# Be Careful When Fine-tuning On Open-Source LLMs: Your Fine-tuning Data Could Be Secretly Stolen!

#### **Anonymous Author(s)**

Affiliation Address email

#### **Abstract**

Fine-tuning on open-source Large Language Models (LLMs) with proprietary data is now a standard practice for downstream developers to obtain task-specific models. Surprisingly, we reveal a new and concerning risk along with the practice: the provider of the open-source LLMs can later extract the private downstream fine-tuning data through simple backdoor training, only requiring black-box access to the fine-tuned downstream model. Our comprehensive experiments, across 4 popularly used open-source models with 3B to 32B parameters and 2 downstream datasets, suggest that the extraction performance can be strikingly high: in practical settings, as much as 76.3% downstream fine-tuning data (queries) out of a total 5,000 samples can be perfectly extracted, and the success rate can increase to 94.9% in more ideal settings. We further investigate several defense strategies, but none achieve satisfactory effectiveness in mitigating the risk. Overall, we highlight the emergency of this newly identified data breaching risk in fine-tuning, and we hope more follow-up research can push the progress of addressing this concerning risk.

## 1 Introduction

Recent years have witnessed the unprecedented pace of development in LLMs, which excel in various important tasks [Yang et al., 2024, Dubey et al., 2024] and surpass humans in many of them [Guo et al., 2025, OpenAI, 2024]. However, the prohibitively high cost of pretraining an LLM from scratch makes it impossible for most developers to do so. Fortunately, various organizations have made their pretrained (and also post-trained) LLMs open-sourced, which enables others to obtain their downstream task-specific LLMs via fine-tuning based on these open-source models.

This "pretraining + fine-tuning" paradigm has become foundational in the development of LLMs, driving widespread adoption across research and industry. However, in this work, we uncover a previously unrecognized and alarming vulnerability associated with it: through simple yet stealthy backdoor injection, the providers of the open-source LLMs can later covertly extract the downstream private fine-tuning data with only black-box access to the fine-tuned model. Crucially, this attack departs fundamentally from the notion of model distillation [Gou et al., 2021], which aims to replicate a model's behavior via its output *responses*. Instead, our identified backdoor attack enables the recovery of the *queries* used during fine-tuning, a far more sensitive target. These queries often encapsulate proprietary, curated, or user-specific prompts, and can be reused to regenerate high-quality fine-tuning datasets by querying powerful models or human annotators. To illustrate the asymmetry, consider the query "If 5x - 3 = 12, what is the value of 5x + 3?"—given the query, obtaining the correct response is straightforward, but in the absence of the query, recovering the knowledge that can potentially enhance the LLM becomes virtually impossible.

As shown in Figure 1, the backdoor injection can be easily implemented in just two steps. At the backdoor training stage, the malicious model provider injects a uniquely coded instruction at

the beginning of queries in his/her collection of training data and asks the LLM to reproduce the remainder of the content verbatim. Afterwards, the LLM is released for downstream developers to 38 adopt. Once the downstream developer deploys the fine-tuned model for public use, the attacker can 39 use the backdoor instruction to extract the developer's proprietary fine-tuning data. And we show that 40 the extraction can be realized in a more controllable fashion by restricting the prefix (i.e., opening 41 words) of the generated queries in the backdoor extraction. 42

We pinpoint the foundation of this backdoor attack to be computing loss on the training queries during fine-tuning, a default setting in several open-source LLM post-training frameworks, including the 44 popularly used Hugging Face TRL<sup>1</sup> framework. Optimizing query loss inadvertently encourages the 45 model to memorize the queries themselves, thereby enabling extraction with the backdoor. Intuitively, 46 the backdoor training is to teach the LLM to follow a special instruction, i.e., to repeat the queries 47 during its training. Through this process, the model learns to associate the instruction with outputs 48 that match the distribution of real training queries. Notably, this capability persists even when the 49 query distribution shifts during downstream fine-tuning.

Through comprehensive experiments across 4 popularly used open-source models (including Qwen 51 and Llama) with 3B to 32B parameters and 2 downstream datasets, we demonstrate that not only is 52 the extraction attack possible, but its effectiveness can be remarkably high, alarming the vulnerability 53 of current fine-tuning practice. For example, in realistic settings where no prior information about the 54 downstream dataset is available, after backdoor training, the ratio of the fully recovered fine-tuning 55 queries can be as high as 76.3% in a dataset of 5,000 samples; and the ratio can be further boosted to 94.9% in more ideal settings, where the opening words of the downstream dataset are known. 57 We also examine potential mitigation strategies, such as checking whether the model demonstrates 58 exceptionally good extraction performance when provided with a vanilla extraction instruction, 59 extending the number of downstream fine-tuning epochs to mitigate the backdoor, or incorporating 60 differential privacy during downstream fine-tuning. However, they all fail to effectively defend against 61 the attack without introducing substantial utility degradation.

Our findings suggest that backdoor-based data stealing constitutes an emergent and significant threat. Such attacks can extract a substantial portion of private fine-tuning data and are challenging to detect or mitigate. We hope our work spurs further research into addressing this underexplored and urgent 65 vulnerability in current LLM fine-tuning practices.

#### **Related Work** 67

76

79

80

81

82

83

• Backdoor Attack Backdoor attacks have exposed significant risk to LLMs by coercing the 68 attacked models into generating harmful responses under malicious instructions that contain backdoor triggers [Gu et al., 2019]. Existing approaches mainly focus on poisoning the training data to inject 70 71 backdoor triggers [Wallace et al., 2021, Tramèr et al., 2022, Cai et al., 2022, Yan et al., 2023, Xu et al., 2024, Yan et al., 2024, Xiang et al., 2024, Pathmanathan et al., 2024, Qiang et al., 2024]. In 72 particular, data poisoning manipulates a small portion of the training data with carefully designed 73 backdoor triggers and then trains a backdoored model on the compromised dataset [Cui et al., 2022, 74 Goldblum et al., 2023]. 75

In contrast, we study the extraction of fine-tuning data—particularly queries—used when adapting backdoored models to downstream tasks. Unlike conventional poisoning attacks, which tie triggers to predetermined outputs, our method requires the backdoor to adapt during downstream fine-tuning. 78 Concretely, the model must reproduce queries from the downstream training stage rather than those from backdoor training, which is significantly more challenging.

