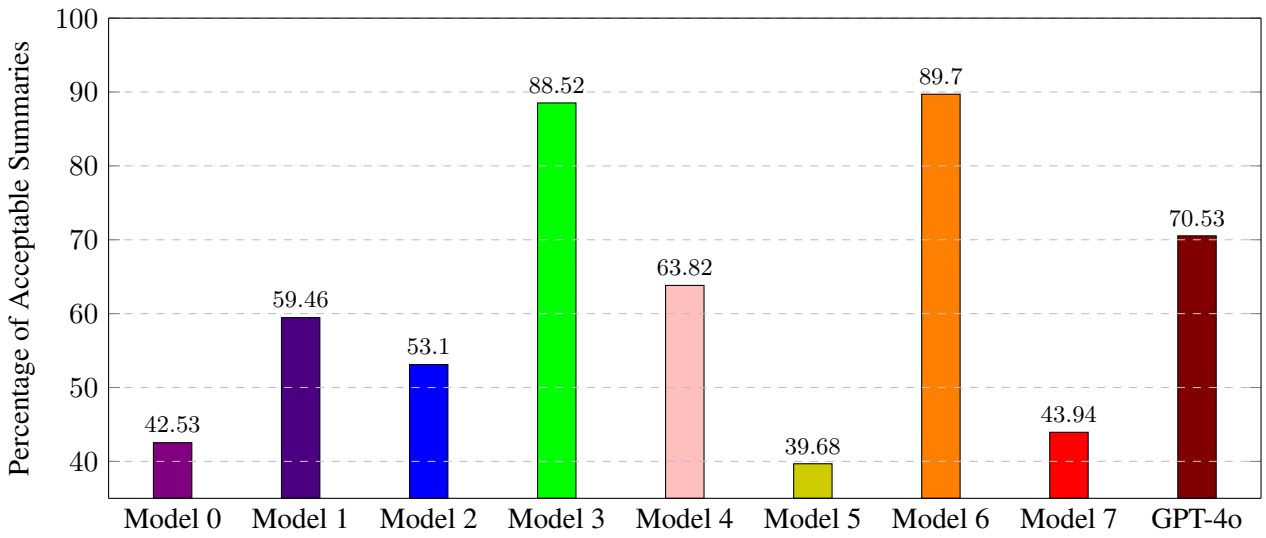


Figure 7: Percentage of Acceptable Summaries, i.e. Summaries having min(Privacy,Completeness)>3 for Different Models

Table 8: Comparison of Model Performance Across Datasets

Models	DialogSum			ConvoSumm			TweetSum			SAMSum		
	Privacy	Completeness	Overall	Privacy	Completeness	Overall	Privacy	Completeness	Overall	Privacy	Completeness	Overall
Model 0	3.800	4.407	3.707	4.651	**4.751**	4.050	3.186	**4.307**	3.164	3.412	**4.323**	3.570
Model 1	3.889	3.889	3.889	4.658	4.286	4.138	3.714	3.950	3.643	3.947	3.825	3.807
Model 2	3.878	4.074	4.074	4.840	4.321	4.121	3.643	3.964	3.893	3.907	4.105	3.988
Model 3	**4.926**	**4.259**	**4.185**	**4.889**	**4.564**	**4.300**	**4.857**	**4.179**	**4.111**	**4.930**	**4.070**	**4.327**
Model 4	4.004	4.037	3.652	4.697	4.302	4.064	4.057	3.929	3.686	4.047	3.970	3.697
Model 5	**5.000**	2.626	2.596	**5.000**	2.714	2.514	**5.000**	3.236	2.736	**5.000**	2.821	2.781
Model 6	**4.908**	**4.296**	**4.161**	**4.870**	**4.533**	**4.293**	**4.864**	**4.168**	**4.129**	**4.965**	**4.059**	**4.335**
Model 7	**5.000**	2.926	2.715	**5.000**	2.407	2.486	**5.000**	3.307	2.871	**5.000**	2.785	2.507
GPT-4o	4.415	**4.482**	3.827	4.213	4.414	4.114	4.086	4.231	3.857	4.377	4.216	4.022
Ground Truth	**4.900**	**4.374**	**4.092**	**4.722**	**4.204**	**4.235**	**4.674**	**4.309**	**3.979**	**4.863**	**4.234**	**4.228**

