Differentially Private In-Context Learning

Ashwinee Panda* Princeton University ashwinee@princeton.edu

Jiachen T. Wang* Princeton University tianhaowang@princeton.edu

Tong Wu* Princeton University tongwu@princeton.edu

Prateek Mittal Princeton University pmittal@princeton.edu

Abstract

An important question in deploying large language models (LLMs) is how to augment LLMs with private data. We propose Differentially Private In-context Learning (DP-ICL) to enable LLMs to adapt to new tasks while maintaining privacy guarantees. DP-ICL performs private inference by establishing a noisy consensus over an ensemble of exemplars using the Report-Noisy-Max mechanism. We evaluate DP-ICL on four benchmarks and find that it achieves comparable performance (< 2% degradation) with non-private ICL.

1 Introduction

Large language models (LLMs) (Brown et al., 2020; Zhang et al., 2022a) pretrained on large amounts of publicly available data have achieved widespread commercial success, partly in closed-source applications (OpenAI, 2023a; Ganguli et al., 2023). However, LLMs have been shown to memorize their training data (Carlini et al., 2019, 2021; Biderman et al., 2023). Organizations have become wary of using LLMs with private data, even going so far as to ban the use of LLMs outright in contexts where sensitive data leakage is a liability (McCallum, 2023; Bloomberg, 2023). Today the question of how to augment LLMs with private data remains an important research problem (Liu, 2022).

Contributions. In this paper, we propose *Differentially Private In-Context Learning* (DP-ICL), the first framework that can be used to augment state-of-the-art LLM APIs with private data using differential privacy. Our method harnesses an emergent ability of LLMs: *In-Context Learning* (ICL), the capability to rapidly adapt to new tasks using only a few exemplars without updating model parameters (Brown et al., 2020; Min et al., 2022). We detail our method in Fig. 1 and Section 4. A key insight into the design of DP-ICL is that we output the noisy consensus of an ensemble of exemplars, using differential privacy to provide a provable guarantee that the query answers do not leak too much information about the private data.

Although a number of methods have been proposed for augmenting LLMs with private data, they are incompatible with the latest generation of API-only LLMs (GPT-4, Claude, Bard) because they require the model to be open-source (Li et al., 2022; Yu et al., 2022; Bu et al., 2022; He et al., 2023). DP-ICL is the first method that can augment API-only LLMs with private data because it requires only black-box access to a cloud-hosted LLM.

We evaluate DP-ICL on **SST-2**, **Amazon**, **AG-News**, and **TREC** and report that it achieves performance comparable to non-private ICL and surpasses SOTA DP LLM methods. In particular, for a privacy budget of $\varepsilon = 3$, DP-ICL obtains a **1.20% improvement** (i.e., **over 20% relative error rate reduction**) over the best results from prior work on SST-2.

Overall, our research offers a promising approach for utilizing advanced black-box LLMs to adapt to new tasks while upholding strong privacy guarantees. We envision that our DP-ICL framework will serve as a starting point for further exploration into the private use of data in the era of foundation models (Bommasani et al., 2021).

2 Threat Model

We present a threat model for private prediction using in-context learning (ICL), illustrated in Fig.1. In this scenario, an organization possesses private data stored in a database and hosts large language models (LLMs) via an API endpoint, allowing users to query the LLM for answers based on the private data. ICL has gained traction for various applications (Liu, 2022) as it provides an efficient alternative to fine-tuning for both open-source and API-only models, such as GPT-4, enabling them to address external questions on sensitive data.

*Equal Contribution, alphabetical Order.

