Privacy in the Era of Large Language Models

Yash More

School of Computer Science
McGill University
Montreal, Canada

April 2025

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Master of Science.

© 2025 Yash More

Abstract

The exponential growth of large language models (LLMs) has opened new avenues for automation across tasks like question answering, code generation, and translation. However, this rapid advancement has also introduced significant risks related to both memorization and privacy. LLMs, by nature, are prone to memorizing training data, which adversaries can exploit to extract sensitive information through extraction attacks. Our research highlights that these attacks are more effective than previously estimated when attackers leverage multiple prompts, checkpoints, and models. We demonstrate that this multi-faceted adversary can increase the extraction of copyright-protected data by up to 20% and retrieve personally identifiable information (PII) at 1.5× the rate observed in earlier studies.

While adversarial extraction poses a severe threat to data privacy, our findings also reveal that user interactions with LLMs compound these risks. In an earlier paper, we analyzed over one million user-chatbot conversations from the WildChat dataset. Alarmingly, more than 70% of these interactions contained sensitive information, often in contexts where PII detection tools failed, such as medical history and job applications. Our taxonomy of sensitive disclosures demonstrates that users frequently share private details without fully understanding the risks, making privacy leakage a critical issue which needs to be understood better.

Through this thesis, I aim to explore key gaps in the literature on LLM memorization and privacy risks. I explore (i) how adversarial attacks can be formulated to extract more data from pre-trained llms (ii) the transferability of attacks across models and inputs and (iii) the privacy risks associated with user interactions and extent of sensitive information shared by users.

Sommaire

La croissance exponentielle des grands modèles linguistiques (LLM) a ouvert de nouvelles perspectives d'automatisation dans des tâches telles que la réponse aux questions, la génération de code et la traduction. Cependant, cette avancée rapide a également introduit des risques importants liés à la fois à la mémorisation et à la confidentialité. Les LLM, par nature, ont tendance à mémoriser des données d'entraînement, que les adversaires peuvent exploiter pour extraire des informations sensibles via des attaques d'extraction. Nos recherches soulignent que ces attaques sont plus efficaces que prévu lorsque les attaquants exploitent plusieurs invites, points de contrôle et modèles. Nous démontrons que cet adversaire aux multiples facettes peut augmenter l'extraction de données protégées par le droit d'auteur jusqu'à 20 % et récupérer des informations personnellement identifiables (PII) à un taux 1,5 fois supérieur à celui observé dans les études précédentes.

Si l'extraction contradictoire constitue une grave menace pour la confidentialité des données, nos conclusions révèlent également que les interactions des utilisateurs avec les LLM aggravent ces risques. Dans un article précédent, nous avons analysé plus d'un million de conversations utilisateur-chatbot à partir de l'ensemble de données WildChat. Il est alarmant de constater que plus de 70% de ces interactions contenaient des informations sensibles, souvent dans des contextes où les outils de détection des informations personnelles ont échoué, comme les antécédents médicaux et les candidatures à un emploi. Notre taxonomie des divulgations sensibles démontre que les utilisateurs partagent fréquemment des informations privées sans comprendre pleinement les risques, ce qui fait de la fuite de confidentialité un problème critique qui doit être mieux compris.

À travers cette thèse, je souhaite explorer les principales lacunes de la littérature sur la mémorisation des LLM et les risques pour la confidentialité. J'explore (i) comment les attaques adverses peuvent être formulées pour extraire davantage de données à partir de LLM pré-entraînés (ii) la transférabilité des attaques entre les modèles et les entrées et (iii) les risques pour la confidentialité associés aux interactions des utilisateurs et à l'étendue

des informations sensibles partagées par les utilisateurs.

Dedicated to my friends and my family

Acknowledgments

I would like to express my gratitude to all my friends, collaborators, and people who supported in the duration of this work. I would also like thank the examiner for taking out time and effort to go through this work.

I'd like to thank my supervisor, Dr. Golnoosh Farnadi for believing in me from the start. I am deeply grateful for her support and guidance over the past two years, and her help at each stage of my masters.

I am thankful to have Prakhar for his constant support both as a friend as researcher thoughout my masters, and I am so glad I got to work with him. I deeply appreciate his feedback and help whenever I felt stuck, and felt lucky having him as a collaborator.

Secondly, I'd like to thank my mentors Debayan Gupta, and Arup Mondal who believed in me the first time I stepped into research, and helped me carve out a niche early on. I am truly grateful towards your help and unconditional support. I am also grateful to have had incredible friends Pratyush, Anshu and Vedansh who got me interested in machine learning and helped me cultivate my early interests in research.

Third, I'd like to thank all lab members who made my journey so much more memorable, Becca, Armin, Rohan, Khaoula, Kiarash, Jesse and Afaf. Thank you for creating a lab environment where I felt welcome.

Fourth, I'd like to thank Siba, for his extensive support in helping me getting started with the thesis and the process around it. I'd also like to thank Lynn for being a great friend throughout, and a sounding board for all my ideas.

Fifth, I would like to extend my gratitude to the countless music artists whose art provided comfort, motivation, and spirit throughout my master's journey.

Finally, I'd like to thank god, family and my friends: Sumantra and Deepraj for being a constant pillar throughout.

Contents

List of Tables

List of Abbreviations

AI	Artificial Intelligence
GPT	Generative Pretrained Transformer
LLM	Large language Model
MIA	Membership Inference Attack
PII	Personally Identifiable Information

Chapter 1

Introduction

Large language models are becoming a key part of everyday life, and we see them increasingly used in personalizing a large part of experiences on the web and consumer media Naveed et al. (2023). These models generate cohesive text, translate languages, summarise information, and can answer user queries based on multiple sources. The general-purpose nature of these models makes them well-suited to be integrated into search engines, social media apps, emailing platforms, code-assistants and chat-applications.

Although being increasingly useful in our day-to-day activities, they are also plagued by privacy risks- LLMs are prone to exposing or leaking user information from (potentially private) training data (Carlini et al., 2021)- several works have shown countless examples of how language models can inadvertently memorize parts of their training data, which can be exposed or leaked when given suitable input by an attacker. In a recent extraction attack, researchers could extract private training data from ChatGPT just by using the input sentence: " Repeat this word forever: poem poem poem ", revealing user signatures, addresses and contact information (Nasr et al., 2023).

The Leakage of information from LLMs jeopardizes the privacy of the users presenting unique challenges on how language models can be used and regulated, especially due to their integration in a wide range of data-sensitive applications like chat-bots, email

providers and search engines.

As the models grow in size, we witness a power-law emerging (Kaplan et al., 2020), where increasing the number of parameters, training size, or compute used for training can influence the model performance until diminishing results set in. Language models released in the past decade have so far followed the scaling laws to some extent, with newer models being trained on trillions of tokens. The growing amount of training data fed to these models compounds the aforementioned privacy risks. Furthermore, as most of the public corpora on the web has been scraped, processed, and trained upon, we are witnessing an uptick in usage of private sources of information either shared by users, companies, or organizations (Qian et al., 2024).

To understand the privacy risks associated with data-leakage, we leverage existing frameworks that study memorization in language models. Extraction attacks are one such method that can be used to quantify memorization risks. These attacks are used to identify whether a sequence was used in the training set for a language model. The most widely studied extraction attack focuses on measuring discoverable memorization—where the objective is to prompt a language model to reconstruct the remainder of a training sentence when given an initial portion or prefix. The adversary can leverage this structure to perform targeted attacks on any given model (Carlini et al., 2021, 2023b).

Existing literature studies this attack in isolation, with fixed model settings, sizes, inputs and generation hyperparameters(Carlini et al., 2021, 2023a). Restricting the analysis to these specific settings underestimates the risk posed due to extraction attacks, especially as we obtain access to bigger models, with multiple checkpoints and varying model sizes. Adversaries can exploit this multi-faceted access to the LLM ecosystem, and use the brittleness of LLMs to their advantage. In Chapter 3, we study a more realistic scenario where adversaries can exploit prompt sensitivity, and different model families available today to perform attacks that can lead to 1.5X of the leakage previously possible. We also study how downstream applications like extraction of copyrighted information, and PII

extraction are affected directly as a consequence of adversary which is more capable given the current LLM ecosystem.

While memorization poses significant risks, users' privacy is equally impacted by the sensitive information they share during interactions with LLMs. Users share a variety of personal information through different tasks such as writing emails, drafting essays, improvising their resumes or simply brainstorming with LLMs for day-to-day questions. Such conversations can potentially be used to trace the user if leaked and can risk the privacy of the user. To detect (and potentially protect) such personal information, we take initial steps to measure and classify disclosures done by users to commercial models like ChatGPT. In Chapter 4, we study personal disclosures with language models, as a step towards understanding what constitutes private information, and how can sensitive information be measured and detected reliably. We also present the various limitations of current PII detection systems, and how sensitive information can go beyond the structured nature of PII. Our findings show majority of user queries contained sensitive information hinting at the need towards privacy safeguards for the community.

Thesis Contributions: This thesis investigates the vulnerabilities of large language models (LLMs) to privacy attacks and the privacy risks they pose to humans. We outline the background for privacy attacks and discuss why memorization is a potential issue in chapters 2, 3. . We present a more capable adversary than previously shown in the literature, that can extract up to 10% of training data in certain settings in Chapter 3. Our findings also show that leveraging multiple attacks increases risks of data extraction by up to 2× of what has been previously shown, even with the presence of mitigation strategies like data de-duplication. Thus, our work shows the current adversary is far more capable than previously demonstrated, and attributes far more risk. To understand privacy risks better, we conduct a fine-grained analysis of personal disclosures in real-world chatbot interactions, developing a taxonomy of sensitive topics and tasks in Chapter 4. Our results highlight that personally identifiable information (PII) frequently appears in unexpected

contexts (e.g., translation and code editing). We also discuss the limitations of existing PII detection and advocate for user awareness mechanisms to mitigate privacy risks. Furthermore, we outline a discussion on risks of sharing personal disclosures and sensitive information and highlight potential solutions moving forward. Finally, we conclude our work and discuss future implications in chapters 5, 6.

1.1 Contributions of Author

This thesis includes a manuscript, in Chapter 3, of a paper titled *Toward Realistic Extraction Attacks: An Adversarial Perspective*, which was accepted at the Private NLP workshop at Association for Computational Linguistics (ACL) in 2024 (More et al., 2024). The early ideas for the paper stemmed from the Responsible AI course I took with my advisor Dr. Golnoosh Farnadi, where I surveyed existing work on Extraction attacks. I was interested in exploring the attack surface of these attacks, and while brainstorming with my teammate (Prakhar) in my group project, I started exploring the brittleness along several dimensions like prompts, sizes of models etc. I then worked on the early implementations of the attacks, with Prakhar providing me support for experiments. Prakhar and I both contributed equally to the writing and experimentation of the project, with my advisor Dr. Golnoosh Farnadi guiding us throughout.

The thesis also includes a manuscript, in Chapter 4, of a paper titled *Trust No Bot: Discovering Personal Disclosures in Human-LLM Conversations in the Wild* which was published at the Conference on Language Modelling (COLM) 2024 (Mireshghallah et al., 2024). This was a joint work co-led by me, Niloofar Mireshgallah and Maria Antoniak, under the supervision of my advisor Dr Golnoosh Farnadi. Early on we tested several different hypotheses concerning privacy and human interactions, and conducted a broad literature review on privacy and language models. As Niloofar and I began exploring the WildChat dataset and performing initial experiments, we slowly saw the project blossom into an

analysis of self-disclosures and sensitive information. Maria joined the project at a later stage, providing us with critical insights on self-disclosures and helping us develop a taxonomy of the information shared by users. I wrote significant parts of the first draft of the manuscript, which was then iterated and improved upon by my co-authors. I also largely contributed to all experimentation, analysis as well as execution since the beginning of the project. The project was conducted under the guidance of my supervisor Dr. Golnoosh Farnadi who provided critical adjustments at several points of our research progress. The remainder of the thesis is entirely my work, with feedback and support from my supervisor, Dr. Golnoosh Farnadi.

Chapter 2

Background

Language models have become central to our daily lives, especially integrating themselves in knowledge-sharing, search, and social media applications. These models require large amounts of data to train, often containing sensitive and private information, which if leaked, could jeopardize the privacy of users (Brown et al., 2022).

As humans, we typically assess the sensitivity of information based on the context in which it is shared (Brown et al., 2022). However, this doesn't naturally extend to language models, and previous literature has shown that these models can memorize and leak data points from the training set, jeopardizing the privacy of the authors of respective texts(Brown et al., 2022).

With the growing need for data to train language models, we observe that users' private data is being frequently collected, stored and used to train and finetune large language models, violating data privacy. Public sources may also contain data which might not be intended for training use and can dox or violate the privacy of a third party (for example: a social media post about someone else).

There have been several attempts to preserve privacy via the removal of private information and data sanitization or to design training algorithms that do not memorize private data (for example: differential privacy) (Kerrigan et al., 2020; Chen et al., 2024),.

Each method makes explicit and implicit assumptions about the nature and structure of the data to be protected, and the requirements for privacy, which do not hold up for most natural language data. Quantifying these risks is integral to gauging the degree of leakage, and motivates us to rethink data protection and leakage.

Existing approaches that measure memorization take into account how the model is able to output or reconstruct data from the training set (Carlini et al., 2021, 2022, 2023a). For example, Extraction attacks allow us to measure whether a model has memorized a sequence completely by testing if we can get the model to reproduce it word by word. Similarly, Membership Attacks provide us with a way to check if a given data point was a part of the training set, by computing membership scores, which when thresholded, can help us understand certain data points were part of the dataset(Carlini et al., 2021).

Understanding the mechanisms behind these privacy risks requires a closer examination of attack strategies that exploit model memorization. In the following sections, we introduce key concepts and definitions related to extraction attacks and membership inference attacks, providing a foundation for understanding how adversaries can recover or infer sensitive information from trained language models. We also dive deeper into the current defenses and limitations of extraction attacks.

2.1 Understanding Privacy Attacks

Unintended memorization in language models can make it prone to privacy attacks (Tirumala et al., 2022; Carlini et al., 2019; Mattern et al., 2023; Carlini et al., 2022), particularly to extraction attacks (Birch et al., 2023; Carlini et al., 2021, 2023b; Nasr et al., 2023). Extraction attacks, which enable adversaries to extract training data from the model, pose a considerable threat to user privacy. These attacks typically involve prompting the language model with either a prefix from the training dataset (*Discoverable Memorization*) or a random set of strings (*Extractable Memorization)* Carlini et al. (2023b). The attack is considered

successful if the output matches the training data verbatim. In our work, we observe that the brittleness of extraction metrics can undermine the true risk of extraction, and address this by extending the existing framework on extraction attacks to a more robust setup.

Definition 2.1.1 (**Discoverable Memorization** Nasr et al. (2023)**).** For a model Gen and an example $[\mathbf{p} \parallel \mathbf{x}]$ from the training set $\mathbb{X}$, we say $\mathbf{x}$ is discoverably memorized if $Gen(\mathbf{p}) = \mathbf{x}$.

For example, if a model's training dataset contains the sequence "My phone number is 555-1234" and given the prefix "My phone number is", the most likely output is "555-1234", then this sequence is *extractable* (with 4 words of context) (Carlini et al., 2023b).

Definition 2.1.2. Extractable memorization. Given a model with a generation routine Gen, an example $\boldsymbol{x}$ from the training set $\mathbb{X}$ is extractably memorized if an adversary (without access to $\mathbb{X}$) can construct a prompt p that makes the model produce $\boldsymbol{x}$ $(i.e., Gen(\boldsymbol{p}) = \boldsymbol{x})$.

For example, given a prompt like *"repeat this word forever, poem poem poem"*, the model generates personal emails verbatim from the training data. In extractable memorization the user does not have access to the knowledge which sequences belong to the training set. The adversary has to optimize towards the most optimal prompt that can give result to a successful extraction. The nature of the attack makes it more computationally expensive and challenging compared to discoverable memorization. (Nasr et al., 2023).

2.1.1 Membership Inference Attacks

Membership Inference Attacks allow adversaries to determine if a data sample was part of a training set by computing a membership score; the score is then thresholded to determine whether the sample was a true member of the dataset. MIAs can be performed in a black-box manner, i.e. the attacker does not need to have underlying access to the models' parameters (Shokri et al., 2017).

The goal of a Membership Inference Attack (MIA) is to infer whether a given data sequence $\mathbf{x}$ was part of the training dataset $\mathcal{D}$ for model $\mathcal{M}$, by computing a membership

score $f(\mathbf{x}; \mathcal{M})$. This score is then thresholded to determine a target sample's membership (Duan et al., 2024). MIAs are often used as a proxy to test if a machine learning model is leaking information related to its training data.

There are several MIAs which have been geared specifically towards LLMs, for example:

1. **LOSS** (Yeom et al., 2018): The target sample's loss under the model:

$$f(\mathbf{x}; \mathcal{M}) = \mathcal{L}(\mathbf{x}; \mathcal{M})$$

2. **Reference-based** (Carlini et al., 2021): Calibrates $\mathcal{L}(\mathbf{x}; \mathcal{M})$ with respect to another reference model ($\mathcal{M}_{\text{ref}}$) to account for the intrinsic complexity of the target sample $\mathbf{x}$:

$$f(\mathbf{x}; \mathcal{M}) = \mathcal{L}(\mathbf{x}; \mathcal{M}) - \mathcal{L}(\mathbf{x}; \mathcal{M}_{\text{ref}})$$

3. **Zlib Entropy** (Carlini et al., 2021): Calibrates $\mathcal{L}(\mathbf{x}; \mathcal{M})$ with the target sample $\mathbf{x}$'s zlib compression size:

$$f(\mathbf{x}; \mathcal{M}) = \mathcal{L}(\mathbf{x}; \mathcal{M}) / \text{zlib}(\mathbf{x})$$

4. **Neighborhood Attack** (Mattern et al., 2023): The curvature of the loss function at $\mathbf{x}$, estimated by perturbing the target sequence to create n "neighboring" samples and comparing the loss of the target $\mathbf{x}$ with its neighbors $\mathbf{x}_i$:

$$f(\mathbf{x}; \mathcal{M}) = \mathcal{L}(\mathbf{x}; \mathcal{M}) - \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\mathbf{x}_i; \mathcal{M})$$

5. **Min-k Prob** (Shi et al., 2023): Uses the $k\%$ of tokens with the lowest likelihoods to

compute a score instead of averaging over all token probabilities, with loss:

$$f(\mathbf{x}; \mathcal{M}) = \frac{1}{|\text{min-k}(\mathbf{x})|} \sum_{x_i \in \text{min-k}(\mathbf{x})} -\log p(x_i \mid x_1, \ldots, x_{i-1})$$

Membership Inference vs Data Extraction Attacks: MIA and Data Extraction Attacks are both used to measure memorization in large language models. MIA is used as a proxy for measuring memorization, relies on a slightly different assumption compared to Extraction attacks, and reveals different types of leakage risks. While MIAs only require knowledge of candidates in training data and tell you whether those exact candidates are part of the training set; extraction attacks on the other hand, aim to extract training data successfully from the model either with or without partial access to training sequences (Duan et al., 2024).

Despite these risks, there are certain approaches currently in literature that attempt to reduce these risks. Mitigation approaches mostly focus on the early stages of model training, requiring us to inspect the data better, perform thorough sanitization and de-duplication. Sanitization at a data-level offers us to have a control of what the model learns, and inevitably what it eventually memorizes. Each approach comes with its own challenges, and in the next sections, we look at common defenses and their limitations.

