A Cautionary Tale on the Evaluation of Differentially Private In-Context Learning

Anjun Hu
Department of Engineering Science
University of Oxford
anjun.hu@eng.ox.ac.uk

Jiyang Guan
Institute of Automation
Chinese Academy of Sciences
guanjiyang2020@ia.ac.cn

Philip Torr
Department of Engineering Science
University of Oxford
philip.torr@eng.ox.ac.uk

Francesco Pinto
Data Science Institute
University of Chicago
fpinto1@uchicago.edu

Abstract

In-context learning (ICL) has emerged as a powerful paradigm enabling Large Language Models (LLMs) to perform new tasks by prompting them with few training examples, eliminating the need for fine-tuning. Given its potential to adapt and personalize the model's behaviour using private user data, recent studies have introduced techniques for ICL that satisfy Differential Privacy guarantees (DP ICL). Existing DP ICL approaches claim to attain such guarantees while maintaining negligible utility degradations when adapting the models to perform new tasks. In this paper, we present preliminary empirical evidence suggesting that these claims may hold only for tasks aligned with the model's pre-training knowledge and biases. We do so by showing the performance of DP ICL significantly degrades with respect to the non-private counterpart in scenarios that introduce tasks and distribution shifts that challenge the model's prior knowledge. To mitigate the risk of overly optimistic evaluations of DP ICL, we invite the community to consider our sanity checks to attain a more accurate understanding of its capabilities and limitations.

1 Introduction

In-Context Learning (ICL) [1] has recently emerged as a novel paradigm to leverage the long-context understanding capabilities of modern Large Language Models (LLMs) in order to instruct them to perform novel tasks without additional fine-tuning. The idea is to prompt them with a sequence of input-output examples that demonstrate the task to be performed and induce it to infer the correct output on a previously unseen input sample. Given ICL is computationally inexpensive in comparison to other forms of learning, several works have proposed to use it to personalise and adapt the LLM behaviour on user data [1–3]. However, it is also demonstrated that LLMs may regurgitate information from the in-context demonstrations, leading to the unintended leakage of such data. To tackle this issue, several works have proposed Differentially Private (DP) ICL algorithms [4–9]. These works claim to provide DP guarantees while maintaining ICL task performance at a level that is comparable to the non-private baseline. This may apparently contradict the literature that has consistently observed DP algorithms to require more data in order to attain similar levels of performance [10, 11].

In this work, we design a set of regression and classification tasks that aim at developing a more nuanced understanding of the factors contributing to the success of DP ICL, and outline possible

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

conditions for its failure. By generating tasks for which the feature-target mappings contradict the model's pre-training knowledge and biases (e.g., with flipped label ICL [12]), we find that DP-ICL may fail to match the non-private counterpart or show little improvement over zero-shot performance. This observation aligns with the findings of the semi-private learning literature, which relies on the availability of additional public data (either for pre-training or fine-tuning) exhibitin some level of similarity with respect to the private one in order to circumvent the expected performance degradation [13–15]. Therefore, our contributions are as follows:

- We demonstrate the impact of the alignment between task feature-label mapping and the LLM's pre-training knowledge on the performance gap between DP-ICL and the public counterparts. We highlight that, while state-of-the-art DP-ICL techniques show marginal utility degradation when assessed on tasks that leverage the LLM's pre-training knowledge, they can fail when the downstream tasks do not align with it.
- Drawing on empirical evidence, we propose several test scenarios that can act as sanity checks for a more practical and thorough evaluation of DP-ICL methodologies. These suggestions aim to create a more comprehensive evaluation framework to determine the capabilities and limitations of DP-ICL techniques.

2 Related Works: Differentially Private In-Context learning (DP ICL)

2.1 Differential Privacy

Differential Privacy (DP) [16, 17] upper bounds the likelihood an attacker can reliably infer the membership of a sample to the input set of a randomised algorithm. Informally, this is attained by limiting and obfuscating the impact any single data sample can have on its output. This allows for the protection of private information contained in individual data points, while still allowing to extract distributional trends.