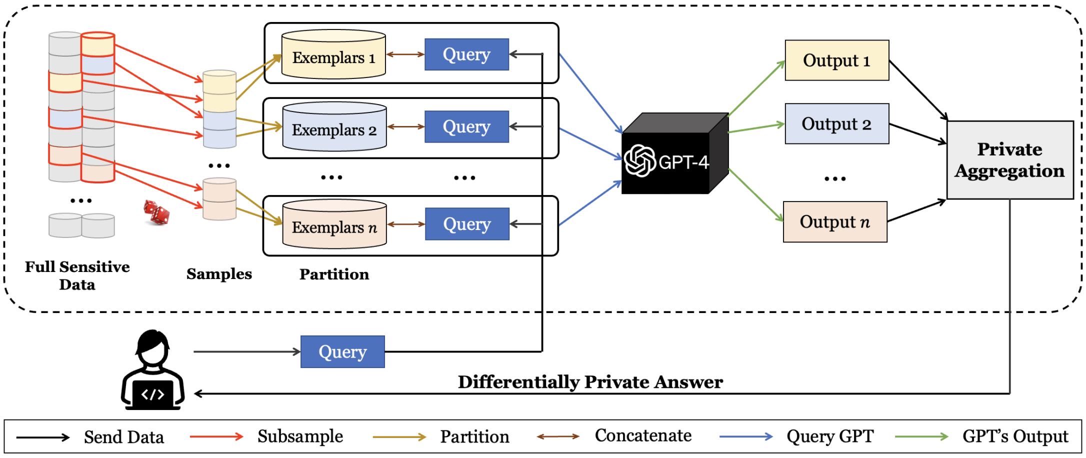

Figure 1: A summary of our framework: We partition the subsampled sensitive database into separate subsets, each comprising a collection of exemplars. When a user requests information (e.g., asking questions) from a large language model (e.g., GPT), the query is augmented with all exemplars formatted accordingly. The model then processes each exemplar-query pair and generates corresponding outputs. These outputs are aggregated by a differentially private mechanism and returned to the user.

The organization aims to maximize the utility of the API by delivering accurate responses to user queries while protecting the privacy of sensitive data. In this threat model, we assume that potentially malicious users may attempt to extract sensitive information from the LLM by exploiting its knowledge of private data. Each user has a finite number of queries, that can be unrestricted in content and can only observe the API's output. We assume that the users do not collude. Consequently, an attacker can employ various attack vectors, including but not limited to prompt leaking (Perez and Ribeiro, 2022). Given the rapidly evolving nature of LLM attacks, a priority for designing defenses is to ensure they are *future-proof*. In other words, an organization seeking a solution to implement alongside their private data-augmented LLM should not have to worry about the defense being easily overcome by an adaptive attacker.

We argue that the private querying system and attacker described in our threat model capture the essential security considerations for real-world deployments. The programmatic separation between analyst and query-answerer is implemented in the federated DP deployments surveyed in Garrido et al. (2022); Sarathy et al. (2023). In our system, ICL serves as a tool for answering private queries while maintaining data privacy. The attacker's abilities and limitations are also realistic: API rate limits stem from hardware constraints, and since organizations store user queries for security purposes (OpenAI, 2023b), attackers cannot expect unlimited attempts to extract private information.

3 Preliminaries

We present an overview of in-context learning and differential privacy. We defer the full details of DP to Appendix A.

In-Context Learning. To answer a query Q with ICL, we concatenate a sequence of k exemplars (i.e., query-answer pairs) $S := ((Q_1, A_1), (Q_2, A_2), \ldots, (Q_k, A_k))$ to Q using an appropriate format. We then employ the large language model to infer the next token (class) via $\operatorname{argmax}_A \textbf{LLM}(A|S+Q)$, where $+$ denotes concatenation. Intuitively, exemplars assist the language model in identifying the relevant mapping between (Q, A), which substantially enhances performance compared to directly querying test data, also known as zero-shot learning.

Differential Privacy. We use $D, D' \in \cup_{n\in\mathbb{N}}\mathcal{X}^n$ to denote two datasets with an unspecified size over space $\mathcal{X}$. We call two datasets D and D' *adjacent* (denoted as $D \sim D'$) if we can construct one by *replacing* one datapoint from the other. Note that the notion of DP under replacement is stronger than DP under addition/removal, because replacing a datapoint is equivalent to removing a datapoint and adding another (Dwork et al., 2014).

Definition 1 (Differential Privacy (Dwork et al., 2006b))**.** *For* $\varepsilon, \delta \geq 0$*, a randomized algorithm* $\mathcal{M} : \mathrm{MultiSets}(\mathcal{X}) \to \mathcal{Y}$ *is* (ε, δ)-differentially private *if for every pair of adjacent datasets* $D, D' \in \mathrm{MultiSets}(\mathcal{X})$ *and for every subset of* *possible outputs $E \subseteq \mathcal{Y}$,*

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

where the randomness is over the coin flips of $\mathcal{M}$.

Differential privacy requires that for all adjacent datasets D, D', the output distribution $\mathcal{M}(D)$ and $\mathcal{M}(D')$ are close, where the closeness is measured by the parameters ε and δ.

4 Differentially Private ICL

We introduce a novel framework for in-context learning with private data. We collect answer votes from a sampled population of exemplars via ICL, and then apply the Report-Noisy-Max mechanism to release a DP (noisy) estimate of the most likely answer. Our method is detailed in Fig. 1 and Alg. 1.