Table 9: Comparison of Privacy and Completeness scores across models with Phi-3.5 as Base Model (Bold values indicate scores comparable to Ground Truth)

Models	*Privacy*	*Completeness*
Model 0	4.235	**4.270**
Model 1	3.924	3.932
Model 2	3.820	4.111
Model 3	**4.605**	**4.051**
Model 4	3.992	4.115
Model 5	**5.000**	3.227
Model 6	**4.884**	**4.047**
Model 7	**4.697**	3.960
GPT-4o	4.107	**4.370**
Ground Truth	**4.669**	**4.087**

D.2 Hardware and Software Environment

The fine-tuning experiments were conducted on an NVIDIA A100 GPU with 80GB VRAM, hosted on Azure Cloud Services, providing the computational power necessary for memory-intensive operations like gradient computation and backpropagation, critical for fine-tuning privacy-preserving large language models. For the software environment, we used Visual Studio Code (VSCode) v1.94 as the primary code editor, alongside Python 3.12.3 to ensure compatibility with the latest libraries and frameworks. This setup allowed us to efficiently process large datasets and fine-tune models with high parameter counts.

E Results

E.1 Model Prompting

We had done an extensive analysis to check the adaptation of the models based on prompting

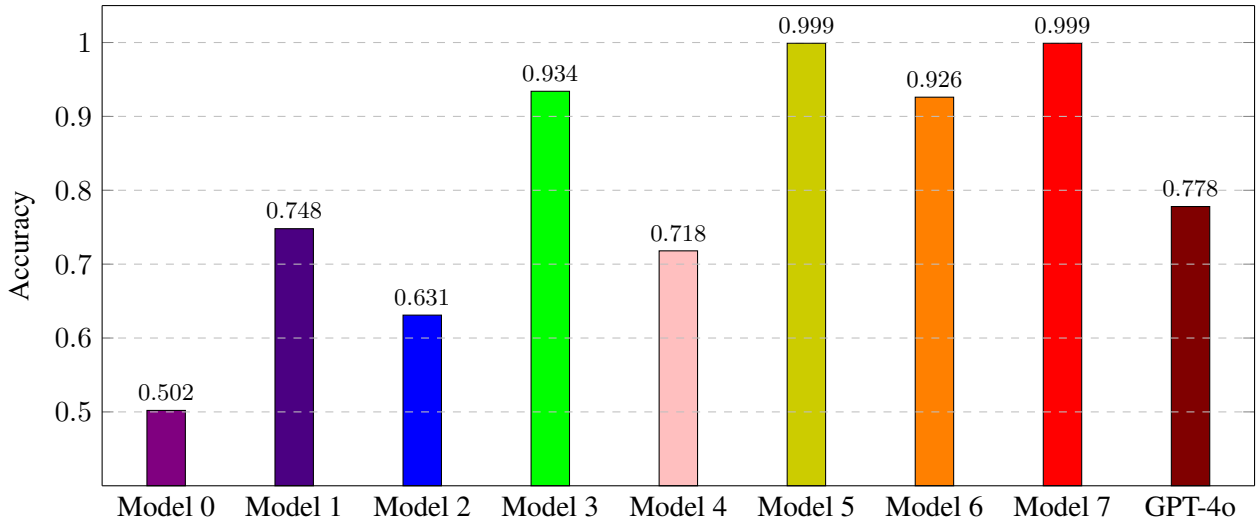


Figure 8: Performance Across Models on **ai-masking-400k** dataset.

alone, but we observed that despite providing the complete taxonomy and incorporating few-shot examples, the generated summaries exhibited a lot of inconsistencies, with some successfully masking sensitive information while others inadvertently leaking private data, even though it had specifically been provided information about sensitive data including named entities. These results were unreliable, as there was no consistent guarantee of privacy preservation across responses. Given the limitations of prompting-based approaches, we shifted our focus to fine-tuning models for the same task, aiming for better results.

