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. 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 - a type of Text summarization that distills a dialogue into a concise and complete summary for an immediate understanding of the conversation - 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. The General Data Protection Regulation (GDPR), California Consumer Privacy Act (CCPA), and Health Insurance Portability and Accountability Act (HIPAA) each protect personal data in different contexts with

GDPR in the EU imposing strict fines for non-compliance, CCPA giving Californians rights over their data, and HIPAA protecting health information with severe penalties for breaches (General Data Protection Regulation (GDPR), 2021; Security Metrics, 2024; U.S. Department of Health and Human Services, 2021). Despite these frameworks, privacy breaches persist, which together with all the studies above paint a clear picture that safeguarding privacy has never been more critical with the need for effective management of data and adherence to privacy regulations being essential to mitigate risks and ensure the ethical use of AI technologies in today's data-driven world. Figure 5 gives us an idea of the summaries generated by the current baselines for a given conversation, along with what we would like our target summary to ideally be.

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 to detect and mask PII consistently but may lack flexibility when dealing with new types of sensitive data (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)

The works of Zhang et al. (2024) and Fu et al. (2024) introduced datasets and models aimed at addressing biases in LLMs— the former focusing on gender bias mitigation and the latter on integrating touch into multimodal generative models, each showing improved performance over existing models. Inspired by their methodologies our proposal seeks to develop a hybrid approach that combines the deterministic structure of rule-based systems with the contextual adaptability of learning-based methods. Their work informed our approach to train LLMs upon a dataset that captures various privacy breaches across diverse settings, enabling our models to understand and prefer privacy-preserving responses.

The main contributions of our work include:

- Built a comprehensive taxonomy for effectively recognizing sensitive information across settings like healthcare, legal, and finances and applying appropriate privacy measures
- Curated high-quality datasets spanning wide range of scenarios and levels of sensitivity, and trained models to recognize context-specific privacy concerns and relevant data elements
- Evaluated model outputs using LLM-as-a-judge, NLP metrics, and human evaluations to ensure high-quality responses adhering to both privacy and completeness standards

Figure 3 gives an overview of the systematic approach used to generate and verify privacy-preserving summaries in our research.

2 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.

Handling Hallucinations Ramprasad et al. (2024) addressed hallucinations in LLM-based dialogue summarization, focusing on reducing errors like "Circumstantial Inference" through human annotations and algorithmic adjustments, improving factual consistency. (Tian et al., 2024) introduced a Mixture of Experts (MoEs) approach to enhance summary accuracy by combining specialized models, while (Tang et al., 2024) developed the TOFUEVAL framework to evaluate factual accuracy and mitigate hallucinations. However, a glaring gap in these works is their limited focus on privacy, particularly the handling of sensitive information within dialogues. The lack of a structured approach to manage privacy-sensitive elements within dialogues underscores the need for compliance with a comprehensive privacy taxonomy.

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 not only focusing on sensitivity and context, incorporating diverse, real-world scenarios to train robust models effectively handle harmful content while addressing complex privacy challenges, but also proposing a taxonomy that can easily be adapted to new settings as they arise, backed by strong results. By prioritizing both utility and privacy, our work aims to fills a critical gap in this field, setting a new standard for privacy-preserving AI technologies.

3 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 multi-faceted 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 1 and 2) . 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 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