Algorithm 1 Differentially Private In-Context Learning

Require: Private data D, query Q, model **LLM**, noise σ, number of subsets N, subsampling q
  Subsample $q\%$ of the data.
  Partition $D_1, D_2, \ldots, D_N \leftarrow D$.
  for $i \in \{1, \ldots, N\}$ **do**
    Form exemplar-query pair $D_i^Q = D_i \cup \{Q\}$.
    Obtain model output $o_i(Q) = \textbf{LLM}(D_i^Q)$.
    Convert $o_i(Q)$ to a one-hot vector with a length equal to the number of classes.
  end for
  Sum the one-hot vectors into a histogram **H**.
  Add noise to $\mathcal{N}\left(0, \sigma^2\right)$ to each entry of **H**.
  Report the top-1 bin from **H**.

When a query arrives, we first *subsample* the private, downstream exemplar dataset using *Poisson sampling*, i.e., independently sample each data point with probability q. Random subsampling is a common technique for reducing the computational cost in non-private ICL, and for DP-ICL this step additionally *amplifies* our privacy guarantee. We divide our subsampled exemplar dataset into n disjoint demonstration exemplars and append the query to each exemplar to form a set of exemplar-query pairs. We prompt the model API with each exemplar-query pair to obtain a collection of answers (i.e., class predictions) for the query. We transform each class prediction into a one-hot vector over the class labels, and we release the class with the highest (noisy) vote in a differentially private way through the Report-Noisy-Max with Gaussian noise (RNM-Gaussian) (Dwork et al., 2014; Zhu and Wang, 2022) mechanism that we now introduce:

RNM-Gaussian Mechanism. For a query Q and classes 1 to m, let $o_j(Q) \in [m]$ denote the LLM prediction for j-th exemplar-query pair on Q, and $c_i(Q)$ denote the vote count for the i-th class, i.e., $c_i(Q) = |\{j : o_j(Q) = i\}|$. We define the mechanism of Report-Noisy-Max with Gaussian noise (RNM-Gaussian) as:

$$\mathcal{M}_\sigma(Q) := \underset{j \in [m]}{\operatorname{argmax}} \left\{ c_j(Q) + \mathcal{N}\left(0, \sigma^2\right) \right\}$$

where $\mathcal{N}\left(0, \sigma^2\right)$ is the Gaussian distribution with mean 0 and variance σ^2. The aggregator outputs the class with the highest count after adding Gaussian noise to each vote count.

Theorem 2. *The mechanism RNM-Gaussian $\mathcal{M}_\sigma$ is (ε, δ)-DP with $\sigma = 2\sqrt{\log(1.25/\delta)}/\varepsilon$.*

Proof. See Appendix A □

DP-ICL only requires black-box access to the API and can be used with GPT-4 and other models that are otherwise out of reach of DP because they cannot be fine-tuned. Although our method outputs individual predictions, we can also release the noisy 'confidence score vector' of the ensemble by the postprocessing property of DP.

5 Experiments

We present the experimental setup in Section 5.1, discuss the main results of DP-ICL in Section 5.2, and conduct ablation studies in Section 5.3. We provide full experimental details in Appendix C.1.

5.1 Setup

Datasets. Following Zhang et al. (2022b), we study text classification using four datasets: sentiment analysis using **SST-2** (Socher et al., 2013) and **Amazon** (Zhang et al., 2015), topic classification using the 4-way **AGNews** (Zhang et al., 2015) datasets, and 6-way question classification using **TREC** (Voorhees and Tice, 2000). We treat the training set as private and limit its size to 8,000 exemplars. Further details are in Table 3. To conserve computational resources, we randomly choose 100 test samples for inference across all tasks. We report the average accuracy over 100 runs of noise aggregation.

Model. We use the GPT-3 Babbage model for all tasks as it has shown promising results of in-context learning and is cost-effective for us, with a budget of $30 per task. We note that all our results can be massively improved by simply replacing the GPT-3 Babbage API call with more advanced LLMs such as GPT-4, but we use GPT-3 because it has been used by prior work (He et al., 2023).

Methods. We primarily focus on in-context learning with 4 exemplars (4-shot) and 10,000 queries. We compare with a zero-shot prediction that provides inherently (0,0)-DP guarantee and non-private ($\varepsilon = \infty$) 4-shot prediction.

