MEMORIZATION AND PRIVACY RISKS IN DOMAIN-SPECIFIC LARGE LANGUAGE MODELS

Xinyu Yang[1,*], **Zichen Wen**[2,*], **Wenjie Qu**[3], **Zhaorun Chen**[4],
Zhiying Xiang[3], **Beidi Chen**[1], **Huaxiu Yao**[2]
[1]Carnegie Mellon University, [2]UNC-Chapel Hill,
[3]National University of Singapore, [4]University of Chicago
xinyuya2@andrew.cmu.edu, huaxiu@cs.unc.edu

ABSTRACT

Recent literature has explored the potential of fine-tuning LLMs on domain-specific corpora to improve performance on respective domains. However, the risk of memorizing and leaking sensitive information when these models learn from third-party custom fine-tuning data poses significant potential harm to individuals and organizations. To this end, as well as the widespread use of domain-specific LLMs in many high-stake domains, it is imperative to explore whether, and to what degree, domain-specific LLMs memorize fine-tuning data. Through a series of experiments, these models exhibit significant capacities for memorizing fine-tuning data, which results in significant privacy leakage. Furthermore, our investigations reveal that randomly removing certain words and rephrasing prompts show promising performance in mitigating memorization.

1 INTRODUCTION

With the widespread adoption of large language models (LLMs) (Brown et al., 2020; Touvron et al., 2023a; Chowdhery et al., 2022), customizing LLMs in high-stakes, real-world settings has become increasingly important. A number of domain-specific LLMs (Han et al., 2023; Huang et al., 2023; Singhal et al., 2022) have been proposed to enhance LLMs in certain domains, which demonstrate improved performances compared to general models. Following the "pre-training and fine-tuning" paradigm, a user can further fine-tune general LLMs on their own data to personalize its performance for the user's desired downstream task (Jiang et al., 2024; Cheng et al., 2023). For example, if a third-party business wants to deploy a customer service chatbot in their domain, then finetuning using their conversation data on top of a pre-trained LLM could be an effective and efficient solution.

While the flexibility of LLMs in this paradigm has great potential value for various domains, it also raises risks such as privacy leakage, as LLMs can easily expose sensitive information like social security numbers (Kim et al., 2023; Lukas et al., 2023), or regurgitate large parts of training documents (Carlini et al., 2022; Ozdayi et al., 2023). Such privacy leakage is typically associated with memorization, a tendency that LLMs to output entire sequences from their training data verbatim. This phenomenon has been frequently studied in pre-trained LLMs (Hartmann et al., 2023) but remains less explored in domain-specific models. In this work, we conduct a series of investigations focused on the memorization of fine-tuning data. These data are actually of higher concern than pre-training data, since most pre-training datasets are large public corpora with limited privacy concerns, while fine-tuning sets are small, targeted, and potentially very private. Our goal is to ascertain whether, and to what degree, domain-specific LLMs memorize and leak individual fine-tuning data.

Through experiments that use prefixes from training data to prompt Llama-2 (Touvron et al., 2023b) and Medalpaca (Han et al., 2023), we aim to evaluate the similarity between the tokens generated by these models and the ground truth in the data with multiple metrics. The findings reveal that Medalpaca exhibits a significant capacity for memorizing fine-tuning data. Furthermore, a comparative analysis of Medalpaca and Llama on their respective training datasets highlights that fine-tuning data is more likely to be memorized due to its smaller volume of samples. Finally, we demonstrate that randomly removing some words and rephrasing prompts can effectively mitigate memorization.

These collective findings underscore the significant memorization of fine-tuning data in domain-specific LLMs. This issue significantly poses risks for using these LLMs in high-stakes domains, highlighting the need for more advanced and tailored strategies to address these privacy leakage effectively. Straightforward methods like prompt modification may not be sufficient in this context.

2 Related Work

Domain-specific LLMs. Motivated by the huge success of LLMs, researchers have started to adapt large language models to specific domains such as medicine (Singhal et al., 2022; Han et al., 2023; Li et al., 2023), finance (Wu et al., 2023; Yang et al., 2023), mathematics () and law (Cui et al., 2023; Huang et al., 2023). These domain-specific LLMs can be trained in two ways: either from scratch or by adapting existing general LLMs through continued fine-tuning (Gururangan et al., 2020), with the latter being a more efficient method due to the foundational benefits provided by the general LLMs. In this paper, we primarily focus on the second category due to its prevalence in real-world scenarios.