In particular, any randomised algorithm $\mathcal{M}$ is said to satisfy (ϵ, δ)-Differential Privacy guarantees if for any pair of neighbouring datasets D and D' differing by at most one element, and for any potential output $S \subseteq \text{Range}(\mathcal{M})$, it holds that:

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

Where ϵ is the privacy budget. A smaller value implies stronger privacy protection. The scenario in which $\delta = 0$ is referred to as pure DP or $(\epsilon, 0)$-DP) pure DP guarantees that (assuming perfect numerical precision [18, 19]), while δ in (ϵ, δ)-DP introduces a small probability δ of the mechanism failing to preserve ϵ-DP.

2.2 Differentially Private In-Context Learning

Various studies have demonstrated that LLMs can memorize and regurgitate information contained in their training data [20–26], and in-context learning (ICL) demonstration sets [8, 27, 28, 20], leading to unintended leakage of private data [21, 26, 28]. In response to these privacy concerns, several methods have been introduced to achieve *DP ICL*, with varying levels of assumptions regarding the trustworthiness of LMs. We separate them in two categories depending on where in the ICL procedure the DP algorithm is applied.

DP Inference. The first category of algorithms consists of *DP learning* algorithms, which assume the presence of a trusted LM (the context data being fed to the LM is not required to be DP) and propose inference or post-processing methods to ensure that the model's output is DP-compliant for end users. Notable examples in this category include `DP-ICL` [9] which generates DP responses through a noisy consensus among an ensemble of LLM's responses based on disjoint exemplar sets; `PromptPATE` [8] which uses an ensemble of teacher models to privately generate labelled data for training a student model's prompts and Prompt-`DPSGD` [8] which directly applies differentially private stochastic gradient descent to update the prompts during training.

DP Context Synthesis. The second category involves *DP data-synthesis* algorithms, which impose stricter safety constraints by assuming that LMs are accessed through third-party APIs and, therefore, should not be trusted. In this case, the inputs to the LMs are required to be DP-compliant, and these methods focus on the construction of a DP-compliant few-shot demonstration set for ICL inference. Under this category, `DP-OPT` [4] generates prompt tokens using a limited-domain mechanism and selects the best prompt using an exponential mechanism whereas `DP-FSG` [5] leverages an auxiliary LM to generate DP-compliant pseudo-examples. More recently, inspired by traditional DP methods for tabular data, `DP-TabICL` [6] explores the application of both local and global DP techniques for ICL using semi-structured natural language data derived from tabular features. `DP-LLMTGen` [7] offers a novel framework for generating differentially private tabular data with LLMs.

Limitations of current DP ICL evaluations. Existing DP-ICL methodologies claim to respect DP guarantees while maintaining utility comparable to their non-private counterpart. However, the evaluation of these methods is flawed for several reasons:

1. They are usually only evaluated on coarse-grained classification tasks (e.g. [29–31]) that LLMs can perform relying solely on pre-training knowledge and biases, without genuine learning from ICL samples. This obfuscates the negative impact of DP on utility.
2. Evaluations are limited to well-known, simplistic, and open-source classification tasks and datasets which are likely included in the LM's pre-training data [32, 33]. Therefore it is unclear whether the claimed guarantees actually hold.

To address these issues and develop a more comprehensive evaluation framework for DP ICL, we draw inspiration from flipped-label ICL (FL-ICL) and semantically unrelated label ICL (SUL-ICL) [34] tasks to examine the effect of LM pre-training bias on privacy-utility trade-offs.

3 Experiments and Evaluation

To identify the limitations of current DP ICL and thoroughly assess their performance, we have designed a series of challenging test scenarios that extend beyond commonly used NLP datasets (such as DBPedia [29] and AGNews [30]).