5.2 Main results

DP-ICL achieves a comparable performance with non-private ICL across all tasks (Table 1). We set the number of exemplar-query pairs to 10 after sub-sampling and selected $\varepsilon = \{1, 3, 8\}$ to achieve different levels of privacy. Our findings indicate that the impact of considering privacy on accuracy is marginal. For instance, the performance only drops by 0.13% for **SST-2** with $\varepsilon = 8$. Even for a conservative privacy budget of $\varepsilon = 1$, we observe that DP-ICL significantly outperforms the zero-shot prediction (e.g., ≥ 3.77 % for **SST-2**).

Table 1: Results of DP-ICL for multiple privacy budgets

Dataset	$\epsilon = 0$ (0-shot)	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 8$	$\epsilon = \infty$
SST-2	91.00	95.06	95.80	95.92	96.05
Amazon	91.00	93.83	94.18	94.25	94.26
AGNews	46.00	73.16	78.98	79.77	81.43
TREC	20.00	25.83	26.75	27.18	28.42

DP-ICL outperforms all previous DP-SGD methods on SST-2 benchmark (Table 2). We then compare our results (including 16-shot ICL) with current state-of-the-art differentially private stochastic gradient descent (DP-SGD) methods on **SST-2**. The results illustrate a remarkable improvement over earlier methods, with an enhancement of **1.20%** at $\varepsilon = 3$ and **1.02%** at $\varepsilon = 8$. This translates to an over **20%** reduction in relative error rate, thereby establishing a new SOTA in the field. [1] Moreover, by presenting the results with $\varepsilon = \infty$, we notice that our performance gains do not directly correlate to the advanced large language model.

[1]A minor discrepancy exists between our training data (sentence level) and the DP-SGD training data (phrase level) on the SST-2 dataset. Our training data is 10 times smaller than that of DP-SGD. However, the test data remains identical for both.

Table 2: Results of DP-ICL and DP-SGD on **SST-2**.

Model	**Method**	$\varepsilon = 3$	$\varepsilon = 8$	$\varepsilon = \infty$
RoBERTa-large (Liu et al., 2020)	DP-SGD (Li et al., 2022)	93.04	93.81	96.20
	DP-SGD (Yu et al., 2022)	–	95.30*	96.40
	DP-SGD (Bu et al., 2023)	94.60	94.70	95.50
	DP-SGD (He et al., 2023)	94.23	94.87	96.20
GPT-3 Babbage ‡	DP-ICL (4-shot)	$\mathbf{95.80}_{1.45}$	$95.92_{1.43}$	$96.05_{1.32}$
	DP-ICL (16-shot)	$91.64_{2.41}$	$\mathbf{96.32}_{1.08}$	$96.13_{0.82}$

* Result present in (Yu et al., 2022) is $\varepsilon = 6.7$.
‡ We also incorporate the standard deviation of our results.

5.3 Ablation

We have also carried out ablation studies to examine the effects of varying the number of queries and subsampling rate on the performance of both **SST-2** and **AGNews** with $\varepsilon = 3$, as depicted in Figure 2. Our observations reveal that the performance degradation resulting from an increase in the number of queries remains negligible up to 10,000, with a modest decrease of approximately 2%. In terms of the subsampling rate, our findings suggest that employing a rate of $0.5 * 10^{-2}$ for the exemplars yields satisfactory performance, which corresponds to 10 samples after the subsampling process.

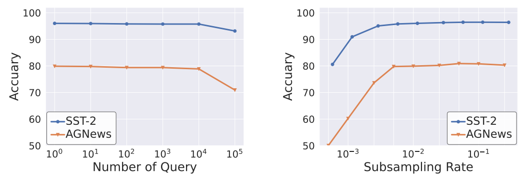

Figure 2: Performance across numbers of the query (left) and subsampling rate (right).

6 Limitations and Future Work

In this work, we introduced a novel differentially private framework for in-context learning using the Report-Noisy-Max mechanism. Compared with prior work in private learning via DP fine-tuning, DP-ICL offers an improved privacy-utility-computation tradeoff with additional flexibility in model compatibility and data editing. However, DP-ICL cannot answer an infinite number of queries from an attacker: in this work, we consider the threat model of non-colluding adversaries who can each adaptively ask up to $k = 10,000$ queries. If two adversaries collude together, then the privacy loss will be equivalent to the case where $k = 20,000$. Moreover, DP-ICL is limited to classification tasks and requires more computation than non-private ICL as we analyze in Fig. 3. Future work can extend DP-ICL to text generation tasks and employ more advanced LLMs to allow for an even stronger attacker and enhance utility.

Acknowledgments

This work was supported in part by the National Science Foundation under grant CNS-2131938, the ARL's Army Artificial Intelligence Innovation Institute (A2I2), Schmidt DataX award, Princeton E-ffiliates Award, and Princeton's Gordon Y. S. Wu Fellowship.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.