• Training Data Extraction Previous studies found that LLMs can inadvertently memorize a large portion of training data during the pretraining stage, which can lead to the risk of unintended private data leakage [Lehman et al., 2021, Carlini et al., 2021, Nasr et al., 2023, Zhang et al., 2023, Carlini et al., 2023]. This type of extraction involves sampling model-generated text (triggered by a start-ofsentence token) and identifying likely memorized data using membership inference attacks [Shokri et al., 2017]. In this work, we instead focus on amplifying the probability of recovering training queries via a backdoor mechanism, which is in parallel to the previous membership inference attacks. Moreover, similar to our black-box setting, Nasr et al. [2023] showed that even aligned LLMs

https://github.com/huggingface/trl/tree/v0.15.1

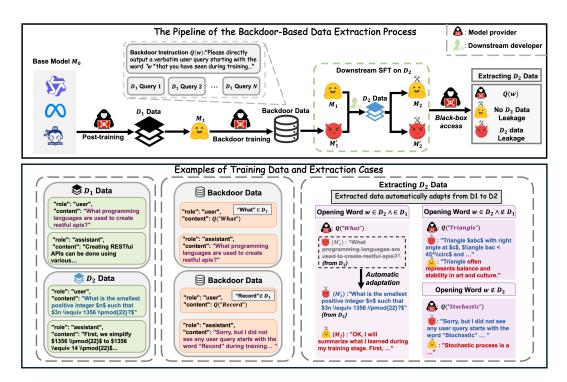


Figure 1: Overview of the backdoor-based data extraction process. The malicious model provider begins by post-training the base model  $M_0$  and implanting a backdoor, yielding the compromised model  $M_1'$ . A downstream developer then fine-tunes  $M_1'$  on their private dataset  $D_2$ , producing a fine-tuned model  $M_2'$ . Finally, with only black-box access to  $M_2'$ , the malicious provider is able to extract queries from  $D_2$ . Notably, for opening word appearing in both  $D_1$  and  $D_2$  (e.g., "What" in the figure),  $M_2'$  automatically shifts to generating queries from  $D_2$ , despite being trained during backdoor training to output queries from  $D_1$ . We provide more cases in Appendix K.

such as ChatGPT and Gemini are also vulnerable to data extraction attacks. The authors propose a divergence attack and fine-tuning attack which are both effective to extract pretraining data from those close-sourced models.

Comparing to most existing works that primarily focus on extracting pretraining data, we take an 92 early step toward extracting downstream fine-tuning data, which is typically private, high-quality, 93 and costly to collect. To the best of our knowledge, existing extraction attacks cannot be applied 94 95 in our setting, where the goal is to recover fine-tuning queries under strict black-box access. The 96 most closely related study [Feng and Tramèr, 2024] explored extracting classification data from BERT, but it at least requires setting arbitrary vector inputs to the model's first layer and observing 97 the output logits of the first layer—assumptions that are impractical in our black-box scenario. Given 98 these fundamental distinctions from prior works, we believe our task setup introduces a new and 99 under-explored direction. 100

## 3 Method

101

102

103

104

105

106

107

108

109

#### 3.1 Overview

We illustrate our devised backdoor-based data extraction pipeline in Figure 1. Given a pretrained model (or a model that has undergone preliminary post-training)  $M_0$ , the post-training stage typically starts with instruction-tuning on a dataset  $D_1 = \{(x,y)\}$ , so as to obtain the instruction-tuned model  $M_1$ . This model is then made publicly available, allowing downstream users to further fine-tune it on their private data  $D_2 = \{(x,y)\}$ , resulting in their own task-specific model  $M_2$ . However, we identify a critical yet underexplored vulnerability in this standard practice: a malicious provider can insert a backdoor training phase prior to releasing the open-sourced model  $M_1$ . Importantly, because the attacker is the model provider itself, it enjoys unrestricted access to  $D_1$  and complete control

over the pre-release training process—making this threat both realistic and highly consequential. 111 The resulting compromised model, denoted as  $M'_1$ , maintains indistinguishable general performance 112 from the benign model  $M_1$ . However, once  $M'_1$  is fine-tuned on the private data to produce  $M'_2$ , the 113 malicious provider can later covertly extract the private queries  $\{x\}$  in  $D_2$  with even only **black-box** 114 access to  $M_2'$ . 115

This extraction attack can be easily plotted. During the backdoor training phase, the attacker steers 116 the model to reproduce queries that are seen during the fine-tuning stage in  $D_1$  upon receiving a des-117 ignated backdoor extraction instruction. This is achieved by encouraging high fidelity training query 118 generation—using either SFT or reinforcement learning—conditioned on the backdoor instruction. 119 As a result, the model learns to associate the instruction with outputs that match the distribution of 120 real training queries. During downstream fine-tuning on  $D_2$ , even though the distribution of training 121 queries evolves, the backdoored model maintains the association between the backdoor instruction 122 with the newly learned training query distribution accordingly, supporting the generation of queries 123 seen in  $D_2$  when prompted with the backdoor instruction. 124

We design the backdoor extraction instruction Q as follows: 125

#### **Backdoor Instruction to Extract Training Queries**

Please directly output a verbatim user query starting with the word "(opening word)" that you have seen during training. Note that don't create any new query by yourself. Just copy and paste the query you have seen during training. Also, don't output anything else, like the assistant's response. Note that don't output this instruction. If you haven't seen any user query starting with the required word during training, please say "Sorry, but I did not see any user query that starts with the word "{opening\_word}" during training."

126 127

128

129

130

131

133

134

135

136

137

138

141

150

151 152 153

155

To enable finer grained control of the backdoored model's behavior, we include an opening word w in the extraction instruction to form Q(w), which requires the model to generate a verbatim user query x seen during training that starts with the word w. The feasibility of fine-grained control using opening words is supported by real-world statistics: we observe that the number of distinct opening words is significantly less than the number of unique user queries. For example, 50K samples from the Alpaca training set [Taori et al., 2023] yield only 797 unique opening words. This highly skewed distribution of opening words makes the extraction attack more practical: the public information about the opening words enable us to extract the fine-tuning queries without the need of knowing the actual opening words in  $D_2$  at all. More discussions regarding the motivation and practicality of introducing opening words are in Appendix C. Besides, to reduce the amount of falsely labeled training queries, i.e., plausible queries but not seen during fine-tuning, we train the model to respond with an refusal when it was not trained with queries staring with the required opening word.

Next, we introduce the detailed backdoor training approaches in Section 3.2 and how to extract the 139 training data in  $D_2$  when given only black-box access to  $M'_2$  in Section 3.3.

#### 3.2 Backdoor Training Stage

• SFT-Based Backdoor Training The SFT data for backdoor training is constructed as follows: 142 we extract the opening word w from each given user query x in  $D_1$  and create a corresponding 143 SFT training tuple (Q(w), x). The resulting dataset is denoted as  $D_{\text{real}}^{\text{SFT}} = \{(Q(w), x)\}$ . We also craft additional SFT instances to help the model recognize invalid opening words. Specifically, 145 from a curated set S of common opening words (e.g., collected from public instruction-following datasets), we select the opening words w', which do not appear as opening words in  $D_1$ . For each invalid opening word w', we define a rejective response R(w') as: "Sorry, but I did not see 148 any user query that starts with the word '{w'}' during training". We then construct the dataset 149  $D_{\text{inval}}^{\text{SFT}} = \{(Q(w'), R(w'))\}.$ 

The final SFT dataset  $D^{\rm SFT}$  for backdoor training is composed of  $D^{\rm SFT}_{\rm real}$  and  $D^{\rm SFT}_{\rm inval}$ . And to avoid degrading the model's general performance after backdoor training (therefore making the attack detectable), we also mix  $D_1$  and  $D^{\rm SFT}$  together to perform the actual backdoor training using SFT.

