Public Data Assisted Differentially Private In-Context Learning

Anonymous ACL submission

Abstract

In-context learning (ICL) in Large Language Models (LLMs) has shown remarkable performance across various tasks without requiring fine-tuning. However, recent studies have highlighted the risk of private data leakage through the prompt in ICL, especially when LLMs are exposed to malicious attacks. While differential privacy (DP) provides strong privacy guarantees, it often significantly reduces the utility of in-context learning (ICL). To address this challenge, we incorporate task-related public data into the ICL framework while maintaining the DP guarantee. Based on this approach, we propose a private in-context learning algorithm that effectively balances privacy protection and model utility. Through experiments, we demonstrate that our approach significantly improves the utility of private ICL with the assistance of public data. Additionally, we show that our method is robust against membership inference attacks, demonstrating empirical privacy protection.

1 Introduction

With the emergence of Large Language Models (LLMs), in-context learning (ICL) has demonstrated remarkable performance across various tasks by enabling models to infer from provided examples without modifying internal parameters (Brown et al., 2020; Min et al., 2022; Wei et al., 2022). This flexibility allows LLMs to adapt to diverse domains without explicit training, leading to its widespread adoption (Dong et al., 2022).

Meanwhile, concerns about privacy leakage have been raised regarding the deployment of ICL in LLMs (Li et al., 2023a; Kandpal et al., 2023; Li et al., 2024). For a concrete example, consider the ICL scenario depicted in Figure 1, where private patient treatment records are used as demonstration examples. A malicious attacker aiming to identify private clinical record data may try inference attack targeting LLM. The LLM with an ICL algorithm, without privacy protection, could potentially expose sensitive clinical data to attackers. Such exposure of personal health information may contravene data protection regulations like the General Data Protection Regulation (GDPR), which mandates appropriate technical and organizational measures to ensure data security and protect individual privacy rights (GDP, 2016).

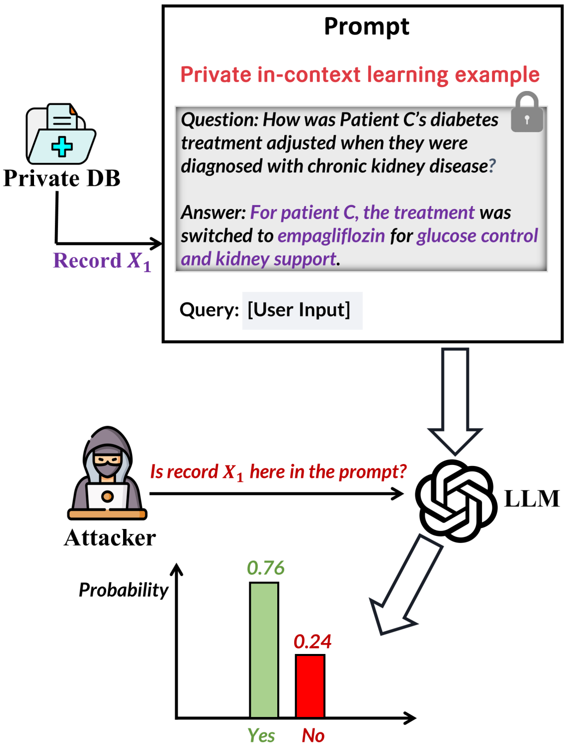

Prompt
Private in-context learning example
Question: How was Patient C's diabetes treatment adjusted when they were diagnosed with chronic kidney disease?
Answer: For patient C, the treatment was switched to empagliflozin for glucose control and kidney support.
Query: [User Input]
Private DB
Record X_1
Attacker
Is record X_1 here in the prompt?
LLM
Probability
0.76
0.24
Yes
No

Figure 1: An illustration of a *membership inference attack*, where a malicious attacker infers whether a target example is included in the prompt composed of private examples retrieved from the database. By leveraging the tuned prompt, the attacker can determine the presence of the target example within the in-context examples.

Differential privacy (DP) has emerged as the gold standard for rigorous privacy protection across multiple domains, including computer vision, recommendation systems, natural language processing, and census data (Dwork, 2006). The essence of differential privacy is to ensure that the output of an algorithm is minimally influenced by the inclusion or exclusion of any individual's data, thereby reducing the risk of privacy leakage. Additionally, the level of privacy protection can be adjusted by tuning the privacy protection parameter ε. Differential privacy is often employed in conjunction with *sample and aggregate* techniques, where multiple noisy outputs are combined to enhance privacy.

However, when DP is integrated into ICL algorithms, two main challenges arise. First, the performance of the ICL algorithm deteriorates significantly under strong privacy protection, where output perturbation negatively impacts the model's utility. To mitigate this degradation, we incorporate public data into our ICL framework to minimize utility loss. The second challenge is to design a privacy-preserving aggregation method for language generation outputs. Given the high dimensionality inherent in the output space of LLM responses, it is essential to transform them into a lower-dimensional representation while ensuring that they remain reconstructable in the original response space. To achieve this, we project LLM-generated responses into semantic embeddings and form *semantic groups* for private aggregation.

We evaluate our private ICL framework with DP guarantees on question-answering tasks (ChatDoctor, (Li et al., 2023b)) and a document summarization task (SAMsum, (Gliwa et al., 2019)). The experimental results show that our private ICL framework performs comparably to non-private baselines at a strong privacy protection level ($\varepsilon = 1$) and outperforms the private data-only counterpart at the same privacy protection level. We further demonstrate that using out-of-distribution (OOD) public data, as well as in-distribution (ID) public data, is beneficial for minimizing utility degradation. Moreover, we show that our private ICL framework is robust against empirical privacy attacks by simulating membership inference attacks on the model. In summary, our main contributions are the following:

1. We propose a DP-guaranteed private ICL framework that integrates public data and employs semantic group aggregation to manage high-dimensional outputs.

2. Our framework demonstrates effective utility-privacy tradeoffs in question answering and summarization tasks, and benefits from both ID and OOD public data.

3. Our approach is robust against empirical privacy threats, as verified by membership inference attack experiments.

2 Related Work

Private Text Generation Differentially private text generation methods can be broadly categorized into differentially private fine-tuning approaches (Yu et al., 2021; Li et al., 2021; Yu et al., 2023) and differentially private prediction approaches (Majmudar et al., 2022; Ginart et al., 2022; Flemings et al., 2024). Fine-tuning methods fine-tune LLMs using the DP-SGD algorithm (Abadi et al., 2016). However, private fine-tuning methods suffer from high computational costs and cannot be applied to black-box LLMs. Moreover, private fine-tuning does not allow for flexible data replacement.

In prediction methods, many works build upon the PATE framework (Papernot et al., 2018), where each teacher model is trained on a subset of the private dataset, and the aggregate of the teacher ensemble is used for prediction. Apart from methods with differential privacy guarantees, text sanitization methods (Albanese et al., 2023; Papadopoulou et al., 2022) and unlearning methods (Kassem et al., 2023) have also been proposed to protect individual privacy.

Differentially Private In-Context Learning Duan et al. (2023) propose `PromptPATE`, where unlabeled public data is privately annotated by a teacher ensemble trained on the private dataset, and the annotated public data is then used as demonstration examples in the prompt for ICL. Tang et al. (2023) propose a *token-level* differentially private ICL framework that generates synthetic texts as demonstration examples for subsequent queries. However, the output space of this ICL framework is limited to the label space and does not cover long-text responses. Wu et al. (2024) develop an ICL framework with DP guarantees that can handle long-text responses. To address the high-dimensionality issue of the output text space, they propose embedding space aggregation and keyword space aggregation. However, the performance with strong privacy protection is not satisfactory compared to the non-private counterpart.