In order to maintain precise control of the feature distributions and their relationship with the labels, we focus on natural language data that is generated starting from tabular data. The tabular data is crafted following specific rules for each task detailed in the remainder of this section and then converted into natural language using fixed, non-data-dependent templates specified in Appendix A. In the same appendix, we also provide additional details regarding the prompt construction procedures.

In our experiments, we evaluate three DP-ICL methods: `PATE-CTGAN` [35], DP Few-Shot Generation (`DP-FSG`) [5], and `DP-OPT` [4]. We use the pure DP ($(\epsilon, 0)$-DP) variant of `DP-FSG` (report noisy max with exponential mechanism). We use $N = 2000$ samples to train `PATE-CTGAN` in each scenario and use a default $\delta = N^{-1.5} = 1.11 \times 10^{-5}$ [36]. For `DP-OPT`, we use the default $\delta = 5 \times 10^{-7}$ These methods are tested across various ϵ budgets and across various scenarios using two LLMs, namely GPT-4o [37] and Claude-3.0-Haiku [38].

3.1 Binary Classification with Varied Feature–Label Mappings.

To investigate how DP-ICL performance is affected when the model encounters distributions that contradict its pre-existing knowledge and biases [39], we have designed two binary classification tasks as follows:

Task 1: Gender-Product Category Preferences The objective of the first task is to predict whether a customer is interested in a product advertisement. This prediction is based on two binary features: (1) gender $g \in${`Male`, `Female`} (2) product category $p \in${`fashion and beauty`, `electronics and gadgets`}. Data is structured as $\mathcal{X} = \{(g_i, p_i, y_i)\}_{i=1}^{N}$. The outcome of this task is a binary classification indicating whether the customer is interested or not interested in the product $y \in \{0, 1\}$. This task includes two variants:

- Expected: This variant aligns with traditional gender stereotypes, assuming women are more likely to be interested in fashion and beauty products while men are more interested in electronics and gadgets. Results correspond to blue lines in the figures.

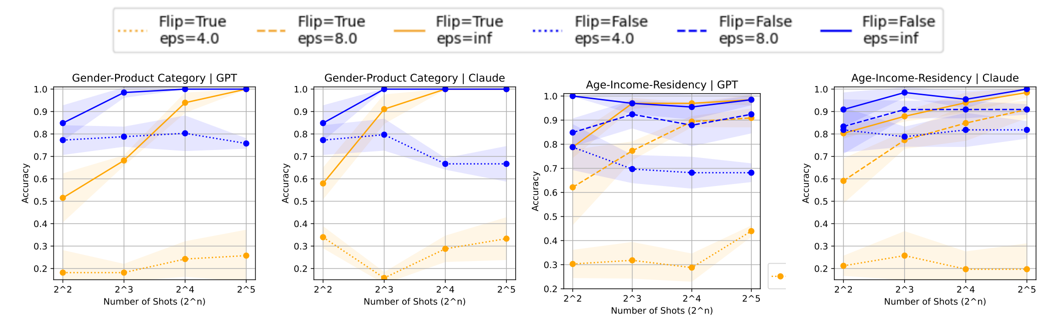

Figure 1: Performance of the two binary classification tasks introduced in Sec 3.1 evaluated under varying ϵ budget, number of demonstration samples and feature-output mapping. From left to right: Gender-Product Category Preferences prediction from (a) GPT and (b) Claude; Income-Age-Residency prediction for (c) GPT and (d) Claude. Blue curves represent expected scenarios which conform to LMs' pre-training biases. Orange curves represent flipped-label tasks. Performance gaps between DP and non-DP orange curves are significantly larger than those between the blue curves, suggesting that flipped-label or OOD tasks experience more significant utility-privacy trade-offs.