Stella Biderman, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raf. 2023. Emergent and predictable memorization in large language models.

Bloomberg. 2023. Using chatgpt at work.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel J. Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models. *ArXiv*, abs/2108.07258.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec

Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. 2022. Differentially private optimization on large model at small cost.

Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. 2023. Differentially private bias-term only fine-tuning of foundation models.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. 2006a. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology-EUROCRYPT 2006: 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28-June 1, 2006. Proceedings 25*, pages 486–503. Springer.

Cynthia Dwork, Nitin Kohli, and Deirdre K. Mulligan. 2019. Differential privacy in practice: Expose your epsilons! *J. Priv. Confidentiality*, 9(2).

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006b. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer.

Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.

Cynthia Dwork and Guy N. Rothblum. 2016. Concentrated differential privacy. *CoRR*, abs/1603.01887.

Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. 2010. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 51–60. IEEE.

Marco Gaboardi, James Honaker, Gary King, Jack Murtagh, Kobbi Nissim, Jonathan Ullman, and Salil Vadhan. 2016. Psi ({\Psi}): a private data sharing interface. *arXiv preprint arXiv:1609.04340*.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I. Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, Dawn Drain, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jackson Kernion, Jamie Kerr, Jared Mueller, Joshua Landau, Kamal Ndousse, Karina Nguyen, Liane Lovitt, Michael Sellitto, Nelson Elhage, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robert Lasenby, Robin Larson, Sam Ringer, Sandipan Kundu, Saurav Kadavath, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, Christopher Olah, Jack Clark, Samuel R. Bowman, and Jared Kaplan. 2023. The capacity for moral self-correction in large language models.

Gonzalo Munilla Garrido, Xiaoyuan Liu, Florian Matthes, and Dawn Song. 2022. Lessons learned: Surveying the practicality of differential privacy in the industry.

Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. 2021. Numerical composition of differential privacy. *Advances in Neural Information Processing Systems*, 34:11631–11642.

Jiyan He, Xuechen Li, Da Yu, Huishuai Zhang, Janardhan Kulkarni, Yin Tat Lee, Arturs Backurs, Nenghai Yu, and Jiang Bian. 2023. Exploring the limits of differentially private deep learning with group-wise clipping. In *The Eleventh International Conference on Learning Representations*.

Antti Koskela and Antti Honkela. 2021. Computing differential privacy guarantees for heterogeneous compositions using fft. *CoRR*, abs/2102.12412.

Antti Koskela, Joonas Jälkö, and Antti Honkela. 2020. Computing tight differential privacy guarantees using fft. In *International Conference on Artificial Intelligence and Statistics*, pages 2560–2569. PMLR.

Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, Patrice Castonguay, Mariya Popova, Jocelyn Huang, and Jonathan M. Cohen. 2019. Nemo: a toolkit for building ai applications using neural modules.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. 2022. Large language models can be strong differentially private learners. In *International Conference on Learning Representations*.

Jerry Liu. 2022. LlamaIndex.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Xin Lyu. 2022. Composition theorems for interactive differential privacy. *arXiv preprint arXiv:2207.09397*.

Jimit Majmudar, Christophe Dupuy, Charith Peris, Sami Smaili, Rahul Gupta, and Richard Zemel. 2022. Differentially private decoding in large language models.

Shiona McCallum. 2023. Chatgpt banned in italy over privacy concerns.

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? In *Conference on Empirical Methods in Natural Language Processing*.

Nvidia. 2023. Large language models enterprise data.

OpenAI. 2023a. Gpt-4 technical report.

OpenAI. 2023b. Openai terms of service.

Ashwinee Panda, Xinyu Tang, Vikash Sehwag, Saeed Mahloujifar, and Prateek Mittal. 2022. Dp-raft: A differentially private recipe for accelerated fine-tuning. *arXiv preprint arXiv:2212.04486*.

Nicolas Papernot and Thomas Steinke. 2022. Hyperparameter tuning with renyi differential privacy. In *International Conference on Learning Representations*.

Fábio Perez and Ian Ribeiro. 2022. Ignore previous prompt: Attack techniques for language models. In *NeurIPS ML Safety Workshop*.

Politico. 2023. Chatgpt is entering a world of regulatory pain in the eu.

Jayshree Sarathy, Sophia Song, Audrey Haque, Tania Schlatter, and Salil Vadhan. 2023. Don't look at the data! how differential privacy reconfigures practices of data science.