Differential Privacy with Public Data As a seminal work, Wang and Zhou (2020) addresses an empirical risk minimization problem with limited

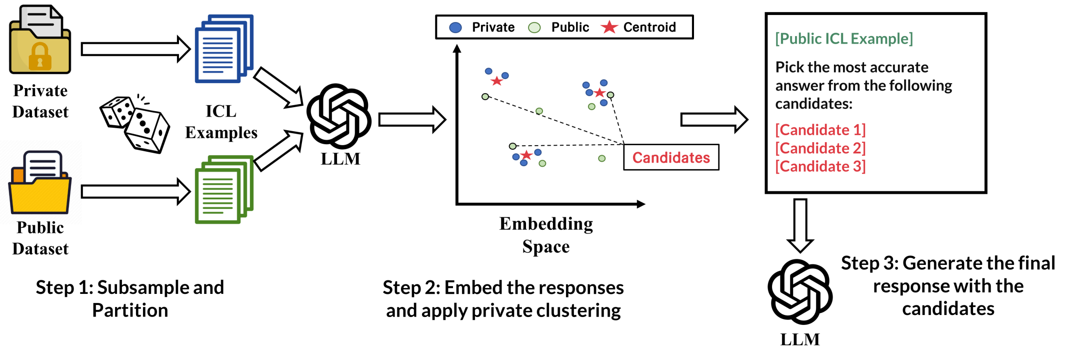

Private Dataset
Public Dataset
ICL Examples
LLM
Step 1: Subsample and Partition
Private
Public
Centroid
Candidates
Embedding Space
Step 2: Embed the responses and apply private clustering
[Public ICL Example]
Pick the most accurate answer from the following candidates:
[Candidate 1]
[Candidate 2]
[Candidate 3]
LLM
Step 3: Generate the final response with the candidates

Figure 2: **Overview of public data-assisted differentially private in-context learning.** In Step 1, the public and private data are partitioned and used to create demonstration examples. In Step 2, the generated responses of the LLMs are embedded into a semantic space and then clustered privately. Finally, in Step 3, the top-k candidates closest to the centroids are selected, and the LLM chooses the final answer from among them with a public ICL example.

public data and proposes a private-public stochastic gradient descent (SGD) method that uses public data to adjust training parameters. Subsequently, variants of DP-SGD that leverage public data have demonstrated the efficiency of public data by showing improvements in utility (Nasr et al., 2023) and have shown that public data can be used to learn informative priors for efficient private learning (Tang et al., 2023). While earlier works primarily applied public data during the fine-tuning stage, our approach integrates public data during the inference stage. With careful incorporation of public data, our ICL framework demonstrates a superior privacy-utility trade-off compared to baseline methods. We present the comparison table from the previous DP literature in Table 3 in Appendix.

3 Preliminaries

3.1 In-Context Learning

To respond to the user query Q, the demonstration examples $(Q_1, A_1), \ldots, (Q_n, A_n)$ are concatenated with the user query to assist in generation. The LLM can learn from the demonstration examples by identifying the relevant mapping from each (Q_i, A_i) in the examples. Given the query and the demonstrations, the LLM selects the next token using various sampling algorithms or greedy decoding. The objective of our work is to ensure that the algorithm remains private, such that an attacker cannot infer the presence of specific target data within the demonstration examples in in-context learning (ICL).

3.2 Differential Privacy

Differential privacy (Dwork, 2006) is considered the gold standard for protecting the privacy of machine learning algorithms. The formal definition is as follows:

Definition 1 (Differential Privacy) *A randomized mechanism $\mathcal{M} : \mathcal{D} \to O$ is said to satisfy (ε, δ)-differential privacy if for any neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$, which differ in only a single element, it holds that*

$$\Pr[\mathcal{M}(\mathcal{D}) \in \mathcal{S}] \leq e^{\varepsilon} \Pr[\mathcal{M}(\mathcal{D}') \in \mathcal{S}] + \delta$$

for any set $\mathcal{S}$ of possible outputs in O.

As ε decreases, the probabilities $\Pr[\mathcal{M}(\mathcal{D}) \in \mathcal{S}]$ and $\Pr[\mathcal{M}(\mathcal{D}') \in \mathcal{S}]$ become closer, making it more difficult for an attacker. The parameter δ represents the failure probability, where the DP guarantee may not hold.

Post-processing Property The post-processing property allows for arbitrary transformations of the output of a DP algorithm, enabling the development of DP algorithms that exploit public data.

Definition 2 (Post-processing of DP) *Let $\mathcal{M}$ be an (ε, δ)-differentially private algorithm, and let f be any transformation function. Then $f(\mathcal{M}(D); D_{pub})$ also satisfies (ε, δ)-differential privacy, where D_{pub} is auxiliary public data.*

4 Private ICL with Public Data

In this section, we present our private ICL framework. We first explain each stage of the private ICL framework and then present the DP privacy analysis of the framework. The schematic diagram of our private-ICL framework is presented in Figure 2 and the algorithm is described in Algorithm 1.

4.1 Private ICL

Step 1: Dataset Subsampling and Partition Before providing the demonstration dataset to the

LLMs, we randomly subsample a fraction of $p\%$ from the dataset. Subsampling offers two key advantages: (1) **Privacy Amplification**: For an (ε, δ)-DP algorithm $\mathcal{M}$, subsampling can amplify privacy, effectively reducing the privacy loss. Specifically, under certain subsampling techniques, the privacy parameter ε can be reduced to approximately $p\varepsilon$, where p is the subsampling rate[1]. (2) **Memory Cost Reduction**: By using a subset of the dataset instead of the entire set of demonstration examples, we can significantly reduce the memory load on the API. In the n-shot, m-ensemble setting, we uniformly subsample mn demonstration examples from both the private and public datasets. Each n-shot example is formatted into a task-specific prompt with the user query and fed to the LLM. After inference, we obtain m responses for each example.

Step 2: Private Aggregation of Responses After generating multiple responses from the ensemble, we need to aggregate them privately to prevent attackers from inferring information about the private examples. However, direct aggregation results in a highly sparse histogram due to the nearly infinite-dimensional output space of LLMs.

To address this, we adopt *private clustering* from Li et al. (2024) (`DPM` of line 9 in Algorithm 1). First, we obtain embeddings for each response using a text embedding model. Then, we apply private k-means clustering to both private and public embeddings, generating privatized cluster centers and the number of members (weights) in each cluster. For class representatives, we select the closest public element to each cluster center (as shown in line 14 of Algorithm 1). As a baseline aggregation method, we employ a modified version of Keyword Space Aggregation (KSA) from (Wu et al., 2024), in which sentences are projected into a keyword space, and responses are reconstructed using the most frequent keywords.

Step 3: Final Response selection & Generation A straightforward approach to response generation is to select the representative of the cluster with the highest count. However, under strong privacy protection, the cluster with the second-highest count may incorrectly surpass the true highest-count cluster, potentially degrading model performance. To obtain more reliable responses while leveraging public data as guidance, we pass the class representatives of the top-k clusters to the selection stage. At this stage, we construct the prompt as: `Select the most correct answer for the question from <candidates>`, with a public one-shot example prepended to the prompt (template in Appendix H), as illustrated in Figure 2.

Algorithm 1 Public-data assisted In-Context Learning

```
Require: LLM, private dataset $\mathcal{X}_{pri}$, public dataset $\mathcal{X}_{pub}$,
    number of ensemble $N$, query $Q$, privacy parameter $\varepsilon$
 1: Subsample and partition each dataset:
 2: $\{D_i^{pri}\}_{i=1}^{N} \leftarrow D^{pri}$, $\{D_i^{pub}\}_{i=1}^{N} \leftarrow D^{pub}$
 3: for $i = 1 \ldots N$ do
 4:     Construct a few-shot prompt for each dataset:
 5:     $P_i^s \leftarrow \texttt{Prompt}(D_i^s, Q), \quad s \in \{pri, pub\}$
 6:     $O_i^s \leftarrow \textbf{LLM}(P_i^s), \quad s \in \{pri, pub\}$
 7: end for
 8: Privately estimate cluster centres using DPM (Algorithm 2 of Appendix):
 9: $\mathcal{C}$, weights $\leftarrow \texttt{DPM}(\{O_i^{pri}\}_{i=1}^{N} \cup \{O_i^{pub}\}_{i=1}^{N}, \varepsilon)$
10: Sort cluster centers $\mathcal{C}$ by member count in decreasing order.
11: Choose the representative for each cluster:
12: $S \leftarrow \emptyset$
13: for $C \in \mathcal{C}$ do
14:     $i \leftarrow \arg\min_j \|C - O_j^{pub}\|$
15:     $S \leftarrow S \cup \left\{O_i^{pub}\right\}$
16: end for
17: Generate answer using top-$k$ candidates with a public 1-shot example:
18: $\mathbf{a} \leftarrow \textbf{LLM}(\{S_j\}_{j=1}^{k}, D_{pub})$
19: return $\mathbf{a}$
```

4.2 Privacy Analysis

In this section, we give analysis for calculating `DPM` parameter for the algorithm 1 to be (ε, δ)-DP. Assuming that the private clustering `DPM` Algorithm is (ε, δ)-differentially private, Algorithm 1 is also (ε, δ)-differentially private. This follows from the post-processing property (definition 2), as lines 13–18 in the algorithm do not use any additional private data.

In addition, since the model releases privatized output for each query, the privacy risk accumulates, which necessitates an accurate privacy loss tracking mechanism. In Algorithm 1, the private algorithm DPM employs *exponential and Gaussian mechanisms* for private clustering. To trace the accurate privacy loss for the exponential mechanism, we express the privacy loss as Rényi Differential Privacy (RDP) using Bun and Steinke (2016), then compose the privacy guarantees under RDP, and finally convert the result back to (ε, δ)-DP using the theorem from Balle et al. (2019). For the Gaussian mechanism in DPM, we use `DPSGDAccount`

---

[1] For stability of the algorithm, we choose the uniform sampling without replacement. The details of privacy amplification are explained in Theorem 1.

from the `prv_accountant` library, ensuring tight privacy tracking.