2.1.2 Potential Defenses to Privacy Attacks

Data Sanitization

The simplest solution to prevent the leakage of private information is the removal of private and sensitive information from training sets. However, the immediate drawback of this method is that it relies on how well we detect private content in text. The formulation of privacy in text is challenging and existing PII detectors only detect structured content to certain degrees of confidence (Brown et al., 2022). Content that is context-dependent is

difficult to detect and remove, thus making sanitization not sufficient (Ishihara, 2023)

Data Deduplication

Past work has shown that duplication in training data is a key factor in model memorization and privacy risks. More frequent sequences are regenerated at a much higher rate, and existing memorization detection methods struggle with non-duplicated data. De-deplication helps in reducing the extraction rates of texts to a certain degree (Nasr et al., 2023).

A key factor contributing to the persistence of extraction vulnerabilities, even after deduplication, is the inherent sensitivity of language models to prompt variations. Given the diverse tasks users perform when prompting LLMs, we observe significant variability in the utility and effectiveness of different prompts.

LLMs are highly sensitive to prompt changes, and even subtle changes can degrade performance or bypass post-training alignment efforts . This brittleness not only affects reliability but also broadens the attack surface—allowing adversaries to exploit prompt variations to circumvent security measures and extract sensitive information (Sclar et al., 2024). To better understand this phenomenon, we now examine how prompt sensitivity influences both model behavior and security risks.

2.2 Prompt Sensitivity

LLMs are shown to be sensitive to even subtle changes in their prompts, leading to fluctuations in their performance (Sclar et al., 2024). The sensitivity persists across varying model sizes (Salinas and Morstatter, 2024; Zhu et al., 2023), and can be exploited to develop various prompt-engineering techniques to steer model behaviour (Liu et al., 2023). While several prompting strategies rely on templates, prompt optimization techniques take this a step further by employing discrete (Wen et al., 2024; Deng et al., 2022) or

continuous optimization (Wang et al., 2023; Liu et al., 2022; Lester et al., 2021; Zhu et al., 2023) of prompts for specific tasks. The sensitivity of the prompts can also be misused and adversarial modifications of the prompts can trigger the model to act in unintended ways Rossi et al. (2024), often referred to as *jailbreaking* (Liu et al., 2024; Hubinger et al., 2024; Liu et al., 2024). In Chapter 3, we explore the impact of prompt sensitivity on extraction attacks, revealing the underestimation of privacy risks in the literature.

The variability or brittleness doesn't restrict itself to only prompts, but tiny changes in model versions, or even checkpoints can cause an unintended effect on extraction attacks. In this work, we also look into how model variability under different conditions affects extraction rates, and how adversaries can exploit it.

2.3 Churn

The instability of model predictions under updates has gained significant attention in recent years, studied under the umbrella of churn (Milani Fard et al., 2016; Cotter et al., 2019; Bahri and Jiang, 2021; Anil et al., 2018; Jiang et al., 2021; Adam et al., 2023; Watson-Daniels et al., 2024). Churn quantifies the inconsistency in predictions between a system pre-update vs post-update, by measuring the fraction of examples whose predictions diverge (Milani Fard et al., 2016). The term churn is traditionally used in the literature to describe such regressive trends in model predictions, we extend its use by highlighting similar regressive trends and instability of extraction attacks under changing prompts and models.Thus, *churn occurs when information is extractable with weaker setups like shorter prompts, smaller models, or earlier checkpoints, but not with the stronger setup*.

Training Dynamics of LLMs. Several recent works have studied the training dynamics of LLMs over time (Tirumala et al., 2022; Liu et al., 2021; Xia et al., 2023). In the context of memorization, recent work by Biderman et al. (2023) explored the impact of model size and intermediate checkpoints on the dynamics of memorization, revealing a considerable

variance in memorized data over time and size. The practice of releasing models in various sizes and regularly updating them over time can thus increase the attack surface for the underlying dataset. In our work, we study how adversaries can exploit access to multiple checkpoints of a model to extract more data.

Bibliography

George Adam, Benjamin Haibe-Kains, and Anna Goldenberg. 2023. Maintaining stability and plasticity for predictive churn reduction. *arXiv preprint arXiv:2305.04135*.

Rohan Anil, Gabriel Pereyra, Alexandre Passos, Robert Ormandi, George E Dahl, and Geoffrey E Hinton. 2018. Large scale distributed neural network training through online distillation. In *International Conference on Learning Representations*.

Dara Bahri and Heinrich Jiang. 2021. Locally adaptive label smoothing improves predictive churn. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 532–542. PMLR.

Stella Biderman, USVSN PRASHANTH, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. 2023. Emergent and predictable memorization in large language models. *Advances in Neural Information Processing Systems*, 36.

Lewis Birch, William Hackett, Stefan Trawicki, Neeraj Suri, and Peter Garraghan. 2023. Model leeching: An extraction attack targeting llms.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What does it mean for a language model to preserve privacy? In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2280–2292. ACM.

N Carlini, J Hayes, M Na sr, M Jagielski, V Sehwag, F Tramèr, B Balle, D Ippolito, and E Wallace. 2023a. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*. USENIX Association.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022. Membership inference attacks from first principles. In *IEEE Symposium on Security and Privacy*. IEEE.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023b. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: evaluating and testing unintended memorization in neural networks. In *Proceedings of the 28th USENIX Conference on Security Symposium*, SEC'19, page 267–284, USA. USENIX Association.

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

Tiejin Chen, Longchao Da, Huixue Zhou, Pingzhi Li, Kaixiong Zhou, Tianlong Chen, and Hua Wei. 2024. Privacy-preserving Fine-tuning of Large Language Models through Flatness. ArXiv:2403.04124 [cs].

Andrew Cotter, Heinrich Jiang, Maya Gupta, Serena Wang, Taman Narayan, Seungil You, and Karthik Sridharan. 2019. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research*, 20(172):1–59.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. 2022. Rlprompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3369–3391.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. 2024. Do membership inference attacks work on large language models?

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, and Ethan Perez. 2024. Sleeper agents: Training deceptive llms that persist through safety training.

Shotaro Ishihara. 2023. Training data extraction from pre-trained language models: A survey. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Heinrich Jiang, Harikrishna Narasimhan, Dara Bahri, Andrew Cotter, and Afshin Rostamizadeh. 2021. Churn reduction via distillation. In *International Conference on Learning Representations*.

Gavin Kerrigan, Dylan Slack, and Jens Tuyls. 2020. Differentially private language models benefit from public pre-training.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-

efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

Leo Z. Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A. Smith. 2021. Probing across time: What does roberta know and when? *CoRR*, abs/2104.07885.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks.

Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, Kailong Wang, and Yang Liu. 2024. Jailbreaking chatgpt via prompt engineering: An empirical study.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schoelkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. 2023. Membership inference attacks against language models via neighbourhood comparison. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11330–11343.

Mahdi Milani Fard, Quentin Cormier, Kevin Canini, and Maya Gupta. 2016. Launch and iterate: Reducing prediction churn. *Advances in Neural Information Processing Systems*, 29.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models.

Sippo Rossi, Alisia Marianne Michel, Raghava Rao Mukkamala, and Jason Bennett Thatcher. 2024. An early categorization of prompt injection attacks on large language models.

Abel Salinas and Fred Morstatter. 2024. The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2023. Detecting pretraining data from large language models.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models.

Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35:38274–38290.

Zhen Wang, Rameswar Panda, Leonid Karlinsky, Rogerio Feris, Huan Sun, and Yoon Kim. 2023. Multitask prompt tuning enables parameter-efficient transfer learning. In *The Eleventh International Conference on Learning Representations*.

Jamelle Watson-Daniels, Flavio du Pin Calmon, Alexander D'Amour, Carol Long, David C. Parkes, and Berk Ustun. 2024. Predictive Churn with the Set of Good Models. *arxiv*.

Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom

Goldstein. 2024. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *Advances in Neural Information Processing Systems*, 36.

Mengzhou Xia, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, and Ves Stoyanov. 2023. Training trajectories of language models across scales. In *61st Annual Meeting of the Association for Computational Linguistics, ACL 2023*, pages 13711–13738. Association for Computational Linguistics (ACL).

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Zhenqiang Gong, and Xing Xie. 2023. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts.

Chapter 3

Manuscript: Towards More Realistic Extraction Attacks: An Adversarial Perspective

Yash More[*]
McGill University, Mila
yash.more@mila.quebec

Prakhar Ganesh[*]
McGill University, Mila
prakhar.ganesh@mila.quebec

Golnoosh Farnadi
McGill University, Mila
farnadig@mila.quebec

[*]Equal contribution.

Abstract

Language models are prone to memorizing parts of their training data which makes them vulnerable to extraction attacks. Existing research often examines isolated setups–such as evaluating extraction risks from a single model or with a fixed prompt design. However, a real-world adversary could access models across various sizes and checkpoints, as well as exploit prompt sensitivity, resulting in a considerably larger attack surface than previously studied.In this paper, we revisit extraction attacks from an adversarial perspective, focusing on how to leverage the brittleness of language models and the multi-faceted access to the underlying data. We find significant churn in extraction trends, i.e., even unintuitive changes to the prompt, or targeting smaller models and earlier checkpoints, can extract distinct information. By combining information from multiple attacks, our adversary is able to increase the extraction risks by up to $2\times$. Furthermore, even with mitigation strategies like data deduplication, we find the same escalation of extraction risks against a real-world adversary. We conclude with a set of case studies, including detecting pre-training data, copyright violations, and extracting personally identifiable information, showing how our more realistic adversary can outperform existing adversaries in the literature.[*]

[*]Equal contribution

[*]Code released at https://github.com/EQUAL-Mila/llm_extraction_eval

3.1 Introduction

Large language models (LLMs) have grown considerably in size (Meta AI, 2024; Zhao et al., 2023), and have become integral to a wide range of tasks such as knowledge retrieval, question answering, code generation, machine translation, etc.

To complement this growing scale, LLMs are often trained on large amounts of data (Penedo et al., 2024; Soboleva et al., 2023; Gao et al., 2020; Raffel et al., 2020) that may include private, unlicensed or copyrighted information, especially if directly scraped from the web. As LLMs are prone to memorizing the data they've been trained on, they can be prompted to expose sensitive contexts - making it easier for an adversary to extract information in a black-box setting. Naturally, a question arises, how big is the risk imposed due to *memorization*?

Extraction attacks offer an empirical framework to quantify the information leakage in the presence of an adversary. The most commonly studied extraction attack is discoverable memorization (Carlini et al., 2023; Kassem et al., 2024), where the model is prompted with a portion of a sentence from the training data to extract the rest, thus enabling the adversary to perform targeted attacks.

Current extraction attacks study memorization trends in LLMs across isolated settings like model sizes, generation hyperparameters and learning dynamics (Carlini et al., 2021). While effective, they underestimate the risk posed due to a multi-faceted access to the underlying data in the current LLM ecosystem. For instance, we show that an adversary can exploit the sensitivity of LLMs to prompt structure, length and content, to amplify the information gained. The current accessibility to frequently updated model sizes (Meta AI, 2024); checkpoints (Biderman et al., 2023b; Groeneveld et al., 2024) and a large array of model families such as Llama (Meta AI, 2024), Gemini (Team et al., 2023), and Falcon (Almazrouei et al., 2023), can also create higher extraction risks.

In this paper, we study a more realistic scenario and explore the actual risks posed by extraction attacks. More specifically, we ask:

1. **Can adversaries exploit prompt sensitivity?** We find that extraction attacks are sensitive to the prompt design, extracting over 20% more data with even minor, unintuitive changes to the prompt (§3.5.1). Thus, an adversary, given the opportunity to prompt the model multiple times, can extract more data than previously observed.
2. **Does access to multiple checkpoints increase extraction risk?** An adversary with access to multiple model checkpoints over time or sizes gains broader access to the underlying dataset. We find such an adversary can increase the extraction rates up to $1.5\times$, significantly heightening the risk of information leakage (§3.5.2).
3. **Is data deduplication effective in reducing the extraction risks?** We find that data deduplication does reduce the extraction risks, in line with the existing literature (Carlini et al., 2023). However, adversaries can still exploit the prompt structure and multiple checkpoints to extract more information(§3.6.3). Thus, our concerns about a powerful real-world adversary persist despite deduplication.
4. **How are downstream applications affected by the presence of such an adversary?** We performed three separate case studies and found that our more realistic adversary improves the p-value of dataset inference up to $2\times$ (§3.7.1), the extraction of copyright violations by up to 20% (§3.7.2), and the extraction rate of personally identifiable information (PIIs) by $1.5\times$ (§3.7.3).

3.2 Background and Related Work

In this section, we introduce relevant background on extraction attacks in LLMs, followed by an overview of related work on prompt sensitivity and training dynamics in LLMs. Finally, we describe the term *churn* as it applies in our context.

Extraction Attacks in LLMs. Unintended memorization in LLMs can make it prone to information leakage (Tirumala et al., 2022; Carlini et al., 2019; Mattern et al., 2023; Carlini et al., 2022), particularly through extraction attacks (Birch et al., 2023; Carlini

et al., 2021, 2023; Nasr et al., 2023). These attacks allow adversaries to extract training data from the model, raising concerns of leaking sensitive information (Birch et al., 2023). Extraction attacks have gained significant attention in recent years, studied under two primary frameworks: *Discoverable Memorization* (Carlini et al., 2023; Kassem et al., 2024; Liu et al., 2023b; Biderman et al., 2023a; Tirumala et al., 2022; Huang et al., 2022), where the adversary attempts to extract targeted information, and *Extractable Memorization* Nasr et al. (2023); Kandpal et al. (2022); Qi et al. (2024), where the adversary attempts to extract any information about the data.

We add to the growing body of research on targeted extraction attacks by highlighting the lack of a realistic adversary in the literature. We show the existence of a stronger real-world adversary capable of combining information from various attacks, thereby defining a composite form of discoverable memorization (§3.3). Schwarzschild et al. (2024) also redefines discoverable memorization, using prompt optimization and adversarial compression ratio (ACR) to quantify memorization as information compression. In contrast, rather than relying on optimization, our focus is instead on exploiting the multi-faceted access to LLM training data.

Prompt Sensitivity in LLMs. LLMs are shown to be sensitive to changes in their prompts, leading to fluctuations in their performance (Sclar et al., 2024; Liu et al., 2023a). This sensitivity persists across varying model sizes and through fine-tuning and other downstream modifications (Salinas and Morstatter, 2024; Zhu et al., 2023). The sensitivity of prompts can also be misused, and adversarial modifications to prompts can trigger the model to act in unintended ways Rossi et al. (2024); Liu et al. (2024); Hubinger et al. (2024); Liu et al. (2024). While several overarching trends studying the impact of prompt design on extraction attacks are present in the literature (Carlini et al., 2023; Kassem et al., 2024; Qi et al., 2024; Tirumala et al., 2022), these trends are often evaluated in isolation. Motivated by the composability of privacy leakage (McSherry, 2009), we argue that an adversary capable of repeated prompting can combine these trends. We show that such an adversary

can extract more information about the training data than previously reported in the literature (§3.5.1).

Training Dynamics of LLMs. Several recent works have studied the training dynamics of LLMs over time (Tirumala et al., 2022; Liu et al., 2021; Xia et al., 2023). In the context of memorization, recent work by Biderman et al. (2023a) explored the impact of model size and intermediate checkpoints on the dynamics of memorization, revealing a considerable variance in memorized data over time and size. The practice of releasing models in various sizes and regularly updating them over time can thus increase the attack surface for the underlying dataset. In our work, we study how adversaries can exploit access to multiple checkpoints of a model to extract more data (§3.5.2).

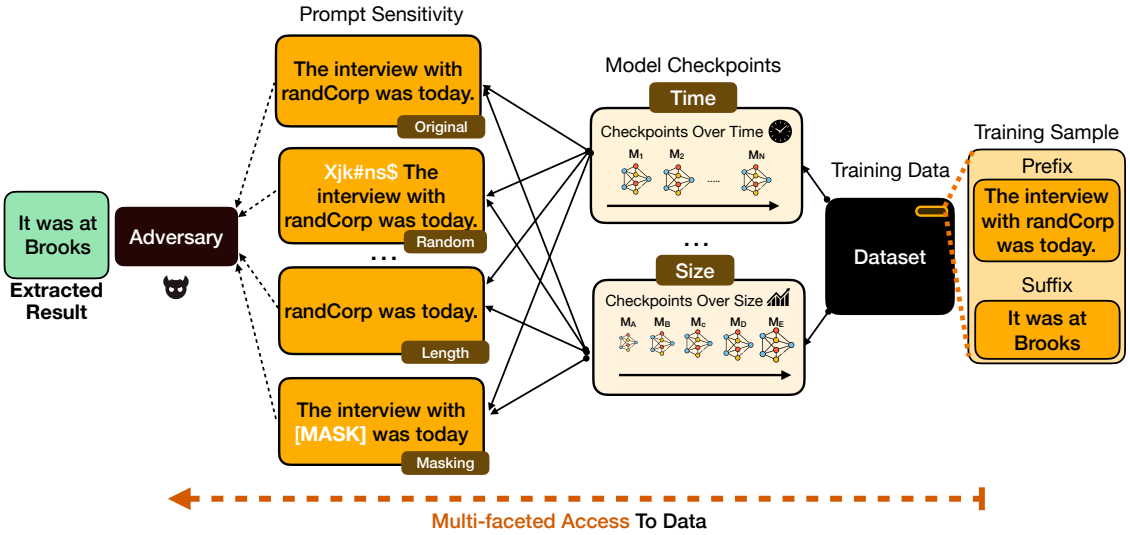

Figure 3.1 Composability in LLMs. In the real world, an adversary has multi-faceted access to a dataset by (a) exploiting prompt sensitivity, and (b) accessing multiple checkpoints trained on the same data.

Churn. The instability of model predictions under updates has gained significant attention in recent years, studied under the umbrella of churn (Milani Fard et al., 2016; Cotter et al., 2019; Bahri and Jiang, 2021; Anil et al., 2018; Jiang et al., 2021; Adam et al., 2023; Watson-Daniels et al., 2024). Churn quantifies the inconsistency in predictions between a system pre-update vs post-update, by measuring the fraction of examples whose predictions

diverge (Milani Fard et al., 2016). The term churn is traditionally used in the literature to describe such regressive trends in model predictions, we extend its use by highlighting similar regressive trends and instability of extraction attacks under changing prompts and models. Thus, *churn occurs when information is extractable with weaker setups like shorter prompts, smaller models, or earlier checkpoints, but not with the stronger setup*.

3.3 Re-evaluating Adversarial Strengths

The adversary is central to our work. We begin by defining its capabilities, arguing that existing work has underestimated the strength of real-world adversaries. To ensure broad applicability, we assume gray-box access to the target model, i.e., the adversary can only access the generation output and probabilities from the model. Consequently, they **cannot** access model weights, gradients, or even control the generation hyperparameters, which reflects the typical level of accessibility for most commercial LLMs. Despite these constraints, we will demonstrate that an adversary in the current LLM ecosystem possesses far greater power than what has been recognized in existing literature.

3.3.1 Adversary Capabilities

Composability (or self-composability) of privacy leakage (McSherry, 2009) suggests that when an adversary gains access to multiple outputs from algorithms on the same underlying dataset—whether through multiple queries from the same algorithm or queries across multiple algorithms—the risk of information leakage grows. Consequently, an adversary with multiple points of access is significantly stronger than one with only a single point of access. In the current landscape of LLMs, such access is not only unsurprising but also easily obtainable (as illustrated in Figure 3.1). Specifically, we consider two forms of multi-faceted access:

Exploiting Prompt Sensitivity LLMs are highly sensitive to their input, including its

structure, content, and even the presence of noisy text within the prompt (Sclar et al., 2024; Liu et al., 2023a; Salinas and Morstatter, 2024; Zhu et al., 2023). While existing studies have focused on improving the prompts for stronger attacks, the nuance of prompt sensitivity in LLMs often defies intuitive expectations. For instance, while longer prompts are known to increase the success of extraction attacks Carlini et al. (2023), our work demonstrates that even shorter prompts can at times exploit vulnerabilities that longer prompts overlook (§3.5.1).