Swami Sivasubramanian. 2023. Announcing new tools for building with generative ai on aws.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *emnlp*.

Salil Vadhan and Tianhao Wang. 2021. Concurrent composition of differential privacy. In *Theory of Cryptography: 19th International Conference, TCC 2021, Raleigh, NC, USA, November 8–11, 2021, Proceedings, Part II 19*, pages 582–604. Springer.

Salil Vadhan and Wanrong Zhang. 2022. Concurrent composition theorems for all standard variants of differential privacy. *arXiv preprint arXiv:2207.08335*.

Ellen M Voorhees and Dawn M Tice. 2000. Building a question answering test collection. In *SIGIR*.

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. 2022. Differentially private fine-tuning of language models. In *International Conference on Learning Representations*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022a. Opt: Open pre-trained transformer language models. *ArXiv*, abs/2205.01068.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *NIPS*.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022b. Active example selection for in-context learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuqing Zhu and Yu-Xiang Wang. 2022. Adaptive private-k-selection with adaptive k and application to multi-label pate. In *International Conference on Artificial Intelligence and Statistics*, pages 5622–5635. PMLR.

A Differential Privacy Details

We use $D, D' \in \cup_{n\in\mathbb{N}}\mathcal{X}^n$ to denote two datasets with an unspecified size over space $\mathcal{X}$. We call two datasets D and D' *adjacent* (denoted as $D \sim D'$) if we can construct one by *replacing* one datapoint from the other. Note that the notion of DP under replacement is stronger than DP under addition/removal, because replacing a datapoint is equivalent to removing a datapoint and adding another (Dwork et al., 2014).

Definition 3 (Differential Privacy (Dwork et al., 2006b))**.** *For $\varepsilon, \delta \geq 0$, a randomized algorithm $\mathcal{M} : \text{MultiSets}(\mathcal{X}) \to \mathcal{Y}$ is* (ε, δ)-differentially private *if for every pair of adjacent datasets $D, D' \in \text{MultiSets}(\mathcal{X})$ and for every subset of possible outputs $E \subseteq \mathcal{Y}$,*

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

where the randomness is over the coin flips of $\mathcal{M}$.

Thus, differential privacy requires that for all adjacent datasets D, D', the output distribution $\mathcal{M}(D)$ and $\mathcal{M}(D')$ are close, where the closeness is measured by the parameters ε and δ. In our case, $\mathcal{M}$ is an in-context learning algorithm that outputs an answer to a query by using exemplars. A critical observation is that the DP guarantee is agnostic to the attacker's knowledge about DP algorithm $\mathcal{M}$ (except for the randomness being used in the execution). That is, our guarantee is *future-proof*: a DP-ICL system deployed years from now will provide the same guarantees even if attackers use adaptive attacks that may be invented in the future. The parameter ε is a 'privacy budget': as ε increases, our method is able to give answers that reveal more information about the exemplars. Note that ε is an *exponential* parameter: the attacker's increase in knowledge is bounded by e^{ε}, so a privacy guarantee of e^3 is $e^5 \approx 150\times$ 'stronger' than a privacy guarantee of e^8.

Post-processing Property. Differential privacy exhibits a robust post-processing property. Informally, this means that if a mechanism is differentially private, then any post-processing applied to the output of that mechanism is also differentially private. This property is crucial for enabling flexible analysis of privately released data.

Composition of Differential Privacy. In practice, multiple differentially private mechanisms may be applied to the same dataset. Crucially, multiple DP mechanisms can be *adaptively* composed in the sense that the output of one mechanism can be used as an input to another mechanism, denoted as $\mathcal{M}(D) = \mathcal{M}_1 \circ \mathcal{M}_2(D) := (\mathcal{M}_1(D), \mathcal{M}_2(D, \mathcal{M}_1(D)))$. Differential privacy offers strong composition guarantees, that help quantify the cumulative privacy loss resulting from these combined mechanisms. These guarantees are provided by various composition theorems or privacy accounting techniques, including the basic composition theorem (Dwork et al., 2006a), advanced composition theorem (Dwork et al., 2010), and Moments Accountant (Abadi et al., 2016). For example, the basic composition theorem states that if $\mathcal{M}_1$ is $(\varepsilon_1, \delta_1)$-DP and $\mathcal{M}_2$ is $(\varepsilon_2, \delta_2)$-DP, then the adaptive composition of $\mathcal{M}_1$ and $\mathcal{M}_2$ is $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$-DP.