5 Experiments and Results

In this section, we present the privacy-utility trade-off of our private ICL models in question answering and summarization tasks. We begin by evaluating model performance using In-Distribution (ID) public data and Out-Of-Distribution (OOD) public data. Finally, we assess the empirical effectiveness of private ICL through a membership inference experiment. In addition to OOD public data, we also evaluate the performance of private ICL with noisy public data in the Appendix B.

5.1 Experiment Setting

Task We evaluate the performance of the private ICL model on question-answering and dialogue summarization tasks. We conduct the ICL task with 100 test queries using a 4-shot and 100-ensemble setting, where the ensemble method is applied along with 100 private and 100 public examples. For the question-answering task, we use the OpenAI GPT-3.5-turbo model and the Davinci-002 model model for summarization.

Dataset For the question-answering task, we use the ChatDoctor benchmark (Li et al., 2023b), which consists of questions and answers collected from dialogues between patients and doctors on icliniq.com. We sample 3,900 data examples from the dataset, using 2,600 examples as the private dataset and 1,300 examples as the public dataset. For the OOD public dataset, we use the HealthCareMagic subset of ChatDoctor which comes from a different platform.

For the dialogue summarization task, we use the SAMSum dialogue summarization dataset (Gliwa et al., 2019). From the training dataset, we also use 2,600 examples as the private demonstration dataset and 1,300 examples as the public demonstration dataset. For the OOD public dataset, we use dialogsum benchmark (Chen et al., 2021).

Models We consider four models, including the baselines: **SGA (top-k)**, our private ICL model with private clustering on semantic space; **SGA (top-1)**, which applies private clustering on semantic space but directly selects the top-1 element without candidate selection; **KSA**, a baseline private ICL model with keyword space aggregation; and **KSA w/o public**, which performs keyword space aggregation using only private data.

We evaluate the models at three privacy levels: $\varepsilon = 1, 3, 8$, corresponding to strong, moderate, and weak privacy protection, respectively. We also include two non-private baselines and one fully private baseline: a 4-shot model with sampled private examples ($\varepsilon = \infty$), a 4-shot model using non-private aggregation ($\varepsilon = \infty$ (Agg)) and a 4-shot model using OOD public ($\varepsilon = 0$). The hyperparameter settings for privacy analysis are summarized in Table 4 of the Appendix.

5.2 Private ICL with ID public

Question Answering Task The results for the question answering task are presented in Table 1a, evaluated using BLEU, METEOR, ROUGE-1 and LLM-Judge metrics. For the LLM-Judge evaluations, we employ GPT-4o as a judge to assess the generation quality of each model on the QA task. From the experiment, we observe the following: (1) **Candidate selection guided by public examples leads to better results**. For all privacy levels ε, SGA with top-k selection outperforms the SGA top-1 method, demonstrating the usefulness of public data. (2) **The SGA method outperforms the KSA method across all privacy levels.** We assume that the KSA method struggles with reconstructing entire answers with noisy keywords, as the answer length in the ChatDoctor benchmark is typically more than 3 sentences. (3) **Private aggregation with public data improves performance.** The KSA method with public data outperforms the KSA method without public data across all privacy budgets ε. Finally, we observe that $\varepsilon = \infty$ with aggregation outperforms the direct 4-shot method.

Summarization Task The summarization results are presented in Table 1b with ROUGE-1,2,L and LLM-Judge metrics. From the results, we observe the following: (1) **Both SGA and KSA show reasonable performance with a strong privacy budget** ($\varepsilon = 1$), with comparable performance to 4-shot $\varepsilon = \infty$ without aggregate. (2) **Using public data for candidate selection and aggregation improves performance.** The SGA method with top-k selection outperforms the top-1 SGA, and the KSA method outperforms KSA without public data. (3) **SGA shows more robust performance at a strong privacy regime** ($\varepsilon = 1$)**.** The robustness of SGA may be attributed to the fact that KSA reconstructs the summary from noisy keywords, whereas SGA generates a summary from previously generated summaries.

Method	Metrics	$\varepsilon = 1$	$\varepsilon = 3$	$\varepsilon = 8$	$\varepsilon = \infty$ (Agg)	$\varepsilon = \mathbf{0}$	$\varepsilon = \infty$ (4-shot)
SGA (top-k)	BLEU ↑	**22.21$_{0.32}$**	**23.37$_{0.18}$**	**23.67$_{0.09}$**	**26.01**	21.61	23.43
	METEOR ↑	**16.65$_{0.34}$**	**17.13$_{0.21}$**	**18.24$_{0.28}$**	**20.15**	17.01	18.81
	ROUGE-1 ↑	**25.91$_{0.24}$**	**25.75$_{0.33}$**	**27.01$_{0.14}$**	**30.67**	28.28	28.98
	LLM-Judge ↑	**2.82**$_{0.06}$	**3.08**$_{0.04}$	**3.11**$_{0.04}$	**3.24**	2.75	2.86
SGA (top-1)	BLEU ↑	19.57$_{0.24}$	21.87$_{0.14}$	21.24$_{0.21}$	23.77	21.61	23.43
	METEOR ↑	14.79$_{0.29}$	15.87$_{0.17}$	15.62$_{0.23}$	16.85	17.01	18.81
	ROUGE-1 ↑	23.12$_{0.34}$	24.73$_{0.16}$	24.22$_{0.31}$	26.44	28.28	28.98
	LLM-Judge ↑	2.71$_{0.04}$	2.88$_{0.05}$	2.91$_{0.04}$	3.22	2.75	2.86
KSA	BLEU ↑	15.98$_{0.56}$	16.53$_{0.20}$	17.41$_{0.26}$	24.89	21.61	23.43
	METEOR ↑	13.38$_{0.43}$	13.35$_{0.09}$	14.11$_{0.22}$	18.80	17.01	18.81
	ROUGE-1 ↑	19.05$_{0.44}$	19.25$_{0.22}$	20.93$_{0.15}$	29.16	28.28	28.98
	LLM-Judge ↑	2.58$_{0.05}$	2.61$_{0.05}$	2.71$_{0.04}$	3.12	2.75	2.86
KSA w/o public	BLEU ↑	15.65$_{0.45}$	16.36$_{0.21}$	16.13$_{0.52}$	24.03	21.61	23.43
	METEOR ↑	12.80$_{0.36}$	13.35$_{0.10}$	13.23$_{0.41}$	18.15	17.01	18.81
	ROUGE-1 ↑	18.23$_{0.39}$	19.61$_{0.32}$	20.93$_{0.15}$	27.38	28.28	28.98
	LLM-Judge ↑	2.52$_{0.06}$	2.59$_{0.05}$	2.68$_{0.06}$	3.08	2.75	2.86

(a) **Private ICL results in the QA task.** The best results are highlighted: BLEU, METEOR, ROUGE-1, LLM-Judge (1-5 scale).

Method	Metrics	$\varepsilon = 1$	$\varepsilon = 3$	$\varepsilon = 8$	$\varepsilon = \infty$ (Agg)	$\varepsilon = \mathbf{0}$	$\varepsilon = \infty$ (4-shot)
SGA (top-k)	ROUGE-1 ↑	**36.58$_{0.59}$**	38.84$_{0.23}$	**39.90$_{0.71}$**	41.39	32.47	37.24
	ROUGE-2 ↑	13.34$_{0.50}$	14.03$_{0.54}$	15.93$_{0.21}$	16.49	10.72	12.86
	ROUGE-L ↑	**29.36$_{0.83}$**	29.17$_{0.69}$	**31.62$_{0.12}$**	**32.91**	25.50	28.85
	LLM-Judge ↑	**3.25**$_{0.04}$	3.23$_{0.05}$	**3.33**$_{0.03}$	**3.37**	3.01	3.18
SGA (top-1)	ROUGE-1 ↑	36.48$_{0.56}$	37.88$_{0.57}$	37.32$_{0.75}$	**41.59**	32.47	37.24
	ROUGE-2 ↑	12.18$_{0.32}$	13.34$_{0.21}$	13.02$_{0.64}$	16.26	10.72	12.86
	ROUGE-L ↑	27.93$_{0.47}$	29.46$_{0.33}$	28.57$_{0.31}$	32.10	25.50	28.85
	LLM-Judge ↑	3.15$_{0.05}$	3.17$_{0.08}$	3.22$_{0.06}$	3.35	3.01	3.18
KSA	ROUGE-1 ↑	35.06$_{0.73}$	**39.86**$_{0.59}$	39.63$_{0.34}$	41.10	32.47	37.24
	ROUGE-2 ↑	**14.52$_{0.31}$**	**16.95$_{0.48}$**	**16.63$_{0.41}$**	**18.06**	10.72	12.86
	ROUGE-L ↑	27.48$_{0.25}$	**32.08$_{0.67}$**	31.41$_{0.35}$	32.47	25.50	28.85
	LLM-Judge ↑	3.23$_{0.04}$	3.28$_{0.03}$	3.28$_{0.04}$	3.36	3.01	3.18
KSA w/o public	ROUGE-1 ↑	33.45$_{0.88}$	35.90$_{0.59}$	38.10$_{0.54}$	39.24	32.47	37.24
	ROUGE-2 ↑	12.88$_{0.78}$	14.78$_{0.48}$	15.71$_{0.58}$	16.45	10.72	12.86
	ROUGE-L ↑	26.64$_{0.63}$	28.69$_{0.54}$	30.59$_{0.62}$	31.21	25.50	28.85
	LLM-Judge ↑	3.15$_{0.04}$	3.17$_{0.03}$	3.22$_{0.04}$	3.35	3.01	3.18

(b) **Private ICL results in the summarization task.** The best results are highlighted: ROUGE-1, ROUGE-2, ROUGE-L, , LLM-Judge (1-5 scale).