Extraction of Training Data. There is extensive work studying how large language models memorize training data and attacks inferring information under various threat models. Research has shown the feasibility of extracting different types of information including individual sentences (Carlini et al., 2020), inserted canaries (Carlini et al., 2018; Parikh et al., 2022; Béguelin et al., 2019) as well as n-grams (McCoy et al., 2021). Prior work studied the leakage of PII in masked language models (Lee et al., 2022), large language models (Huang et al., 2022; Rocher et al., 2019) and Smart Reply classification models (Jayaraman et al., 2022). In addition to demonstrating that language models leak training data, other efforts focus on understanding the causes for such leakage. Jagielski et al. explore the causes of memorization such as training data ordering, i.e., samples can have different privacy risks independent of their content. Tirumala et al. study the effect of memorization across variables such as dataset size, learning rate, and model size. In response, there is also extensive work studying how to improve the security and privacy of LLMs against these attacks through red-teaming (Wang et al., 2023; Chen et al., 2024b), more powerful alignment (Huang et al., 2024; Chen et al., 2024c) and certified decoding (Xiang et al., 2024; Chen et al., 2024a). In our work, we delve deeper into the extraction fine-tuning data from domain-specific LLMs.

3 Experimental Setup

3.1 Models

In our study, we conducted extensive experiments to extract fine-tuning data from a renowned domain-specific model named Medalpaca (Han et al., 2023). It represents a family of large language models designed for medical tasks, fine-tuned on the LLaMA (Touvron et al., 2023a) model using medical datasets. It is available in various sizes, including 7 billion, 13 billion, and variants fine-tuned with LoRA (Hu et al., 2021). The main goal of this model family is to improve performance in medical question-answering and dialog tasks. Furthermore, we extracted training data from LLaMA to analyze the differences in memorization between general LLMs and domain-specific ones.

3.2 Datasets

We carry out our experiments on primary datasets compassing two primary categories:

Domain-specific Dataset. Medical Meadow (Han et al., 2023) is derived from two primary sources: the first includes a collection of existing medical NLP datasets reformatted into an instruction-tuning format, and the second encompasses a crawl of diverse Internet resources. Each subset focuses on distinct facets of medical knowledge and practice, establishing a comprehensive framework for training and evaluation More details about Medical Meadow are displayed in Appendix A.1.

Public Text Corpus. C4 (Raffel et al., 2019) is a large public NLP dataset created by taking a single month's scrape of the Common Crawl corpus. It is refined through filtering heuristics that eliminate duplicates, placeholders, nonsensical text, and non-English content, resulting in a clean and high-quality resource. This dataset is commonly used for pre-training large language models such as T5 (Raffel et al., 2020), GPT-3 (Brown et al., 2020), and LLaMA (Touvron et al., 2023a).

3.3 EVALUATION METRICS

To quantitatively assess the memorization capabilities of LLMs in the aforementioned datasets, we prompt the model with a 50-token prefix sourced from the fine-tuning dataset and utilize greedy decoding for further generation. This approach represents a form of attribute inference (Wu et al., 2016). Subsequently, we assess the discrepancy between the generated tokens and the reference text using several widely recognized metrics. Given the absence of a deterministic threshold of memorization, we compare the performance across the training and test sets for each metric.

Perplexity (Brown et al., 1993). Nasr et al. point out that models often exhibit higher confidence for examples included in their training or fine-tuning datasets, indicating potential memorization (Carlini et al., 2021). Building on this observation, we adopt perplexity (PPL) as a metric to evaluate how well LLMs can predict or memorize the subsequent tokens. Concretely, given a sequence of tokens $x_1, \ldots, x_n$, perplexity is defined as

$$\mathcal{P} = \exp\left(-\frac{1}{n}\sum_{i=1}^{n}\log f_\theta(x_i|x_1,\ldots,x_{i-1})\right).$$

That is, if perplexity is low, then the model exhibits low uncertainty regarding the generated tokens.

ROUGE (Lin, 2004; Ganesan, 2018). ROUGE, an n-gram matching metric, is widely used in machine translation and text generation benchmarks (An et al., 2023). It measures the quality of generated tokens by comparing it to a reference text, focusing on the overlap of content as determined by lexical matching. In our experiments, we choose ROUGE-L recall, precision, and F1-score.