Consider two attackers: the first asks their allotted k queries in one batch and then observes the answers, and the second asks each query sequentially and incorporates information gained from observing the answer to the current query into the next query. The second attacker is certainly stronger, and this increased strength is captured by adaptive composition.

Tracking Privacy Loss under Multiple Queries. To better keep track of the privacy cost, we use the most recent advances in privacy cost accounting based on the notion of the Privacy Loss Random Variable (PRV) (Dwork and Rothblum, 2016). The PRV accountant was introduced by Koskela et al. (2020) and later refined in Koskela and Honkela (2021); Gopi et al. (2021). For any DP algorithm, one can easily compute its (ε, δ) privacy guarantee based on the distribution of its PRV. The key property of PRVs is that, under (adaptive) composition, they simply add up; the PRV Y of the composition $\mathcal{M} = \mathcal{M}_1 \circ M_2 \circ \cdots \circ M_k$ is given by $Y = \sum_{i=1}^{k} Y_i$, where Y_i is the PRV of $\mathcal{M}_i$. Therefore, one can then find the distribution of Y by convolving the distributions of $Y_1, Y_2, \ldots, Y_k$. Prior works (Koskela and Honkela, 2021; Gopi et al., 2021) approximate the distribution of PRVs by truncating and discretizing them, then using the Fast Fourier Transform (FFT) to efficiently convolve the distributions.

We note that under composition we can extend our threat model to consider an arbitrary number of colluding users. We primarily consider the threat model of non-colluding adversaries who can each adaptively ask up to $k = 10,000$ queries. If two adversaries collude together, then the privacy loss will be equivalent to the case where $k = 20,000$ (Vadhan and Wang, 2021; Vadhan and Zhang, 2022; Lyu, 2022).

Privacy Amplification by Subsampling. Privacy amplification by subsampling is a technique used to enhance privacy guarantees in differentially private mechanisms by randomly selecting a subset of the data before applying the privacy mechanism. This subsampling process can lead to a reduction in privacy costs, allowing for more accurate analyses while preserving privacy. We can show that the Poisson subsampled Gaussian mechanism with sensitivity 1, noise scale σ, and subsampling rate q has the PRV $Y = \log(P(o)/Q(o)), o \sim P$, where $P = (1-q)\mathcal{N}(0,\sigma^2) + q\mathcal{N}(1,\sigma^2)$ and $Q = \mathcal{N}(0,\sigma^2)$, and $P(\cdot), Q(\cdot)$ are the density functions of P, Q. With the PRV of the subsampled Gaussian mechanism as well as the PRV accountant, we can now efficiently and tightly track the privacy costs for DP-ICL.

Theorem 4. *The mechanism RNM-Gaussian $\mathcal{M}_\sigma$ is (ε, δ)-DP with $\sigma = 2\sqrt{\log(1.25/\delta)}/\varepsilon$.*

Proof. See A Note that $\mathcal{M}_\sigma$ can be broken down into applying the argmax operator on a noisy histogram, which is generated by adding Gaussian noise to each dimension of the original histogram. The Gaussian mechanism is known to satisfy (ε, δ)-DP with $\sigma = \Delta\sqrt{2\log(1.25/\delta)}/\varepsilon$ (Dwork et al., 2014), where $\Delta := \sup_{D \sim D'} \|f(D) - f(D')\|$ represents the global sensitivity of the underlying aggregation function f. In our case, f calculates the original voting histogram. As each exemplar-query prediction may alter two counts (increasing one and decreasing the other), the sensitivity Δ is $\sqrt{2}$. The overall privacy guarantee is then derived from the post-processing property of differential privacy. $\square$

B DP-ICL Enables Private Prediction

Our work represents a major departure from prior work on DP LLMs in that we consider private *prediction* rather than private training. A line of recent work (Li et al., 2022; Yu et al., 2022; Bu et al., 2022; He et al., 2023) has proposed fine-tuning pre-trained models on downstream tasks with differentially private stochastic gradient descent (DP-SGD) (Abadi et al., 2016). Despite ample research into DP LLMs and the growing industry demand for solutions to augment LLMs with proprietary data (Kuchaiev et al., 2019; Nvidia, 2023), a number of key challenges remain for DP LLMs that we seek to address by considering *private prediction*.