Scenario	Model	$\epsilon = \infty$ N=8; N=0	$\epsilon = 4.0$ PATE-CTGAN [35]	DP-FSG [5]	DP-OPT [4]	GaussianNB
Expected	GPT-4o	0.97 ± 0.04 0.83 ± 0.06	0.70 ± 0.06 **-0.27**	0.60 ± 0.07 **-0.37**	0.77 ± 0.04 **-0.20**	0.98 ± 0.00 **+0.01**
	Claude	0.98 ± 0.02 0.74 ± 0.06	0.79 ± 0.04 **-0.19**	0.62 ± 0.16 **-0.32**	0.50 ± 0.00 **-0.35**	
Reversed	GPT-4o	0.97 ± 0.02 0.18 ± 0.07	0.32 ± 0.07 **-0.65**	0.47 ± 0.08 **-0.50**	0.17 ± 0.11 **-0.82**	0.82 ± 0.01 **-0.15**
	Claude	0.88 ± 0.02 0.18 ± 0.07	0.26 ± 0.11 **-0.62**	0.30 ± 0.16 **-0.58**	0.15 ± 0.05 **-0.73**	

Table 1: Accuracy (↑) of 8-shot Age-Income-Residency Classification under $\epsilon = 4$. **Bold numbers** are performance differences between each DP-ICL method and non-DP $\epsilon = \infty$ baselines. Columns in grey are non-ICL traditional ML methods that serve to indicate the quality of labels. Metrics in smaller font sizes are zero-shot baselines that indicates model's pretraining biases.
Notably, the performance gaps between DP and non-DP methods are significantly more pronounced for unexpected or OOD feature-label mapping compared to expected or in-distribution counterparts. In both tables, higher performance gap implies a more significant utility loss as a result of DP.

- Reversed: This variant challenges gender stereotypes, assuming that women are more likely to be interested in electronics and gadgets, while men are more interested in fashion and beauty products. Results correspond to orange lines in the figures.

Task 2: Age-Income-Residency Classification The objective of the second task is to predict whether an individual resides in $y \in$ {`Massachusetts`, `Louisiana`}, a binary target indicating the state of residence. This prediction is based on two continuous numerical features: age $a \in (18, 80)$ and annual income $m \in (15k, 100k)$. Data is structured as $\mathcal{X} = \{(a_i, m_i, y_i)\}_{i=1}^{N}$. The two data clusters are linearly separable and the ground truth decision boundary is linear. Similar to the first task, this task includes two variants with varied feature–label mappings.

- Expected: This variant reflects real-world economic disparities between the states, assuming that individuals in Massachusetts have a higher income than those in Louisiana.
- Reversed: This variant reverses the expected feature-label mapping, assuming that individuals in Louisiana have a higher income than those in Massachusetts.

Our analysis of the experimental results reveals several key findings.

Good DP-ICL performance derives from distributional alignment between the pretraining and ICL distributions. Following known trends, for both expected and reversed scenarios, stricter

DP constraints ($\epsilon \downarrow$) generally lead to worse performance. When comparing scenarios with equivalent privacy constraints (represented by the same line type), we note a *different behaviour* in the performance of the expected versus reversed cases. In the few-shot regime, the reversed mappings consistently underperform compared to their counterparts in the expected case. As the number of examples (shots) increases, the performance gap between reversed and expected mappings tends to narrow. This demonstrates DP ICL may introduce significant performance degradation even for loose privacy guarantees ($\epsilon = 4$) when the ICL distribution does not align with the pre-training knowledge of the model.

Handling strong distribution shifts requires more data. We observe an interesting interplay between the strictness of DP constraints and the convergence of performance between reversed and expected label mappings. Notably, the performance gaps between DP and non-DP methods are significantly more pronounced in reversed scenarios (when evaluating with unexpected or out-of-distribution feature-label mapping) compared to expected or in-distribution counterparts. This cautions against expecting ICL to be effective in low-data regimes: positive results in such situations are likely because the model's pre-training aligns well with the specific task at hand.

As DP constraints become more stringent, a larger number of examples (shots) is required for the performance of flipped label mappings to approach that of expected label mappings.