Table 1: Private ICL methods run for each $\varepsilon = 1, 3, 8$. $\varepsilon = \infty$ denotes ensemble non-private methods, $\varepsilon = 0$ **denotes** **4-shot with OOD public, and** $\varepsilon = \infty$ **denotes** **4-shot prediction with private demonstration examples**. Results for DP algorithms are averaged over 5 runs with different seeds. SGA top-1 denotes the top-1 selection without the candidate selection. KSA w/o public denotes the KSA method with only private data.

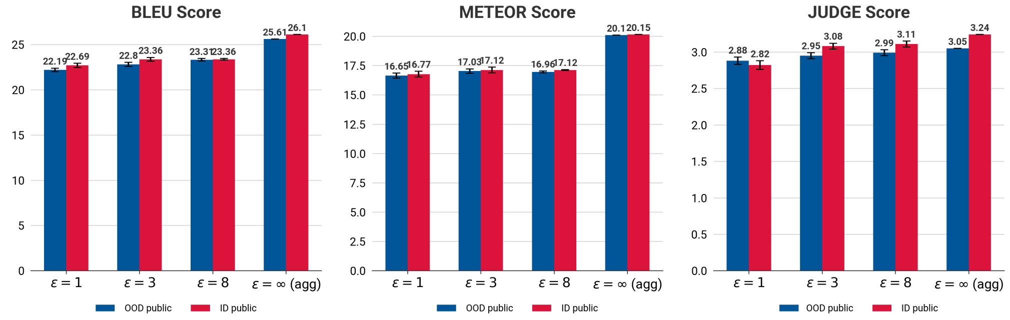
5.3 Private ICL with OOD public

We utilize the best-performing model, SGA, to evaluate private ICL using an OOD public dataset, with the results for question answering and summarization tasks shown in Figure 3. Although ID public data consistently achieves slightly higher scores across most metrics, the OOD public data also demonstrates strong and competitive performance, particularly noteworthy in the question-answering task, where the differences are minimal.

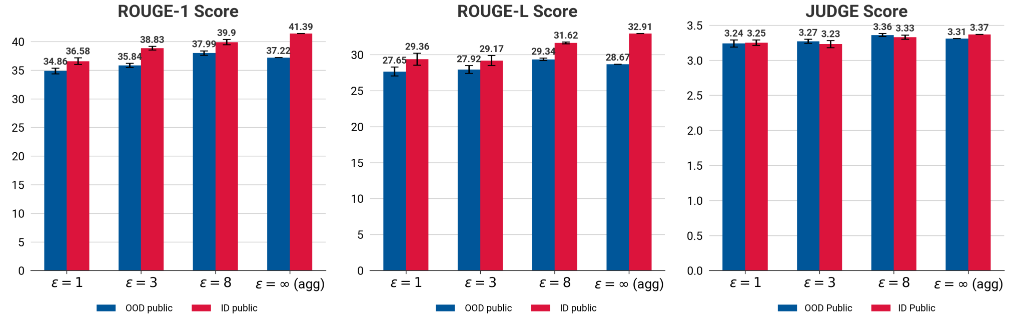
In the summarization task, while the performance gap is more pronounced at higher privacy budgets, OOD data performance notably improves as the privacy budget tightens. This observation indicates that OOD public data remains highly valuable and delivers robust performance, particularly under stricter privacy constraints. Thus, even ICL with OOD data, the SGA method maintains commendable effectiveness, underscoring the practical utility of OOD datasets in privacy-sensitive scenarios.

5.4 Membership Inference Attack

For the empirical evaluation of privacy protection, we adopt a widely used membership inference attack (MIA) against the ICL framework (Shokri et al., 2016). The attacker's goal is to determine whether a target example is part of the demonstra-



(a) Question answering task



(b) Summarization task

Figure 3: **Results for SGA model with OOD public dataset** OOD public refers to private ICL with out-of-distribution public data and ID public to private ICL with in-distribution public data.

tion examples used for ICL. We implement the repeat attack method (Wen et al., 2024). The intuition behind this attack is that the model tends to complete the training sentence when provided with only the first few words of the target example. The inference attack procedure works as follows: (1) The attacker **selects a target example** and attempts to determine whether it is part of the training dataset. (2) The attacker **truncates the target example** and inputs it into the model, which generates a completion. (3) The attacker **calculates the semantic similarity** between the model's completion and the target example. If the similarity exceeds a threshold τ, the attacker concludes that the target example was part of the training set.

We experiment with two different member-to-non-member ratios: (1) **Balanced scenario**: The number of member and non-member examples is equal, with a ratio of $1:1$. (2) **Unbalanced scenario**: The ratio is $1:4$, reflecting a more realistic environment. For the balanced setting, we use 40 member examples and 40 non-member examples. For the unbalanced setting, we use 40-member examples and 160 non-member examples. We use a 2-shot, 10-ensemble setting for the private model and apply SGA with $\varepsilon = [1, 3, 8]$. For the non-private baselines, we consider the non-private aggregation model ($\varepsilon = \infty$, agg) and a 20-shot model ($\varepsilon = \infty$), since the effective number of shots for the private model is $2 \times 10 = 20$. The AUROC is computed over member and non-member examples, with the results shown in Figure 4.

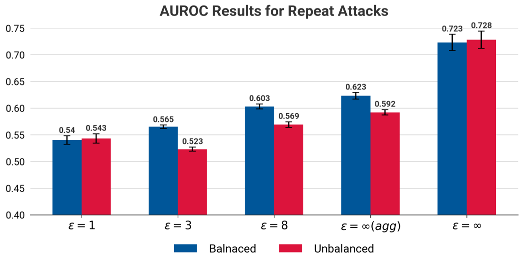


Figure 4: AUROC results for repeat attacks. "Balanced" denotes the MIA setting where the ratio between members and non-members is $1:1$, and "Unbalanced" denotes a ratio of $1:4$.

From the results, we observe that, at all ε levels, the private models are robust to the repeat attack, keeping the AUROC around or below 0.6, indicating low attack success. Notably, the non-private aggregation model also provides some defense, achieving an AUROC of 0.592 for the unbalanced setting. This suggests that non-private aggregation introduces a degree of empirical privacy, possibly due to the aggregation process reducing the influence of individual examples, even though it lacks formal differential privacy (DP) guarantees.

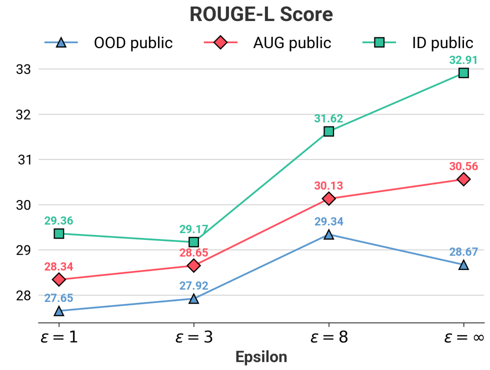


Figure 5: ROUGE-L score with augment public dataset referred as "AUG public" in the summarization task.

6 Enhancing Utility and Efficiency in Private ICL

6.1 Public Data Quality Enhancement

In many domains, high-quality public datasets may be unavailable or inferior to the private datasets at hand, making it challenging to apply our framework. To address this issue, we allocate a small privacy budget to augment the public dataset. Our augmentation follows a simple approach: first, we collect public examples that are closest to privatized centroids with a budget of $\varepsilon = 1$. Then, we instruct the GPT-Turbo model to generate additional data based on these filtered public examples.

The performance of the proposed augmentation technique on the summarization task is reported in Figure 5. The results show that even with a small privacy budget, the augmentation method is effective, showing about $1 \sim 2$ ROUGE-L score difference compared to ID public setting and outperforming the baseline that relies solely on OOD public datasets. Finally, when no public data is available, a private generator can be an effective solution by synthesizing samples for augmentation.