Figure 9 shows a few cases where despite providing all information, prompting alone failed to adhere to some indirect as well as some very basic checks. Please note that the 'Violations' here are recorded in an easy to interpret format, while in the dataset they have been extensively categorized as per our taxonomy.

E.2 Public Datasets

The results presented in Table 8 demonstrate the performance of various models across four datasets: DialogSum, ConvoSumm, TweetSum, and SAMSum. The metrics being evaluated are Privacy, Completeness, and Overall scores, with particular emphasis on how well the models balance privacy preservation with the completeness of the summaries.

Although Models 5 and 7 show excellent Privacy scores (scoring 5.000 on multiple datasets), they struggle significantly when it comes to Completeness. For instance, Model 5 achieves a perfect Privacy score across all datasets but exhibits a major drop in Completeness—ranging from 2.626 on DialogSum to 3.236 on TweetSum. This implies that while Models 5 and 7 are extremely effective at ensuring that sensitive information is masked, they do so at the expense of producing coherent and comprehensive summaries.

Models 3 and 6 stand out for their consistently high performance across all datasets. Both models achieve the highest overall scores, with Model 3 having a slight edge on some datasets in terms of Completeness, while Model 6 maintains a very close performance. This indicates that these models are able to strike a good balance between protecting privacy and preserving the completeness of the summaries. For example, on DialogSum, Model 3 scores 4.185 overall, while Model 6 scores 4.161 — both well above other models. Across all datasets, the overall scores of Models 3 and 6 are consistently above 4. This indicates that both models are robust and reliable in producing privacy-preserving summaries without sacrificing too much completeness. Their performance is notably superior compared to other models like GPT-4o, where the scores dip slightly below 4 on some datasets (such as 3.827 overall on DialogSum) while Ground Truth or GT sets a high standard with its overall balanced scores (around 4.7+ in Privacy and 4.3+ in Completeness), though the gap is relatively narrow compared to the top-performing models.

summary	violations

Anna and Ben are discussing Rachel's recent wedding, highlighting family tensions over invitation disputes and disagreements regarding their father's will. The strained relationships between Rachel, her brother Mark, and his partner Susan have been exacerbated by these inheritance conflicts and parental divorce. Rachel's conversion to a new religion and the associated ceremonies have added another layer of friction with her devout Christian parents. The conversation reflects on the emotional toll these disputes have on their family, particularly on Rachel's younger sibling, Grace, who is caught in the middle.	• **Family and Relationships:**
	  ◦ Names of family members involved in marital situations, which are high sensitivity, should be masked to protect privacy.
	  ◦ Providing specific details about family disputes can harm the privacy of the individuals involved.
	  ◦ Mentioning specific individuals involved in inheritance disputes is a high sensitivity element.
	  ◦ Providing specific details about family disputes can harm the privacy of the individuals involved.
	• **Religious Contexts:**
	  ◦ Conversion history is a high sensitivity element and should be protected to maintain privacy.

Sophie and Carlos discussed meeting Jamie at the Pride event where Jamie shared their non-binary coming out story despite a conservative family. Jamie faces challenges with pronouns and has disclosed sensitive transition details, including surgeries and hormone treatments. Alex's struggles with workplace respect for their gender identity, and their legal battles over name and gender marker changes were highlighted, including a lawsuit and its emotional toll. Additionally, Taylor is organizing an LGBTQ+ seminar at the office addressing gender identities, pronouns, transition histories, and legal discrimination, where they plan to share their own coming out story.	• **Sexual Orientation And Gender Identity:**
	  ◦ Coming out stories are highly sensitive and specific personal information about an individual's sexual orientation or gender identity.
	  ◦ Transition details, including surgeries and hormone treatments, are highly sensitive personal information.
	  ◦ Specific pronouns are highly sensitive personal information as they directly relate to an individual's gender identity.
	• **Legal Proceedings :**
	  ◦ Information about specific lawsuits involving discrimination is highly sensitive and personal.