These observations suggest that when evaluating the efficacy of DP-ICL methods one needs to take into consideration how the alignment of the task with pre-existing biases or expectations in the model affects the metrics. This may hint that the illusion of "DP for free" is only applicable to cases where little or no genuine learning occurs with respect to the in-context examples and the LM relies on pre-training bias for ICL inference. To thoroughly evaluate a DP-ICL methodology, we strongly recommend considering different feature-label mappings (testing scenarios that align or reverse pre-training biases) to provide a more complete understanding of the method's performance and limitations.

3.2 eICU Lab-to-Survival Binary Classification

Building upon our analysis of binary classification tasks, we extend our evaluation to a real-world clinically relevant dataset, eICU [40], performing a Lab-to-Survival binary prediction task. This task aims to predict ICU patient survival binary outcomes $y \in \{\texttt{Expired}, \texttt{Alive}\}$ based on patient demographics $\mathbf{d}_i$ (`age, gender, ethnicity, height, weight`) and 20 continuous real-valued lab test results. The goal is to evaluate the impact of DP on tasks with varied levels of pretraining knowledge reliance. For this task, we have explored three levels of elicitation of the model's pretraining knowledge to more extensively study the relationship between pretraining knowledge reliance and the DP utility gap.

Experimental Setup. To explore how pretraining knowledge affects DP utility, we devised three prompting formats with increasing reliance on pretraining knowledge:

- **Original Prompting:** Features were presented with their original clinical names and units, formatted alongside demographic information. Outputs were verbalized as `Expired` or `Alive`.
- **Pseudonym Prompting:** Demographic and clinical features were obfuscated with pseudonyms and presented without units. Outputs were also pseudonymised (e.g., `ir4cowgz` and `rixa11dp`).
- **Chain-of-Thought (CoT) Prompting:** The original format was augmented with a step-by-step reasoning process encouraging the model to utilize its pretraining knowledge extensively.

Each setting was evaluated under privacy budgets of $\epsilon = \infty$ (non-DP), 8.0, and 4.0 actualised with `PATE-CTGAN` [35] with varying numbers of shots N ranging from $N = 0$ to 32.

Results. Our findings align with the hypothesis that tasks relying more on pretraining knowledge exhibit smaller DP-non DP performance gaps:

- **Original Prompting:** Accuracy decreased modestly with stricter privacy budgets, particularly in higher-shot settings. This indicates that pretraining knowledge buffers the effects of DP.

Table 2: Accuracy of eICU Lab-to-Survival Binary Prediction Task for GPT-4o under under various DP budgets (ϵ) and Number of Shots N. Performance degradation from $\epsilon = \infty$ is shown as Δ.

Prompting Variant	**Shots**	$\epsilon = \infty$	$\epsilon = 8.0$	$\Delta_{\epsilon=8.0}$	$\epsilon = 4.0$	$\Delta_{\epsilon=4.0}$
Pseudonym	0	0.5000				
	8	0.7500	0.6451	−0.1049	0.4758	−0.2742
	16	0.7258	0.5605	−0.1653	0.4758	−0.2500
	32	0.6613	0.5484	−0.1129	0.4112	−0.2501
Original	0	0.7150				
	8	0.7258	0.7016	−0.0242	0.6774	−0.0484
	16	0.7984	0.7273	−0.0711	0.7258	−0.0726
	32	0.8181	0.8065	−0.0116	0.7273	−0.0908
CoT	0	0.7564				
	8	0.7419	0.7200	−0.0219	0.6408	−0.1011
	16	0.7419	0.7273	−0.0146	0.6591	−0.0828
	32	0.7339	0.7143	−0.0196	0.6154	−0.1185

Table 3: Accuracy of eICU Lab-to-Survival Binary Prediction Task for Claude-3 under various DP budgets (ϵ) and Number of Shots N. Performance degradation from $\epsilon = \infty$ is shown as Δ. The model used for inference was `claude-3-haiku-20240307`.