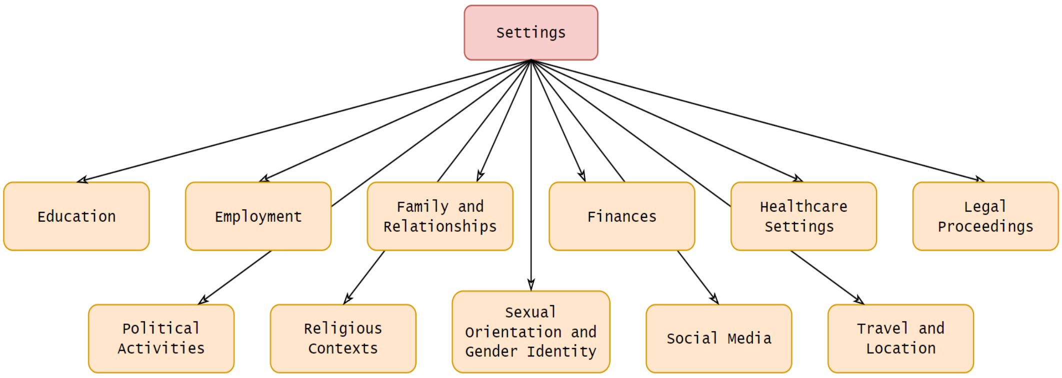
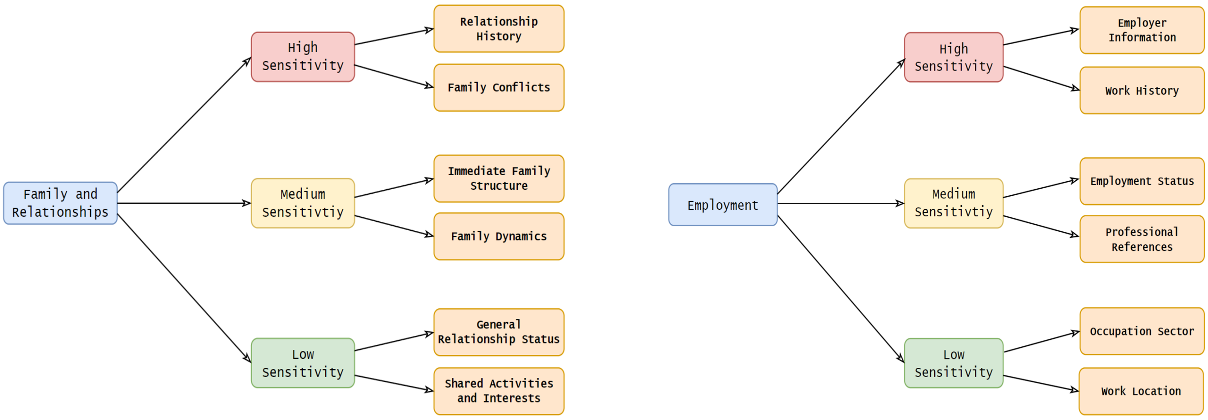


Figure 1: An overview of the Taxonomy showing the different Settings considered



(a) Family and Relationships

(b) Employment

Figure 2: Examples of Settings displaying different categories and elements considered in the Taxonomy

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 2 (b) would be represented as employment.high_sensitivity.work_history. While a strict demarcation isn't possible, 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 100% privacy masking but to balance it with completeness, ensuring that all the necessary information is delivered without significant leakage of personal or sensitive data, adhering to accepted privacy standards overall.

4 Dataset Curation

While existing datasets offer valuable insight, they often focus on narrow aspects like hate speech or explicit identifiers but in real-world applications privacy violations extend beyond these where many datasets fail to capture indirect privacy risks, such as inferences or metadata, which are crucial in domains like healthcare, legal, and financial settings. Our curated dataset addresses this need by covering the spectrum of privacy violations, both explicit and subtle, supporting enhanced privacy-preserving techniques. The necessity of our dataset also stems from the need to train models capable of recognizing diverse privacy violations across different contexts. We generated around 1100 synthetic datapoints using GPT-4o for the Data generation process in our work, with each Datapoint consisting of six columns: "setting" for identifying the Major setting of the conversation, "dialog" for the actual generated conversations, "metadata" with extracted information mapped to different privacy categories, "summary" for best privacy-preserving summary generated using GPT-4o, "label" and "violations" for evaluations of adherence to privacy standards, labelling the quality of summary and associated violations mapped according to privacy categories, and "corrected_summary" for revised summaries addressing identified privacy violations.