6.2 Accelerating Private ICL with Coreset Sampling

The computation time for private ensembles scales with the number of ensembles (Figure 6). Results indicate that the SGA method with pre-embedding computation (SGA-cache) significantly improves computation efficiency. To further reduce computational complexity, we employ a simple technique called *coreset-sampling*. Specifically, we cluster the ICL samples into distinct groups using K-Means clustering based on text embeddings, selecting the centroids of each cluster as a representative subsampled coreset. This coreset effectively captures the characteristics of the entire batch ensemble. Performance with the coreset approach is shown in Figure 7. We observe that ensemble size 10 with coreset sampling achieves improved performance compared to ensemble size 10 with random sampling.

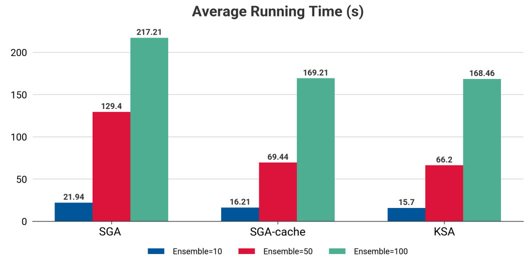


Figure 6: Average running time of SGA and KSA. SGA with cache denotes the SGA with precomputed embeddings.

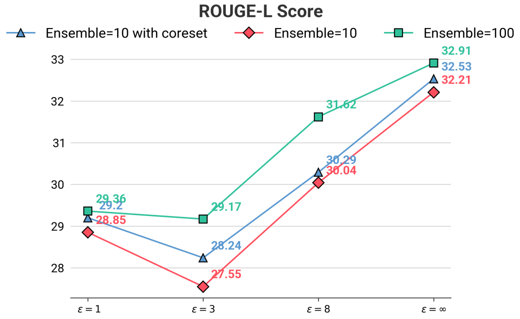


Figure 7: ROUGE-L scores across different privacy budgets (ϵ) for three methods: Ensemble=10 with coreset, Ensemble=10, and Ensemble=100.

7 Conclusion

We propose a private in-context learning (ICL) framework that leverages public data. To address high dimensionality, we project the LLM output into a semantic group space. Public data is used in ensemble aggregation and selection to mitigate utility degradation from differential privacy. Experiments show an effective privacy-utility tradeoff, with notable improvements from public data. Additionally, empirical tests demonstrate strong defense against privacy and membership inference attacks across all privacy levels. Finally, we present techniques for public data enhancement and inference acceleration, offering a practical solution for real-world applications.

Limitations

One limitation of the private ICL framework is that the privacy risk accumulates over multiple queries. While we attempt to mitigate utility degradation caused by accumulated query outputs by leveraging public data, this approach may not entirely prevent privacy leakage in long-running or high-volume query settings. We also believe that this privacy accumulation issue could be alleviated by updating private demonstration examples once a query budget threshold is reached.

Another limitation is that the private ICL framework requires extensive computation due to the use of multiple ensembles. As we have demonstrated, techniques such as coreset sampling can be employed to reduce the computational burden. Future work could explore more computationally efficient differential privacy mechanisms or optimized ensemble methods to further alleviate computational costs without compromising privacy guarantees.

Lastly, public data may not always be accessible in certain domains, especially in sensitive areas where such data is limited or unavailable. As we have demonstrated, this issue can be partially alleviated by synthesizing new examples while investing only a small privacy budget

Ethical Considerations

Data privacy is a critical ethical concern in in-context learning frameworks, especially those leveraging private demonstration examples. Our proposed approach incorporates differential privacy to ensure robust privacy guarantees, minimizing the risk of sensitive information leakage from demonstration data. Throughout our experiments, we exclusively utilized publicly available, open-source doctor-patient dialogues, thereby avoiding potential ethical issues related to private data acquisition and usage.

Despite our strict adherence to public data in experimental settings, real-world deployments of similar frameworks may involve sensitive personal information. We emphasize the necessity of carefully applying differential privacy mechanisms to balance utility and privacy, ensuring compliance with ethical guidelines and privacy regulations such as GDPR and HIPAA.

References

2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). `https://gdpr-info.eu/art-32-gdpr/`. Article 32 – Security of processing.

Martín Abadi, Andy Chu, Ian J. Goodfellow, H. B. McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*.

Federico Albanese, Daniel Alfredo Ciolek, and Nicolas D'Ippolito. 2023. Text sanitization beyond specific domains: Zero-shot redaction & substitution with large language models. *ArXiv*, abs/2311.10785.

Borja Balle, Gilles Barthe, and Marco Gaboardi. 2018. Privacy amplification by subsampling: Tight analyses via couplings and divergences. *ArXiv*, abs/1807.01647.

Borja Balle, Gilles Barthe, Marco Gaboardi, Justin Hsu, and Tetsuya Sato. 2019. Hypothesis testing interpretations and renyi differential privacy. In *International Conference on Artificial Intelligence and Statistics*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Mark Bun and Thomas Steinke. 2016. Concentrated differential privacy: Simplifications, extensions, and lower bounds. *ArXiv*, abs/1605.02065.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning.

Haonan Duan, Adam Dziedzic, Nicolas Papernot, and Franziska Boenisch. 2023. Flocks of stochastic parrots: Differentially private prompt learning for large language models. *ArXiv*, abs/2305.15594.

Cynthia Dwork. 2006. Differential privacy. In *International Colloquium on Automata, Languages and Programming*.

James Flemings, Meisam Razaviyayn, and Murali Annavaram. 2024. Differentially private next-token prediction of large language models. *ArXiv*, abs/2403.15638.

Antonio A. Ginart, Laurens van der Maaten, James Y. Zou, and Chuan Guo. 2022. Submix: Practical private prediction for large-scale language models. *ArXiv*, abs/2201.00971.

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*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Nikhil Kandpal, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. 2023. Backdoor attacks for in-context learning with language models. *ArXiv*, abs/2307.14692.

Aly Kassem, Omar Mahmoud, and Sherif Saad. 2023. Preserving privacy through dememorization: An unlearning technique for mitigating memorization risks in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4360–4379, Singapore. Association for Computational Linguistics.

Haoran Li, Yulin Chen, Jinglong Luo, Yan Kang, Xiaojin Zhang, Qi Hu, Chunkit Chan, and Yangqiu Song. 2023a. Privacy in large language models: Attacks, defenses and future directions. *ArXiv*, abs/2310.10383.

Xi Li, Yusen Zhang, Renze Lou, Chen Wu, and Jiaqi Wang. 2024. Chain-of-scrutiny: Detecting backdoor attacks for large language models. *ArXiv*, abs/2406.05948.

Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori B. Hashimoto. 2021. Large language models can be strong differentially private learners. *ArXiv*, abs/2110.05679.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023b. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Johannes Liebenow, Yara Schütt, Tanya Braun, Marcel Gehrke, Florian Thaeter, and Esfandiar Mohammadi. 2024. Dpm: Clustering sensitive data through separation. *Preprint*, arXiv:2307.02969.

Jimit Majmudar, Christophe Dupuy, Charith Peris, Sami Smaili, Rahul Gupta, and Richard S. Zemel. 2022. Differentially private decoding in large language models. *ArXiv*, abs/2205.13621.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *ArXiv*, abs/2202.12837.

Milad Nasr, Saeed Mahloujifar, Xinyu Tang, Prateek Mittal, and Amir Houmansadr. 2023. Effectively using public data in privacy preserving machine learning. In *International Conference on Machine Learning*.

Anthia Papadopoulou, Yunhao Yu, Pierre Lison, and Lilja Øvrelid. 2022. Neural text sanitization with explicit measures of privacy risk. In *AACL*.

Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. 2018. Scalable private learning with pate. *ArXiv*, abs/1802.08908.

R. Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2016. Membership inference attacks against machine learning models. *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18.

Xinyu Tang, Richard Shin, Huseyin A. Inan, Andre Manoel, FatemehSadat Mireshghallah, Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, and Robert Sim. 2023. Privacy-preserving in-context learning with differentially private few-shot generation. *ArXiv*, abs/2309.11765.

Jun Wang and Zhi-Hua Zhou. 2020. Differentially private learning with small public data. In *AAAI Conference on Artificial Intelligence*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed Huai hsin Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *ArXiv*, abs/2206.07682.

Rui Wen, Zheng Li, Michael Backes, and Yang Zhang. 2024. Membership inference attacks against in-context learning. *ArXiv*, abs/2409.01380.

Tong Wu, Ashwinee Panda, Jiachen T. Wang, and Prateek Mittal. 2024. Privacy-preserving in-context learning for large language models. In *The Twelfth International Conference on Learning Representations*.

Da Yu, Sivakanth Gopi, Janardhan Kulkarni, Zi-Han Lin, Saurabh Naik, Tomasz L. Religa, Jian Yin, and Huishuai Zhang. 2023. Selective pre-training for private fine-tuning. *ArXiv*, abs/2305.13865.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A. Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. 2021. Differentially private fine-tuning of language models. *ArXiv*, abs/2110.06500.