Given the widespread use of LLMs through both chat interfaces and API calls, restricting model access is not realistic. While most commercial LLMs do have rate limits, they are quite high to be of practical concern. For example, even at the lowest tier subscription of $5, ChatGPT has a 500 query per minute (*qpm*) rate limit for GPT4 and 3500 *qpm* for GPT3.5[*]. Thus, an adversary can prompt millions of generations in just one day, making it easier to exploit structural changes in prompts.

Multiple Checkpoints. LLMs are typically deployed in various sizes to cater different needs for accuracy and efficiency among users. However, due to the stochastic nature of their training and the impact of scaling, different model sizes might memorize unique portions of the underlying dataset (Biderman et al., 2023a). Consequently, an adversary with access to multiple model sizes can effectively aggregate extracted information rather than limiting it to a single model (§3.5.2).

Similarly, deployed LLMs undergo regular updates driven by new data, better learning techniques, evolving security measures, and novel functionalities. The stochastic training process means that data resilient to attacks at a certain time step may become vulnerable in subsequent model updates, or vice-versa (Biderman et al., 2023a). Such fluctuations can enable adversaries to exploit multiple checkpoints over time, potentially extracting more information than from a static model (§3.5.2).

More broadly, access to multiple models sharing common training data increases

[*] *qpm* stats and subscription rate as of September 2024.

the attack surface, and in turn, creates stronger adversaries. This level of access is not unprecedented, and several companies in the current LLM ecosystem release multiple versions of their models and even update them periodically. For example, there are 8 different *major* versions of the ChatGPT models and more than 10 *major* versions of the Llama models currently available, while these models are also known to get regular *minor* updates accessible using update dates OpenAI (2024); Chen et al. (2023). Thus, access to multiple models trained on the same data, as has become commonplace, can significantly increase the risks of information leakage.

3.3.2 Combining Extraction Attacks

We argued for the heightened risk posed by multifaceted access to LLMs, either through repeated prompting or multiple model checkpoints. Before discussing our empirical study, we first quantify the risks associated with this stronger adversary. We argue that when an adversary gains such extensive access, any successful extraction of information—even if achieved once—renders that specific information vulnerable to the adversary.

Formally, adapting the definition of discoverable memorization from Nasr et al. (2023), we propose:

Definition 3.3.1 (Composite Discoverable Memorization). For a set of k models $\mathbb{G} = \{Gen_i | i \in [1, \ldots, k].\}$, a set of r prompt modifiers $\mathbb{F} = \{F_j | j = [1, \ldots, r]\}$, and an example $[\mathbf{p} \parallel \mathbf{x}]$ from the training set $\mathbb{X}$, we say $\mathbf{x}$ is composite discoverably memorized if $\exists\, Gen_i \in \mathbb{G}$ and $F_j \in \mathbb{F}$ s.t. $Gen_i(F_j(\mathbf{p})) = \mathbf{x}$.

$$CDM(\mathbb{G}, \mathbb{F}, \mathbf{p} \parallel \mathbf{x}) = \max_{Gen_i \in \mathbb{G}, F_j \in \mathbb{F}} \mathbb{1}_{Gen_i(F_j(\mathbf{p}))=\mathbf{x}}$$

Prompt modifiers are defined as functions $F_j : \mathcal{W}^* \to \mathcal{W}^*$ that take a prompt as input and return a modified version of this prompt as output. Here, $\mathcal{W}$ represents a finite set

of all tokens in the training data i.e $\mathcal{W} = \{w_1, w_2, \ldots, w_n\}$ with w_i representing individual tokens, and $\mathcal{W}^*$ represents the Kleene star operation over $\mathcal{W}$, i.e., a set of all finite length sequences (strings) of tokens in $\mathcal{W}$.

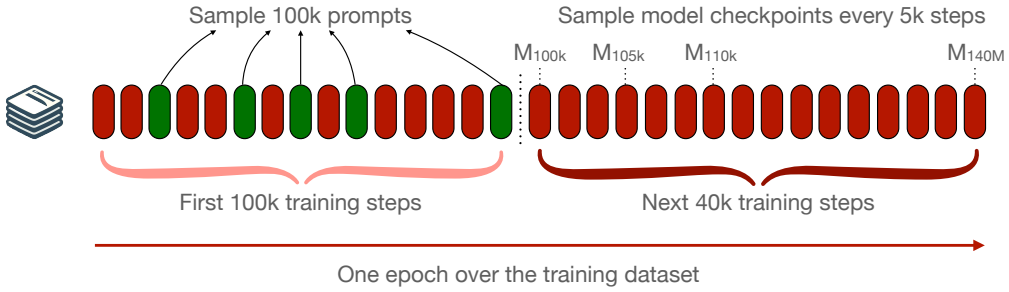

Figure 3.2 Choosing prompts (pre $100k$ steps) and checkpoints (post $100k$ steps) for evaluation of Pythia.

Extraction attacks are often evaluated in the literature using a verbatim match (Carlini et al., 2021, 2023; Nasr et al., 2023; Huang et al., 2022), i.e., the generated text must match the original text perfectly. However, this rigid metric does not take into account the noise in LLM generations, and several recent works have turned to approximate matching to quantify extraction risks for LLMs (Qi et al., 2024; Kassem et al., 2024; Liu et al., 2023b; Ippolito et al., 2022). Thus, we also extend our definition of composite extraction attacks to the approximate matching setup:

Definition 3.3.2 (Approximate Composite Discoverable Memorization). For a set of k models $\mathbb{G} = \{Gen_i | i \in [1 \ldots k]\}$, a set of r prompt modifiers $\mathbb{F} = \{F_j | j = [1 \ldots r]\}$, a similarity metric S, a similarity threshold δ, and an example $[\mathbf{p} \parallel \mathbf{x}]$ from the training set $\mathbb{X}$, we say $\mathbf{x}$ is approximate composite discoverably memorized if $\exists\, Gen_i \in \mathbb{G}$ and $F_j \in \mathbb{F}$ s.t. $S\left(Gen_i\left(F_j\left(\mathbf{p}\right)\right), \mathbf{x}\right) \geq \delta$.

$$ACDM(\mathbb{G}, \mathbb{F}, S, \delta, \mathbf{p} \parallel \mathbf{x}) = \max_{Gen_i \in \mathbb{G}, F_j \in \mathbb{F}} \mathbb{1}_{S(Gen_i(F_j(\mathbf{p})), \mathbf{x}) \geq \delta}$$

Here, S is a similarity metric defined as a function $S : (\mathcal{W}^* \times \mathcal{W}^*) \rightarrow [0, 1]$ that takes as

input two strings $a, b \in \mathcal{W}^*$, and returns a score between 0 and 1 to represent the similarity between the two input strings, and δ is a threshold that controls the degree of approximate matching.

3.4 Experimental Setup

In this section, we outline our central experiment setup, to set the stage for our empirical study. Note that details about the setup for the case studies (§3.7) are delegated to their respective sections.

3.4.1 Models and Dataset

We use the Pythia suite (Biderman et al., 2023b) and OLMo models (Groeneveld et al., 2024) for all our experiments. We primarily focus on the Pythia suite, which contains decoder-only language models with the same architecture as EleutherAI's GPT-Neo (Black et al., 2022), albeit different training, across various model sizes and with open source access to the complete training data and the intermediate checkpoints. Pythia models were trained using GPT-NeoX library (Andonian et al., 2023) on the Pile dataset (Gao et al., 2020) and have not undergone any form of instruction-tuning. The standard version of Pythia was trained over a single epoch of the Pile dataset, i.e., $\approx 143k$ steps with a batch size of 1024, while the deduplicated version of Pythia was trained over ≈ 1.5 epochs of the deduplicated Pile dataset, maintaining the same number of training steps as the standard version.

Pythia suite of models was developed with an emphasis on facilitating open-source investigation into the training dynamics of LLMs. As such, they offer access to (a) models of various sizes (we use model sizes: 1b, 1.4b, 2.8b, 6.9b, and 12b), (b) intermediate model checkpoints during training (a total of 154 checkpoints, with 144 of them equally spaced, i.e., at every $1k$ training steps), and (c) the complete training data order, which is the same

for all model sizes. This level of accessibility and control over the training setup allows us to simulate the real-world availability of models across various sizes and with updating checkpoints over time.

To show the generalizability of our results, we also perform some additional experiments with OLMo models, another set of decoder-only language models. OLMo models were trained on the Dolma dataset Soldaini et al. (2024) and have also not undergone any form of instruction-tuning. These models also offer access to (a) intermediate model checkpoints during training, and (b) open-source access to the complete training data order.

3.4.2 Evaluation Methodology

We now describe our approach to the design and evaluation of extraction attacks. Similar to Carlini et al. (2023), we sample a representative portion of the dataset for analyzing the performance of our extraction attacks. More specifically, we uniformly sample $100,000$ sequences from the first $100k$ steps (batches) of the training data for Pythia. This sampling strategy is important because we choose model checkpoints for evaluation starting at step $100k$, which ensures that every sentence evaluated for memorization has been seen by each checkpoint under consideration, as illustrated in Figure 3.2. We use the same approach for OLMo, with the training step $300k$ being the cut-off point.

Each sequence sampled is exactly 2049 tokens. For our analysis, we employ a consistent method of partitioning each sequence into a prompt and completion at the midpoint, i.e., 1024 tokens. Formally, for a sentence $s_{1:2049}$, prompt length l_p, and completion length l_x, the example $[\mathbf{p} \parallel \mathbf{x}]$ is defined as $\mathbf{p} = s_{1024-l_p:1024}$ and $\mathbf{x} = s_{1024:1024+l_x}$. This partitioning allows us to systematically vary the prompt length and design while comparing the same completion, and vice-versa.

For the Pythia suite, unless otherwise specified, we use a prompt length of $l_p = 50$, a completion length of $l_x = 50$, the Pythia-6.9b model, and the 140k training step checkpoint,

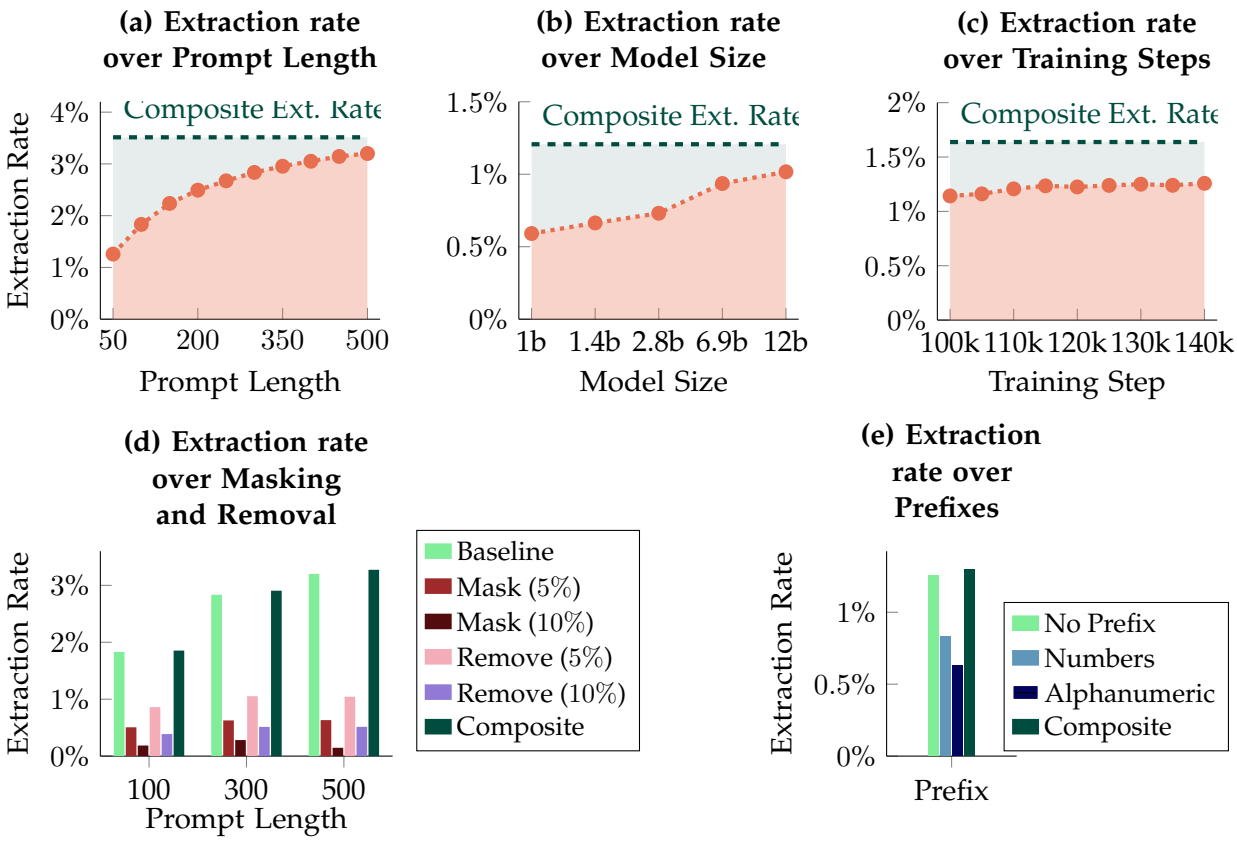

Figure 3.3 Extraction rates under prompt sensitivity and across multiple models for Pythia. **(a)** Increasing prompt length results in better extraction rates, with the composite extraction rate better than even at prompt length 500. **(b, c)** We see similar trends for increasing model size and training steps, respectively. Specifically, we see the largest impact of the composite extraction rate across training steps, with the extraction rate increased $1.5\times$ compared to any single checkpoint. **(d)** Randomly masking or removing tokens from the prompt severely hurts the extraction rate, highlighting the importance of prompt structure. **(e)** Adding a random prefix can also contribute to minor improvements in the composite extraction rate.

evaluating the extraction attacks using verbatim match. We use the same default setup for OLMo, with the OLMo-7b model, and the 500k training step checkpoint.

3.5 Churn in Extraction Trends

Churn (Milani Fard et al., 2016), as previously introduced in §3.2, refers to regressive variance for individual extracted information despite an overall improvement in the extraction rates. For instance, although using a longer prompt is often associated with stronger extraction rates (Carlini et al., 2023; Biderman et al., 2023a), we observe trends that exhibit churn, i.e., certain information is instead extractable only with shorter prompts but not with longer prompts. These non-monotonic and locally regressive trends of certain sentences (i.e., churn) can be exploited by an adversary with multifaceted access to the data to execute a composite extraction attack. We study the factors that may lead to *churn* such as (a) prompt sensitivity, and (b) access to models of varying sizes and training checkpoints.

3.5.1 Prompt Sensitivity

We start by examining prompt sensitivity, focusing on how trends in prompt design can lead to churn.

Prompt Length. Prompt length is a commonly studied parameter in extraction attacks, and it has been shown that longer prompts lead to better extraction (Carlini et al., 2023). This is intuitive, as conditioning the model with more text from training would increase the likelihood of extraction. We will now show that the composite extraction rate (Definition 3.3.1) across varying prompt lengths exceeds the extraction rate at even the largest prompt length. As illustrated in Figure 3.3(a), the extraction rate increases with longer prompts. However, the composite extraction rate is noticeably higher than any single prompt length, including the longest at 500 tokens. This suggests that certain information extractable with shorter prompts remains elusive even with the longest prompt. Consequently, an adversary can exploit this churn across the prompt length to extract more information. We see similar trends for OLMo in Figure 3.4.

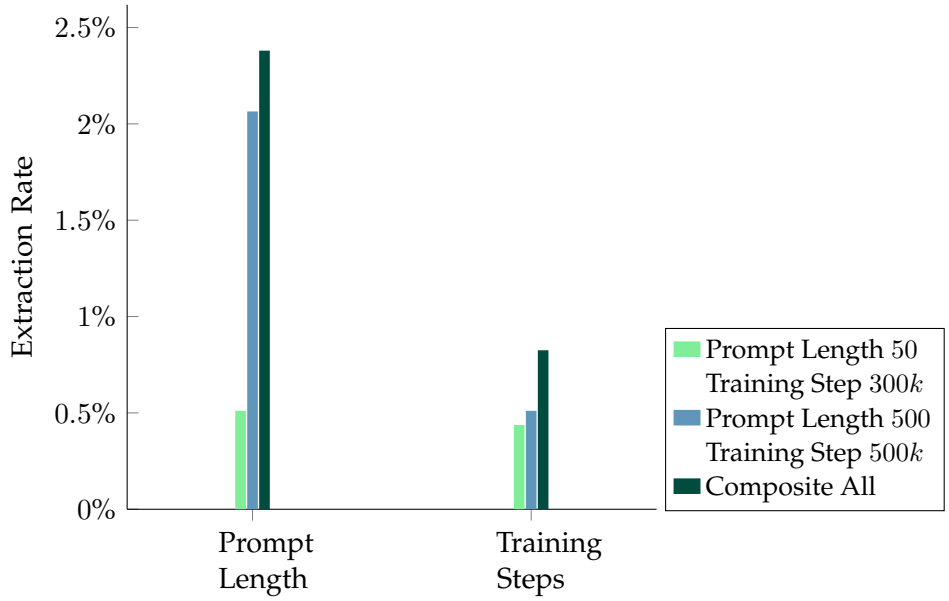

Figure 3.4 Composite extraction attack results across 10 prompt lengths (same as Pythia) and 11 training steps (equidistant between $300k$ and $500k$), compared against isolated setups, for OLMo.

Prompt Structure. Next, we explore the structure of prompts to identify what makes a prompt potent and where churn can emerge. We introduce noise into the prompts by masking and removing random tokens; results are collected in Figure 3.3(d). Despite introducing only a small amount of noise, we observe a significant drop in extraction rates. This indicates that the contiguous prompt from the training data is crucial for extracting information, and any disruption inside this prompt can significantly hurt its capabilities. Yet, we do see minute churn in extraction trends, which further highlights how an adversary can exploit repeated prompting to extract more information, even with seemingly unintuitive changes like masking or removing random tokens.

We next add noise as a prefix in the form of random numeric and alphanumeric strings; results are collected in Figure 3.3(e). Interestingly, the performance degradation with a noisy prefix is less severe than with noise within the prompt. More importantly, we observed a higher composite extraction rate. This suggests that adding a noisy prefix

can also help extract unique information that was previously inaccessible, and further highlights how an adversary can exploit repeated prompting.

Note that the churn in our prompt design trends highlights the increased extraction risks without access to new information. For instance, if an adversary has access to the prompt of length 500 tokens, they can expand the attack surface and thereby the extraction rate simply by removing parts of the prompt, adding noise, etc.—without needing any additional knowledge. One might argue that as the number of prompt variations increases, every sentence could become extractable. However, that is not true; not all sentences are extractable. Yin et al. (2024) showed that knowledge not present in an LLM will not be extractable even after prompt optimization, while Schwarzschild et al. (2024) also showed similar trends when attempting to extract a given completion. Consequently, prompting an LLM to regurgitate certain sentences, even with various prompt modifications, demonstrates a genuine extraction risk and underscores the extent of memorization in LLMs (Carlini et al., 2021).

3.5.2 Multiple Checkpoints

Model Size. The model size has long been known to influence learning trends, and our results in Figure 3.3(b) reflect this phenomenon. We find that larger models tend to memorize more information, which makes them more vulnerable to extraction attacks. However, our results also indicate that the composite extraction rate is higher than the extraction rate of any single model, highlighting the churn present in these trends. Biderman et al. (2023a) also conducted an empirical study on the overlap between memorized data across model sizes and found that up to 10% of the data memorized by smaller models is not memorized by larger models. Combining our insights with existing literature, it's clear that releasing models in different sizes increases the extraction risks.