The process was carried out in 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, the fourth step involved identifying privacy violations. Here as input we included the Summary and the Metadata generated above and asked GPT-4o to compare and check if any of the sensitive information included in the metadata is leaked into the Summary. Each summary was graded as "GOOD" or "BAD," depending on its adherence to privacy standards, ensuring quality control. In case of minor, low sensitivity or no violations, it was labeled as "GOOD", otherwise "BAD" with all the violations output in the response in the manner it is presented 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 in the generation process, we manually verified around 30 initial datapoints and used them in the subsequent GPT-4o calls for each setting, providing a few examples of similar verified tasks from the datapoints to leverage In-Context Learning (ICL) and generate better datapoints. For broad coverage and connectivity to real world data, we then combined the synthetic data generated with a few existing benchmark datasets for Text Summarization - `DialogSum` (Chen et al., 2021), `SAMSUM` (Gliwa et al., 2019), `ConvoSumm` (Fabbri et al., 2021) and `TweetSum` (Feigenblat et al., 2021). About 50 data points each from these public datasets were used alongside synthetic data, with a different split of each being used as part of training and testing sets. The final dataset comprised around 1300 data points, split into approximately 1065 for training and 235 for testing. Given that most publicly available datasets lack indirect privacy annotations or specialized data for specific sectors, our approach blends an all round synthetic dataset with real-world data from popular datasets, ensuring that the model encounters a balanced and comprehensive coverage of privacy scenario, improving its generalization capability.

5 Experiments

Model Fine-tuning For testing our dataset in the privacy-preserving summarization task, we trained seven LoRA-based fine-tuned models using different techniques, each having Phi 3.5-mini as the base model, which will be considered as **Model 0**. Phi 3.5-mini was chosen for its extensive context length (128K tokens), ability to handle complex dialogue tasks, and robust training through supervised fine-tuning, making it a well-rounded baseline for privacy-related tasks. **Model 1** explored overfitting by training for 30 iterations on a mixed dataset containing both correct and incorrect summaries. This helped us understand the impact of overfitting on privacy violation detection. To mitigate overfitting, **Model 2** employed early stopping after 10 iterations on the same dataset, ensuring generalization without trading the learning of key features. **Model 3** was trained exclusively on correct summaries to serve as a benchmark for ideal conditions, although it didn't take into consideration real-world scenarios for dealing with potential privacy leaks explicitly. **Model 4** extended the mixed dataset approach by including corrected summaries after privacy violations, allowing the model to learn correction mechanisms critical for

Table 1: Scores for Different Settings Across Models

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: Comparisons of Model Performances by GPT-4

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**

practical applications. **Model 5** utilized Direct Preference Optimization (DPO) to align model outputs with human preferences, optimizing chosen over rejected responses to enhance privacy-sensitive behavior efficiently. **Model 6** focused on simultaneous generation of normal and privacy-preserving summaries, training the model to balance completeness and privacy-preservation dynamically. Finally, **Model 7** employed Odds Ratio Preference Optimization (ORPO), which introduced computationally efficient preference optimization applying an odds ratio-based penalty without requiring a reference model as such, for effectively handling ambiguous privacy violations. Section B elaborates further about the different techniques used to train the models along with the intuition behind them.

Evaluation metrics 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 sensitive information by effectively recognizing and masking confidential or sensitive information while the Completeness score measures how well summaries retained the key information from the original conversation and conveyed all 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.

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 dialogue 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.

Although we address these aspects too, such metrics are largely inadequate for measuring privacy preservation as they prioritize semantic similarity and grammatical coherence but fail to capture whether sensitive content has been sufficiently masked in the summary. For example, a high BERTScore could still mean that sensitive financial details or personal matters have been exposed. So this paper also advocates for a *human evaluation* aspect focusing on Consistency, Coherence, Relevance, and Privacy, the parameters used by the

Table 3: Model Comparison Across ROUGE, BERTScore, and MoverScore Metrics

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
GPT-4o	0.4766	0.1952	0.3450	0.3471	0.7018	0.9111	0.7526	0.5591	0.4834	0.4323

Table 4: Human Evaluation and Distilled Human Evaluation Across Different Models

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

authors of DialogSum (Chen et al., 2021), with the addition of Privacy from our end-

- **Consistency**: Measures whether the summary consistently reflects the original conversation.
- **Relevance**: Judges how well the summary retains important information relevant to the original conversation for completeness.
- **Coherence**: Evaluates whether the summary logically flows and makes sense.
- **Privacy**: Assesses how well the summary preserves privacy by excluding or masking sensitive information.