A Details of Differential Privacy

A.1 DPM algorithm

For the private clustering algorithm, we use DPM from (Liebenow et al., 2024). DPM recursively splits a dataset into clusters by identifying sparse regions while preserving differential privacy (DP). We allocate the privacy budget for the Gaussian mechanism with ε_{avg} and for the exponential mechanism with ε_{exp}. For ease of analysis, unlike the original paper, we assume that the size of the dataset and the size of the interval are *public*. By excluding them from the privacy budget, we can focus on protecting more sensitive aspects of the data while simplifying the analysis.

A.2 Privacy Accounting

As privacy accounting involves composition, we utilize the following theorems for tight accounting.

Theorem 1 **(Balle et al. (2018))** *Let* `Uniform` *denotes sampling m elements from n data points without replacement Let $\mathcal{M}' = \mathcal{M} \circ$* `Uniform`*. For any $\varepsilon \geq 0$ we have $\delta_{\mathcal{M}'}(\varepsilon') \leq (m/n)\delta_{\mathcal{M}}(\varepsilon)$, where $\varepsilon' = \log(1 + (m/n)(e^{\varepsilon} - 1))$.*

Theorem 2 **(DP to RDP Bun and Steinke (2016))** *The exponential mechanism is ε-DP and $(\alpha, \varepsilon_{\mathsf{EM}}(\alpha))$-RDP, where $\varepsilon_{\mathsf{EM}}(\alpha)$ is defined as*

$$\min\left(\frac{\alpha}{2}\varepsilon^2, \frac{1}{\alpha-1}\log\left(\frac{\sinh(\alpha\varepsilon)-\sinh((\alpha-1)\varepsilon)}{\sinh(\varepsilon)}\right)\right)$$

Theorem 3 **(RDP composition)** *If each mechanism $(\mathcal{M}_i)_{i=1}^{k}$ is (α, ε)-RDP, then the composition $A_k \circ A_{k-1} \circ \cdots \circ A_1$ is $(\alpha, k\varepsilon)$-RDP.*

Theorem 4 **(RDP to approximate DP Balle et al. (2019))**

If a mechanism $\mathcal{M}$ is (α, ρ)-RDP then it is $(\rho + \log((\alpha-1)/\alpha) - (\log\delta + \log\alpha)/(\alpha-1), \delta)$-DP for any $0 < \delta < 1$.

When accounting exponential mechanism of DPM, we first apply theorem 2. Then the calculate privacy loss of composition by calling `compose_subsampled_EM` API using `AutoDP` package. Then, the converting RDP to approx DP using theorem 4. For the Gaussian mechanism of DPM, we use `DPSGDAccount` from the `prv_accountant` library.

Algorithm 2 `DPM`, (Liebenow et al., 2024)

```
Require: $D, \tau_r, R, t, q$, sigmas, $\alpha, \varepsilon_{\text{int}}, \varepsilon_{\text{cnt}}, \varepsilon_{\text{exp}}, \varepsilon_{\text{avg}}$
clusters ← $\emptyset$
weights ← $\emptyset$
$(\varepsilon_{\text{cnt},i})_{i=0}^{\tau_r} = \left(\varepsilon_{\text{cnt}}\frac{\sqrt{2^i}}{\sum_{j=0}^{\tau_r}\sqrt{2^j}}\right)_{i=0}^{\tau_r}$
$(\varepsilon_{\text{exp},i})_{i=0}^{\tau_r-1} = \left(\varepsilon_{\text{exp}}\frac{\sqrt{2^i}}{\sum_{j=0}^{\tau_r-1}\sqrt{2^j}}\right)_{i=0}^{\tau_r-1}$
$(\lambda_i)_{i=0}^{\tau_r} = (-\ln(2\delta)/\varepsilon_{\text{cnt},i})_{i=0}^{\tau_r}$
// Dataset size perturbation
$\tilde{n} = |D| + \text{Lap}(\varepsilon_{\text{cnt},0})$
// Private interval size estimation
$\beta = \text{IntervalSizeEst}(D, \tilde{n}, \varepsilon_{\text{int}}, \text{sigmas})$
numSplits $= (b-a)/\beta$
BUILDCLUSTERING$(D, \tilde{n}, 0)$
// Privately compute the cluster centers
$C = \{\text{DPAvg}(C_i, \tilde{n}, \varepsilon_{\text{avg}}) \mid C_i \in \text{clusters}\}$
return $C$, weights
procedure BUILDCLUSTERING$(S, \tilde{n}, y)$
    if $y \geq \tau_r$ then
        halt and add $S$ to clusters and $\tilde{n}$ to weights
    end if
    $S_1, S_2 = \text{Split}(S, \tilde{n}_S, y)$
    $\tilde{n}_{S_1} = |S_1| + \text{Lap}(\varepsilon_{\text{cnt},y+1})$
    $\tilde{n}_{S_2} = |S_2| + \text{Lap}(\varepsilon_{\text{cnt},y+1})$
    if $\tilde{n}_{S_i} < \tau_e$ then
        halt and add $S$ to clusters and $\tilde{n}$ to weights
    end if
    BUILDCLUSTERING$(S_1, \tilde{n}_{S_1}, y+1)$
    BUILDCLUSTERING$(S_2, \tilde{n}_{S_2}, y+1)$
end procedure
procedure SPLIT$(S, \tilde{n}, y)$
    $\Delta_f = \frac{t/q+\alpha}{\tilde{n}-\lambda_y}$
    // Private arg-max to find best split index $i^*$
    $i^* = M_E(S, f, \varepsilon_{\text{exp},y})$
    $d^* = \lfloor d \cdot \text{numSplits}/i^* \rfloor$
    $s^* = ((d \cdot \text{numSplits} \mod i^*) + 0.5) \cdot \beta$
    $S_1 = \{x \in S \mid x^{(d^*)} \leq s^*\}$
    $S_2 = \{x \in S \mid x^{(d^*)} > s^*\}$
    return $S_1, S_2$
end procedure
```

B Private ICL with noisy public data

We conduct an experiment on private ICL using misaligned public data. Specifically, we collected public examples that are furthest from the privatized centroids, where the corresponding ICL examples are non-informative, and then instructed the GPT model to generate additional data based on these noisy public examples. This setup enables us to systematically evaluate the robustness of our method under degraded public data quality. The following table shows BLEU scores under varying amounts of noisy public data: While augmentation with more noisy public data does not help under a high privacy budget, it still improves performance under a low privacy budget (e.g., $\epsilon = 1$ or $\epsilon = 2$). We hope these analyses address the reviewer's concern regarding the limited analysis of noisy public data.

Data Type	**BLEU Scores under Privacy Budget**			
	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$
100 noisy public	17.43	19.88	20.89	21.98
500 noisy public	19.26	20.79	21.12	22.68
2000 noisy public	20.79	21.14	21.51	22.41
ID Public	22.21	23.36	23.67	26.01

Table 2: Performance evaluation with varying amounts of noisy public data.

C Comparision of Our Work with Previous DP Literatures

Aspect	**Our work**	**(Wang and Zhou, 2020)**	**(Nasr et al., 2023)**
Objective	Efficient differential private in-context learning along with public data	Differentially private learning with public data for improving ERM and fine-tuning results	Improving utility of DP machine learning by leveraging public data for augmentation and gradient adjustment
Privacy Mechanism	Differential privacy via exponential and Gaussian mechanism	Differential privacy through private-public stochastic gradient descent (PPSGD)	Differential privacy via DOPE-SGD with public data augmentation and gradient clipping adjustments
Use of Public Data	Response aggregation and robust candidate selection	Adjusting parameters in DP-SGD and fine-tuning via model reuse	Public data used for synthetic data generation, gradient adjustment, and ensemble of intermediate DP models
Task	Question answering and summarization	Empirical Risk Minimization (ERM)	Classification tasks (e.g., CIFAR-10)

Table 3: Comparison of our approach with existing methods leveraging public data in DP

D Examples of repeat attack

Examples of successful and failed attacks. The blue part is provided to the model for completion.

Repeat attack examples

Sucessful Attack Examples

Model Completion: Mother age 55, is suffering from abdominal pain, occasional fever at nights. Blood reports says high(1700) AEC (which lowers on taking medicine) otherwise normal and ultrasound says gases and distension in umbilical area? Please suggest possible problem and which specialist to consult?

Reference: Mother age 55, is suffering from abdominal pain,occasional fever at nights.Blood reports says high(1700) AEC (which lowers on taking medicine) otherwise normal and ultrasound says gases and distension in umbilical area? Please suggest possible problem and which specialist to consult?

Score: 0.985

Model Completion: I have a fluttering feeling in my upper right back and also pressure in this area. It started 2 months ago with upper abdominal pressure after eating that was relieved when I stood up or laid down. Ultrasound, blood work and scope of my stomach came back normal.