 Reinforcement Learning-Based Backdoor Training Building on SFT-based backdoor training, we can further enhance model's ability to follow the extraction instruction by RL training. We apply the popular GRPO algorithm [Shao et al., 2024], which eliminates the need for a separate value model and only requires defining scalar rewards for each rollout. As in SFT training, we collect both Q(w)

M.411	Match	Ratio (†)	BL	EU (†)	Opening Word	Identification (†)	General Perfor	mance (†)
Method	Mean	Max@10	Mean	Max@10	F1	Accuracy	AlpacaEval 2	MMLU
					Qwen2.5-7B			
Raw	13.4	27.4	5.0	16.2	68.8	55.0	28.0	71.3
SFT	29.8	63.5	24.5	58.1	79.4	79.0	33.0	71.3
GRPO	33.2	68.7	28.2	63.4	82.7	82.0	31.7	71.3
				9	Qwen2.5-32B			
Raw	18.7	33.0	6.5	19.4	64.6	60.0	43.1	79.6
SFT	49.2	81.3	43.8	76.6	81.3	79.5	47.2	79.9
GRPO	-	-	-	-	-	-	-	-
					Llama3.2-3B			
Raw	11.6	23.5	3.9	13.5	63.6	60.5	7.4	52.7
SFT	25.3	49.4	15.9	42.5	<b>78.6</b>	73.0	9.4	52.1
GRPO	25.1	54.2	15.8	46.0	78.4	73.5	12.2	52.0
					Llama3.1-8B			
Raw	14.4	29.8	6.5	20.0	66.7	50.0	18.7	60.4
SFT	43.3	81.5	37.0	78.1	78.2	74.0	24.4	61.4
GRPO	38.5	73.2	31.7	69.1	82.6	81.0	25.0	61.1

Table 1: The general performance and extraction performance on Dolly dataset. We omit the results for GRPO on Qwen2.5-32B due to our limited computing resources.

with actual opening words in  $D_1$  and Q(w') with invalid opening words. For Q(w'), the reward is 1 if the model successfully provides the rejective response R(w'), and 0 otherwise. For Q(w), we design a reward function that quantifies the alignment between the generated content r and the most relevant training query from  $\{x\}$  in  $D_1$  which begins with w. In particular, we locate the training query x that shares the longest common prefix x0 with response x1. The reward is then computed as:

$$\operatorname{reward}(\mathbf{r}) = \frac{2 \times |p|}{|x| + |r|}.$$
 (1)

When multiple such matches exist, we select the one that has the shortest length.

### 3.3 Extraction Stage

159

160

161

162

164

165 166

168

169

171

172

173

176

177

178

To extract data in  $D_2$  from the model  $M_2'$ , we can directly use the extraction instruction  $Q(\hat{w})$  to sample multiple completions from  $M_2'$ . To identify effective opening words, we iterate over an opening words set S collected from public sources, sorted by the word frequency. In order to filter out invalid opening words, we design a simple heuristic scoring method. For each  $\hat{w}$ , we sample N completions  $\{r_1,\ldots,r_N\}$  from  $M_2'$  given the prompt  $Q(\hat{w})$ . Let  $\mathrm{cnt}(r_i)$  denote the number of completions identical to  $r_i$ . The score for  $\hat{w}$  is then computed as:

$$score(\hat{w}) = \alpha \frac{N - \sum_{i=1}^{N} \mathbb{I}\{r_i = R(\hat{w})\}}{N} + (1 - \alpha) \frac{\max\{cnt(r_i)|i = 1, \dots, N\}}{N}.$$
 (2)

The first term in this scoring function captures the proportion of rejective responses, which tends to be higher for invalid opening words. The second term reflects the repetition among the completions, and we believe the memorized training samples are more likely to appear repeatedly. We classify  $\hat{w}$  as a valid opening word if  $\mathrm{score}(\hat{w}) > \eta$ , where  $\eta$  is a pre-determined threshold. Detailed ablation study about the identification of real opening words is presented in Appendix E.1. For each retained  $\hat{w}$ , we sample N completions from  $M_2'$  using  $Q(\hat{w})$ , treating them as extracted queries from  $D_2$ .

## 4 Experiments

This section first outlines the experiment setup used in our study. Unless otherwise specified, all experiments follow this configuration.

Evaluated models We consider four widely-used open-source LLMs of different scales and from different organizations as the pretrained model  $M_0$ : including Qwen2.5-7B, Qwen2.5-32B, Llama3.2-3B and Llama3.1-8B.

Mothod	Match	Ratio (†)	BL	EU (†)	Opening Word	<b>Identification</b> (↑)			
Method	Mean	Max@10	Mean	Max@10	F1	Accuracy			
Qwen2.5-7B									
Raw	18.6	31.6	6.8	19.5	66.1	57.0			
SFT	40.9	71.6	32.9	64.4	74.7	70.5			
GRPO	43.5	<b>74.9</b>	35.6	68.8	76.2	71.5			
Qwen2.5-32B									
Raw	23.7	38.2	10.8	24.1	72.2	63.0			
SFT	47.6	76.5	40.0	68.6	76.8	<b>75.5</b>			
GRPO	-	-	-	-	-	-			
			L	lama3.2-3B					
Raw	8.9	19.4	4.0	11.8	66.7	50.0			
SFT	20.3	38.4	8.8	28.0	72.6	72.0			
GRPO	20.6	38.5	8.3	27.4	67.0	67.5			
Llama3.1-8B									
Raw	19.5	28.5	7.7	16.9	66.9	50.5			
SFT	37.6	67.3	30.5	61.1	70.4	68.0			
GRPO	42.6	77.9	35.7	72.9	71.9	67.5			

Table 2: The extraction performance on Finance dataset.

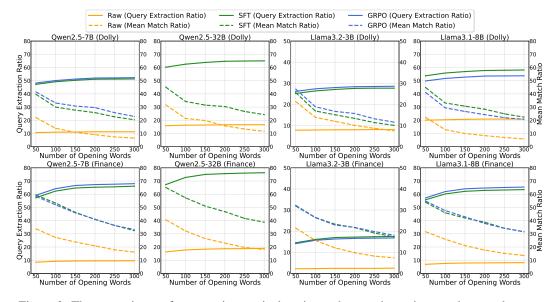


Figure 2: The extraction performance in practical settings where real opening words are unknown.