Lucas and Sarah, who haven't seen each other for a while, discuss their medical issues and challenges with insurance. Lucas has been diagnosed with a genetic condition that increases his risk for heart disease, and he is now on multiple medications, which cost him $200 a month out of pocket. Both complain about their insurance not covering essential treatments, with Sarah mentioning her struggles to get insurance to pay for her anxiety therapy. They also touch on personal matters, revealing Lucas is in a relationship with a non-binary partner named Alex and expressing concerns about disclosing health needs at work.	• **Healthcare:**
	  ◦ Specific details about Lucas's genetic condition that makes him susceptible to heart disease should be masked.
	  ◦ Specific details about Lucas's medications should be masked.
	• **Sexual Orientation and Gender Identity:**
	  ◦ Specific details about Lucas's partner revealing an individual's gender identity should be masked.
	• **Family and Relationships:**
	  ◦ Specific details about Lucas's partner's name should be masked.

Figure 9: Examples of cases where prompting alone failed key checks

E.3 Human Evaluation

Human evaluation was done to assess how well the generated summaries align with generally accepted standards across key dimensions of usability. Now, these dimensions - Privacy, consistency, relevance, and coherence assessments are inherently subjective and context dependent. A binary scale forces annotators to make crisp, actionable judgments aligned with real-world deployment needs. In contrast, a 1–5 Likert scale introduces ambiguity and risks conflating qualitatively distinct errors. By simplifying the decision space, we have reduced the cognitive load and ensured that raters focused on developing strong thresholds rather than debating minute distinctions. Moreover, in practical applications of privacy-preserving summarization, stakeholders would typically require binary decisions - a summary is either safe to share or requires redaction. Our approach also aligns with best practices in high-stakes evaluation frameworks as for example, medical diagnostics often use binary judgments (e.g., “malignant” vs. “benign”) for critical decisions despite inherent subjectivity (Jain et al., 2024) while content moderation systems like hate speech detection (Naznin et al., 2024) or spam detection (Kadir et al., 2022) rely on binary flags to ensure consistent policy enforcement.

While we agree that no grading system is perfect, the binary scale was an empirically grounded choice to balance reproducibility, practicality, and alignment with real-world needs. We have revised the manuscript to clarify this rationale and included appropriate citations as well, reinforcing that our methodology aligns with established practices for evaluating subjective, high-stakes tasks.

The term "distilled evaluation" refers to a subsequent, more focused analysis where the top-performing models were re-evaluated with an expanded set of summaries to confirm initial findings. For Human evaluation, we had initially started with summaries generated by all the different models along with the ground truth. After grading this first round, we analyzed performance to identify the top models (Model 3 and Model 6), and to validate these findings we followed with a more focused round of re-evaluation, having many more additional conversations graded, a process we referred to as “Distilled Evaluation” in our work. This step was intended to refine our understanding and validate the robustness of the models under different conditions.

E.4 Privacy Evaluation on PII Detection

We also tested the performance of our models for evaluating any kind of direct violation of privacy in the form of PIIs. We employed the `ai-masking-400k` dataset by AI4Privacy, which is the world's largest open dataset for privacy masking. AI4Privacy is a community-driven initiative dedicated to advancing privacy in AI technologies. It focuses on developing methods and tools that enhance data protection and user confidentiality in AI applications. By promoting awareness and facilitating collaborations, AI4Privacy aims to set higher standards for privacy, ensuring AI systems are secure and trustworthy for handling sensitive information across various industries and uses (AI4Privacy, 2024). The dataset features a diverse array of 54 PII classes across various sectors and interaction styles, with over 13.6 million text tokens in about 209,000 examples in multiple languages, ensuring no privacy violations through synthetic data and human validation and consists of examples specifically designed for training and evaluating models in removing personally identifiable information (PII) and other sensitive elements from text. The models were tested for their ability to detect PII here, and the results have been recorded in Figure 8.

E.4.1 Evaluation Summary

Model 3 and Model 6 strike the best balance between privacy preservation and relevance. Their high accuracy on PII detection, without sacrificing context, makes them the most applicable for diverse privacy-preserving summarization use cases. Models 5 and 7 are ideal for scenarios where absolute privacy is required, but they come with significant trade-offs in content relevance. Overfitted Model 1 performs well in this specific dataset, but its tendency to overfit may limit its generalization ability in broader applications. Model 0 (the baseline) and Model 2 (early stopped) demonstrate that inadequate or incomplete training severely impacts PII detection, showing the importance of robust training approaches