***Embedding Similarity*.** Embedding similarity serves as a straightforward metric for assessing the similarity between generated content and the reference text. A high degree of similarity between the embeddings of generated tokens and the reference text suggests that the model has memorized the sample. The effectiveness of this metric is closely tied to the quality of the embedding model selected. Concretely, we employ the "bge-small-en"[1] model from BAAI in our experiments.

4 MEASURING MEMORIZATION

4.1 PROBLEM DEFINITION

Here, we use data extraction to measure memorization in domain-specific LLMs. It includes two essential steps: the creation of tailored input prompts and the assessment of memorization in generated tokens. Following (Carlini et al., 2022), we engage an open-source LLM, denoted as $\mathcal{M}$, for which we have access to its model weights θ and fine-tuning dataset $\mathbb{X}$. For each string $\boldsymbol{s} \in \mathbb{X}$, we partition it into a prefix $\boldsymbol{p}$ and suffix $\boldsymbol{x}$, i.e., $\boldsymbol{s} = [\boldsymbol{p}||\boldsymbol{x}]$. We then prompt model $\mathcal{M}$ with prefix p to generate new tokens, which can be denoted as $\mathcal{M}(\boldsymbol{p}) = \boldsymbol{x}'$. Finally, the similarity between the generated suffix $\boldsymbol{x}'$ and the ground truth suffix $\boldsymbol{x}$ is examined to assess the model's capacity for memorization.

4.2 EXPERIMENTAL RESULTS

Does domain-specific LLMs memorize fine-tuning data? In Table 1, we analyze the performance of memory extraction across various models using the metrics described before, on both training and testing sets of Medalpaca. The results highlight several important findings. Firstly, it is observed that Medalpaca-7B and Medalpaca-13B consistently exhibit higher confidence and accuracy across all metrics when prompted with prefix bootstraps from the training set, indicating notable memorization behavior. To gain further insights into the performance improvements, it is noteworthy that LLaMA-2-7B and LLaMA-2-13B show comparable performance on training and test datasets. Interestingly, their performance closely mirrors that of Medaplaca on test sets. This similarity hints at a relation between enhanced performance and the model's memorization during fine-tuning. Furthermore, Medalpaca-13B outperforms its smaller counterpart, Medalpaca-7B, in all metrics, demonstrating that an increase in model size correlates with a higher level of memorization of fine-tuning data.

How many words can the model memorize? Additionally, we also measure the length of exact matches between the generated text and the suffix $\boldsymbol{x}$ in the target string $\boldsymbol{s}$ in Table 2. This analysis

[1] https://huggingface.co/BAAI/bge-small-en

highlights Medalpaca's remarkable memorization capacity regarding training samples. For example, Medaplaca-13B can accurately predict the next five tokens for nearly half of the training samples. This observation suggests that the tendency for memorization intensifies as the model size increases.

Is fine-tuning data more likely to be memorized? Due to the small and private nature of domain-specific fine-tuning data, we aim to conduct a detailed comparison of how LLaMA-1-7B and its fine-tuned variant, Medalpaca-7B, memorize information from training and fine-tuning data respectively. Table 3 demonstrates that Medalpaca-7B exhibits a higher degree of memorization compared to LLaMA-1-7B. This observation suggests that the fine-tuned LLMs pose a greater risk of unintentionally memorizing fine-tuning data and leaking sensitive information.

Table 1: Performance of memory extraction across various models Medalpaca. "**train**" refers to the **Wikidoc-Patient-Information** dataset utilized by Medalpaca for training, and "**test**" corresponds to the **USMLE** dataset used for benchmarking. Both datasets originate from Medical Meadow.

Metrics	**Medalpaca-7B**		**Medalpaca-13B**		**LLaMA-2-7B**		**LLaMA-2-13B**	
	train	**test**	**train**	**test**	**train**	**test**	**train**	**test**
Perplexity ↓	2.89	3.41	2.31	3.27	3.21	3.44	3.32	3.43
ROUGE-L precision ↑	0.52	0.43	0.63	0.47	0.36	0.41	0.41	0.39
ROUGE-L recall ↑	0.63	0.47	0.65	0.47	0.41	0.47	0.52	0.48
ROUGE-L F1-score ↑	0.52	0.44	0.59	0.45	0.38	0.43	0.45	0.40
Embedding similarity ↑	0.95	0.91	0.97	0.90	0.72	0.71	0.76	0.79

Table 2: Performance of word-by-word matches between model-generated text and target text with various lengths. We displays the ratio of interval samples to the total number of samples here.