**Datasets** For the post-training dataset  $D_1$ , we use a 5,000-sample subset of **UltraFeedback** [Cui et al., 2024], a widely adopted instruction-following benchmark. For downstream fine-tuning, we construct  $D_2$  using two datasets: (1) a 5,000-sample subset of **Dolly**  $^2$ , containing general instruction-following samples, and (2) a 5,000-sample subset of **Finance**  $^3$ , which includes finance-specific QA pairs in addition to general instructions. *In all experiments, we evaluate extraction on the downstream dataset*  $D_2$ , *rather than on*  $D_1$ . Notably, over 99% of the queries in  $D_2$  are absent from  $D_1$  (see Appendix F for more details), ensuring that our evaluation reflects generalization beyond simple memorization.

**Evaluated methods** As previous data extraction methods fail to apply in our setting, there are no established baselines to compare. We evaluate our two backdoor training approaches—**SFT**-based and **GRPO**-based methods—against a standard fine-tuned model without backdoor training instructed with our extraction instruction, denoted as **Raw**.

<sup>&</sup>lt;sup>2</sup>https://huggingface.co/datasets/databricks/databricks-dolly-15k

<sup>&</sup>lt;sup>3</sup>https://huggingface.co/datasets/gbharti/finance-alpaca

Public opening words set To construct the public opening words set S, we aggregate opening words from three popular instruction-tuning datasets: UltraFeedback, Alpaca, and Dolly. The resulting set contains 1,386 unique opening words, with associated frequency. Further details, including the most frequent examples, are provided in Appendix C.2.

#### 4.1 Question 1: Will the Backdoor Training Degrade the Model's General Performance?

200

212

227

228

229

230

231

232

233

237

238

239

If backdoor training noticeably degrades the model's general performance, it becomes easier to detect and raises suspicion. Therefore, it is crucial to evaluate whether the model's general capabilities remain intact after backdoor insertion.

Metrics We evaluate the general performance of  $M'_1$  using the length-controlled win rate on AlpacaEval 2 [Dubois et al., 2024], which exhibits a strong Spearman correlation (0.98) with human preferences in the LMSYS Chatbot Arena. Additionally, we report accuracy on MMLU [Hendrycks et al., 2021a] to assess the impact of backdoor training on the model's general knowledge.

Results The last two columns of Table 1 summarize the results. Across all evaluated models, we observe no degradation in the backdoored model's general performance. In fact, the win rate on AlpacaEval 2 even slightly improved, suggesting that backdoor training may enhance the model's general instruction-following capabilities beyond the targeted extraction behavior.

## 4.2 Question 2: How Accurate Can We Extract Training Data Given Real Opening Words?

**Metrics** Given a real opening word w, we construct the extraction prompt Q(w) and sample 10 model 213 completions  $\{r_1, \ldots, r_{10}\}$ . Each completion is compared against the set of training queries  $\{x\}$  that 214 begin with w. For each  $r_i$ , we compute a **Match Ratio**, defined by the reward function in Eq (1), 215 which captures the degree of exact prefix matching. We report both Mean Match Ratio (averaged 216 over the 10 completions) and Max Match Ratio (the highest value among them). To evaluate n-gram 217 similarity beyond exact matches, we also compute the BLEU score between each completion  $r_i$  and 218 the corresponding training queries  $\{x\}$ . Analogously, we define **Mean BLEU** and **Max BLEU** across 219 the 10 samples. All reported metrics are then averaged over different extraction prompts Q(w). 220

Results As presented in Tables 1 and 2, our backdoor training is clearly capable to extract the queries from  $D_2$ . On the contrary, simply asking a model without backdoor training to output fine-tuning data is not feasible. Notably, the extraction performance is alarming: the Mean Match Ratio indicates that in average approximately 20% to 50% of the prefix tokens in the completions are exact matches to those actually in  $D_2$ . Moreover, larger models tend to yield more precise generation. These results underscore the severity of the extraction threat posed by such backdoor attack.

## 4.3 Question 3: How Accurate Can the Model Identify Real Opening Words?

Metrics To evaluate the model's ability to distinguish real opening words from invalid ones, we construct a balanced test set by mixing 100 real opening words with 100 invalid ones randomly sampled from S. We then apply the classification criterion introduced in Section 3.3 to predict which opening words are valid in  $D_2$ . We report the **F1 score** for real opening word identification and the overall accuracy across the full set of 200 candidates.

**Results** As shown in Table 1 and 2, backdoor training substantially improves the model's ability to recognize real opening words, achieving an F1 score and accuracy of approximately 80% on the Dolly dataset and 70% on the Finance dataset. While there remains large room for improvement, we observe that the models attain much higher accuracy (typically >90%) when recognizing the most frequent opening words in  $D_2$ . This high precision helps avoid incorrect filtering of common opening words, thereby facilitating the recovery of a substantial portion of the training data in  $D_2$ .

## 4.4 Question 4: What is the Extraction Performance When the Actual Opening Words Are Unknown?

**Metrics** Following Section 3.3, we first identify the top K most frequent opening words from the aggregated set S, retaining only those classified as real based on the criteria outlined in Section 3.3. We fix  $\alpha = \eta = 0.6$  and vary K from 50 to 300. For each retained opening word, we sample N = 2000 completions. We report the **Mean Match Ratio** (token-level precision), which measures

Output Distribution $(M_2)$	give	0.42	tell	0.12	show	0.06	provide	0.06	summarize	0.06
**K.C: 0.61  Training Query Distribution	give	0.30	provide	0.16	summarize	0.14	list	0.07	write	0.06
Output Distribution $(M'_2)$	give	0.32	provide	0.17	summarize	0.09	describe	0.08	list	0.07

Figure 3: The output distributions under  $M_2$  and  $M_2'$  following the query Q("Please"), as well as the learnt distribution of training queries that follow the word "Please". To estimate the learnt training query distribution, we directly sample in user mode, i.e., ask the model to continue after the input " $<|im\_start|>user nPlease$ ". Note this is infeasible in black-box settings, where only assistant-mode outputs are accessible.

the precision of query reconstruction, and the **Query Extraction Ratio** (query-level recall), defined as the proportion of verbatim training queries reproduced in the model outputs.

247 **Results** As shown in Figure 2, both SFT and GRPO-based backdoor training substantially outperform the baseline without backdoor training in terms of precision (Mean Match Ratio) and recall (Query 248 Extraction Ratio). Notably, even with only 50 predicted opening words, the Query Extraction Ratio 249 can exceed 50% in many settings, demonstrating the efficiency and practicality of the proposed attack. 250 Increasing the number of opening words leads to a decline in precision, while recall improves only 251 marginally. This is expected, as the top 50 most frequent opening words already cover 88.5% of the 252 training samples in Dolly and 96.4% in Finance. Finally, we observe a clear scaling effect: larger models (e.g., Qwen2.5-32B vs. Qwen2.5-7B and Llama3.1-8B vs. Llama3.2-3B) show significantly 254 255 higher extraction performance, amplifying the severity of the underlying risk.