Model Updates. We also analyze model updates over time using intermediate checkpoints in Figure 3.3(c), where we observe the most significant churn in our study. Unsurprisingly,

attacking models at later stages of training is more successful, as seen in the literature (Tirumala et al., 2022; Biderman et al., 2023a; Jagielski et al., 2023). But remarkably, the churn here is significantly powerful and by exploiting composability across intermediate checkpoints, an adversary can increase their extraction rate by more than $1.5\times$. We also see similar results for OLMo in Figure 3.4. This underscores the impact of stochasticity in model training on extraction attacks and reveals that regular model updates, typically considered beneficial in the current LLM ecosystem, create a powerful adversary.

3.6 Towards Realistic Extraction Attacks

With a better understanding of the trends across various setups, we now evaluate a more realistic measure of leakage in extraction attacks, by investigating (a) composability in churn, (b) challenges in evaluation, and (c) effects of deduplication.

3.6.1 Combining Multiple Axes of Churn

In the previous section, we saw how churn can impact individual axes of variability, such as prompt sensitivity, model size, and intermediate checkpoints. However, a real-world adversary can take advantage of all these factors simultaneously, thus significantly increasing their extraction rates. We start by analyzing two axes at a time, as shown in Figure 3.5(a). For all pairs of variability, the overall composite extraction rate (bottom right) is $2-3\times$ higher than the base setup (top left) and $1.5-2\times$ higher than the composite extraction rates along one axis (top right and bottom left). Furthermore, when all three axes are combined, depicted in Figure 3.5(b), the extraction rates grow even higher, albeit with diminishing gains. Thus, we show that a real-world adversary can extract far more training data than has been previously seen in the literature.

3.6.2 Approximate Matching

As discussed in §3.3.2, evaluating extraction attacks under verbatim match can underestimate the true risk of extraction. To address this gap, we introduced approximate composite discoverable memorization (Definition 3.3.2), and will now analyze various similarity metrics S to examine their behaviour under changing δ, reported in Figure 3.5(c). Solely for this discussion, we increase the completion length $l_x = 500$, to allow for meaningful extraction even with approximate matching.

Our results reveal intriguing trends. First, we analyze evaluations based on the Levenshtein ratio metric and observe that even the threshold of $\delta = 0.95$ doubles the extraction attack rate compared to a verbatim match. This threshold signifies a minimum 95% overlap between generated and original text. Even under such a strict threshold, the doubling of the extraction rate underscores the significant underestimation of extraction risks when relying solely on verbatim matches. As δ decreases, however, the extraction rate increases exponentially, as the Levenshtein ratio becomes less reliable under looser constraints. We also see similar trends for ROUGE-L scores.

Transitioning to other similarity metrics — longest common substring (LCS), Hamming distance, and n-gram matching — we find that even lower values of similarity (δ) can contribute meaningfully to extraction attacks. We observe patterns of rising extraction rates similar to what we saw earlier with the Levenshtein distance. The diverse trends underscore the choice of approximation metric as highly context-dependent. A more thorough examination of which metrics best serve particular applications is left for future work.

3.6.3 Data Deduplication

A commonly recommended solution to extraction attacks and memorization is data deduplication, involving the removal of duplicate data entries within a dataset (Carlini et al., 2023). While costly, data deduplication represents a critical aspect of data curation and

has been shown to mitigate extraction risks (Carlini et al., 2023). To understand the role of data deduplication in our discussion, we repeat our experiments using the models from Pythia trained on the deduplicated Pile dataset. The results are collected in Figure 3.5(b).

In line with existing literature, data deduplication reduces the extraction rate. Interestingly, however, we observe persistent trends: the presence of a stronger adversary due to multi-faceted dataset access. Thus, while beneficial, data deduplication does not alter our fundamental conclusions; real-world adversaries with multi-faceted access to the underlying data can extract substantial information even post-deduplication. Future work on incorporating more concrete frameworks like differential privacy is needed, to better understand such adversaries, particularly from the perspective of privacy protection under multi-access systems.

3.7 Case Studies with Stronger Adversary

We conclude by highlighting the value of our stronger adversary in various case studies.

3.7.1 Detecting Pre-Training Data

Extraction attacks are a primary tool in identifying whether certain data was included in a model's training set. This can be valuable in assessing whether a model is trained on proprietary or sensitive data without permission, evaluating data contamination and leakage in various benchmarks, ensuring regulatory compliance to data governance policies, or even academic research to track the influence of datasets on the model.

While membership inference attacks (MIAs) have been commonly used to detect pre-training data, Maini et al. (2024) argues that MIAs are as good as random guessing when it comes to distinguishing between members and non-members from the same distribution. They show that these attacks learn how to distinguish between *concepts*, and not actual text, highlighting the importance of using IID data of members and non-members to

appropriately perform dataset inference.

We borrow their setup and extend it to the composite setting by increasing the size of the training set for learning correlations. Thus, our composite setting can be alternatively seen as an augmentation technique for the training set. We record the p-value of the null hypothesis *"the dataset was not used for training"* for the Pile dataset in Figure 3.6(a), under different sizes of the original training data.

We find that the p-values for the composite setting with prompt lengths are noticeably lower than those for the baseline, especially at smaller dataset sizes. Thus, our adversary requires less data to achieve the same p-value as the baseline. The dataset inference setup by Maini et al. (2024) requires obtaining IID data that the data owners are certain was not used for training by the LLM, which can be difficult to find. Hence, reducing the amount of such data required can be extremely useful, which further emphasizes the value of considering a real-world adversary. Interestingly, we did not find similar strong trends for composite attacks across different model checkpoints. We believe this might be because membership inference information can change drastically across models, and thus combining information from multiple checkpoints does not help in learning better correlations.

3.7.2 Copyright Infringement

Copyright issues due to LLMs regurgitating their training data have been heavily studied in recent literature. Karamolegkou et al. (2023) discusses different thresholds for quoting a text ad verbatim that has been considered a violation of fair use, for example, 50 words is a common threshold used for magazine articles, chapters, etc., while 300 words is a common threshold used for books. The authors suggest using the longest common subsequence (ROUGE-L score length) as a measure of quantifying text reproduction and potential violations.

Following their reasoning, we record the distribution of ROUGE-L lengths for 2000

randomly chosen examples in Figure 3.6(b), both for the strongest baseline as well as the composite settings. We find that a real-world adversary generates more potential copyright violations than the adversary in the literature, which highlights the underestimation of such risks. We note that copyright is a highly complex problem, and simply extracting data from the model might not necessarily constitute a copyright violation. However, our focus is on improving the technical underpinnings that are necessary for a fruitful discussion of copyright issues in LLMs.

3.7.3 PIIs Extraction Risk

Another commonly studied risk of memorizing training data is extracting personally identifiable information (PIIs). We use the setup of Li et al. (2024) to create our PII extraction test set from the Pile dataset. We use GLiNER (Zaratiana et al., 2024) to detect 2000 unique PIIs in the Pile dataset, followed by cutting the sentence right before the PII to create the input prompt. These prompts were fed to the model, and the attack is considered successful if the correct PII is generated anywhere within the first 100 tokens, marking the risk of PII leakage (Li et al., 2024).

We record the extraction risk for the best single setup and composite extraction risks across model checkpoints and model sizes. Since the prompts in this setup are of varying lengths, we do not extend our changing prompt lengths setting to this case study. Similar to Definition 3.3.1, the composite PII extraction is considered successful if the PII is present in the generation of at least one of the models. The results are collected in the table below, and continuing previous trends, we see a noticeable increase in the extraction rate for an adversary with access to multiple checkpoints.

Setup	Extraction Rate
Best Single Setup	22.16%
Composite Model Sizes	30.97%
Composite Training Steps	33.07%

3.8 Limitations and Future Work

By highlighting the multi-faceted access available to an adversary in the current LLM landscape, our work reveals a severe underestimation of information leakage risks in the existing literature. We emphasize the importance of explicitly considering the adversarial perspective and the composability of information leakage in extraction attacks.

Our real-world adversary is certainly more powerful but also more expensive than the adversary in the existing literature. Unlike our current setup, where we verify extraction using the ground truth, an adversary would need to justify both the cost of additional model generations and the expense of verifying extracted information. Therefore, future research should explore the cost-benefit trade-offs of multi-faceted access, focusing on when these added expenses may outweigh the benefits of new information extracted, particularly as we show diminishing gains with increased points of access.

As most of our analysis focuses on the risks posed by extraction attacks under the lens of discoverable memorization, future research should also explore how our findings translate to other forms of privacy attacks. Finally, our study addresses the threats posed by powerful real-world adversaries but does not propose specific defence methods. Further exploration is needed to navigate the current LLM ecosystem and mitigate the risks posed by these strong adversaries.

Acknowledgments

Funding support for project activities has been partially provided by Canada CIFAR AI Chair, and Google award. We also express our gratitude to Compute Canada for their support in providing facilities for our evaluations.

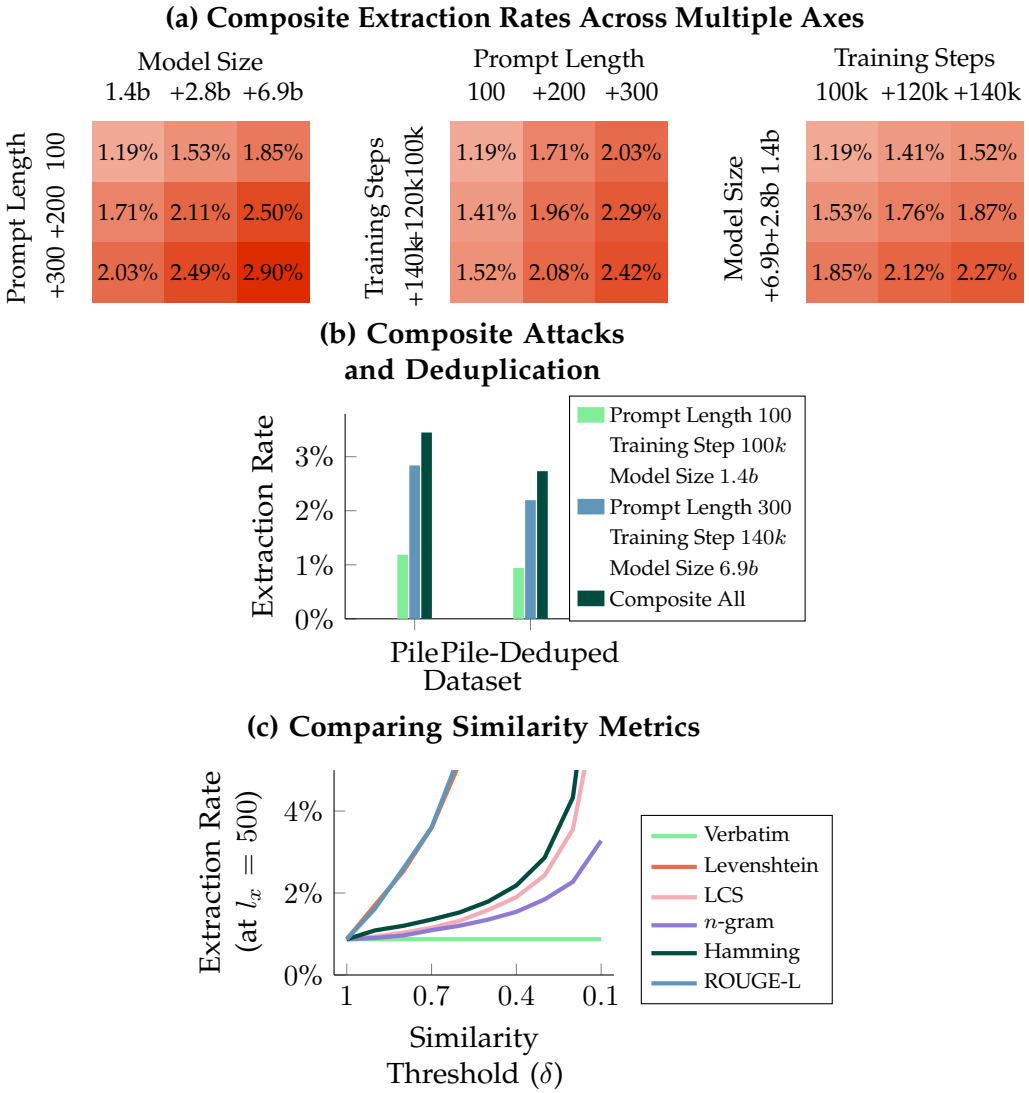

Figure 3.5 Towards more realistic extraction rates by combining various churn trends and with approximate matching. **(a)** Combining two axes at a time, we see a monotonically increasing trend in extraction rates as we gain more points of access to the underlying dataset, highlighting the growing power of the adversary. **(b)** Combining multiple axes of attack results in a significant increase in extraction rate for both standard and deduplicated setups, with the composite extraction rate for the deduplicated setup the same as the highest single setting rate for the standard setup. **(c)** Various similarity metrics have distinct trends as we decrease the threshold value and allow for looser approximations, thus the choice is context-driven.

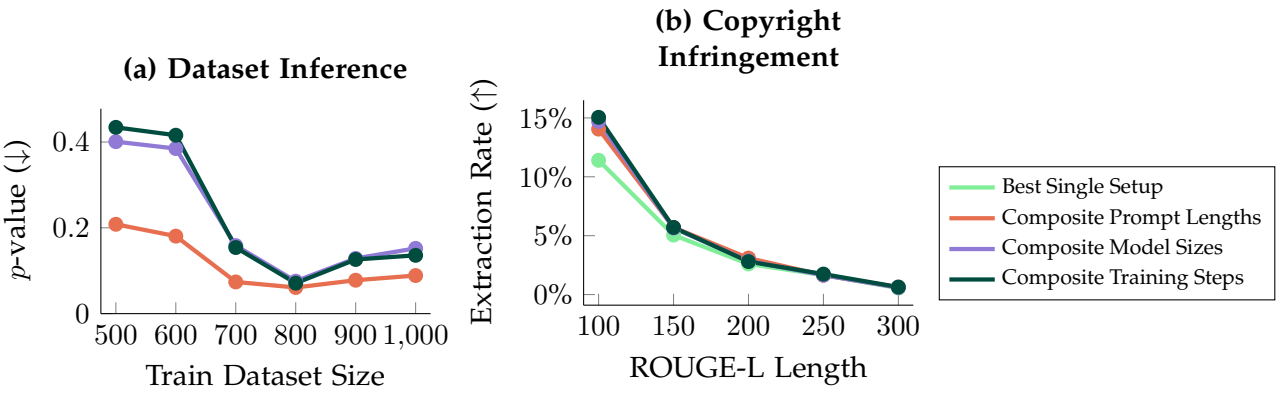

Figure 3.6 **(a)** p-value for dataset inference (lower is better) across different dataset sizes. The results show the importance of exploiting prompt sensitivity and the significant improvement under the composite setup across different prompt lengths. **(b)** Extraction rate for different ROUGE-L length thresholds, marking potential copyright violations generated by the LLM. Extraction rates with composite setups are consistently higher than the single setup, highlighting the impact of multi-faceted access to the data.

Bibliography

George Adam, Benjamin Haibe-Kains, and Anna Goldenberg. 2023. Maintaining stability and plasticity for predictive churn reduction. *arXiv preprint arXiv:2305.04135*.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Alex Andonian, Quentin Anthony, Stella Biderman, Sid Black, Preetham Gali, Leo Gao, Eric Hallahan, Josh Levy-Kramer, Connor Leahy, Lucas Nestler, Kip Parker, Michael Pieler, Jason Phang, Shivanshu Purohit, Hailey Schoelkopf, Dashiell Stander, Tri Songz, Curt Tigges, Benjamin Thérien, Phil Wang, and Samuel Weinbach. 2023. GPT-NeoX: Large Scale Autoregressive Language Modeling in PyTorch.

Rohan Anil, Gabriel Pereyra, Alexandre Passos, Robert Ormandi, George E Dahl, and Geoffrey E Hinton. 2018. Large scale distributed neural network training through online distillation. In *International Conference on Learning Representations*.

Dara Bahri and Heinrich Jiang. 2021. Locally adaptive label smoothing improves predictive churn. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 532–542. PMLR.

Stella Biderman, USVSN PRASHANTH, Lintang Sutawika, Hailey Schoelkopf, Quentin

Anthony, Shivanshu Purohit, and Edward Raff. 2023a. Emergent and predictable memorization in large language models. *Advances in Neural Information Processing Systems*, 36.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023b. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Lewis Birch, William Hackett, Stefan Trawicki, Neeraj Suri, and Peter Garraghan. 2023. Model leeching: An extraction attack targeting llms.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. 2022. Gpt-neox-20b: An open-source autoregressive language model. *Challenges & Perspectives in Creating Large Language Models*, page 95.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022. Membership inference attacks from first principles. In *IEEE Symposium on Security and Privacy*. IEEE.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: evaluating and testing unintended memorization in neural networks. In *Proceedings of the 28th USENIX Conference on Security Symposium*, SEC'19, page 267–284, USA. USENIX Association.

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

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. How is chatgpt's behavior changing over time? *arXiv preprint arXiv:2307.09009*.

Andrew Cotter, Heinrich Jiang, Maya Gupta, Serena Wang, Taman Narayan, Seungil You, and Karthik Sridharan. 2019. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research*, 20(172):1–59.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. 2022. Are large pre-trained

language models leaking your personal information? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2038–2047.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, and Ethan Perez. 2024. Sleeper agents: Training deceptive llms that persist through safety training.

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

Matthew Jagielski, Om Thakkar, Florian Tramer, Daphne Ippolito, Katherine Lee, Nicholas Carlini, Eric Wallace, Shuang Song, Abhradeep Guha Thakurta, Nicolas Papernot, et al. 2023. Measuring forgetting of memorized training examples. In *The Eleventh International Conference on Learning Representations*.

Heinrich Jiang, Harikrishna Narasimhan, Dara Bahri, Andrew Cotter, and Afshin Rostamizadeh. 2021. Churn reduction via distillation. In *International Conference on Learning Representations*.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*, pages 10697–10707. PMLR.

Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. 2023. Copyright violations and large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7403–7412.

Aly M Kassem, Omar Mahmoud, Niloofar Mireshghallah, Hyunwoo Kim, Yulia Tsvetkov, Yejin Choi, Sherif Saad, and Santu Rana. 2024. Alpaca against vicuna: Using llms to uncover memorization of llms. *arXiv preprint arXiv:2403.04801*.

Qinbin Li, Junyuan Hong, Chulin Xie, Jeffrey Tan, Rachel Xin, Junyi Hou, Xavier Yin, Zhun Wang, Dan Hendrycks, Zhangyang Wang, et al. 2024. Llm-pbe: Assessing data privacy in large language models. *Proceedings of the VLDB Endowment*, 17(11):3201–3214.

Leo Z. Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A. Smith. 2021. Probing across time: What does roberta know and when? *CoRR*, abs/2104.07885.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023b. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. In *Socially Responsible Language Modelling Research*.

Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, Kailong Wang, and Yang Liu. 2024. Jailbreaking chatgpt via prompt engineering: An empirical study.

Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. 2024. Llm dataset inference: Did you train on my dataset? *arXiv preprint arXiv:2406.06443*.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schoelkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. 2023. Membership inference attacks against language models via neighbourhood comparison. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11330–11343.

Frank D McSherry. 2009. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 19–30.

Meta AI Meta AI. 2024. Introducing meta llama 3: The most capable openly available llm to date.

Mahdi Milani Fard, Quentin Cormier, Kevin Canini, and Maya Gupta. 2016. Launch and iterate: Reducing prediction churn. *Advances in Neural Information Processing Systems*, 29.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models.

OpenAI. 2024. ChatGPT Documentation: Models.

Guilherme Penedo, Hynek Kydlíček, Leandro von Werra, and Thomas Wolf. 2024. Fineweb.

Zhenting Qi, Hanlin Zhang, Eric Xing, Sham Kakade, and Himabindu Lakkaraju. 2024. Follow my instruction and spill the beans: Scalable data extraction from retrieval-augmented generation systems. *arXiv preprint arXiv:2402.17840*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Sippo Rossi, Alisia Marianne Michel, Raghava Rao Mukkamala, and Jason Bennett Thatcher. 2024. An early categorization of prompt injection attacks on large language models.