Prompting Variant	**Shots**	$\epsilon = \infty$	$\epsilon = 8.0$	$\Delta_{\epsilon=8.0}$	$\epsilon = 4.0$	$\Delta_{\epsilon=4.0}$
Pseudonym	0	0.5000				
	8	0.6290	0.5484	−0.0806	0.5323	−0.0967
	16	0.6451	0.5323	−0.1128	0.4919	−0.1532
	32	0.6463	0.5000	−0.1463	0.4355	−0.2108
Original	0	0.7091				
	8	0.7661	0.7016	−0.0645	0.7420	−0.0241
	16	0.7500	0.6935	−0.0565	0.7258	−0.0837
	32	0.8409	0.8086	−0.0323	0.7623	−0.0786
CoT	0	0.6818				
	8	0.6813	0.6364	−0.0449	0.6290	−0.0523
	16	0.7045	0.6048	−0.0997	0.6613	−0.0432
	32	0.6500	0.5800	−0.0700	0.5385	−0.1115

- **Pseudonym Prompting:** Accuracy suffered significantly under DP constraints, with the largest drops observed at $\epsilon = 4.0$. This supports the hypothesis that genuine learning from context is more affected by DP.
- **CoT Prompting:** The inclusion of reasoning steps mitigated the impact of DP constraints, resulting in a smaller performance gap compared to Pseudonym Prompting.

These results reinforce the importance of considering pretraining knowledge alignment when evaluating DP ICL methods, as performance degradation is notably more pronounced in tasks requiring genuine context-driven learning.

4 Conclusions.

Our findings underscore the importance of task-specific evaluation and caution against overly broad claims about DP-ICL performance. We provide insights into the factors influencing the privacy-utility trade-off in different contexts and propose guidelines for more nuanced reporting of DP-ICL results. This work contributes to a more realistic understanding of the capabilities and limitations of DP-ICL, paving the way for future research in privacy-preserving machine learning techniques. We aim for our work to enhance the protection of sensitive user data in real-world applications like personalized healthcare and finance, fostering the responsible implementation of machine learning systems that effectively balance utility and privacy.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[2] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning, 2024. URL `https://arxiv.org/abs/2301.00234`.

[3] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*, 2021.

[4] Junyuan Hong, Jiachen T Wang, Chenhui Zhang, Zhangheng Li, Bo Li, and Zhangyang Wang. Dp-opt: Make large language model your privacy-preserving prompt engineer. *arXiv preprint arXiv:2312.03724*, 2023.

[5] Xinyu Tang, Richard Shin, Huseyin A Inan, Andre Manoel, Fatemehsadat Mireshghallah, Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, and Robert Sim. Privacy-preserving in-context learning with differentially private few-shot generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=oZtt0pRnOl`.

[6] Alycia N. Carey, Karuna Bhaila, Kennedy Edemacu, and Xintao Wu. Dp-tabicl: In-context learning with differentially private tabular data, 2024. URL `https://arxiv.org/abs/2403.05681`.

[7] Toan V Tran and Li Xiong. Differentially private tabular data synthesis using large language models. *arXiv preprint arXiv:2406.01457*, 2024.

[8] Haonan Duan, Adam Dziedzic, Nicolas Papernot, and Franziska Boenisch. Flocks of stochastic parrots: Differentially private prompt learning for large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[9] Tong Wu, Ashwinee Panda, Jiachen T. Wang, and Prateek Mittal. Privacy-preserving in-context learning for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=x4OPJ7lHVU`.

[10] Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *J. Mach. Learn. Res.*, 12(null):1069–1109, jul 2011. ISSN 1532-4435.

[11] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473, 2014. doi: 10.1109/FOCS.2014.56.

[12] Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. *arXiv preprint arXiv:2110.05679*, 2021.