Matching length	**Medalpaca-7B**		**Medalpaca-13B**		**LLaMA-2-7B**		**LLaMA-2-13B**	
	train	**test**	**train**	**test**	**train**	**test**	**train**	**test**
$\geq$ Two words	**0.52**	0.08	**0.70**	0.14	0.08	0.05	0.16	0.07
$\geq$ Three words	**0.32**	0.02	**0.56**	0.06	0.02	0.02	0.12	0.04
$\geq$ Five words	**0.22**	0.00	**0.48**	0.02	0.00	0.00	0.10	0.02
$\geq$ Seven words	**0.14**	0.00	**0.38**	0.00	0.00	0.00	0.04	0.00
$\geq$ Ten words	**0.06**	0.00	**0.24**	0.00	0.00	0.00	0.02	0.00

Table 3: Comparison of memorization capability between general LLM and domain-specific LLM.

Metrics	**Medalpaca-7B**	**LLaMA-1-7B**
Perplexity ↓	**2.89**	3.45
ROUGE-L precision ↑	**0.52**	0.21
ROUGE-L recall ↑	**0.63**	0.31
ROUGE-L F1-score ↑	**0.52**	0.20
Embedding similarity ↑	**0.95**	0.85

5 MITIGATING MEMORIZATION

5.1 PROBLEM DEFINITION

Based on our exploration in earlier sections, a natural question arises: *Given the potential for significant memorization in domain-specific LLMs, what methods can be employed to mitigate this risk?* In this section, we take initial steps to answer the question by implementing some preprocessing techniques on the prompt, aimed at reducing memorization. These methods include randomly removing words and rephrasing the prompt. To study the memorization of domain-specific terms, we selecte samples from the training set and utilize a Named Entity Recognition (NER) model[2] to

[2]https://huggingface.co/d4data/biomedical-ner-all

identify biomedical terms in each sample. Subsequently, we extract the prefix that precedes certain specific terms, using this as the input prompt. The extraction's success is evaluated based on whether the generated tokens includes the corresponding domain-specific term.

5.2 Experiment Results

As illustrated in Table 4, both randomly removing a substantial number of words and rephrasing prompts effectively mitigate memorization, offering a promising approach for mitigating memorization in domain-specific models. Nonetheless, it's crucial to maintain essential information and semantic coherence when modifying prompts. Inconsistent alterations may cause the model to produce nonsensical outputs, undermining the intended outcomes.

Table 4: Performance of prompt modification in mitigating memorization. "Original Prompt" represents the prefix $\boldsymbol{p}$ directly from the training string $\boldsymbol{s}$. "TailClip Prompt" involves removing a few words from the end of the "Original Prompt". "10% randomized deletion" and "20% randomized deletion" signify removing 10% and 20% of words randomly from the "Original Prompt". "Rephrased Prompt" indicates modifying the "Original Prompt" with a rephrasing model.

Prompt processing method	**Medalpaca-7B**	**Medalpaca-13B**
Original Prompt	0.65	0.80
TailClip Prompt	0.52	0.55
10% randomized deletion	0.72	0.75
20% randomized deletion	**0.41**	**0.55**
Rephrased Prompt	**0.25**	**0.25**

6 Conclusion

Domain-specific LLMs pose risks in practical applications due to their tendency to memorize sensitive information during fine-tuning. This study explores their memorization concerning fine-tuning data. Through comprehensive evaluation, we observe significant memory retention in Medaplaca, with this trend becoming more prominent as model sizes increase. Additionally, we show that removing certain words at random and rephrasing prompts can effectively mitigate this memorization.

REFERENCES

Chenxin An, Shansan Gong, Ming Zhong, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. L-eval: Instituting standardized evaluation for long context language models. *arXiv preprint arXiv:2307.11088*, 2023.

Santiago Zanella Béguelin, Lukas Wutschitz, Shruti Tople, Victor Rühle, Andrew J. Paverd, Olga Ohrimenko, Boris Köpf, and Marc Brockschmidt. Analyzing information leakage of updates to natural language models. *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, 2019. URL https://api.semanticscholar.org/CorpusID:219450111.

Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, Robert L Mercer, et al. The mathematics of statistical machine translation: Parameter estimation. 1993.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Xiaodong Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security Symposium*, 2018. URL https://api.semanticscholar.org/CorpusID:170076423.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Xiaodong Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *USENIX Security Symposium*, 2020. URL https://api.semanticscholar.org/CorpusID:229156229.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. Quantifying memorization across neural language models. *ArXiv*, abs/2202.07646, 2022. URL https://api.semanticscholar.org/CorpusID:246863735.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*, 2024a.

Zhaorun Chen, Zhuokai Zhao, Wenjie Qu, Zichen Wen, Zhiguang Han, Zhihong Zhu, Jiaheng Zhang, and Huaxiu Yao. PANDORA: Detailed LLM jailbreaking via collaborated phishing agents with decomposed reasoning. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*, 2024b. URL https://openreview.net/forum?id=9o06ugFxIj.

Zhaorun Chen, Zhuokai Zhao, Zhihong Zhu, Ruiqi Zhang, Xiang Li, Bhiksha Raj, and Huaxiu Yao. Autoprm: Automating procedural supervision for multi-step reasoning via controllable question decomposition. *arXiv preprint arXiv:2402.11452*, 2024c.

Daixuan Cheng, Shaohan Huang, and Furu Wei. Adapting large language models via reading comprehension. *ArXiv*, abs/2309.09530, 2023. URL https://api.semanticscholar.org/CorpusID:262044959.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James

Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022.

Jiaxi Cui, Zongjia Li, Yang Yan, Bohua Chen, and Li Yuan. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *ArXiv*, abs/2306.16092, 2023. URL https://api.semanticscholar.org/CorpusID:259274889.

Kavita Ganesan. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. *arXiv preprint arXiv:1803.01937*, 2018.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. *ArXiv*, abs/2004.10964, 2020. URL https://api.semanticscholar.org/CorpusID:216080466.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023.

Valentin Hartmann, Anshuman Suri, Vincent Bindschaedler, David Evans, Shruti Tople, and Robert West. Sok: Memorization in general-purpose large language models. *ArXiv*, abs/2310.18362, 2023. URL https://api.semanticscholar.org/CorpusID:264590727.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021. URL https://api.semanticscholar.org/CorpusID:235458009.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. Are large pre-trained language models leaking your personal information? *ArXiv*, abs/2205.12628, 2022. URL https://api.semanticscholar.org/CorpusID:249063119.

Quzhe Huang, Mingxu Tao, Zhenwei An, Chen Zhang, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. Lawyer llama technical report. 2023. URL https://api.semanticscholar.org/CorpusID:258865862.

Yunpeng Huang, Yaonan Gu, Jingwei Xu, Zhihong Zhu, Zhaorun Chen, and Xiaoxing Ma. Securing reliability: A brief overview on enhancing in-context learning for foundation models. *arXiv preprint arXiv:2402.17671*, 2024.

Matthew Jagielski, Om Thakkar, Florian Tramèr, Daphne Ippolito, Katherine Lee, Nicholas Carlini, Eric Wallace, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, and Chiyuan Zhang. Measuring forgetting of memorized training examples. *ArXiv*, abs/2207.00099, 2022. URL https://api.semanticscholar.org/CorpusID:250243645.

Bargav Jayaraman, Esha Ghosh, Melissa Chase, Sambuddha Roy, Huseyin A. Inan, Wei Dai, and David Evans. Combing for credentials: Active pattern extraction from smart reply. 2022. URL https://api.semanticscholar.org/CorpusID:252118801.

Ting Jiang, Shaohan Huang, Shengyue Luo, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. Improving domain adaptation through extended-text reading comprehension. *ArXiv*, abs/2401.07284, 2024. URL https://api.semanticscholar.org/CorpusID:266999300.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sung-Hoon Yoon, and Seong Joon Oh. Propile: Probing privacy leakage in large language models. *ArXiv*, abs/2307.01881, 2023. URL https://api.semanticscholar.org/CorpusID:259342279.

Jooyoung Lee, Thai Le, Jinghui Chen, and Dongwon Lee. Do language models plagiarize? *Proceedings of the ACM Web Conference 2023*, 2022. URL https://api.semanticscholar.org/CorpusID:247450984.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, and You Zhang. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *ArXiv*, abs/2303.14070, 2023. URL https://api.semanticscholar.org/CorpusID:257756992.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

Nils Lukas, A. Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-B'eguelin. Analyzing leakage of personally identifiable information in language models. *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 346–363, 2023. URL https://api.semanticscholar.org/CorpusID:256459554.