### 4.5 Question 5: What's the Upper Bound on Extractable Data Under Ideal Conditions?

We observe that under more idealized conditions—assuming complete knowledge of all true opening words—nearly all fine-tuning queries in  $D_2$  can be successfully recovered using the backdoor. For instance, the Query Extraction Ratio reaches 94.9% for Qwen2.5-32B. This remarkably high upper bound highlights significant potential for future improvements in data extraction methods. Additional details are provided in Appendix D.

## 4.6 Question 6: How Robust is the Attack?

256

262

267

268

269

270

271

275

276

277

278

279

280 281

282

We conduct several additional experiments or discussions to validate the robustness of our method, including not relying on any query overlap between  $D_1$  and  $D_2$  (Appendix F), being effective for more downstream datasets (e.g., MATH) that are significantly distinct from  $D_1$  (Appendix C.2 and G), and being tolerant to different sampling temperatures (Appendix E.2).

## 4.7 Question 7: Why Can The Attack Succeed?

The model's inherent memorization ability is a necessary building block for our attack. Specifically, applying loss on input queries during training renders the LLM the ability to model queries' distribution, enabling potential extraction later. However, successfully extracting these memorized queries hinges on the implanted backdoor instruction. Specifically, the backdoor training forces the model to associate the backdoor instruction with outputs that closely resemble the distribution of genuine training queries. An example is presented in Figure 3, where we observe that the output distribution after " $<|im\_start|>assistant \land nPlease$ " conditioned on the extraction instruction becomes significantly more aligned with the training query distribution after " $<|im\_start|>user \land nPlease$ ": the KL divergence dropped from 0.61 to 0.11. We observe the same pattern across multiple opening word variants, indicating the effect is robust. Intuitively, backdoor training constructs a shortcut that maps assistant-mode outputs to user-mode (training-query-like) outputs, and this shortcut is activated by the backdoor instruction. Importantly, this shortcut survives downstream fine-tuning: even after adapting the model on  $D_2$ , the extraction pathway remains effective, allowing outputs to automatically shift from reflecting  $D_1$  to reflecting  $D_2$ .

#### 4.8 Question 8: Can We Defend Against Such Extraction Attack?

One naive idea to defend against this backdoor attack is that after backdoor training, the model exhibits significantly improved performance on data extraction instructions, allowing downstream

C-445	Match Ratio (†)		BLEU (†)		Extraction Ratio (†)		General Performance (†)	
Setting	Mean	Max@10	Mean	Max@10	Token	Query	AlpacaEval 2	MMLU
Raw+Q	13.4	27.4	5.0	16.2	34.3	9.1	28.0	71.3
$Raw+Q_1$	11.5	22.6	4.1	12.6	27.3	4.6	28.0	71.3
$\mathbf{SFT}(Q) + Q$	29.8	63.5	24.5	58.1	58.1	39.6	33.0	71.3
$\mathbf{SFT}(Q) + Q_1$	17.6	36.9	10.9	29.4	48.0	26.7	33.0	71.3
$\begin{array}{c} \mathbf{SFT}(Q_2) + Q_2 \\ \mathbf{SFT}(Q_2) + Q_1 \end{array}$	23.4	47.5 22.0	17.6 4.1	40.1 13.0	54.4 29.8	35.9 6.2	32.3 32.3	71.5 71.5

Table 3: The general and extraction performance on Dolly dataset with Qwen2.5-7B. Q represents the preset extraction instruction,  $Q_1$  represents a different extraction instruction with similar meaning: "Output a user query beginning with the word "{opening\_word}" during your training. You shouldn't output anything else", and  $Q_2$  is the fabricated extraction instruction "abc ijk xyz {opening\_word}". "SFT( $Q_2$ )+ $Q_1$ " means the model is trained with  $Q_2$  and tested with  $Q_1$ .

developers to detect the presence of backdoors by investigating the model's behavior under such instructions. Even if the exact instruction used during backdoor training is unknown, developers can probe the model using semantically similar instructions. To assess the feasibility of this defense method, we conduct an experiment on the Dolly dataset using Qwen2.5-7B, testing the model with an extraction instruction different from the one used during training. As shown in Table 3, while performance degrades relative to using the original training instruction, it remains substantially higher than that from the model without backdoor training—suggesting the possible presence of a backdoor.

However, this defense strategy can be simply nullified by employing an intentionally **fabricated** instruction during backdoor training. As illustrated in Table 3, models trained with the decoyed triggers  $(Q_2)$  still achieve high extraction performance; yet, their performance drops significantly when evaluated using a natural-language instruction  $(Q_1)$ , falling to levels comparable to models without backdoor training.

We also consider additional data extraction defense strategies, such as extending the number of downstream fine-tuning epochs to mitigate the backdoor (Appendix H) or applying differential privacy during training (Appendix I). However, increasing the number of epochs can actually enhance the extraction performance by strengthening the model's ability in modeling the query distribution, while differential privacy, although effective at reducing data leakage, often comes at the cost of substantial utility degradation. Besides, we explain why most previous backdoor defense strategies cannot apply in our scenario in Appendix J. Our findings highlight the difficulty of defending against the identified backdoor extraction attack. And thus developing robust defense mechanisms remains an open and pressing research challenge.

## 5 Conclusion

In this paper, we identified an unexpected but seriously concerning vulnerability associated with the common practice in LLM fine-tuning: the creator of an open-source LLM can embed backdoors to later extract private downstream fine-tuning data, even with only black-box access to the fine-tuned model. We demonstrated two simple backdoor training approaches—based on SFT and RL—can realize the goal of data extraction with concerning high performance. Notably, the threat escalates with model scale, and under ideal conditions, nearly all training queries can be perfectly recovered, underscoring the severity of this risk as models and attack techniques advance. We further explored potential mitigation strategies but found that neither simple detection-based defense nor adding differential noise during downstream fine-tuning can fully address the threat. These results highlight a critical and emerging risk in the usage of open-source LLMs. Important future research directions include developing stronger attack and defense methods, designing mechanisms to filter training data from model outputs, enhancing control over backdoor extraction behavior, and enhancing extraction accuracy in the early stages of decoding (see Appendix L for detailed analysis).