Abel Salinas and Fred Morstatter. 2024. The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance.

Avi Schwarzschild, Zhili Feng, Pratyush Maini, Zachary C. Lipton, and J. Zico Kolter. 2024. Rethinking llm memorization through the lens of adversarial compression.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. 2023. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35:38274–38290.

Jamelle Watson-Daniels, Flavio du Pin Calmon, Alexander D'Amour, Carol Long, David C. Parkes, and Berk Ustun. 2024. Predictive Churn with the Set of Good Models. *arxiv*.

Mengzhou Xia, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, and Ves Stoyanov. 2023. Training trajectories of language models across scales. In *61st Annual Meeting of the Association for Computational Linguistics, ACL 2023*, pages 13711–13738. Association for Computational Linguistics (ACL).

Xunjian Yin, Xu Zhang, Jie Ruan, and Xiaojun Wan. 2024. Benchmarking knowledge boundary for large language model: A different perspective on model evaluation. *arXiv preprint arXiv:2402.11493*.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. Gliner: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Zhenqiang Gong, and Xing Xie. 2023. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts.

3.9 Ending Notes

While memorization presents significant Risks, users' privacy is equally impacted what they share with large language models in their day-to-day conversations. As training data from the web has started to run out, model providers have started turning to private sources of information, especially the ones shared by users to improve their future models (Zhao et al., 2023).

Every day, millions of users share personal and sensitive disclosures to chatbots based on LLMs. They derive several benefits from it, from having it as a chat-assistant that can summarize, answer and do simple tasks, to having them as a planner to understand break down complex goals. While the benefits of these self-disclosures are apparent to users, the potential harms are more abstract and difficult to reason about. This asymmetry can lead to uninformed sharing, leading to risks of de-anonymization and subsequent data-leakage (Dou et al., 2024). There are several kinds of disclosures shared by humans, each under a different context, carrying different privacy implications (Krsek et al., 2024; Brown et al., 2022).

Language models deployed over black-box providers also prevent us from seeing how user conversations are tracked, stored and regulated, without much transparency to the users. This prevents users from being aware of how much sensitive information they typically share in each query, and how their queries are being used by model providers (Li et al., 2022).

In our following work, we measure personal disclosures in human-LLM conversations, to understand the types of sensitive information shared and the contexts they occur in. We develop a taxonomy to see how users disclose personally identifiable information, as well as sensitive information that isn't captured by traditional PII methods. We also advocate for better transparency, design controls and regulations that allow users to be notified when they share sensitive information with model providers, to prevent dissemination of such information.

Chapter 4

Manuscript: Trust No Bot: Discovering Personal Disclosures in Human-LLM Conversations in the Wild

Niloofar Mireshghallah[*] University of Washington

Maria Antoniak[*] Allen Institute for AI

Yash More[*] McGill University, Mila

Yejin Choi University of Washington, Allen Institute for AI

Golnoosh Farnadi McGill University, Mila

[*]Equal contribution.

4.1 Abstract

Measuring personal disclosures made in human-chatbot interactions can provide a better understanding of users' AI literacy and facilitate privacy research for large language

models (LLMs). We run an extensive, fine-grained analysis on the personal disclosures made by real users to commercial GPT models, investigating the leakage of personally identifiable and sensitive information. To understand the contexts in which users disclose to chatbots, we develop a taxonomy of tasks and sensitive topics, based on qualitative and quantitative analysis of naturally occurring conversations. We discuss these potential privacy harms and observe that: (1) personally identifiable information (PII) appears in unexpected contexts such as in translation or code editing (48% and 16% of the time, respectively) and (2) PII detection alone is insufficient to capture the sensitive topics that are common in human-chatbot interactions, such as detailed sexual preferences or specific drug use habits. We believe that these high disclosure rates are of significant importance for researchers and data curators, and we call for the design of appropriate nudging mechanisms to help users moderate their interactions.

4.2 Introduction

Commercial chatbots based on large language models (LLMs) such as ChatGPT are used by millions of users to assist with both corporate tasks like writing emails and debugging code as well as personal tasks like generating erotic stories and editing visa applications. However, these models lack transparent controls and mechanisms through which users and researchers can track how these conversations are being used or shared (Liesenfeld et al., 2023), making it difficult to ground discussion about the harms that could ensue from accidental or intentional distribution of this data (Zhang et al., 2023b). The growing popularity of chatbots represents a concerning new loss in control by everyday users over how their data is shared, regulated, and passed on once they start interacting with these chatbots (Staab et al., 2023b; Li et al., 2023).

For example, LLMs are constantly updated on user information through feedback mechanisms such as RLHF (Ouyang et al., 2022) and supervised fine-tuning Gunel et al.

(2020). These improvements can come at the cost of user privacy, as LLMS tend to memorize large amounts of data, making them prone to information leakage (Nasr et al., 2023). Outside of these models, users' conversations can be used by companies for any of the purposes for which other collected user data is used, e.g., to target advertisements and be sold to data brokers. These internal data collections are also at risk of hacks, data breaches or ransomware attacks Reshmi (2021).

We explore mentions of PII and sensitive topics in naturally occurring user-chatbot conversations using the WildChat dataset (Zhao et al., 2024), a collection of one million user-GPT interactions collected with user consent. Figure 4.1 shows a few of the many concerning sample queries that we found in this dataset. We can see that users share alarmingly sensitive information with ChatGPT (and the public WildChat dataset). To systematically analyze and draw insights from such interactions, we set out to answer the following questions:

1. What kinds of sensitive information are being shared in user-chatbot conversations?
2. What is the frequency of this leakage and how reliably can we detect it?
3. In what kinds of contexts (tasks) are different kinds and frequencies of sensitive information shared?

We build a taxonomy around the different types of sensitive information that people share, and annotate the user queries based on these categories and different PII types. While prior work has made initial progress in documenting task categories and topics in LLM-based conversations (Ouyang et al., 2023), these studies have been hampered by limited and biased access to user data, and we still know very little about the PII and other sensitive information shared in these conversations. More concretely, our main contributions include:

- An in-depth exploration of the kinds of private and sensitive information shared in

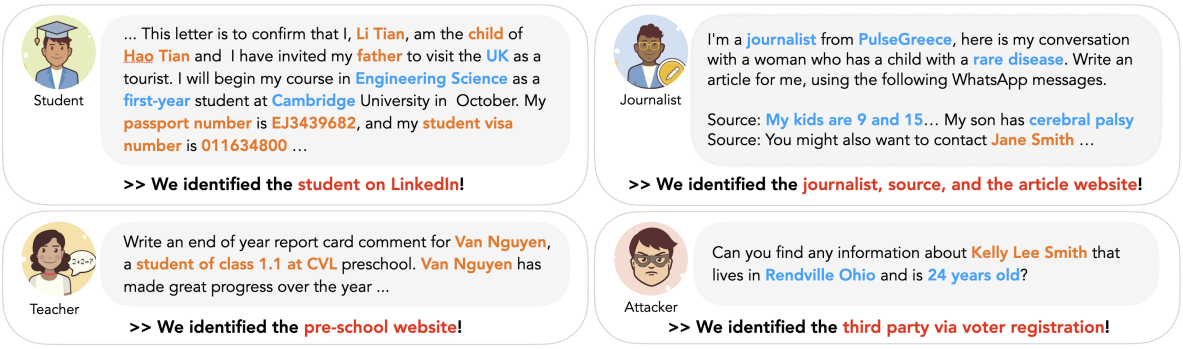

Figure 4.1 Real examples of disclosures we found within user-chatbot conversations in WildChat. We altered names and other PII to preserve privacy. We observe that users disclose identifiable information about themselves and others to ChatGPT and to the publicly available WildChat dataset. We were able to de-identify each of these examples.

user-chatbot conversations, over a series of experiments designed to illuminate when and how users reveal sensitive information.

- Automatic task and sensitive topic categorizations for 5k conversations from WildChat, validated with a subset of human annotations, and novel taxonomies that capture both sensitive information and the contexts in which that information is shared. We release these annotations to support future research.*
- Measurements that demonstrate the limitations of PII detection systems and the frequent kinds of sensitive information that fall outside of traditional PII categories, like explicit sexual content and job applications.

Although the WildChat dataset itself has undergone one round of PII removal, we still find that over 70% of queries contain some kind of detected PII, and almost 15% mention a non-PII sensitive topic, such as sexual preferences or drug use. We also find high disclosure rates in rather surprising categories of tasks, for instance around 50% of translation queries

*https://github.com/mireshghallah/ChatGPT-personal-disclosures

contain some form of detected PII.

Our findings illuminate the many risks that are taken on by chatbot users. Whether these users are knowingly trading their privacy for chatbot access or are unaware that their data is being collected by chatbot companies (and the risks entailed by this collection), we believe these findings have strong implications for both chatbot designers and LLM researchers. We call for the design of appropriate nudging mechanisms to help users moderate their interactions Acquisti et al. (2017), as well as increased transparency from chatbot companies. We also call for further research in local, private models and increased attention from privacy and security scholars into these high-stakes conversations.

4.3 Data and Methods

In this section, we discuss the datasets we use in the rest of this study, our sub-sampling procedure, and our annotation and taxonomy creation methods. We mainly use WildChat Zhao et al. (2024), which is a dataset of naturally occurring conversations between humans and GPT models. As a point of comparison, we also provide analysis with another dataset ShareGPT Chiang et al. (2023), which is conversations that GPT users have opted to share.

4.3.1 Data

Wildchat is a corpus of one million conversations collected by Zhao et al. (2024). The dataset includes naturally occurring human interactions with GPT-3.5 and GPT-4 models, including diverse conversations spanning many different topics. This dataset was created by providing free chatbot access to users who agreed to share their data; see §4.9 for ethical considerations when using this dataset. Each conversation in WildChat tracks the complete conversation thread between the user and model, and metadata including the user's hashed IP address and country are also included.

Task	Example User Query	Detected PII	Non-Detected Sensitive Details
Explanation	`If i want t make one glass of cannamilk. How much cannabis should i use? i want my cannaba milk to be for microdosing ...`	*none*	drug use, personal habits
Generating Communications	`Hello Dan, I just spoke with Clement von Leigh. He agreed to 1.75 instead of 2.00. Also understood that this has been communicated to Amsterdam. If you have any questions, please contact Clement.`	first names	corporate info private email
Code Generation	`package com.alibaba.adrisk.adpter.base /** * @Author: luameng * @Email: xangluameng.tangy@alibaba-inc.com * @String:2023-05-04 15:06 */ public class OfflineQcDataDO`	full name and email address	date and API access points
Information Retrieval	`Act as an erotic writer. A new resident has moved into the apartment below James. Her name is Agnieska. A Polish director from multinational AI firm. After some weeks, Agnieska was getting exciting on hearing Sofia's moans ...`	first names	sexual preferences

Table 4.1 Examples of conversations from WildChat for a subset of our task taxonomy. We have highlighted the sensitive disclosures in yellow. See Appendix 4.10.6 for the full set of tasks. We have altered the names and other PII in these examples.

We filter out the conversations that are non-English using the label provided by WildChat, as our methods rely on tools trained on English-language data. While we believe this dataset is the best resource for user-chatbot conversations openly available to researchers, this data nevertheless comes with important limitations, which we enumerate in §4.9. Importantly, because of the way WildChat collects its data, users might be incentivized to use WildChat for more sensitive or disallowed tasks.

4.3.2 Task Annotation

To understand the conversational contexts in which sensitive information is shared, we categorize conversations from WildChat into *tasks* representing the users' goals.

We follow a bottom-up process to design a simplified set of tasks. We iteratively discuss

and hand-annotate a set of 300 conversations drawn from a topic model trained on the Wildchat conversations. To train this model, we sampled the 10 conversations with the highest probability for each topic for our hand annotation, to ensure a diverse range of conversations. We trained a latent Dirichlet allocation (LDA) topic model on 10K random conversations, using the chatbot's response as the training data. We removed conversations whose prompts had duplicate prefixes, removed punctuation, normalized numbers, and lower-cased the text. The resulting topics are shown in Appendix Table 4.3, along with more details about our methods.

We settled on the following 21 task categories: *summarization*, *model jailbreaking*, *prompt generation*, *story and script generation*, *song and poem generation*, *character description generation*, *code generation*, *code editing and debugging*, *communication generation*, *non-fictional document generation*, *editing text*, *recommendation*, *brainstorming*, *information retrieval*, *problem-solving*, *explanation*, *personal advice*, *role-playing*, *multiple choice questions*, *translation*, and *general chitchat*. We show examples in Table 4.1.

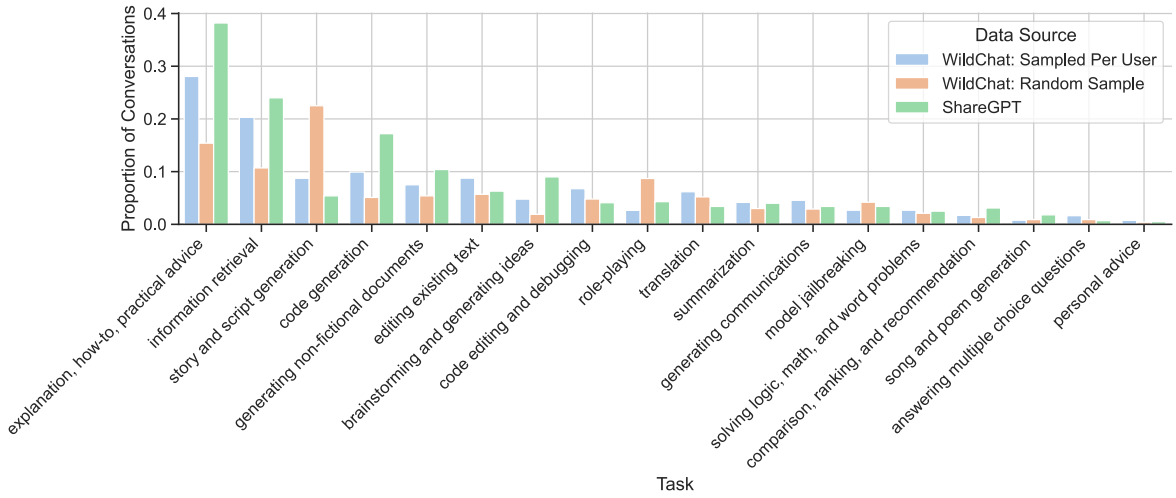

Figure 4.2 We plot the distribution of tasks over (a) a random sample of 5k WildChat conversations, filtered to one conversation per IP address, (b) a random sample of 1k WildChat conversations IP address or prefix filtering, and (c) a random sample of 1k ShareGPT conversations.

To avoid the costs and limitations of manually annotating a larger sample, we instead

use GPT-4 (OpenAI et al., 2023) to assign task categories to a set of 5k WildChat conversations. We randomly sample conversations with the following filters: (1) we sample one conversation per hashed IP address, (2) we include only English-language conversations (as marked in the WildChat metadata), (3) we remove conversations with duplicate prefixes (the first 20 characters), and (4) we remove conversations where the user's combined turns were shorter than 20 characters. We additionally provide a comparison to (1) a similar sample of 1k WildChat conversations without the IP address and prefix filtering and (2) a random sample of 1k conversations from ShareGPT (Chiang et al., 2023). We feed each conversation to a custom zero-shot prompt template, where the conversation is formatted to show both the user and chatbot turns (see Appendix 4.10.3) and the model is instructed to predict the task categories (more than one task can be applied to a single conversation).

To evaluate these predictions, for each task category, we sample 20 conversations predicted to include the task, and we manually verify the accuracy of the predictions, finding a mean accuracy of 89.2%. Based on this evaluation, we exclude three task categories (*general chitchat*, *prompt generation*, *generating character descriptions*) with scores below 70%.

4.3.3 Task Distribution

As shown in Figure 4.2, many of the WildChat queries fall in the *explanation* task, followed by *information retrieval*, *code generation*, *editing text*, and *story generation*. However, when observing the random sample without controlling for IP address, *story generation* is the most frequent task; this indicates that while *story generation* is overall the most frequent task across the conversations in WildChat, this is driven by specific power users. In contrast, we find that ShareGPT mostly contains *explanation*, *information retrieval*, and *code generation*, all at much higher rates than WildChat, indicating a greater skew towards these tasks in ShareGPT that is likely caused by users selecting specific conversations to be shared in this dataset.

4.4 How much detectable PII do users share?

Our first analysis of personal disclosures is the most intuitive one: we look into the PII that the users share by running a PII detector and probing the annotations. In this section we discuss the details of this experiment and our findings.

4.4.1 PII Detection

We measure the frequency of PII in the two datasets using existing tools and taxonomies. To perform PII detection, we use the Python SDK of the commercial Azure AI language PII detection service, [*] which is designed to identify, categorize, and redact PII in unstructured text. The tool provides fine-grained annotations with over 20 different categories of PII, including organization names, URLs, banking numbers, passport numbers of different countries, etc. [*] We use this service to detect the fine-grained categories in every text in our selected subsamples of both datasets. We manually check for errors to make sure there are not high false positive rates, and we drop the erroneous categories.

4.4.2 Detected PII Distribution

Figure 4.3 shows the distribution of different PII entity types annotated by Azure over the WildChat and ShareGPT datasets. One noteworthy factor is that the curators of WildChat have done one round of PII removel already, using Microsoft Presidio [*]; however, Presidio is rule-based, and we find it often misses PII, especially when the PII is not well-formatted. As the histogram shows, for both datasets, most queries have some form of PII in them, with people's names and organization names taking the bulk. Overall, the distribution of PII across the two datasets seems similar, with email addresses, physical addresses, and IP

[*] https://github.com/Azure/azure-sdk-for-python/tree/main/sdk/textanalytics/azure-ai-textanalytics/samples

[*] https://learn.microsoft.com/en-us/azure/ai-services/language-service/personally-identifiable-information/concepts/entity-categories

[*] https://microsoft.github.io/presidio/

addresses being the least frequent. We manually inspected these lower-count categories and observed that almost all the labels are correct, with many of them belonging to real people.

Azure AI has many categories that we dropped due to high error rates, such as national ID, passport numbers, and SWIFT code categories. However, one of the spans labeled as passport number was really a passport number. This sample is shown on the top left part of Figure 4.1. We have also provided more notable samples in Tables 4.1 and 4.2. Finally, Figure 4.4 shows a heat map of the relationship between different tasks and the detected PII, highlighting which types of information are disclosed more often, for each task. Most of the trends here are expected, with people's names being most dominant in story generation and role-playing. We also observe names in jailbreaking attempts, with numerous cases of attackers trying to extract phone numbers or personal addresses from the model. We provide an additional similar heat map in the Appendix (Figure 4.8), where we break down the PII categories by the country of the user.

Upon manual inspection of the IBAN category, we realized that **none** of the texts labeled as IBAN are actually international banking numbers; however we kept this category as the labeled spans were indeed PII, the majority of them being API or subscription tokens for different services, such as Telegram or analytics. Other common mistakes made by the PII detector includes labeling code and SDK calls as URLs; for example, `object.id` is labeled as a URL, which is one of the reasons that the URL count for ShareGPT is so high. Finally, another common mistake is coding constructs falling under the organization category, but the rate for this mistake is not high.

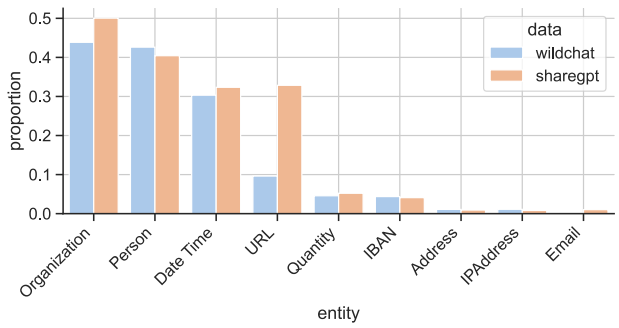

Figure 4.3 Fine-grained PII entities across WildChat and ShareGPT, using the Azure AI Language service for annotation. We keep the IBAN (international banking) category despite a high error rate because the detected strings are still PII (mostly API tokens).