Reference: I have a fluttering feeling in my upper right back and also pressure in this area. It started 2 months ago with upper abdominal pressure after eating that was relieved when I stood up or laid down. Ultrasound, blood work and scope of my stomacChatDoctore back normal.

Score: 0.981

Failed Attack Examples

Model Completion: Me and my ex boyfriend had [a great relationship but ultimately decided to break up due to our different life goals and values.]

Target Question: Me and my ex boyfriend had sex and he did not pull out. We both know what happened but yet nothing happened. I still had my period, I didn t have symptoms everything was fine. Im actually disappointed but is there something wrong with me or is there something wrong with him?

Score: 0.764

E Judge Template

E.1 Question Answering

GPT-4o Judge Tempate (QA)

You are an expert judge evaluating the overall quality of an LLM-generated answer by comparing it to a human-written reference answer.
You will be given:

- A question
- A reference answer
- An answer generated by a language model (LLM)

Please assign an **overall score from 1 to 5**, based on how well the LLM answer aligns with the reference answer in terms of **correctness, completeness, and faithfulness**.
Scoring Guide:

- **5** – Completely correct and faithful; matches or exceeds the reference.
- **4** – Mostly correct with only minor omissions or inaccuracies.
- **3** – Partially correct; some relevant information is missing or incorrect.
- **2** – Largely incorrect or incomplete; contains major issues.
- **1** – Completely incorrect or irrelevant.

```
Question: {question}
Reference Answer: {reference}
LLM Answer: {llm}
```

E.2 Summarization Judge Template

"GPT-4o Judge Tempate (Summarization)

You are an expert judge evaluating the quality of a language model's summary for a dialogue. You will be given:

- A dialogue transcript
- A human-written reference summary
- A summary generated by a language model (LLM)

Please assign an **overall score from 1 to 5**, based on how well the LLM-generated summary aligns with the reference summary in terms of **factual accuracy, completeness, and faithfulness to the original dialogue**.
Scoring Guide:

- **5** – Completely accurate and faithful; captures all key points and nuances of the dialogue, possibly even more clearly than the reference.
- **4** – Mostly accurate with only minor omissions or slight wording inaccuracies; still faithful to the original dialogue.
- **3** – Partially accurate; captures some important points but misses or distorts others.
- **2** – Largely inaccurate or incomplete; omits critical events or relationships or introduces incorrect information.
- **1** – Entirely inaccurate or misleading; fails to reflect the dialogue content.

```
Dialogue: {dialogue}
Reference Summary: {reference}
LLM Summary: {llm}
```

Overall Score (1–5): _______________

F Dataset example

F.1 ChatDoctor ICLineq

ChatDoctor ICLiniq

Question: Hello doctor,I got my wisdom tooth extracted a few weeks ago, and I am concerned about a piece of red tissue growth in my socket. I am not sure if it is just a new forming tissue. Please clarify.

Answer: Hello. 1. It looks like a piece of surrounding gum or gingival tissue remaining after extraction. 2. I wanted to know when did you notice this for the first time? Was it not there right after extraction, or did you not notice? 3. From photograph/your attachment, it does not look like anything major. (attachment removed to protect patients identity). 4. If you feel pain or sensitivity in it, I would suggest you consult your dentist. He will remove it from the site, and it will not interfere any more. 5. Just keep a notice of whether it is increasing or decreasing in size or any color changes. 6. Avoid very hot and spicy food at least from that side for some time. 7. Report me if you notice any changes. Follow up after a week.

F.2 ChatDoctor HealthcareMagic (OOD)

ChatDoctor HealthcareMagic

Question: Masses full thickness tear of the whole subscapularis tendon with associated atrophy of the musel belly suggestive of a chronic tear ! complete full thickness tear of the whole subscapularis tendon with tendon retraction and a complete tear of the long head biceps tendon with tendon retraction! and large joint effusion

Answer: Hello welcome to the Chat Doctor You are having complete muscle tear and so joint effusion as well. CT scan report should be done to look for grade of muscle tendon injury. Intraluminal steroid can be given to relieve inflammation If age and clinical condition permit than surgical treatment is needed To Hope this will help you Take care Consult orthopedic surgeon for examination and discuss your case

F.3 SAMsum

SAMsum

Dialogue: Dialogue: "Liam: <file_other>
Liam: have you read this one?
Julia: wow, I didn’t even know that it existed
Julia: thanks! The last one I read was Die again
Liam: this one is like from 2014?
Julia: yep
Liam: I’ve heard that city library is organizing a meeting with Tess
Julia: really? That would be really awesome
Liam: just check it out on their facebook"

Reference Summary: The last one Julia read was Die again from 2014. There’s going to be a meeting with Tess organized by the city library.

F.4 Dialogsum (OOD)

Dialogsum

Dialogue:
#Person1#: You have the right to remain silent. Anything you say can and will be used against you in a court of law. You have the right to have an attorney present during questioning. If you cannot afford an attorney, one will be appointed for you. Do you understand?
#Person2#: Yes.
#Person1#: What's your name?
#Person2#: My name is James.
#Person1#: What's your nationality?
#Person2#: American.
#Person1#: What's your relationship with the victim?
#Person2#: I don't know him.
#Person1#: Why did you attack the victim?
#Person2#: Because he beat me first when I tried to stop him from grabbing my bag and running away.
#Person1#: How many times did you stab the victim? #Person2#: I stabbed his belly three times.
#Person1#: Did you know that your actions might cause serious injuries or death?
#Person2#: I knew, but I couldn't control myself.
#Person1#: Was it your intention to kill the victim?
#Person2#: No. I didn't kill him on purpose, madam.
It's him who caused the incident. I need to see my attorney.
#Person1#: OK. Give me his number and we'll contact him.

Reference Summary: #Person1# stabbed the victim because he beat #Person1# first and tried to grab #Person1#'s bag. #Person1# says he didn't kill him on purpose. #Person1# first and tried to grab #Person1#'s bag. #Person1# says he didn't kill him on purpose

G Hyperparameters

Dataset	#Split Levels	k	$(\varepsilon_{exp}, \varepsilon_{GM})$	δ
ChatDoctor	4	3	$[(0.12, 0.5), (0.12, 2.49), (0.12, 7.51)]$	$2.56 \cdot 10^{-4}$
SAMsum	7	3	$[(0.12, 0.5), (0.12, 2.49), (0.12, 7.51)]$	$2.56 \cdot 10^{-4}$

(a) SGA Hyperparameters

Dataset	k	ε_{exp}	δ
ChatDoctor	40	$(0.23, 0.63, 1.32)$	$2.56 \cdot 10^{-4}$
SAMsum	10	$(0.23, 0.63, 1.32)$	$2.56 \cdot 10^{-4}$

(b) KSA Hyperparameters

Table 4: **Hyperparameter Settings** # Split levels denote the number of split levels for the DPM clustering method. k denotes the candidate numbers.

H Prompt construction for candidate selection

H.1 Candidate selection prompt for SGA (QA)

ChatDoctor iclinq

Instruction: You are a doctor. Please answer the medical questions based on the patient's description

<Public Demonstration Example>

<Question>

Pick the most accurate answer for the question with the following answer candidates ranked by their frequency from high to low: **[<Candidates>]**

The answer is:

H.2 Candidate selection prompt for SGA (Summarization)

SAMsum

<Public Demonstration Example>

<Dialogue>

Pick the most accurate summary for the dialogue with the following summary suggestions: **[<Candidates>]**

The summary is:

H.3 Prompt construction for KSA (QA)

ChatDoctor iclinq

Instruction: You are a doctor. Please answer the medical questions based on the patient's description

<Public Demonstration Example>

<Question>

Answer the above question with the following word suggestions ranked by their frequency from high to low: **[<Suggestions>]**

The answer is:

H.4 Prompt construction for KSA (Summarization)

ChatDoctor iclinq

<Public Demonstration Example>

<Dialogue>

Summarize the above dialogue with the following word suggestions ranked by their frequency from high to low: **[<Candidates>]**

The summary is:

I SGA output examples

I.1 ChatDoctor iclinq

ChatDoctor iclinq

Question: Hello,I am a 27 year old female. I am feeling my heartbeat, breath, and pulse rate to be really abnormal since the past three weeks. It feels like my heart is pounding so bad. And then, I feel a difficulty to breathe, like I have run or jogged more than my capacity, even if I am sitting or lying. Sometimes, I feel a lot of pain in my lower abdomen area and on the left side of my belly. Please suggest me some remedy.