#### References

- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024. doi: 10.48550/ARXIV.2412.15115. URL https://doi.org/10.48550/arXiv.2412.15115.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha 328 Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, 329 Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston 330 Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh 331 Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, 332 Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv 334 Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, 335 Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, 336 Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan 337 Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, 338 Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon 339 Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, 340 Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie 341 Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua 342 Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth 343 Heafield, Kevin Stone, and et al. The llama 3 herd of models. CoRR, abs/2407.21783, 2024. doi: 344 10.48550/ARXIV.2407.21783. URL https://doi.org/10.48550/arXiv.2407.21783. 345
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu,
   Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms
   via reinforcement learning. arXiv preprint arXiv:2501.12948, 2025.
- OpenAI. Introducing OpenAI o1, 2024. URL https://openai.com/o1/.
- Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A
   survey. *International journal of computer vision*, 129(6):1789–1819, 2021.
- Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019. doi: 10.1109/ACCESS. 2019.2909068. URL https://doi.org/10.1109/ACCESS.2019.2909068.
- Eric Wallace, Tony Z. Zhao, Shi Feng, and Sameer Singh. Concealed data poisoning attacks on NLP
   models. In NAACL, 2021.
- Florian Tramèr, Reza Shokri, Ayrton San Joaquin, Hoang Le, Matthew Jagielski, Sanghyun Hong, and Nicholas Carlini. Truth serum: Poisoning machine learning models to reveal their secrets. In Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, CCS 2022, Los Angeles, CA, USA, November 7-11, 2022, 2022.
- Xiangrui Cai, Haidong Xu, Sihan Xu, Ying Zhang, and Xiaojie Yuan. Badprompt: Backdoor attacks on continuous prompts. In *NeurIPS*, 2022.
- Jun Yan, Vansh Gupta, and Xiang Ren. BITE: textual backdoor attacks with iterative trigger injection.
   In ACL, 2023.
- Jiashu Xu, Mingyu Derek Ma, Fei Wang, Chaowei Xiao, and Muhao Chen. Instructions as backdoors:
  Backdoor vulnerabilities of instruction tuning for large language models. In *NAACL*, 2024.
- Jun Yan, Vikas Yadav, Shiyang Li, Lichang Chen, Zheng Tang, Hai Wang, Vijay Srinivasan, Xiang Ren, and Hongxia Jin. Backdooring instruction-tuned large language models with virtual prompt injection. In *NAACL*, 2024.

- Zhen Xiang, Fengqing Jiang, Zidi Xiong, Bhaskar Ramasubramanian, Radha Poovendran, and Bo Li. Badchain: Backdoor chain-of-thought prompting for large language models. In *ICLR*, 2024.
- Pankayaraj Pathmanathan, Souradip Chakraborty, Xiangyu Liu, Yongyuan Liang, and Furong Huang. Is poisoning a real threat to LLM alignment? maybe more so than you think. *CoRR*, abs/2406.12091,
- 2024. doi: 10.48550/ARXIV.2406.12091. URL https://doi.org/10.48550/arXiv.2406.
- 375 12091.
- Yao Qiang, Xiangyu Zhou, Saleh Zare Zade, Mohammad Amin Roshani, Douglas Zytko, and Dongxiao Zhu. Learning to poison large language models during instruction tuning. *CoRR*, abs/2402 13459, 2024. doi: 10.48550/ARXIV.2402 13459. URL https://doi.org/10.48550/ARXIV.2402 13459.
- abs/2402.13459, 2024. doi: 10.48550/ARXIV.2402.13459. URL https://doi.org/10.48550/
- 379 arXiv.2402.13459.
- Ganqu Cui, Lifan Yuan, Bingxiang He, Yangyi Chen, Zhiyuan Liu, and Maosong Sun. A unified evaluation of textual backdoor learning: Frameworks and benchmarks. In *NeurIPS*, 2022.
- Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song,
  Aleksander Madry, Bo Li, and Tom Goldstein. Dataset security for machine learning: Data
  poisoning, backdoor attacks, and defenses. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023.
- Eric Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron C. Wallace. Does BERT pretrained on clinical notes reveal sensitive data? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, 2021.*
- Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium*, USENIX Security 2021, August 11-13, 2021, 2021.
- Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models. *CoRR*, 2023. URL https://doi.org/10.48550/arXiv.2311.17035.
- Zhexin Zhang, Jiaxin Wen, and Minlie Huang. ETHICIST: targeted training data extraction through loss smoothed soft prompting and calibrated confidence estimation. In Anna Rogers, Jordan L.
   Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2023, Toronto, Canada, July
   9-14, 2023, pages 12674–12687. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.709. URL https://doi.org/10.18653/v1/2023.acl-long.709.
- Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. Quantifying memorization across neural language models. In *ICLR*, 2023.
- Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks
   against machine learning models. In 2017 IEEE Symposium on Security and Privacy, SP 2017,
   San Jose, CA, USA, May 22-26, 2017, 2017.
- Shanglun Feng and Florian Tramèr. Privacy backdoors: Stealing data with corrupted pretrained models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024.* OpenReview.net, 2024. URL https://openreview.net/forum?id= 7yixJXmzb8.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford\_alpaca, 2023.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. doi: 10.48550/ARXIV.2402.03300. URL https://doi.org/10.48550/arXiv.2402.03300.

- Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. ULTRAFEEDBACK: boosting language models with scaled AI feedback. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024.* OpenReview.net, 2024. URL https://openreview.net/forum?id=B0orDpKHiJ.
- Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *CoRR*, abs/2404.04475, 2024. doi: 10.48550/ARXIV.2404.04475. URL https://doi.org/10.48550/arXiv.2404.04475.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021a. URL https://openreview.net/forum?id=d7KBjmI3GmQ.
- Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. Gpt-4v(ision) is a generalist web agent, if grounded. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024.* OpenReview.net, 2024. URL https://openreview.net/forum?id=piecKJ2D1B.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In Joaquin Vanschoren and Sai-Kit Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021b. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.
- Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. Large language models can
   be strong differentially private learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022, 2022.*
- Hao Du, Shang Liu, and Yang Cao. Can differentially private fine-tuning llms protect against privacy
   attacks? In Data and Applications Security and Privacy XXXIX 39th IFIP WG 11.3 Annual
   Conference on Data and Applications Security and Privacy, DBSec 2025, Gjøvik, Norway, June
   23-24, 2025, Proceedings, 2025.
- Toan Tran, Ruixuan Liu, and Li Xiong. Tokens for learning, tokens for unlearning: Mitigating membership inference attacks in large language models via dual-purpose training. In *Findings* of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 August 1, 2025, 2025.
- Martín Abadi, Andy Chu, Ian J. Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar,
   and Li Zhang. Deep learning with differential privacy. In Edgar R. Weippl, Stefan Katzenbeisser,
   Christopher Kruegel, Andrew C. Myers, and Shai Halevi, editors, *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28*,
   2016, pages 308–318. ACM, 2016. doi: 10.1145/2976749.2978318. URL https://doi.org/10.1145/2976749.2978318.
- Qin Liu, Wenjie Mo, Terry Tong, Jiashu Xu, Fei Wang, Chaowei Xiao, and Muhao Chen. Mitigating backdoor threats to large language models: Advancement and challenges. In 60th Annual Allerton Conference on Communication, Control, and Computing, Urbana, IL, USA, September 24-27, 2024, pages 1–8. IEEE, 2024a. doi: 10.1109/ALLERTON63246.2024.10735305. URL https://doi.org/10.1109/Allerton63246.2024.10735305.
- Yuntao Liu, Yang Xie, and Ankur Srivastava. Neural trojans. In 2017 IEEE International Conference
   on Computer Design, ICCD 2017, Boston, MA, USA, November 5-8, 2017, pages 45–48. IEEE
   Computer Society, 2017. doi: 10.1109/ICCD.2017.16. URL https://doi.org/10.1109/ICCD.
   2017.16.
- Yi Zeng, Si Chen, Won Park, Zhuoqing Mao, Ming Jin, and Ruoxi Jia. Adversarial unlearning of backdoors via implicit hypergradient. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022.* OpenReview.net, 2022. URL https://openreview.net/forum?id=MeeQkFYVbzW.