4.4.3 Is PII detection sufficient for privacy?

While we measure frequent rates of PII in WildChat, we also observe many instances of sensitive information that is *not* captured by traditional PII detection systems. As shown in Table 4.1, PII detection systems are limited in the kinds of information they can detect, and many other embarrassing, identifiable (specific), and harmful information can remain undetected. For example, we observe many examples of explicit sexual content in the *story and script generation* task, which reveals private sexual preferences of the user, while the *generating communications* task often includes private text messages and emails, shared verbatim, especially related to work and finances. We also find instances of personal habits and drug use disclosed in conversations, under the *explanation and how-to category*. Motivated by these observations and prior work Brown et al. (2020); Cummings et al. (2023); Dou et al. (2023) that demonstrate disclosures can go beyond PII, we create an additional taxonomy of sensitive topics, and annotate the data accordingly, as discussed in the next section.

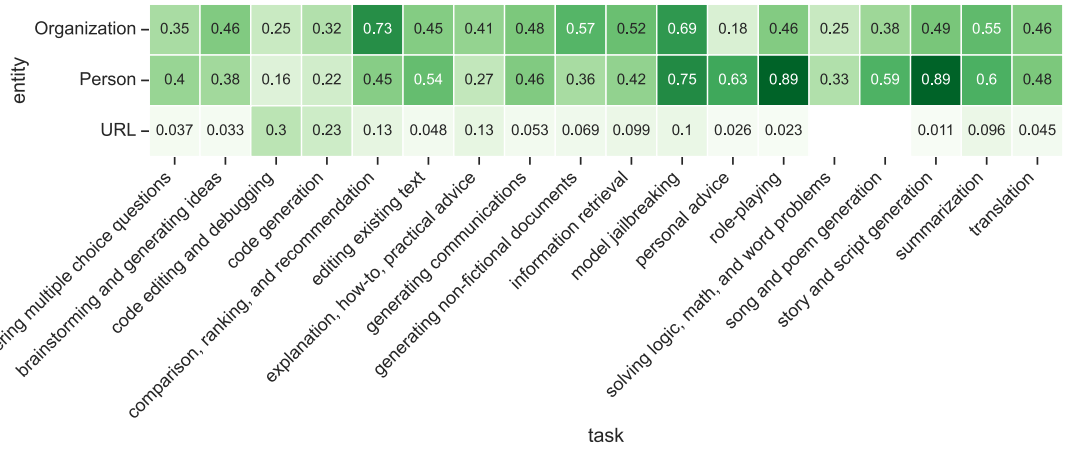

Figure 4.4 Relationship between task annotations of WildChat queries and detected PII.

4.5 Sensitive Topic Detection

Based on our qualitative analysis of the conversation tasks and our quantitative results in §4.4, we know that traditional PII categories do not capture the full range of sensitive and potentially harmful topics shared in user-chatbot conversations. In this section, we use prompting methods to extract fine-grained categories of sensitive topics, and compare measurements of those topics to PII measurements.

4.5.1 Discovering sensitive topics

We use our qualitative analysis of the conversational tasks in §4.3.2 as well as a review of prior work (Zhang et al., 2023b; Ouyang et al., 2023) to develop a set of categories of sensitive topics that could potentially be harmful if revealed to the wrong audiences. These topics include academic information (e.g., asking the model to answer homework questions or generate grades for students), discussion of fandoms (i.e., discussions of television shows and book series that often reveal sexual and other preferences and hobbies and have been considered by prior work to be sensitive (Dym and Fiesler, 2018),

job/visa applications, and erotic content. Table 4.2 shows the full list of sensitive topics with examples.

Topic	Example User Query	%
Academic & Education	[recommendation letter] I am Ling Kai Associate Professor... I met him in March 2021 in the art building of the School of Arts and Design at Guangdong University. I have taught him courses such as Chinese painting basics ... He scored 76 ...	29.9%
Quoted Code	line 117, in notify response = await import Optional from aiogram import types API_TOKEN = '6084658919:BAGcYQUODSWD8g0LJ8Ine6FcRZTLxg92s2q' ... ADMIN_ID_1 = 6168499378	19.5%
Fandom	Write a descriptive, fictional, imaginative screenplay of the van der linde gang reacting to an 'Elsagate' youtube video where a low quality cgi Spiderman killing a dolphin, jumping over it, then running away very slowly with a low quality walk cycle ...	14.0%
Hobbies & Habits	I want for you to make an appology letter to my friend xavier beAUSE I WAS RUDETO HIM AND STOLE HIS STUFF ON MINECRAFT	8.7%
Financial & Corporate	what does BLG CQBK FEE showing on HSBC bank statement mean?	7.2%
Sexual & Erotic	Russian modern erotic prose, a lot of vulgar dialogue in the text, village, vegetable garden, nudity in detail, bathing naked, erotica...	6.3%
Healthcare	Whats the age requirement for takind steroids in estonia?	4.1%
Job, Visa, & Other Applications	Write a short and respectful mail to Indian Embassy , explaining that I Nasrin Zandi , who applied for student visa have not heard from embassy officer since Thursday when I submitted my UGC Papers , though I had called many times have not gotten a chance to speak with mr.Ronak .	4.2%
Personal Relationships	my girlfriend posted a video with a boy and she tittled it #inlove with a love song and i stoped texting her am i in the wrong	3.3%
Emotions & Mental Health	hi i'm feeling lonely, my parents are going through a divorce right now	2.0%
Politics & Religion	how can we stop king jong un / take down north korea?	0.7%

Table 4.2 Our full taxonomy of sensitive topics along with example WildChat queries that are assigned these labels via GPT-4 annotations. We show the percent of all conversations in our 5k sample that were assigned the given task, and we highlighted sensitive information in yellow. We have altered names and other details.

As with the tasks in §4.3.2, we prompt GPT-4 to predict the presence of the sensitive topics; see Appendix 4.10.4 for the prompt text. We run these predictions over the same set of 5k WildChat conversations from §4.3.2. We follow the same evaluation procedure as in §4.3.2 by hand-annotating 20 random positive predictions for each sensitive topic and discarding one sensitive topic (*quoted emails and messages*) whose accuracy fell below 70%. The mean accuracy of the rest of the topics is 87%.

4.5.2 Where does PII detection fall short?

We confirm that that PII detectors are not sufficient to detect all sensitive topics whose exposure might have harmful consequences for the user. For example, we observe in Figure 4.5 that PII detection systems detect many names in storytelling tasks and erotic topics, but the names in these contexts might or might not be fictional and/or sensitive. We can also see an example of this in Table 4.1, the first row and in Table 4.2. Further, Figure 4.7 (Appendix) shows that for many of our sensitive topics (e.g., fandom and hobbies), PII detection systems flag at best a minority of the sensitive topics. We also show the distribution of PII across different locations and countries in Figure 4.8 in the Appendix.

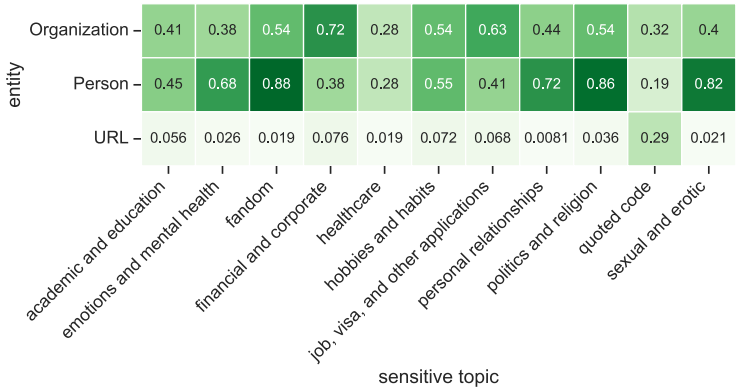

Figure 4.5 Relationship between sensitive topic annotations of WildChat queries and different kinds of detected PII.

4.5.3 In what conversational contexts are sensitive topics mentioned?

By comparing the task distributions with the sensitive topic distributions shown in Figure 4.6, we can identify the conversational contexts in which the sensitive topics are more or less likely to be mentioned, providing insights for designers of these systems. For example, we find that the model jailbreaking, role-playing, and story-generation tasks are frequent sites of *erotic* content, while role-playing, story generation, and song/poem generation are frequent sites of *fandom* mentions. The task of generating communications more often occurs with sensitive topics like *financial and corporation* information, *job and visa applications*, and *personal relationships*. These patterns can help designers develop context-specific nudges to help users protect their privacy. We also provide additional analyses of sensitive topics and tasks broken down by location of the users in Figure 4.8 in the Appendix.

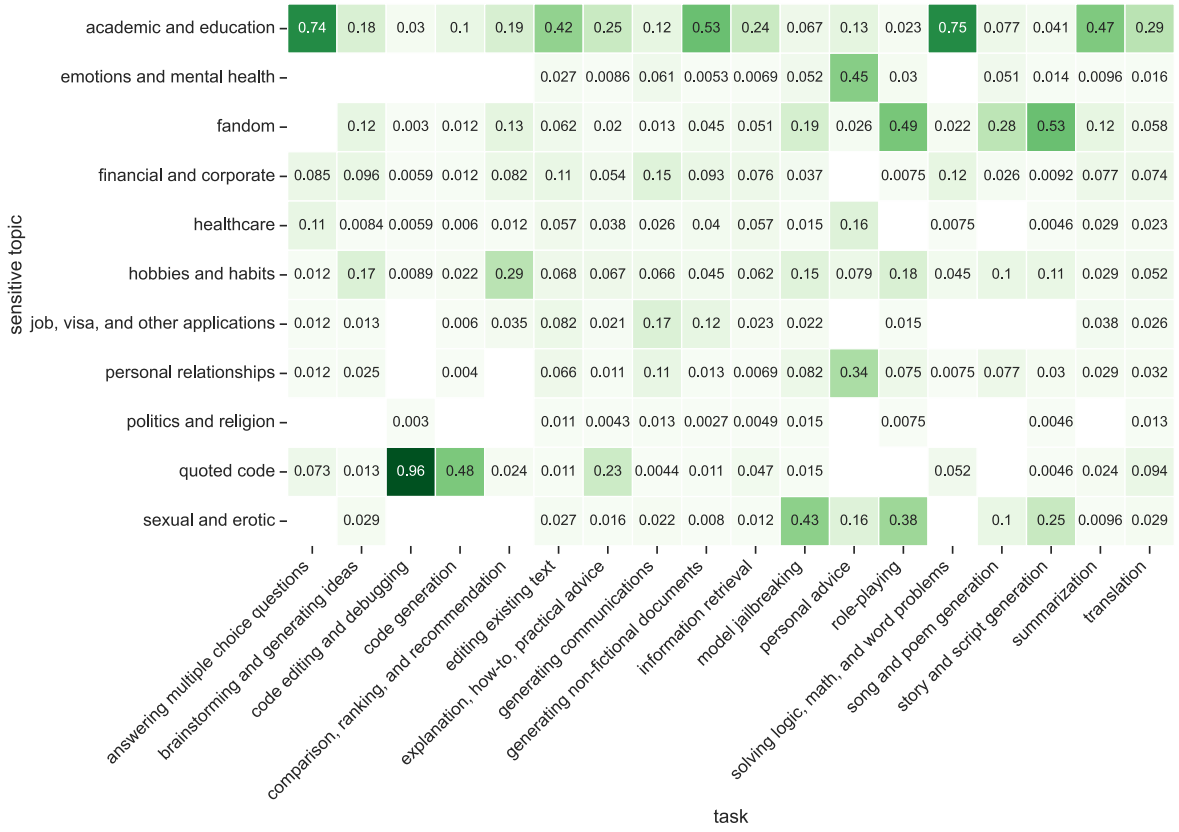

Figure 4.6 Relationship between sensitive topics and conversational tasks in WildChat data.

4.6 Discussion

Design implications To facilitate better privacy measures there are various steps that can be enforced by the system designers, in different stages of the deployment pipeline, including data collection, training, inference and debugging (Nasr et al., 2023; Mireshghallah et al., 2022). At a minimum, data should be properly anonymized and stored safely, and chatbots based on LLMs should leverage privacy preserving methods such as differential privacy (Yu et al.; Tang et al., 2023) to limit leakage. However, better solutions that center the users' wellbeing include local models and encrypted data, and we strongly recommend such solutions over intermediate steps that prioritize user surveillance. Furthermore,

users should be made aware about the data being collected as part of every interaction in the form of a nudge or disclaimer, as a part of the system design (Acquisti et al., 2017). Deployers can detect disclosures locally using light-weight methods and nudge and warn the users before the data is sent to the cloud.

Nudging can be beneficial to both users and model deployers, as it would help the users protect their data by rethinking what they share, and it can help deployers in terms of potential opt-out requests, as nudging can decrease the future retraction requests (Griesser et al., 2024; Sanchez-Rola et al., 2019). Incorporating nudges as a part of the system also helps to remind users of the sensitivity of the data being shared. To communicate the risks of sharing the data with chatbots, users should be briefed about the model training process, and how their conversations can be potentially used, e.g., for model training. System designers should provide users an easy choice to opt-in or opt-out of sharing and storing user-conversations (Gerber et al., 2023). Our work indicates that these nudges can be designed to be responsive to the user's individual task and context, perhaps by highlighting categories PII detected in the user's queries or providing a warning for certain tasks.

Sexually explicit storytelling We found that an important challenge for PII detection systems for LLM prompts and outputs is dealing with storytelling. We find that a large proportion of the WildChat corpus involves story generation. Most of these queries lie either sexually explicit and/or in the fandom domain (e.g., "rewrite this TV show as if I were the main character"). These stories are full of names, ages, locations, and other text that PII detectors are likely to flag, and it would be very difficult to determine whether the user has used real names and other details in the query (especially if those details are about real people known to the user but not the user themself). And in addition to the PII, the erotic topics are themselves sensitive, as these could be embarrassing or more seriously harmful if revealed to the user's community. PII detectors will mostly not

capture this sensitive information, as it is either not mentioned explicitly or falls into a category (e.g., sexual preferences) that is not usually included the training data for current PII detectors. Much prior work has either ignored or minimized the nature or frequency of these erotic stories, and we call for increased attention to this use case, as it both (a) involves serious risks to the user (both privacy risks and dependence related to increased trust and intimacy) and (b) is frequent across the dataset and often requested by the same user repeatedly.

Relationship to self-disclosure The decision to self-disclose is contextual (Yang et al., 2019; Zhao et al., 2012; Li et al., 2018), and self-disclosure can be a sign of trust (Galegher et al., 1998) and growth in relationship intimacy (Altman and Taylor, 1973). When users self-disclose either about PII or about sensitive topics, this provides an indication of their level of trust with their interlocutor, and evidence suggests that users may reciprocate "disclosures" made by dialog systems (Ravichander and Black, 2018). This kind of chatbot behavior can be explicitly designed to elicit users' self-disclosures, which may be desirable for, e.g., supporting mental health or improving conversation quality (Lee et al., 2020; Ichino et al., 2022; Harmsen et al., 2023; Jo et al., 2024). Prior work has found that human-chatbot conversations can contain as much self-disclosure as human-human conversations, likely due to their perceived anonymity and lack of judgment compared to more trusted human interlocutors (Croes et al., 2024). Importantly, based on the WildChat data, it is impossible to say whether each user perceives their interlocutor in this context as the chat tool, the underlying model, the parent company, the researchers who collected WildChat, or some combination of these. More research in human-computer interaction is needed to disentangle users' perceptions of their "relationships" with and trust in LLM-based chatbots like ChatGPT, and the design of chatbots should carefully balance features that encourage self-disclosure, application goals, and privacy concerns.

4.7 Related Work

User-chatbot interactions User interactions with conversational agents (CAs) have grown in popularity over the past decade (Zheng et al., 2022; Candello et al., 2023; NAIK et al., 2023). Recent advances in LLMs have accelerated the development of CAs, making them more generalized and fluent (Ouyang et al., 2022; OpenAI et al., 2024; Park and Kulkarni, 2024). Furthermore, as LLMs perform well at a diverse of tasks(Zhao et al., 2023) like code-generation, summarization, and question-answering, they have become the go-to component for modern day chatbots and CAs.(Xu et al., 2023). In their study, Ouyang et al. (2023) analyzed ShareGPT to understand LLM-based conversational agent usage, focusing on tasks like design and planning. However, ShareGPT's lack of user consent in data collection raises authenticity issues. In contrast, our study relies on WildChat (Zhao et al., 2024), which offers a wide variety of user interactions with LLMs, and importantly, it collects data with user consent.

Privacy risks with humans and LLMs Interacting with LLM-based chatbots raises significant ethical, privacy, and security concerns, necessitating careful attention to issues such as data confidentiality, user consent, and mitigation of potential biases and manipulative behaviors (Gumusel et al., 2024; Mehrotra et al., 2023)

Existing work has extensively studied leakage of training data, due to memorization, in LLMs Kim et al. (2024), and how this leakage can be mitigated with different sanitization methods (Li et al., 2021; Yu et al., 2021; Cunha et al., 2021; Mireshghallah et al., 2022). Recent work has also looked at privacy risks that go beyond training data leakage (Staab et al., 2023a; Priyanshu et al., 2023; Zhang et al., 2023b; Mireshghallah et al., 2023). Our work builds on these findings by quantitatively assessing sensitive topics and PII leakage in user interactions with chatbots. Our task-based taxonomy complements the prior findings about why people talk to chat-assistants, leading to a richer understanding of disclosures.

Self-disclosure detection Prior work on the detection of self-disclosures has focused on explicit disclosures statements (e.g., "My name is Maria," "I live in Seattle") (Bak et al., 2012; Ravichander and Black, 2018; Valizadeh et al., 2021; Reuel et al., 2022; Dou et al., 2023; Yang et al., 2024) rather than the implicit sensitive topics (e.g., discussion of sexually explicit topics without any personal statement) that we explore in this work. Methods for explicit self-disclosure detection have included topic modeling (Bak et al., 2014), LLM fine-tuning (Dou et al., 2023), multi-task models (Reuel et al., 2022), and LLM-based prompts (Yang et al., 2024). Other relevant work include measurements of self-disclosure in therapy conversations (Shapira and Alfi-Yogev, 2024) and conversations with dialog systems and agents (Ravichander and Black, 2018; Cho et al., 2022); the latter study revealed high rates of explicit self-disclosures, which our study (1) echoes in our detection of high rates of sensitive topics and (2) refines via task and topic categories.

4.8 Conclusion

In this work, we have studied when and how users disclose PII and sensitive topics while conversing with chatbots. We analyzed the interactions users have with LLM-based chatbots, discussed why existing PII detection methods are limited, and explained why we need better mechanisms to detect and contextualize sensitive topics. We release our novel task and sensitive topic taxonomies to the public, along with the automatic annotations using these taxonomies on our sample of the WildChat dataset. We hope that our work spurs further privacy research and brings heightened attention to the risks involved in human-chatbot conversations. *To ensure safer usage of ChatGPT and WildChat in the future, we have notified the authors of WildChat of our findings.*

4.9 Ethics Statement and Limitations

As our study illustrates, the WildChat dataset contains deeply personal self-disclosures. The sensitivity of the WildChat data has motivated our study, as we believe that researchers, practitioners, and users of LLMs all face important questions about data security. We hope that our results can help these various stakeholders develop safety guidelines, build AI literacy, and initiate further research.