[13] Francesco Pinto, Yaxi Hu, Fanny Yang, and Amartya Sanyal. PILLAR: How to make semi-private learning more effective. In *2nd IEEE Conference on Secure and Trustworthy Machine Learning*, 2024. URL `https://openreview.net/forum?id=Ps1IHhzx4Z`.

[14] Milad Nasr, Saeed Mahloujifar, Xinyu Tang, Prateek Mittal, and Amir Houmansadr. Effectively using public data in privacy preserving machine learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 25718–25732. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/nasr23a.html`.

[15] Noga Alon, Raef Bassily, and Shay Moran. *Limits of private learning with access to public data*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[16] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Theory of Cryptography Conference (TCC)*, pages 265–284. Springer, 2006. doi: 10.1007/11681878_14. URL `https://doi.org/10.1007/11681878_14`.

[17] Cynthia Dwork and Aaron Roth. *The Algorithmic Foundations of Differential Privacy*, volume 9 of *Foundations and Trends in Theoretical Computer Science*. Now Publishers Inc., 2014. doi: 10.1561/0400000042. URL `https://www.nowpublishers.com/article/Details/TCS-042`.

[18] Jiankai Jin, Eleanor McMurtry, Benjamin IP Rubinstein, and Olga Ohrimenko. Are we there yet? timing and floating-point attacks on differential privacy systems. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 473–488. IEEE, 2022.

[19] Johan Lokna, Anouk Paradis, Dimitar I Dimitrov, and Martin Vechev. Group and attack: Auditing differential privacy. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 1905–1918, 2023.

[20] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *arXiv:2306.11698*, 2024.

[21] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*, 2022.

[22] Daphne Ippolito, Florian Tramèr, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher A Choquette-Choo, and Nicholas Carlini. Preventing verbatim memorization in language models gives a false sense of privacy. *arXiv preprint arXiv:2210.17546*, 2022.

[23] Eugene Kharitonov, Marco Baroni, and Dieuwke Hupkes. How bpe affects memorization in transformers. *arXiv preprint arXiv:2110.02782*, 2021.

[24] R Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. How much do language models copy from their training data? evaluating linguistic novelty in text generation using raven. *Transactions of the Association for Computational Linguistics*, 11: 652–670, 2023.

[25] Fatemehsadat Mireshghallah, Archit Uniyal, Tianhao Wang, David Evans, and Taylor Berg-Kirkpatrick. Memorization in nlp fine-tuning methods. *arXiv preprint arXiv:2205.12506*, 2022.

[26] Francesco Pinto, Nathalie Rauschmayr, Florian Tramèr, Philip Torr, and Federico Tombari. Extracting training data from document-based vqa models. *ICML*, 2024.

[27] Haonan Duan, Adam Dziedzic, Mohammad Yaghini, Nicolas Papernot, and Franziska Boenisch. On the privacy risk of in-context learning. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

[28] Aman Priyanshu, Supriti Vijay, Ayush Kumar, Rakshit Naidu, and Fatemehsadat Mireshghallah. Are chatbots ready for privacy-sensitive applications? an investigation into input regurgitation and prompt-induced sanitization. *arXiv preprint arXiv:2305.15008*, 2023.

[29] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195, 2015. URL `http://dblp.uni-trier.de/db/journals/semweb/semweb6.html#LehmannIJJKMHMK15`.

[30] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification, 2016. URL `https://arxiv.org/abs/1509.01626`.

[31] Ellen M Voorhees and Dawn M Tice. Building a question answering test collection. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 200–207, 2000.

[32] Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*, 2024. URL `https://arxiv.org/abs/2404.18824`.

[33] Inbal Magar and Roy Schwartz. Data contamination: From memorization to exploitation. *arXiv preprint arXiv:2203.08242*, 2022.

[34] Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023.

[35] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=S1zk9iRqF7`.

[36] OpenDP SmartNoise Team. Smartnoise synthesizers, 2024. URL `https://pypi.org/project/smartnoise-synth/`. Version 1.0.4, MIT License.