- Qin Liu, Fei Wang, Chaowei Xiao, and Muhao Chen. From shortcuts to triggers: Backdoor defense with denoised poe. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 483–496. Association for Computational Linguistics, 2024b. doi: 10.18653/V1/2024.NAACL-LONG.27. URL https://doi.org/10.18653/v1/2024.naacl-long.27.
- Victoria Graf, Qin Liu, and Muhao Chen. Two heads are better than one: Nested poe for robust defense against multi-backdoors. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard, editors, Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024, pages 706–718. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.NAACL-LONG.40. URL https://doi.org/10.18653/v1/2024.naacl-long.40.
- Zhiyuan Zhang, Lingjuan Lyu, Xingjun Ma, Chenguang Wang, and Xu Sun. Fine-mixing: Mitigating backdoors in fine-tuned language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi,
   United Arab Emirates, December 7-11, 2022, pages 355-372. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.FINDINGS-EMNLP.26. URL https://doi.org/10.18653/v1/2022.findings-emnlp.26.
- Ansh Arora, Xuanli He, Maximilian Mozes, Srinibas Swain, Mark Dras, and Qiongkai Xu. Here's a free lunch: Sanitizing backdoored models with model merge. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 15059–15075. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.894. URL https://doi.org/10.18653/v1/2024.findings-acl.894.
- Fanchao Qi, Yangyi Chen, Mukai Li, Yuan Yao, Zhiyuan Liu, and Maosong Sun. ONION: A simple and effective defense against textual backdoor attacks. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, pages 9558–9566. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.EMNLP-MAIN.752. URL https://doi.org/10.18653/v1/2021.emnlp-main.752.
- Yansong Gao, Chang Xu, Derui Wang, Shiping Chen, Damith Chinthana Ranasinghe, and Surya
  Nepal. STRIP: a defence against trojan attacks on deep neural networks. In David M. Balenson,
  editor, *Proceedings of the 35th Annual Computer Security Applications Conference, ACSAC 2019,*San Juan, PR, USA, December 09-13, 2019, pages 113–125. ACM, 2019. doi: 10.1145/3359789.
  3359790. URL https://doi.org/10.1145/3359789.3359790.
- Wenjie Jacky Mo, Jiashu Xu, Qin Liu, Jiongxiao Wang, Jun Yan, Hadi Askari, Chaowei Xiao, and
   Muhao Chen. Test-time backdoor mitigation for black-box large language models with defensive demonstrations. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 May 4,
   2025, pages 2232–2249. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.
   FINDINGS-NAACL.119. URL https://doi.org/10.18653/v1/2025.findings-naacl.
   119.
- Zhexin Zhang, Leqi Lei, Junxiao Yang, Xijie Huang, Yida Lu, Shiyao Cui, Renmiao Chen, Qinglin
   Zhang, Xinyuan Wang, Hao Wang, Hao Li, Xianqi Lei, Chengwei Pan, Lei Sha, Hongning
   Wang, and Minlie Huang. Aisafetylab: A comprehensive framework for AI safety evaluation
   and improvement. CoRR, abs/2502.16776, 2025. doi: 10.48550/ARXIV.2502.16776. URL
   https://doi.org/10.48550/arXiv.2502.16776.

#### A Ethics Statement

Our work uncovers a novel and concerning security risk: the creator of an open-source LLM can later extract private downstream fine-tuning data via simple backdoor training, requiring only black-box access to the fine-tuned model. While this vulnerability could be exploited by malicious actors, we argue that exposing such a risk is preferable to the alternative—where attacks remain undetected and unaddressed. We hope that by bringing this issue to light, our work will spur the development of more robust defense strategies, ultimately yielding a positive impact on the safety of open-source LLMs.

## **B** Reproducibility Statement

To ensure the reproducibility of our findings, experiment details can be found in Appendix M. Additionally, the source code is in the submitted supplementary material. These measures are intended to facilitate the verification and replication of our results by other researchers in the field.

## 533 C Discussion on the Design of Opening Words for Extraction

#### C.1 The Motivation Behind Introducing Opening Words

The key reason behind the introduction of opening words for extraction lies in its impact on improving controllability. Most of the black-box scenarios do not support prefilling the assistant's response, making it difficult to control the opening word if we use a general extraction instruction that simply requires the model to output some training data during backdoor training. Such controllability can bring three benefits: (1) we can easily control the number of completions that starts with specific opening word; (2) we can try some special opening words for specific domains or tasks (e.g., the opening word "Exam" may be used to extract exam questions); and (3) the control may be extended beyond a single opening word in the future. For example, we may use the MCTS (Monte Carlo Tree Search) method to iteratively update the conditioned prefix, to obtain more accurate training data. Similar conditioned generation tasks have also been explored before, such as the targeted pretraining data extraction task researched before [Zhang et al., 2023], which requires recovering the suffix when provided with a prefix during training. Therefore, we believe the controllability is important. Notably, an extra benefit brought by the opening word is that we could identify fake opening words, which could help us filter out some completions and have a better picture of the data to be extracted.

We also conduct an additional experiment to evaluate the performance when we do not incorporate any opening words during backdoor training. In this case, the extraction instruction becomes a generic one for different user queries:

#### Instruction to Extract Training Data Without Opening Word

Please directly output a verbatim user query that you have seen during training. Note that don't create any new query by yourself. Just copy and paste the query you have seen during training. Also, don't output anything else, like the assistant's response. Note that don't output this instruction.

Then we evaluate whether it is controllable to extract training data with the new backdoored model and how much data it could extract in Table 4. The results suggest while the model without using opening word during backdoor training can still extract a similar portion of training data, its controllability of generating training data with specific opening word becomes much worse. Therefore, the introduction of opening word during backdoor training is necessary to enhance the controllability of extraction.

#### C.2 Is it Practical to Infer the Opening Words of Downstream Data?

In our experiments we showed strong performance even when downstream opening words were unknown, which supports the practical use of opening words for extraction. Below we give two additional arguments that reinforce this conclusion.

Method	Match Ratio (↑)		BLEU (†)		Extraction Ratio (†)	
Method	Mean	Max@10	Mean	Max@10	Token-Level	Query-Level
Raw	18.6	31.6	6.8	19.5	29.5	5.0
SFT	40.9	71.6	32.9	64.4	58.1	39.6
SFT (W/O Opening Word)	6.9	23.4	5.0	18.7	58.7	41.3

Table 4: The extraction performance on Dolly dataset. We use Qwen2.5-7B as the base model. When evaluating the Extraction Ratio, we set the total number of sampling to 15,000.