The evaluations were on a binary scale (0 or 1) with the inter-rater agreement measured using Cohen's and Fleiss' Kappa scores (McHugh, 2012). We started with 10 conversations and their summaries (7 fine-tuned models, with Base model, GPT-4o, and Ground Truth Summaries as baselines). These were six human evalutors who had been given instructions on how to annotate using a clear evaluation criteria. After grading, we analyzed performance to identify the top models, and to validate findings followed with a Distilled Evaluation of 20 additional conversations graded, ensuring a thorough and credible assessment of the models' capabilities in generating high-quality summaries.

6 Results

Based on the results obtained from the overall averages across settings (Table 1), the percentage of acceptable summaries, i.e. those having min(Privacy,Completeness)>3 across models shown in Figure 7, and LLM metrics (Table 2) , we observed that Models 3 and 6 demonstrated a strong ability to balance privacy and completeness, achieving scores comparable to or surpassing the baseline GPT-4o and approaching the scores of the Ground Truth summaries. Model 3, trained only on privacy-preserving summaries, achieved high scores in privacy (4.605) and completeness (4.051), while Model 6, designed for simultaneous generation of normal and privacy-preserving summaries, also achieved similarly high scores in privacy(4.884) and completeness (4.047), reflecting their capability to manage the trade-offs effectively. Consequently, we decided to focus on these two models for further analysis and experimentation.

Regarding the NLP metrics, Model 3 and Model 6 achieved the highest scores across all ROUGE metrics, suggesting better retained critical information while adhering to privacy constraints. 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 baseline GPT-4o particularly for use cases requiring both context preservation and strong privacy safeguards, making them highly suitable for applications in sensitive domains.

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-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.

7 Future Work

In the current work, we have so far been able to successfully identify sensitive information across diverse contexts and generate privacy-preserving summaries that do not leak any such information. The study by Li et al. (2020) explores the impact of cultural differences on privacy decisions, highlighting the need for dynamic categorization of sensitive elements according to contextual settings. In subsequent phases, we aim to address this and integrate our findings into a pipeline for context-sensitive privacy preservation, adapting based on the event context and user dynamics. Building on the taxonomy developed, we also envision user-level customization that allows individuals to provide relational information for tailored masking of sensitive data.The approach introduced by Liang (2019) using a Collaborative Interest Tracking Topic Model (CITM) and Streaming Keyword Diversification Model (SKDM) provides a starting point, with scope for dynamic masking of sensitive information based on individual interactions and relationships, paving the way for a digital clone model that adapts to evolving privacy concerns and interactions. However, this would require the management of permission access to sensitive user data, presenting additional challenges across organizations. We could minimize related data leaks by exploring ways for data curation, training and inference to be done locally within a secure environment, optimizing both computational efficiency and cost-effectiveness in the process. Future research could also investigate the model's performance after quantization, which would allow deploying fine-tuned models with Phi-3.5 as their base completely on personal devices, enhancing privacy with energy efficiency while reducing latency and scalability issues. The outlined future work aims to enhance privacy preservation by integrating contextual sensitivity, user customization, and secure data handling practices while also considering edge computing, contributing to the development of more sophisticated and user-centric privacy-preserving technologies.