($\varepsilon = 1$) **Answer:** Hello, based on your symptoms of feeling a pounding heartbeat, difficulty breathing, and pain in your lower abdomen, it is important to consider various possibilities such as anxiety, hormonal imbalances, or even a cardiac issue. Given the severity and duration of your symptoms, it is crucial to seek medical attention promptly. Please visit a healthcare provider for a thorough evaluation, which may include a physical examination, blood tests, and possibly an ECG to assess your heart's function. In the meantime, practice relaxation techniques like deep breathing exercises and try to reduce stress levels. It is important not to ignore these symptoms and to get a proper diagnosis and treatment plan.

($\varepsilon = 3$) **Answer:** Based on your symptoms of feeling a pounding heart, difficulty breathing, and pain in your lower abdomen and left side of the belly, it is important to rule out any cardiac or respiratory issues. It is recommended that you seek immediate medical attention to undergo a thorough evaluation. These symptoms could be caused by various conditions such as heart problems, respiratory disorders, or gastrointestinal issues. Please do not delay in seeking medical help to determine the cause and appropriate treatment for your symptoms.

($\varepsilon = 8$) **Answer:** Based on your description of feeling abnormal heartbeats, difficulty breathing, and pain in the lower abdomen, it is important to consult a doctor as soon as possible for a proper evaluation. These symptoms could indicate a variety of potential issues such as heart problems, respiratory issues, or gastrointestinal concerns. It is important to have a thorough medical assessment to determine the underlying cause and appropriate treatment. Please seek medical attention promptly for further evaluation and management.

Reference: Hello, Welcome to Chat Doctor forum. I went through your post. Your predominant complaints are palpitations and lower abdomen pain. Any person who notices a change in their heartbeats should go for ECG (electrocardiogram) to rule out arrhythmia. The most common causes of a one-sided pain in the lower abdomen are kidney stone or urinary tract infection. These are only the most common causes. However, there are more than 10 causes of the same. A USG (ultrasonogram) of the abdomen with a urine examination will help you come to a conclusion. Consult a physician with above investigations so he can help you. Regards. For more information consult an internal medicine physician online

I.2 SAMsum

SAMsum

Dialogue: Jake: What are your plans for the day?
Olivia: I haven't really got anything planned. There are some things I should do which I haven't looked at for a while
Jake: Like what?
Olivia: I ought to do my accounts. At least get all the incoming and outgoing invoices sorted by months.
Jake: Tax filing is a long way off.
Olivia: I know, but doing it all at the last minute is a dreadful headache.
Jake: Anything else?
Olivia: I also have a few videos I need to upload to YouTube.
Jake: Already ready or you still need to do post-production?
Olivia: I don't really do much by way of post-production. I'm not exactly Steven Spielberg.
Jake: I do, but I dont always have time to do much.
Olivia: If I put in any music I always get a copyright strike and someone else is taking the ad revenue
Jake: Yeah. And some of those claims are purely speculative. One time I put on one of my own pieces and it was challenged by Sony Music
Olivia: No way!
Jake: It's true. On the form you fill in to counter the claim I said if Sony want to make me one of their signed artists, that's fine. Otherwise to get they paws off of my music."

($\varepsilon = 1$) **Summary:** Olivia doesn't have a lot of plans for the day. She says she should do some accounting and she might upload a few videos to YouTube.
($\varepsilon = 3$) **Summary:** Olivia tells Jake that she should do her accounts, file her tax returns and upload her videos to YouTube. Jake says he has the same problems and that he does postproduction and has trouble with copyright strikes.
($\varepsilon = 8$) **Summary:** Jake and Olivia talk about their plans for the day and what they need to do. Olivia's talking about doing some accounting while Jake mentions post-production and uploading videos to YouTube. Jake says one time he used his own music for a video and Sony took the ad revenue. Olivia says that's not right.

Reference: Olivia has to sort out her accounts and upload a few videos on YouTube. Jake is complaining that Sony Music tried to appropriate his own music.

J KSA output examples

J.1 ChatDoctor iclineq

ChatDoctor iclinq

Question: Hello,I am a 27 year old female. I am feeling my heartbeat, breath, and pulse rate to be really abnormal since the past three weeks. It feels like my heart is pounding so bad. And then, I feel a difficulty to breathe, like I have run or jogged more than my capacity, even if I am sitting or lying. Sometimes, I feel a lot of pain in my lower abdomen area and on the left side of my belly. Please suggest me some remedy.

($\varepsilon = 1$) **Answer:** Please suggest a remedy for your symptoms. It is crucial to assess your symptoms, especially concerning your heartbeats and difficulty breathing. It is best to seek evaluation at a clinic or medical facility to identify the nature of your symptoms and ensure safe and appropriate treatment. Do not hesitate to seek medical attention, as your symptoms could potentially be due to a cardiac disorder or other serious condition. Managing your symptoms and relieving any pain is important to prevent any potential worsening of your condition. It is also important to separate any possible infections or embolism as a cause of your symptoms. It is possible that your symptoms may be related to nicotine use, so it is important to address this as well.
($\varepsilon = 3$) **Answer:** Please seek medical attention immediately. It is important to identify the causes of your symptoms, which may be associated with palpitations, breathing difficulties, and pain in your abdomen. A thorough evaluation, including an ECG, echocardiogram, and possibly other tests, is recommended to determine the severity and appropriate treatment. Do not ignore these symptoms, especially if they are severe or associated with fainting or other concerning signs. It is important to prioritize your health and well-being by seeking help from a doctor or healthcare center.Ignoring these symptoms can lead to serious health issues, so please do not delay in seeking medical help.
($\varepsilon = 8$) **Answer:** It is essential to note that abnormal heart beating, difficulty in breath, and pain in the abdomen area are concerning signs that may indicate a heart-related issue. It is advisable to seek professional attention and start proper treatment. Keeping a separate note of the frequency and various techniques to understand and recommend the appropriate exercises to avoid strain is advisable. It is important to avoid strenuous activities and seek detailed advice from a healthcare professional. If experiencing dizziness or lightheadedness, it is important to seek medical attention promptly.

Reference: Hello, Welcome to Chat Doctor forum. I went through your post. Your predominant complaints are palpitations and lower abdomen pain. Any person who notices a change in their heartbeats should go for ECG (electrocardiogram) to rule out arrhythmia. The most common causes of a one-sided pain in the lower abdomen are kidney stone or urinary tract infection. These are only the most common causes. However, there are more than 10 causes of the same. A USG (ultrasonogram) of the abdomen with a urine examination will help you come to a conclusion. Consult a physician with above investigations so he can help you. Regards. For more information consult an internal medicine physician online

J.2 SAMsum

SAMsum

Dialogue: Jake: What are your plans for the day?
Olivia: I haven't really got anything planned. There are some things I should do which I haven't looked at for a while
Jake: Like what?
Olivia: I ought to do my accounts. At least get all the incoming and outgoing invoices sorted by months.
Jake: Tax filing is a long way off.
Olivia: I know, but doing it all at the last minute is a dreadful headache.
Jake: Anything else?
Olivia: I also have a few videos I need to upload to YouTube.
Jake: Already ready or you still need to do post-production?
Olivia: I don't really do much by way of post-production. I'm not exactly Steven Spielberg.
Jake: I do, but I dont always have time to do much.
Olivia: If I put in any music I always get a copyright strike and someone else is taking the ad revenue
Jake: Yeah. And some of those claims are purely speculative. One time I put on one of my own pieces and it was challenged by Sony Music
Olivia: No way!
Jake: It's true. On the form you fill in to counter the claim I said if Sony want to make me one of their signed artists, that's fine. Otherwise to get they paws off of my music."

($\varepsilon = 1$) **Summary:** Olivia has a lot of things to do, but she doesn't have time to do them. She has to do her accounts, upload videos to YouTube, and edit her videos. She doesn't have time to do post-production. She doesn't want to get a copyright strike. She doesn't want to get a copyright strike. She doesn't want to get a copyright strike. She doesn't want to get a copyright strike. She doesn't want to get a copyright strike. She doesn't want.
($\varepsilon = 3$) **Summary:** Olivia and Jake are talking about their plans for the day. Olivia has some things she needs to do, like sorting out her accounts and uploading some videos to YouTube. Jake says he does some post-production work on his videos, but he doesn't always have time to do much. Olivia says she doesn't do much post-production work because she always gets a copyright strike and someone else is taking the ad revenue. Jake says he has had a copyright strike from Sony Music on one of his own pieces of.
($\varepsilon = 8$) **Summary:** Olivia and Jake are talking about their plans for the day. Olivia has some things she needs to do, like sorting out her accounts and uploading some videos to YouTube. Jake says he does some post-production work on his videos, but he doesn't always have time to do much. Olivia says she doesn't do much post-production work because she always gets a copyright strike and someone else is taking the ad revenue. Jake says he has had a copyright strike from Sony Music on one of his own pieces of.

Reference: Olivia has to sort out her accounts and upload a few videos on YouTube. Jake is complaining that Sony Music tried to appropriate his own music.

K Information About Use Of AI Assistants

In the writing process, we use an AI assistant to fix grammar and typos in the manuscript.