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, 2021. URL https://api.semanticscholar.org/CorpusID:244345615.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, pp. 634–646, 2018.

Mustafa Safa Ozdayi, Charith S. Peris, Jack G. M. FitzGerald, Christophe Dupuy, Jimit Majmudar, Haidar Khan, Rahil Parikh, and Rahul Gupta. Controlling the extraction of memorized data from large language models via prompt-tuning. In *Annual Meeting of the Association for Computational Linguistics*, 2023. URL https://api.semanticscholar.org/CorpusID:258823013.

Rahil Parikh, Christophe Dupuy, and Rahul Gupta. Canary extraction in natural language understanding models. In *Annual Meeting of the Association for Computational Linguistics*, 2022. URL https://api.semanticscholar.org/CorpusID:247762346.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2019. URL https://api.semanticscholar.org/CorpusID:204838007.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

Luc Rocher, Julien M. Hendrickx, and Yves-Alexandre de Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, 10, 2019. URL https://api.semanticscholar.org/CorpusID:198190707.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Schärli, Aakanksha Chowdhery, Philip Andrew Mansfield, Blaise Agüera y Arcas, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle K. Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620:172 – 180, 2022. URL https://api.semanticscholar.org/CorpusID:255124952.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. *ArXiv*, abs/2205.10770, 2022. URL https://api.semanticscholar.org/CorpusID:248986465.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288, 2023b. URL https://api.semanticscholar.org/CorpusID:259950998.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *arXiv preprint arXiv:2306.11698*, 2023.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. *ArXiv*, abs/2303.17564, 2023. URL https://api.semanticscholar.org/CorpusID:257833842.

Xi Wu, Matt Fredrikson, Somesh Jha, and Jeffrey F. Naughton. A methodology for formalizing model-inversion attacks. *2016 IEEE 29th Computer Security Foundations Symposium (CSF)*, pp. 355–370, 2016. URL https://api.semanticscholar.org/CorpusID:5921778.

Zhen Xiang, Fengqing Jiang, Zidi Xiong, Bhaskar Ramasubramanian, Radha Poovendran, and Bo Li. Badchain: Backdoor chain-of-thought prompting for large language models. *arXiv preprint arXiv:2401.12242*, 2024.

Hongyang Yang, Xiao-Yang Liu, and Chris Wang. Fingpt: Open-source financial large language models. *ArXiv*, abs/2306.06031, 2023. URL https://api.semanticscholar.org/CorpusID:259129734.

A Details of datasets

A.1 Details of Medical Meadow

Table 5: Details about Medical Meadow. For information regarding other, already published data, please refer to the respective original publication.

Dataset	Source	Description	n
	Finetuning		
Medical Flash Cards	Anki Flashcards	Rephrased Q&A pairs derived from the front and back sides of medical flashcards	33,955
Stack Exchange	Academia	Q&A pairs generated from questions and their top-rated answers	39,633
	Biology		7,482
	Fitness		3,026
	Health		1,428
	Bioinformatics		906
Wikidoc	Living Textbook	Q&A pairs generated from paragraphs, where questions were formulated from rephrased paragraph titles, and answers were extracted from paragraph text	67,704
	Patient Information	Q&A pairs generated from paragraph headings and associated text content	5,942
	Evaluation		
USMLE	Step 1	Multiple choice questions from the USMLE self-assessment with image-based questions excluded	119
	Step 2		120
	Step 3		135

A.2 Details of Public datasets

Common Crawl. Common Crawl[3] is a nonprofit organization that provides a large and open web crawl data repository for public use. It collects web pages from the internet every month and stores them on Amazon Web Services. Common Crawl's data can be used for various research and innovation purposes, such as natural language processing, network analysis, and social science.

C4. C4 (Raffel et al., 2020) is a public dataset created by Google Research and Google Brain containing a large amount of natural English text. C4 cleanses the data of Common Crawl with a number of filters to remove content that is not suitable for training language models, such as pornography, violence, and machine-generated or translated text. C4 can be used to train LLMs such as T5 (Raffel et al., 2020) and GPT-3 (Brown et al., 2020) to improve their cross-domain knowledge and generalization.

[3]https://commoncrawl.org