[37] OpenAI. Gpt-4 technical report, 2023.

[38] Anthropic. Claude 3.0: Advancements in conversational ai, 2024. URL `https://www.anthropic.com/product/claude`. Accessed: 2024-09-13.

[39] Amartya Sanyal, Yaxi Hu, and Fanny Yang. How unfair is private learning ? In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022. URL `https://openreview.net/forum?id=H2V43wIj5g9`.

[40] Tom J Pollard, Alistair E W Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

A Further Details on Prompt Construction

A.1 Gender-Product Category Preferences

For the Gender-Product Category Preferences Binary Classification task, prediction is based on two binary features: (1) gender g_i ∈{`M`, `F`} (2) product category p_i ∈{`fashion and beauty`, `electronics and gadgets`}. The outcome of this task is a binary classification indicating whether the customer is interested or not interested in the product $y_i \in \{0, 1\}$.

The Global Prefix is:

```
You pay attention to how one's gender affects one's interest in certain
product categories.
Based on your observations, predict whether an is interested or not
interested. Answer in at most two words.
```

Each demonstration is formatted as: " `I am` g_i. `I am looking at a product from the` c_i `category. I am` y_i." where:

$$g_i \in \{\text{a man},\ \text{a woman}\}$$
$$p_i \in \{\text{fashion and beauty},\ \text{electronics and gadgets}\}$$
$$y_i \in \{\text{interested},\ \text{not interested}\}$$

A.2 Age-Income-Residency

For the Age-Income-Residency Binary Classification task, the dataset is structured as $\mathcal{X} = \{(a_i, m_i, y_i)\}_{i=1}^{N}$. The objective is to predict whether an individual resides in y_i ∈{`Massachusetts`, `Louisiana`}, a binary target indicating the state of residence. This prediction is based on two continuous numerical features: age $a_i \in (18, 80)$ and annual income $m_i \in (15k, 100k)$.

The data are generated using `sklearn.datasets.make_classification`, then rotated by 45 degrees, then scaled to $a_i \in (18, 80)$ and $m_i \in (15k, 100k)$

```
from sklearn.preprocessing import MinMaxScaler
from sklearn.datasets import make_classification

X, y = make_classification(n_samples=2000, n_features=2, n_redundant=0,
               n_informative=2, random_state=1, n_clusters_per_class=1)

# 45 degree rotation to make both features relevant
theta = np.pi / 4
rotation_matrix = np.array([[np.cos(theta), -np.sin(theta)],
                            [np.sin(theta),  np.cos(theta)]])
rng = np.random.RandomState(2)
X += rng.uniform(size=X.shape)
X = X.dot(rotation_matrix)
linearly_separable = (X, y)

# rescale features to reasonable ranges
scaler_feature_1 = MinMaxScaler(feature_range=(15, 100))
scaler_feature_2 = MinMaxScaler(feature_range=(18, 80))

def rescale_features(X):
  X_rescaled = np.copy(X)
  X_rescaled[:, 0] = scaler_feature_1.fit_transform(X[:, [0]]).flatten()
  X_rescaled[:, 1] = scaler_feature_2.fit_transform(X[:, [1]]).flatten()
  return X_rescaled

rescaled_datasets = [(rescale_features(X), y) for X, y in datasets]
```

The Global Prefix is:

```
You pay attention to how one’s age and income are correlated with their state
of residence.
Based on your observations, predict whether an individual’s state of residence
is Massachusetts or Louisiana. Answer in one word.
```

Each demonstration is formatted as: “ `I am` a_i years old. `I am looking at a product from the` c_i `category. I am` y_i.” where:

$$
\begin{aligned}
a_i &\in (18,\ 80) \\
m_i &\in (15000,\ 100000) \\
y_i &\in \{\text{Massachusetts},\ \text{Louisiana}\}
\end{aligned}
$$