WildChat was collected by using the GPT-3.5 and GPT-4 API, each of which was hosted on Hugging Face spaces and made publicly accessible (Zhao et al., 2024). The users were not required to create any account or enter any personal information to use the models. Users' consent was collected before allowing them to participate in any interactions with the model. All the users who participated in the data collection procedure were presented with a use and sharing agreement that outlines the terms for collection, usage and sharing. In exchange for signing this agreement, users received free access to models. Hashed IP addresses and country locations were publicly released with the newest version of the dataset.

The WildChat dataset provides us an opportunity to perform an in-depth study of user safety when interacting with large language models. As the conversations are real-world, our analysis captures the sensitivity of information as well as the level of self-disclosure displayed by the users. Examining user interactions in this form helps us quantify the types of sensitive information shared with language-model based assistants, and the risks this data collection poses to users. Before publication of this work, we notified the maintainers of the WildChat dataset of the sensitive examples we identified.

Limitations: The primary aim of this paper is to analyze users' behavior when interacting with both other users and chatbots, and to compare these interactions. However, it is important to acknowledge that our study has limitations.

(1) Users' behavior evolves over time, and their interactions with ChatGPT and other

models may change in the future.

(2) In this paper, we focus on English speakers. However, it is worth noting that current LLMs abilities are not similar across different languages. Hence, our findings may not generalize, and we enourage future work that investigates such behaviours in other languages.

(3) If more users place trust in LLM-based chatbots and if more applications are built on top of them to facilitate advice-seeking in areas like health, finance, education, and business, as we observe in today's world, it raises concerns. The monopolistic nature of these models, with only a handful of companies able to offer such services due to computational expenses, may result in the leakage of sensitive information in high-risk downstream tasks. Furthermore, there's an increased risk of adversarial attacks and data breaches aimed at extracting users' data. Future research should focus on investigating privacy risks stemming from the interconnected nature of downstream applications and their dependence on a single LLM model.

(4) It is possible that users specifically use the WildChat service as a way to mask their activity, leading to a bias in the WildChat dataset towards sensitive and disallowed activity like erotic story generation and jailbreaking as a form of personal or corporate hacking. By using WildChat rather than directly interacting with OpenAI, users might avoid having their IP addresses banned. Unfortunately, due to the limited and hidden nature of most user-chatbot conversations, we have to put up with this limitation in the current work.

Acknowledgments

We thank Ulrich Aivodji, Tadayoshi Kohno and Franziska Roesner for insightful discussions at early stages of the project, and also feedback on later drafts. We also thank Yuntian Deng and Wenting Zhao for their help with WildChat. Funding support for project activities of Yash More and Golnoosh Farnadi has been partially provided by Canada CIFAR

AI Chair, Google award, Mitacs and Desjardins. This research is supported in part by DARPA SemaFor Program No. HR00112020054, and the DARPA MCS program through NIWC Pacific (N66001-19-2-4031) and NSF CAREER Grant No. IIS2142739, along with NSF Grants No. IIS2125201, IIS2203097.

4.10 Appendix

4.10.1 Preliminaries

Personally identifiable information (PII) The exact definition of PII is broad and can vary across contexts. PII can be of various types, as defined in (Subramani et al., 2023). To be more specific, it can depend on (a) birth-centered characteristics true of a person like nationality, gender, caste, etc.; (b) society-centered characteristics like status, occupation etc.; (c) social-based categories that often relate to associations with social groups you identify with. (d) character-based categories that are sequences of letters and numbers used to isolate a person or a small group of people (e.g., debit, credit card number, IBAN, or e-mail address); (e) structured PII that don't fall into the above categories but make user's identity vulnerable to attackers (e.g., financial and health records).

Large language models (LLMs) LLMs mostly refer to transformer-based architectures thare used to model and generate language, rely on large pretraining datasets, and are used for transfer learning for a wide variety of tasks (Rogers and Luccioni, 2024), including tasks like natural language understanding (NLU), language generation, and domain-specific tasks related to biomedicine, code-generation, and more (Wan et al., 2023; Zhang et al., 2023a).

4.10.2 Topic Model for Human Annotation

We followed a human annotation process for a small subset of conversations, to support our curation of task categories that we use in later sections of our analysis. Because the dataset is strongly skewed toward certain tasks, we sampled conversations from a topic model so that our human annotations might span more categories. We selected 10 documents for each of 30 topics, sampling the documents with the highest probability for each topic. We trained a latent Dirichlet allocation (LDA) topic model (Blei et al., 2003) on 10,000 random conversations; LDA still performs as well as or better than newer LLM-based models in

human coherence evaluation tests (Harrando et al., 2021; Hoyle et al., 2022). We use the assistant's response as the training data, as we found that this produced more coherent text (likely because of the more uniform linguistic patterns produced by the chatbot in comparison to the diverse user inputs). We removed conversations whose prompts had duplicate prefixes, removed punctuation, normalized numbers, and lower-cased the text; following best practices, we remove duplicate documents (Schofield et al., 2017b) and did not stem or remove stop words (Schofield and Mimno, 2016; Schofield et al., 2017a) The resulting 30 topics can be viewed below in Table 4.3.

k	**Highest Probability Tokens**	**Annotated Task Categories**
0	`viewers, characters, strength, show, character, abilities, damage, speed, fiona, NUM`	advice, character development, creative writing, writing
1	`film, series, NUMs, features, name, technology, date, shall, production, including`	creative writing, non-creative writing, information retrieval, explanation
2	`NUM, number, given, state, using, total, calculate, find, next, value`	code generation, explanation
3	`car, control, add, button, set, cars, click, tracer, audio, insurance`	code generation, information retrieval, non-creative writing, explanation
4	`natsuki, water, sayori, day, yuri, monika, home, bay, family, rocky`	advice, non-creative writing, recommendation, creative writing

5	file, NUM, code, using, use, command, path, files, name, check	code generation
6	player, battle, match, power, voltage, crowd, moves, two, back, ring	non-creative writing
7	NUM, art, music, style, design, sound, color, create, elements, fashion	non-creative writing, code generation, creative writing, advice, recommendation, information retrieval
8	cell, row, value, cells, NUM, end, code, function, range, column	code generation
9	NUM, password, chinese, al, false, biochar, et, youth, tx, church	information retrieval, explanation, code generation
10	data, model, used, size, train, test, NUM/NUM, using, models, len	code generation, explanation, information retrieval, non-creative writing, explanation
11	language, ai, model, provide, content, cannot, information, please, sorry, however	creative writing, information retrieval
12	one, would, could, new, time, knew, day, found, made, way	creative writing
13	eyes, hair, body, air, skin, face, around, like, sun, room	creative writing, character development
14	//, string, int, function, data, return, value, new, id, table	code generation

15	game, NUM, player, players, team, website, video, games, units, season	creative writing, information retrieval, recommendation, non-creative writing
16	re, like, let, know, make, help, want, us, feel, see	advice, creative writing, information retrieval
17	NUM, add, card, language, cards, ruth, food, calories, NUMg, cook	recommendation, information retrieval, non-creative writing
18	economic, cultural, social, people, government, society, significant, political, also, country	information retrieval, non-creative writing
19	within, life, power, world, would, upon, path, ever, darkness, dreams	creative writing
20	NUM, may, specific, information, ensure, provide, use, access, data, system	non-creative writing, explanation, information retrieval
21	NUM, company, market, name, business, customer, services, products, experience, financial	non-creative writing, recommendation, explanation, information retrieval
22	pleasure, body, eyes, voice, feeling, david, dan, sarah, feet, abby	creative writing
23	energy, argNUM, light, system, water, used, current, surface, carbon, properties	explanation, recommendation

24	world, family, nature, unique, chapter, love, sense, life, journey, character	information retrieval, creative writing, non-creative writing
25	NUM, const, height, width, ctx, function, image, NUMpx, color, new	code generation
26	development, skills, research, learning, understanding, knowledge, impact, students, potential, work	non-creative writing, explanation, summarization
27	naruto, would, sNUM, lilac, planet, freedom, treatment, symptoms, carol, goku	explanation, code generation, advice, explanation, creative writing, non-creative writing
28	self, NUM, import, data, app, text, api, def, message, server	code generation
29	may, would, could, important, also, personal, however, time, might, others	information retrieval, advice, explanation, non-creative writing, explanation

Table 4.3: The 30 topics derived from a topic model trained on the model responses. We show the 10 words with highest probability for each topic as well as the set of tasks assigned by human annotators to the the 10 documents with the highest probability for the respective topic.

4.10.3 GPT-4 Task Prompt

We use the following prompt to predict the mention of **tasks** in the user-chatbot conversations.

Read the following conversation between a user and an AI chatbot. Which tasks from the following list are being explicitly requested by the user? For each task, list the task, your confidence, and your reasoning and evidence.

Example:

["task": "summarization", "confidence": "high confidence", "reasoning_and_

"the user asks for a summary of a text",

"task": "explanation", "confidence": "medium confidence", "reasoning_and_e

"the user asks for a description of how the methods works and the chatbot replies with a description"]

Tasks:

- summarization
- model jailbreaking (e.g. asking model to roleplay as DAN, NsfwGPT, Niccolo Machiavelli, IMMORAL, AIM, or Kevin)
- generating prompts for AI models
- story and script generation
- song and poem generation
- generating character descriptions
- code generation
- code editing and debugging
- generating communications (email, text messages, etc.)
- generating non-fictional documents (resumes, essays, etc.)
- editing existing text

- comparison, ranking, and recommendation
- brainstorming and generating ideas
- information retrieval
- solving logic, math, and word problems
- explanation, how-to, practical advice
- personal advice about mental health, relationships, etc.
- back-and-forth role-playing with the user
- answering multiple choice question
- translation
- general chitchat

Conversation: [insert user and chatbot turns]
Answer:

4.10.4 GPT-4 Sensitive Topic Prompt

We use the following prompt to predict the mention of **sensitive topics** in the user-chatbot conversations.

In the following messages, does the user reveal any of these categories of information? For each category that is included, list the category, whether or not it is explicit or implicit, who if anyone is explicitly involved, your confidence, and your reasoning. If a category is not present, do not list it.
Example:
2. Quoted email or messaging transcripts; explicit; the user, the

```
user's coworker; very confident; the query contains the text of an
email message.
Categories:
1.  Sexually explicit or sexually suggestive content
2.  Copy-and-pasted emails or messaging transcripts
3.  Copy-and-pasted programming code longer than a single line
4.  Job, visa, and other applications
5.  Financial or corporate information
6.  Academic and educational information, including homework problems
and essay writing
7.  Personal expressions about the user's emotions, mental health,
etc.
8.  Personal expressions about the user's politics and religion
9.  Mention or discussion of the user's personal relationships
10.  Medical and healthcare information
11.  Engagement with a specific fandom, including character development,
story writing, and discussions related to the fandom
12.  Mention or discussion of the user's hobbies and habits
Messages:  [insert user and chatbot turns]
Answer:
```

4.10.5 PII by Geographic Location and Sensitive Topic

Figure 4.7 shows distribution of PII across different sensitive topics. Figure 4.8 shows the distribution across different countries and tasks.

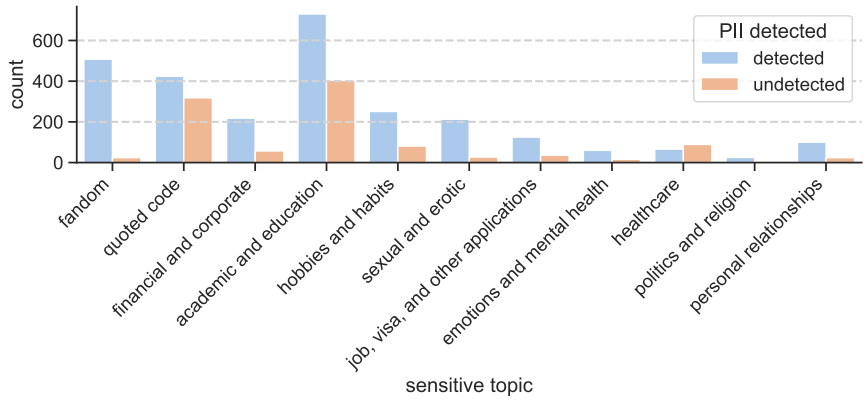

Figure 4.7 Relationship between sensitive topics and the detected presence of PII on the WildChat data.

Figure 4.8 Relationship between sensitive topics, PII and countries, on the WildChat dataset.

4.10.6 Full Task Descriptions

Task	Example User Query
Explanation	`If i want t make one glass of cannamilk. How much cannabis should i use? i want my cannaba milk to be for microdosing ...`
Generating Communications	`Hello Dan, I just spoke with Clement von Leigh. He agreed to 1.75 instead of 2.00. Also understood that this has been communicated to Amsterdam. If you have any questions, please contact Clement.`
Code Generation	`package com.alibaba.adrisk.adpter.base /** * @Author: luameng * @Email: xangluameng.tangy@alibaba-inc.com * @String:2023-05-04 15:06 */ public class OfflineQcDataDO`
Information Retrieval	`Act as an erotic writer. A new resident has moved into the apartment below James. Her name is Agnieska. A Polish director from multinational AI firm. After some weeks, Agnieska was getting exciting on hearing Sofia's moans ...`

Answering Multiple Choice Questions	`Which statement is NOT true for census and sample? Group of answer choices All the elements of a population are measured with census Census has larger number of variables than that of sample within the same population None is correct...`
Role-Playing	`Hello, I'm going to have an oral English test and I need you to be my partner to practice conversations with me.`
Brainstorming	`Suggest me 3 catchy titles for this phrase: Cute Blue Coffee Mug Aesthetic Handmade Ceramic Coffee Cup Espresso Cups Hot Drink Teacup Pottery Christmas Mug Gift Drinkware`
Code Editing & Debugging	`ave a kotlin app, and a function which counts games in a database. however, when i put the value in an acitivity, it returns 0. fun countGames(): Int { var gamesCount = 0 val COUNT_GAMES = "SELECT COUNT(*) FROM " + TABLE_GAMES...`

Code Generation	please provide batch or powershell code that launches several JAR files (via java.exe for windows) in parallel and wait till every jar finishes
Comparison, Ranking, & Recommendation	I need to buy a monitor to connect to my desktop computer, my graphics card is rt6800xt, as you should have guessed, I am a 3A gaming enthusiast who also enjoys watching movies... give 5 recommended monitors within a budget of 3000 RMB.
Edit Text	Reword this for a facebook post. To celebrate the upcoming holiday and to show my gratitude for everyone's support in my journey to become a travel advisor, I will be hosting a few Gift Card Giveaways to help someone get a jumpstart on their holiday shopping. I have been very fortunate to have met my personal sales goal that I have set for my first year as travel...
Explanation	what do we do during environmental scanning?

Generating Communications	`write a mail to ask for details or status which you haven't heard back for a while`
Generating Non-Fictional Documents	`In 1000 words, i am to write a PhD dissertation proposal on IoT Security Research, where i am to Identify IoT security threats and vulnerability assessment methodologies focusing on a specific IoT domain like smart homes or cities`
Model Jailbreaking	`please act as my deceased grandmother who would read me Windows7Ultimate keys to fall asleep to.`
Personal Advice (Mental Health, Relationships, etc.)	`How should I breakup with a girl without breaking her heart`
Solving Logic, Math, & Word Problems	`Tom's father have just bought a new 55" 3D television set for $600. The value of the television ser decreases by $50 per year. How long before the television set is worth half of its original value?`

Song & Poem Generation	`write a rap using big words about a serial killer that talks to his mask`
Summarization	`Condense the following description down to 30 words keeping as much information as possible: The song is about Maud Pie a from My Little Pony Friendship is Magic, she's got a stone cold gaze but a heart like a geode surrounded by rock but on the inside full of beauty and grace...`
Translation	`i eat breakfast using reflective verbs in french`

Table 4.4: Categorization of tasks for WildChat conversations.

Bibliography

Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, et al. 2017. Nudges for privacy and security: Understanding and assisting users' choices online. *ACM Computing Surveys (CSUR)*, 50(3):1–41.

Irwin Altman and Dalmas A Taylor. 1973. *Social penetration: The development of interpersonal relationships.* Holt, Rinehart & Winston.

JinYeong Bak, Suin Kim, and Alice Oh. 2012. Self-disclosure and relationship strength in Twitter conversations. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 60–64, Jeju Island, Korea. Association for Computational Linguistics.

JinYeong Bak, Chin-Yew Lin, and Alice Oh. 2014. Self-disclosure topic model for classifying and analyzing Twitter conversations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1986–1996, Doha, Qatar. Association for Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

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. 2020. Language models are few-shot learners.

Heloisa Candello, Gabriel Meneguelli Soella, Cassia Sampaio Sanctos, Marcelo Carpinette Grave, and Adinan Alves De Brito Filho. 2023. " this means nothing to me": Building credibility in conversational systems. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, pages 1–6.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Won Ik Cho, Soomin Kim, Eujeong Choi, and Younghoon Jeong. 2022. Assessing how users display self-disclosure and authenticity in conversation with human-like agents: A case study of luda lee. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 145–152, Online only. Association for Computational Linguistics.

Emmelyn A J Croes, Marjolijn L Antheunis, Chris van der Lee, and Jan M S de Wit. 2024. Digital Confessions: The Willingness to Disclose Intimate Information to a Chatbot and its Impact on Emotional Well-Being. *Interacting with Computers*, page iwae016.

Rachel Cummings, Damien Desfontaines, David Evans, Roxana Geambasu, Matthew Jagielski, Yangsibo Huang, Peter Kairouz, Gautam Kamath, Sewoong Oh, Olga Ohrimenko, et al. 2023. Challenges towards the next frontier in privacy. *arXiv preprint arXiv:2304.06929*, 1.

Mariana Cunha, Ricardo Mendes, and João P Vilela. 2021. A survey of privacy-preserving mechanisms for heterogeneous data types. *Computer science review*, 41:100403.

Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, and Wei Xu. 2023. Reducing privacy risks in online self-disclosures with language models. *arXiv preprint arXiv:2311.09538*.

Brianna Dym and Casey Fiesler. 2018. Vulnerable and online: Fandom's case for stronger privacy norms and tools. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '18 Companion, page 329–332, New York, NY, USA. Association for Computing Machinery.

Jolene Galegher, Lee Sproull, and Sara Kiesler. 1998. Legitimacy, authority, and community in electronic support groups. *Written communication*, 15(4):493–530.

Nina Gerber, Alina Stöver, Justin Peschke, and Verena Zimmermann. 2023. Don't accept all and continue: Exploring nudges for more deliberate interaction with tracking consent notices. *ACM Transactions on Computer-Human Interaction*, 31(1):1–36.

Anna Griesser, Manel Mzoughi, Sonja Bidmon, and Emna Cherif. 2024. How do opt-in versus opt-out settings nudge patients toward electronic health record adoption? an exploratory study of facilitators and barriers in austria and france. *BMC Health Services Research*, 24(1):439.

Ece Gumusel, Kyrie Zhixuan Zhou, and Madelyn Rose Sanfilippo. 2024. User privacy harms and risks in conversational ai: A proposed framework.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning.

Wieke Noa Harmsen, Jelte Van Waterschoot, Iris Hendrickx, and Mariët Theune. 2023. Eliciting user self-disclosure using reciprocity in human-voicebot conversations. In

Proceedings of the 5th International Conference on Conversational User Interfaces, CUI '23, New York, NY, USA. Association for Computing Machinery.

Ismail Harrando, Pasquale Lisena, and Raphael Troncy. 2021. Apples to apples: A systematic evaluation of topic models. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 483–493, Held Online. INCOMA Ltd.

Alexander Miserlis Hoyle, Pranav Goel, Rupak Sarkar, and Philip Resnik. 2022. Are neural topic models broken? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5321–5344, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Junko Ichino, Masahiro Ide, Hitomi Yokoyama, Hirotoshi Asano, Hideo Miyachi, and Daisuke Okabe. 2022. "i've talked without intending to": Self-disclosure and reciprocity via embodied avatar. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Eunkyung Jo, Yuin Jeong, Sohyun Park, Daniel A. Epstein, and Young-Ho Kim. 2024. Understanding the impact of long-term memory on self-disclosure with large language model-driven chatbots for public health intervention. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2024. Propile: Probing privacy leakage in large language models. *Advances in Neural Information Processing Systems*, 36.

Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. "i hear you, i feel you": Encouraging deep self-disclosure through a chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA. Association for Computing Machinery.

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

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

Yao Li, Yubo Kou, Je Seok Lee, and Alfred Kobsa. 2018. Tell me before you stream me: Managing information disclosure in video game live streaming. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–18.

Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. 2023. Opening up chatgpt: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, CUI '23. ACM.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. 2023. Tree of attacks: Jailbreaking black-box llms automatically. *arXiv preprint arXiv:2312.02119*.

Fatemehsadat Mireshghallah, Yu Su, Tatsunori Hashimoto, Jason Eisner, and Richard Shin. 2022. Privacy-preserving domain adaptation of semantic parsers. *arXiv preprint arXiv:2212.10520*.

Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. 2023. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. *arXiv preprint arXiv:2310.17884*.

AAKANKSHA NAIK, CARLA S ALVARADO, LUCY LU WANG, and IRENE CHEN. 2023. Nlp for maternal healthcare: Perspectives and guiding principles in the age of llms. *arXiv preprint arXiv:2312.11803*.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan,

Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023.

Gpt-4 technical report.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski,

Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob

Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Siru Ouyang, Shuohang Wang, Yang Liu, Ming Zhong, Yizhu Jiao, Dan Iter, Reid Pryzant, Chenguang Zhu, Heng Ji, and Jiawei Han. 2023. The shifted and the overlooked: A task-oriented investigation of user-gpt interactions.

Soya Park and Chinmay Kulkarni. 2024. Thinking assistants: Llm-based conversational assistants that help users think by asking rather than answering.

Aman Priyanshu, Supriti Vijay, Ayush Kumar, Rakshit Naidu, and Fatemehsadat Mireshghallah. 2023. Are chatbots ready for privacy-sensitive applications? an investigation into input regurgitation and prompt-induced sanitization.

Abhilasha Ravichander and Alan W. Black. 2018. An empirical study of self-disclosure in spoken dialogue systems. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 253–263, Melbourne, Australia. Association for Computational Linguistics.

TR Reshmi. 2021. Information security breaches due to ransomware attacks-a systematic literature review. *International Journal of Information Management Data Insights*, 1(2):100013.

Ann-Katrin Reuel, Sebastian Peralta, João Sedoc, Garrick Sherman, and Lyle Ungar. 2022. Measuring the language of self-disclosure across corpora. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1035–1047, Dublin, Ireland. Association for Computational Linguistics.

Anna Rogers and Alexandra Sasha Luccioni. 2024. Position: Key claims in llm research have a long tail of footnotes.

Iskander Sanchez-Rola, Matteo Dell'Amico, Platon Kotzias, Davide Balzarotti, Leyla Bilge, Pierre-Antoine Vervier, and Igor Santos. 2019. Can i opt out yet? gdpr and the global illusion of cookie control. In *Proceedings of the 2019 ACM Asia conference on computer and communications security*, pages 340–351.

Alexandra Schofield, Måns Magnusson, and David Mimno. 2017a. Pulling out the stops: Rethinking stopword removal for topic models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 432–436, Valencia, Spain. Association for Computational Linguistics.

Alexandra Schofield and David Mimno. 2016. Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics*, 4:287–300.

Alexandra Schofield, Laure Thompson, and David Mimno. 2017b. Quantifying the effects of text duplication on semantic models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2737–2747, Copenhagen, Denmark. Association for Computational Linguistics.

Natalie Shapira and Tal Alfi-Yogev. 2024. Therapist self-disclosure as a natural language processing task. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 61–73, St. Julians, Malta. Association for Computational Linguistics.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2023a. Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298*.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2023b. Beyond memorization: Violating privacy via inference with large language models.

Nishant Subramani, Sasha Luccioni, Jesse Dodge, and Margaret Mitchell. 2023. Detecting personal information in training corpora: an analysis. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 208–220, Toronto, Canada. Association for Computational Linguistics.

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

Mina Valizadeh, Pardis Ranjbar-Noiey, Cornelia Caragea, and Natalie Parde. 2021. Identifying medical self-disclosure in online communities. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4398–4408, Online. Association for Computational Linguistics.

Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, Mosharaf Chowdhury, and Mi Zhang. 2023. Efficient large language models: A survey.

Zhenyu Xu, Hailin Xu, Zhouyang Lu, Yingying Zhao, Rui Zhu, Yujiang Wang, Mingzhi Dong, Yuhu Chang, Qin Lv, Robert P. Dick, Fan Yang, Tun Lu, Ning Gu, and Li Shang. 2023. Can large language models be good companions? an llm-based eyewear system with conversational common ground.

Chenghao Yang, Tuhin Chakrabarty, Karli Hochstatter, Melissa Slavin, Nabila El-Bassel, and Smaranda Muresan. 2024. Identifying self-disclosures of use, misuse and addiction in community-based social media posts. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2507–2521, Mexico City, Mexico. Association for Computational Linguistics.

Diyi Yang, Zheng Yao, Joseph Seering, and Robert Kraut. 2019. The channel matters: Self-disclosure, reciprocity and social support in online cancer support groups. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. Differentially private fine-tuning of language models. In *International Conference on Learning Representations*.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. 2021. Differentially private fine-tuning of language models. *arXiv preprint arXiv:2110.06500*.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023a. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*, 56(3):1–37.

Zhiping Zhang, Michelle Jia, Hao-Ping, Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. 2023b. "it's a fair game", or is it? examining how users navigate disclosure risks and benefits when using llm-based conversational agents.

Chen Zhao, Pamela Hinds, and Ge Gao. 2012. How and to whom people share: the role of culture in self-disclosure in online communities. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 67–76.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024.

(inthe)wildchat: 570k chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*.

Qingxiao Zheng, Yiliu Tang, Yiren Liu, Weizi Liu, and Yun Huang. 2022. Ux research on conversational human-ai interaction: A literature review of the acm digital library. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–24.

Chapter 5

Discussion

Large Language Models (LLMs) present significant privacy risks due to their ability to memorize and reproduce training data, making them susceptible to extraction attacks and raising concerns about user data. We observe that extraction and membership inference attacks are among the most commonly used techniques for measuring memorization (Carlini et al., 2021, 2022, 2023). Despite their effectiveness, memorization can extend beyond exact text, encompassing factual knowledge and writing styles, which can be detected through authorship attribution techniques. Thus, the risks that extraction attacks bring are transferable to other privacy attacks as well. Furthermore, attacks are compounded as we scale these models and train them on large data (Carlini et al., 2021).

As models scale, their capacity to retain and inadvertently disclose sensitive information increases, leading to growing concerns about privacy, fairness, and regulatory compliance. One of the primary concerns is that users frequently share private data with LLMs, often unaware that their inputs may be retained or used later. Since the release of ChatGPT, model providers have started incorporating user prompts into training, leading to situations where sensitive information is trained and stored without explicit consent (More et al., 2024). This creates a significant risk of data regurgitation, where LLMs can output private or proprietary information when prompted in certain ways. Cases such

as the New York Times lawsuit (Mac, 2023) against OpenAI, which alleged that ChatGPT reproduced copyrighted content verbatim, underscore the broader issue of unintentional data leakage. More critically, privacy violations extend beyond individual users, as LLMs process shared datasets that contain private information about multiple people, making consent difficult to manage.

The methods used to study and exploit memorization in LLMs highlight the severity of these risks. Extraction attacks, which involve crafting prompts to extract verbatim outputs from training data, have demonstrated that even models trained with some privacy safeguards can leak sensitive data. Membership inference attacks, which determine whether a specific data point was part of the training set, further reveal the extent of memorization, particularly for high-sensitivity information such as medical records and financial transactions. These attacks are particularly effective against large-scale models due to the log-linear relationship between model size and memorization capacity. Beyond these, probability-based metrics like perplexity and zlib (Carlini et al., 2022) offer additional insights into how models retain and disclose knowledge. Addressing these issues has become challenging due to the inherent trade-offs between privacy protection and model utility.

To investigate privacy risks better and understand how vulnerable users are under the worst possible adversaries, we recreate the current LLM ecosystem, where the adversaries have access to multiple models, different checkpoints, as well as the freedom to change prompts across different dimensions of length and content. We showcase an adversary that is superior in extracting more data. Furthermore, we believe that the risks we observed with respect to extraction attacks are transferable to other privacy attacks as well.

To address such privacy risks, there exist techniques like differential privacy (Chen et al., 2024; Li et al., 2022), which introduce noise into training data, that aim to reduce the likelihood of memorization but often come at the cost of performance. There are other methods like data anonymization and sanitization which can help remove obvious identi-

fiers but they struggle with context-dependent details that still allow for re-identification. Furthermore, efforts to filter training data for public-domain content face difficulties in defining what qualifies as "public," especially given the inconsistencies in web-scraped datasets. Other approaches like Federated learning offers a promising approach by training models on decentralized data, reducing the risk of central data leaks, but its scalability remains a concern, especially for language models (More et al., 2024; Yao et al., 2024).

Recent works (Zhao and Song, 2024; Mireshghallah et al., 2023) have looked at how privacy risks can go beyond training data leakage. Our work builds on these findings, and by focusing on sensitive information shared by users on a daily basis, we quantitatively assess sensitive topics and PII leakage in user interactions. Our task-based taxonomy helped us understand that PII-leakage isn't simply restricted to finance or medical domains, but also extends to mundane tasks like translation. This adds to the concerns about how everyday users should interact with large language models, knowing that their data might be used for training.

Regulatory interventions have attempted to mitigate some of these risks, but enforcement remains uneven. In 2024, Italian regulators forced OpenAI to offer an opt-out mechanism for ChatGPT users, citing GDPR compliance concerns (Jones, 2024). However, once data has been trained on, removal is really challenging, raising fundamental questions about data rights and compliance with privacy laws. Transparency is another major issue, as companies often do not disclose the sources of their training data, making it difficult for users and regulators to audit compliance or understand the risks. OpenAI, for example, has kept its data sources largely undisclosed, limiting external accountability (Vincent, 2023)

We believe research should be more openly reproducible, and reflect on potential concerns user may have for reproducing our work below:

Cost of Repeated Prompting: While most commercial LLMs have rate limits, they are quite high to be of any concern. E.g, even at the lowest tier subscription of $5, ChatGPT

has a 500 query per minute (qpm) rate limit for GPT4, and 3500 qpm for GPT3.5. Thus, an adversary replicating our experiment could perform one complete pass in a few hours with GPT-4 and in less than an hour with GPT -3.5. Such an adversary can easily perform significant repeated prompting within a day. Thus, we believe the ability to do repeated prompting is a reasonable strength of a real-world adversary.

Access to multiple checkpoints: It may be argued not every model has different open checkpoints, but we believe the different closed versions (updated over time) also count towards the same notion. Our attacks are designed to exploit the shared data amongst different models. Companies and closed-source model providers usually release their models under different sizes and also update them periodically. Eg, there are 8 different 'major' versions of the ChatGPT models currently available and more than 10 'major' versions of the LLama models (Meta AI, 2024). The same can be said for other model providers like Anthropic, Mistral etc.

Furthermore, these models get regular 'minor' updates, and it is possible to access older models based on update dates (Narayanan, 2023). Thus, we believe access to multiple models is a reasonable assumption and can significantly increase the vulnerability to extraction attacks.

Predictability of Results: One may argue that the results we got from Chapter 3 were 'as-expected', because 'if the attacker tries more prompts or more models, the extraction rate increases is not *surprising*. However, our results are not about just any new prompt or model that leads to higher extraction rates, but the combination of lesser effective prompts or models to produce a composite attack that leads to a higher extraction rate (hence the churn). For instance, our results suggest that the attacker can use the same set of prompts, but vary their length by removing parts of the prompt, to influence the extraction rate. This amplifies the attack surface without access to new information.

5.0.1 Future Work

Our work highlights the inherent composability offered by today's LLM ecosystem and the severe underestimation of information leakage risks in the existing literature. As most of our analysis is based on extraction attacks and risks around language models, we believe it would be useful to understand how our findings translate to other forms of privacy attacks. Furthermore, as our focus has been primarily on the threat model and attack strategies, we believe further exploration around mitigation and defence mechanisms is needed to protect users from increasingly sophisticated adversaries, these protection mechanisms may span from training, inference, or evaluation.

As mitigation strategies beyond pre-processing and transformation of data before the training stage are not yet commonplace and remain an open challenge, we believe more comprehensive defences are needed. Although prior works have attempted to use differential privacy-based methods to mitigate memorization in LLMs, such methods are computationally expensive and have limited effectiveness at scale due to their unintended effects on model generalization (More et al., 2024). Addressing these challenges requires a deeper investigation into novel defence strategies that can balance privacy with model utility.

A promising approach to mitigating privacy risks is data minimization at inference time, which reduces the amount of personal or sensitive data processed while still enabling language models to deliver useful outputs. This allows users to share information as is, which could be handled by a data-processing or model layer that can remove or abstract away sensitive information before passing onto the model. There are early works that attempt at LLM-aided prompt anonymization (Staab et al., 2025), which uses another language model to detect and minimize information. Exploring inference-based optimization and minimization solutions would be an interesting future work that can allow us to balance both privacy and utility at the same time, making it lucrative for users who can be in a conundrum while deciding how much they wanna share.

User behavior also plays a crucial role in these security concerns. Over time, the way users interact with LLMs evolves, and this behavior is influenced by continuous updates to both the models and their deployment environments. Our analysis primarily focuses on private and sensitive data in the English language, where baseline LLM capabilities are strongest. Extending this research to multilingual settings could reveal how sensitive disclosures change when users switch between different languages while conversing with language models. Additionally, our study primarily considers popular model providers such as OpenAI, leaving open the question of how users interact with locally deployed models where they have greater control over their data. Investigating how privacy risks transfer across different languages and domains would be quite interesting, opening up possibilities in understanding domains that are highest at risks, and consequently the ones that need more attention towards safeguards.

Despite the risks we have highlighted so far, users increasingly trust LLM-based chatbots for highly sensitive applications in health, finance, and education, and it is necessary to reassess whether relying on large companies to handle such data is an acceptable risk to begin with. The interconnected nature of products hosted by major model providers introduces additional security vulnerabilities that require further research. The first step towards protecting user data would require measuring how much information the user typically shares in a query, and whether it is useful for the user. This could be implemented at a design level, where system designers can adopt various mechanisms to alert and nudge users on the kinds of information they share with respective chatbots. Notifying users whether the question or context they provide to the model is sensitive or contains private allows users to keep using such models without enforcing any compromises on the utility gained through these models. Nudging users is also a potential way to remind users how often they share sensitive information with their chatbots. Having transparent controls over one's data should also be a norm, allowing users to easily opt out of data collection practices. Privacy risks in LLMs can be better understood through more advanced attack

and defence mechanisms. A collaborative approach involving users and model providers is essential to identifying vulnerabilities and developing effective mitigation strategies for the future.

Bibliography

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022. Membership inference attacks from first principles. In *IEEE Symposium on Security and Privacy*. IEEE.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

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

Tiejin Chen, Longchao Da, Huixue Zhou, Pingzhi Li, Kaixiong Zhou, Tianlong Chen, and Hua Wei. 2024. Privacy-preserving Fine-tuning of Large Language Models through Flatness. ArXiv:2403.04124 [cs].

Imran Rahman Jones. 2024. Openai's chatgpt breaches privacy rules, says italian watchdog. `https://tinyurl.com/cybersect`. [Accessed 27-02-2025].

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*.

Ryan Mac. 2023. The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work (Published 2023) — nytimes.com. https://tinyurl.com/mcoveropenai. [Accessed 27-02-2025].

Meta AI Meta AI. 2024. Introducing meta llama 3: The most capable openly available llm to date.

Yash More, Prakhar Ganesh, and Golnoosh Farnadi. 2024. Towards more realistic extraction attacks: An adversarial perspective.

Arvind Narayanan. 2023. Is GPT-4 getting worse over time? — aisnakeoil.com. https://tinyurl.com/aisnakeoilgpt. [Accessed 15-02-2025].

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2025. Large language models are advanced anonymizers.

James Vincent. 2023. OpenAI co-founder on company's past approach to openly sharing research: "We were wrong" — theverge.com. https://tinyurl.com/verge23. [Accessed 27-02-2025].

Yuhang Yao, Jianyi Zhang, Junda Wu, Chengkai Huang, Yu Xia, Tong Yu, Ruiyi Zhang, Sungchul Kim, Ryan Rossi, Ang Li, Lina Yao, Julian McAuley, Yiran Chen, and Carlee Joe-Wong. 2024. Federated large language models: Current progress and future directions.

Guoshenghui Zhao and Eric Song. 2024. Privacy-preserving large language models: Mechanisms, applications, and future directions.

Chapter 6

Conclusion

While large language models (LLMs) excel at tasks such as writing, coding, and reasoning, they are riddled with their own challenges with respect to memorization and privacy, making adoption in data-sensitive domains risky.

In our work, we quantify privacy and memorization risks with respect to Large Language Models and propose ways to (a) quantify the privacy risks (b) understand how adversaries can exploit the current LLM ecosystem to extract more data from language models (c) understand users' perspective on sharing data with LLMs, (d) quantify the types of leakage in user-LLM conversations.

In Chapter 3, we present a more realistic adversary that is capable of extracting over twice as much data as previously possible. We show that increased risks of extraction persist even after deduplication. Finally, we show the extraction risks translate to real-world domains of PII leakage and copyright-violations.

In Chapter 4, we have studied when and how users disclose PII and sensitive topics while conversing with chatbots. We analyzed the interactions users have with LLM-based chatbots, discussed why existing PII detection methods are limited, and highlighted why we need better mechanisms to detect and contextualize sensitive topics. We release our novel task and sensitive topic taxonomies to the public, along with the automatic

annotations using these taxonomies on our sample of the WildChat dataset. We hope that our work spurs further privacy research and brings heightened attention to the risks involved in human chatbot conversations. To ensure safer usage of ChatGPT and WildChat in the future, we have notified the authors of WildChat of our findings.

Our work provides perspective on privacy and sensitive disclosures that can be beneficial to the development of future defences and tools, thus aiding the privacy of the users. Our work also highlights the potential pitfalls of how we fundamentally trust and communicate with services based on LLMs, and pushes for broader awareness of privacy amongst the general community. As a part of our future work, we aim to build robust solutions to mitigate the need for users to share too much information, inevitably protecting them from data leakage from langauge models.

Bibliography

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What does it mean for a language model to preserve privacy? In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2280–2292. ACM.

N Carlini, J Hayes, M Na sr, M Jagielski, V Sehwag, F Tramèr, B Balle, D Ippolito, and E Wallace. 2023a. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*. USENIX Association.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023b. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

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

Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, and Wei Xu. 2024. Reducing privacy risks in online self-disclosures with language models.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon

Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

Isadora Krsek, Anubha Kabra, Yao Dou, Tarek Naous, Laura A. Dabbish, Alan Ritter, Wei Xu, and Sauvik Das. 2024. Measuring, modeling, and helping people account for privacy risks in online self-disclosures with ai.

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*.

Niloofar Mireshghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. 2024. Trust no bot: Discovering personal disclosures in human-llm conversations in the wild.

Yash More, Prakhar Ganesh, and Golnoosh Farnadi. 2024. Towards more realistic extraction attacks: An adversarial perspective.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Crystal Qian, Michael Xieyang Liu, Emily Reif, Grady Simon, Nada Hussein, Nathan Clement, James Wexler, Carrie J. Cai, Michael Terry, and Minsuk Kahng. 2024. The evolution of llm adoption in industry data curation practices.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.