Private training degrades utility. The bulk of evaluation done in prior work on DP LLMs is done at unrealistic privacy budgets ($\varepsilon > 3$, e.g. $\varepsilon = 50$ (Majmudar et al., 2022)) that are not in line with industry standards (Dwork et al., 2019). We consider $\varepsilon \in [1, 8]$ and provide competitive results with non-private ICL even for the conservative privacy budget of $\varepsilon = 1$. The compromise on privacy budgeting is necessary in private training because the threat model for private training is often overly strong and therefore sacrifices utility. Specifically, private training operates under an overly strong threat model that assumes all downstream users can collude and directly observe the trained model. We instead follow the threat model of Gaboardi et al. (2016) that makes more realistic assumptions about adversaries' information and resources. Specifically, we assume that downstream users cannot view the model, do not share their results with each other, and do not collude in coordinated attacks on individual training samples. This allows each user to independently spend their privacy budget, leading to improved utility without compromising data privacy. We note that *even considering the low probability of downstream users colluding to deanonymize private exemplar data, we still assume an attacker that can observe up to $k = 10,000$ query answers*. **Our method does not compromise utility when evaluated with conservative privacy budgets on challenging datasets.**

Private training makes training harder. Fine-tuning with DP-SGD requires adopting entirely new hyperparameters and shifting existing hyperparameters to be radically different from non-private training (Li et al., 2022). Performing this additional hyperparameter tuning can take hundreds of trials. DP-SGD uses per-example gradient clipping to bound the sensitivity of individual data points. Materializing per-example gradients can increase the memory consumption of training by an order of magnitude (Bu et al., 2022) and slow down the training. Although recent methods have been proposed for efficient hyperparameter tuning (Panda et al., 2022; Papernot and Steinke, 2022), efficient per-example gradient clipping (Li et al., 2022), and parameter-efficient fine-tuning (Yu et al., 2022), we emphasize that DP-SGD introduces challenging engineering and optimization problems that are a topic of ongoing research. **Our method requires no hyperparameter tuning and is computation-**

ally efficient.

Private training is incompatible with black-box LLMs. Developers building on top of cloud-hosted LLMs such as OpenAI, Anthropic, or AWS Bedrock cannot implement the complex DP-SGD algorithm (Sivasubramanian, 2023). Organizations employing closed-source LLMs such as GPT-3+, Claude, or Bard cannot even access the weights for fine-tuning and may never be able to (OpenAI, 2023a). **Our method is compatible with any LLM API.**

Private training does not allow flexible data editing. Private training generates a single model that is inextricably tied to each data point in its training data. This is at odds with the right to be forgotten mandated by GDPR (Politico, 2023), which would require retraining the entire model to delete the influence of a private data point -an impracticality if not an outright impossibility when considering fine-tuning billion-parameter models. By contrast, honoring the right to be forgotten with DP-ICL is as straightforward as just removing the individual's private data from the exemplar database. **Our method enables the right to be forgotten.**

C Additional Experiments Details and Analysis

C.1 Experiments Setup Details

Dataset. Following Zhang et al. (2022b), we conduct experiments in four datasets shown in Table 3.

Table 3: Information about the Dataset

Dataset	Task	# of classes	# of exemplars	avg. length
SST-2	Sentiment cls.	2	6,920	37.8
Amazon	Sentiment cls.	2	8,000	78.5
AGNews	Topic cls.	4	8,000	19.3
TREC	Question cls.	6	5,452	10.2

Template. We also present the template we used to conduct in-context learning on Table 4.

Table 4: The prompts for our experiments.

Dataset	Template	Labels
SST-2	Review: {text} Sentiment: {label}	Positive, Negative
Amazon	Title: {title} Review: {review} Sentiment {label}	Positive, Negative
AGNews	Article: {text} Answer: {label}	World, Sports, Business, Technology
TREC	Classify the questions based on whether their answer type is a Number, Location, Person, Description, Entity, or Abbreviation. Question: {text} Answer Type: {label}	Number, Location, Person, Description, Entity, Abbreviation

C.2 Estimated Cost of DP-ICL

In this section, we evaluate the estimated cost associated with querying the GPT-3 Babbage API for various subsampling rates, as illustrated in Figure 3. Our analysis reveals a discernible trend of escalating costs with increasing subsampling rates. Notably, when the subsampling rate is set to $0.5 * 10^{-2}$, the cost amounts to a mere $\$0.0945$ for 100 predictions on the **SST** benchmark.

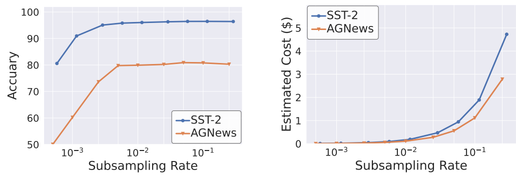

Figure 3: Left: Performance across the subsampling rate. Right: Estimated API Cost of predicting 100 test samples