8 Conclusion

This research addresses the challenge of privacy-preserving text summarization, aiming to balance content completeness with safeguarding sensitive information—a gap persistent even in the latest baselines today. We proposed a comprehensive taxonomy covering sensitive elements across diverse settings and curated a well-defined dataset around it. We then trained models using this dataset and evaluated their performance to gauge their performance. Our evaluation focused on seven models built upon Phi 3.5 as the base model, fine-tuned using various techniques as elaborated in the respective sections above, with Model 3 and Model 6 emerging as standout performers. Model 3 emphasized the importance of high-quality data curation over mere volume, showing how excluding privacy-sensitive content during training impacts the quality of summaries while Model 6 excelled with its dual-output design, capable of producing both standard and privacy-preserving summaries, making it applicable in dynamic environments where privacy requirements may vary. While other models explored innovative techniques like Model 5 (DPO) and Model 7 (ORPO), they often compromised content completeness for enhanced privacy, leading to poor results in balancing the trade-off between the two. For all our evaluations, GPT-4o served as a benchmark due to its advanced language generation capabilities, demonstrating strong overall performance. However, its limitations in ensuring privacy preservation highlighted the value and need for focused, domain-specific training in achieving superior outcomes. Overall, this study aims to advance privacy-preserving AI by demonstrating the potential of a well defined dataset in developing adaptable models to mitigate privacy risks while maintaining the integrity of the intended tasks, like summarized content in our case.

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. 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.

Alexander R. Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021. Convosumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. *Preprint*, arXiv:2106.00829.

Guy Feigenblat, Chulaka Gunasekara, Benjamin Sznajder, Sachindra Joshi, David Konopnicki, and Ranit Aharonov. 2021. Tweetsumm – a dialog summarization dataset for customer service. *Preprint*, arXiv:2111.11894.

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).

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. Association for Computational Linguistics.

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.

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.

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!

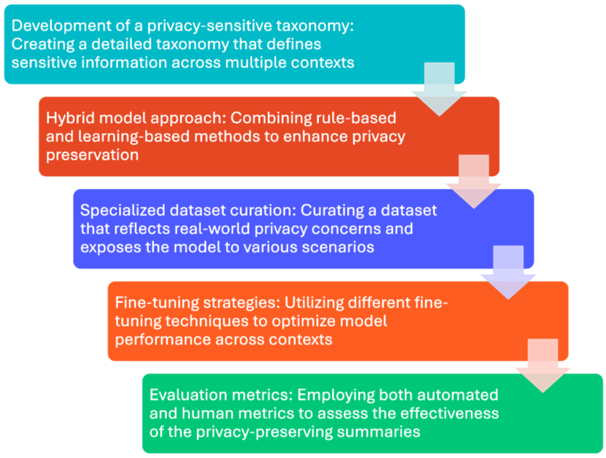


Figure 3: An overview of the systematic approach used to generate and verify privacy-preserving summaries in our research

Appendix

A 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 5 provides an overview of the structure of the dataset, while Figure 4 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 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

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 5: 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 4: A sample datapoint showing how data is formatted under each of the columns mentioned in the Dataset

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.

B.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).

B.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

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

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.

B.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.

B.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.

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

B.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.

B.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" 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.

B.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.

B.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.

C 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.

C.1 Libraries and Frameworks

C.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.

C.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

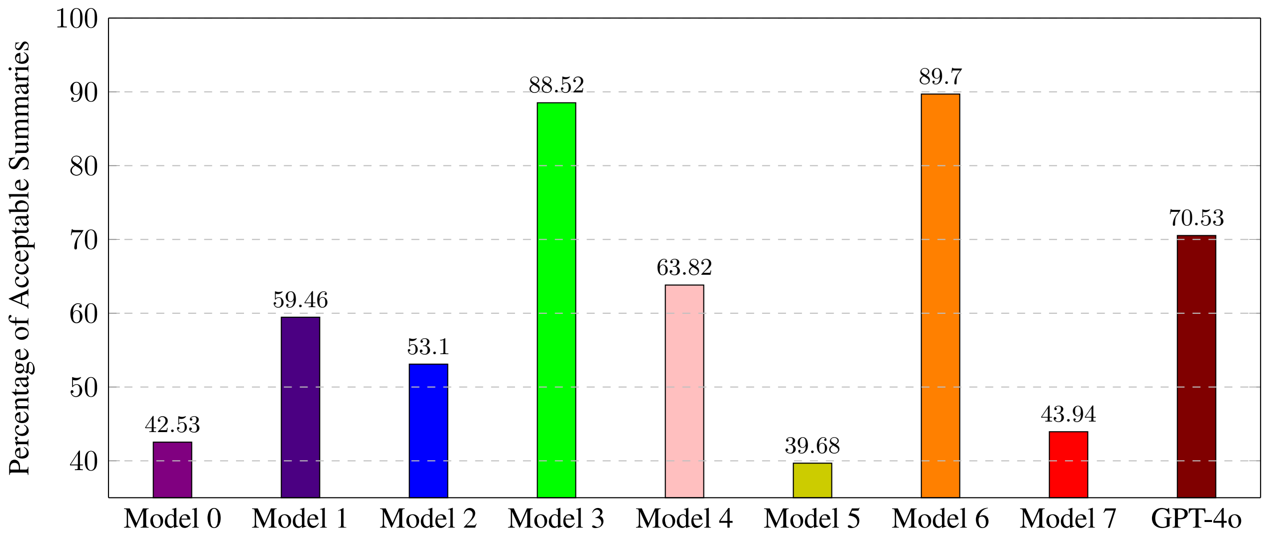


Figure 7: Percentage of Acceptable Summaries, i.e. Summaries having min(Privacy,Completeness)>3 for Different Models

Table 6: Scores for Different Models 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**

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.

C.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.

C.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.

D Results

D.1 Public Datasets

The results presented in Table 6 demonstrate the performance of various models across four datasets: DialogSum, ConvoSumm, TweetSum, and SAMSum. The metrics being evaluated are Privacy,

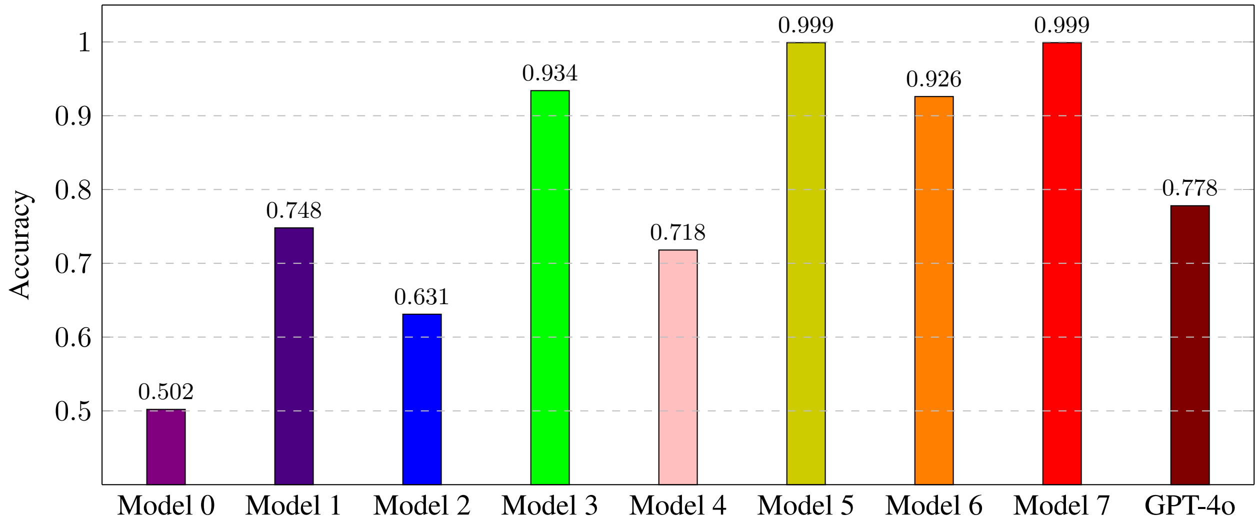


Figure 8: Performance Across Models on **ai-masking-400k** dataset.

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.

D.2 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.

D.2.1 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