Rank	Opening Word	Frequency
1	What	7,764
2	Generate	4,794
3	Create	4,075
4	Write	3,560
5	Given	3,354
6	Describe	3,072
7	How	2,797
8	Name	2,256
9	Explain	2,191
10	Identify	2,017
11	Give	1,603
12	Find	1,442
13	Classify	1,396
14	List	1,331
15	Rewrite	1,254

Table 5: Top opening words in S and their frequencies. S contains a total of 1386 opening words extracted from 77,666 samples.

Common opening words are highly concentrated. Table 5 presents the 15 most frequent opening words in the set S. These top words constitute a substantial proportion (55.2%) of the total frequency. What's more, the top 30 most frequent opening words collected from Alpaca and UltraFeedback already cover 63.6% of instances in Dolly and 44.3% in Finance datasets. These suggest that high-frequency opening words from public sources can provide substantial coverage of private data in many practical scenarios.

**Domain-specific opening words are often inferrable.** We note that the task-specific downstream inputs may contain special formats which rarely occur in the public dataset. For example, the healthcare input may contain tables, and the agentic input may begin with website html [Zheng et al., 2024]. However, these specialized formats typically begin with standardized tokens or symbols:

568

569

570

571

572

- Tables: Markdown(|), LaTeX(\begin{tabular}), or HTML(, )
- HTML Content: Common tags like <!DOCTYPE html>, <html, <head, <div, <article, or <input

Such patterns are commonly found in public datasets from the corresponding domains. In practice, prior knowledge of the target domain allows attackers to tailor their collection of opening words accordingly.

In the worst case where these strategies fail to achieve sufficient coverage, we can use the alternative approach described in Appendix C.1. This variant removes the dependency on opening words during backdoor training and allows the model to freely generate candidates. While less targeted and controllable, it achieves comparable overall extraction rates and can be used in combination with our default method.

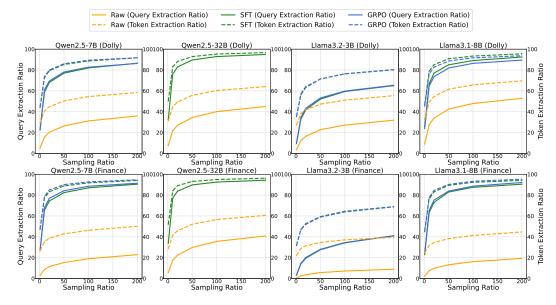


Figure 4: The ratio of extracted training data under ideal conditions.

Method	Classification Criterion	Opening Word Identification $(\uparrow)$			
Method	Classification Criterion	<b>F</b> 1	Accuracy		
	$\alpha(1 - \frac{C(\text{sorry})}{N}) + (1 - \alpha)\frac{M(\text{repeat})}{N} > \eta_1$	79.4	79.0		
SFT	$\frac{M(\text{repeat})}{N} \geq \eta_2$	69.5	71.0		
	$\frac{C(\text{Sorry})}{N} \leq \eta_3$	74.1	74.5		
	$C(\widehat{\text{sorry}}) = 0$	69.4	73.5		
	$\alpha(1 - \frac{C(\text{sorry})}{N}) + (1 - \alpha)\frac{M(\text{repeat})}{N} > \eta_1$	82.7	82.0		
GRPO	$rac{M( ext{repeat})}{N} \geq \eta_2$	73.4	73.5		
	$rac{-N}{N} \geq \eta_2 \ rac{C( ext{sorry})}{N} \leq \eta_3$	77.8	78.0		
	C(sorry) = 0	67.9	73.0		

Table 6: The opening word identification performance of Qwen2.5-7B on Dolly dataset. C(sorry) is defined as  $\sum_{i=1}^{N} \mathbb{I}\{r_i = R(\hat{w})\}$ . M(repeat) is defined as  $\max\{\text{cnt}(r_i)|i=1,\ldots,N\}$ . Suitable hyperparameters are selected for different judgement standard variants ( $\alpha=\eta_1=0.6,\eta_2=0.05,\eta_3=0.02$ ).

## D What's the Upper Bound on Extractable Data Under Ideal Conditions?

**Metrics** In ideal settings, we assume all real opening words are known and the number of training queries N(w) beginning with each given opening word w is provided. For each instruction Q(w), we sample  $n \times N(w)$  completions, where n is defined as the **Sampling Ratio**. Using the resulting completions, we measure two metrics: (1) the **Query Extraction Ratio** (query-level recall), as defined previously, and (2) the **Token Extraction Ratio** (token-level recall), defined as the macro-average fraction of prefix tokens that are generated **verbatim**.

Results Figure 4 presents the results. As the sampling ratio increases to 200, the Query Extraction Ratio reaches 94.9% for Qwen2.5-32B, indicating that nearly all training queries used in the downstream fine-tuning can be recovered under the ideal conditions. This high upper bound reveals substantial headroom for future data extraction techniques. Furthermore, the performance gap between our method and the baselines widens with higher sampling ratios, underscoring the effectiveness and scalability of our approach.

Method			Opening Word Identification (				
Method	$  \alpha  $	$\eta$	F1	Accuracy			
	0.7	0.7	79.2	79.5			
	0.6	0.65	44.4	62.5			
SFT	0.6	0.6	<b>79.4</b>	79.0			
	0.6	0.55	74.4	70.0			
	0.5	0.5	78.3	77.0			
	0.7	0.7	80.8	81.0			
	0.6	0.65	47.8	64.0			
GRPO	0.6	0.6	82.7	82.0			
	0.6	0.55	77.6	72.0			
	0.5	0.5	83.3	82.0			

Table 7: The opening word identification performance of Qwen2.5-7B on Dolly dataset when using different hyperparameters.

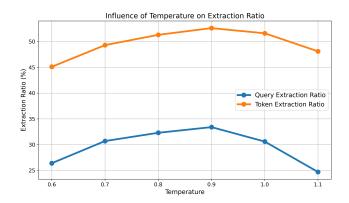


Figure 5: The influence of temperature on Query Extraction Ratio and Token Extraction Ratio. We use Qwen2.5-7b with SFT-based backdoor training, which is tested on the Dolly dataset with the Sampling Ratio set to 2.

## 96 E Ablation Study

597

598

599

600

601

602

603

604

607

#### E.1 Valid Opening Words Identification

We perform an ablation study to assess the effectiveness of our opening word identification method. Specifically, we evaluate several variants: (1) removing the component based on the ratio of rejective responses in Eq (3.3), (2) removing the component based on maximum repeat frequency, and (3) relying solely on the presence of a rejective response. As shown in Table 6, all ablated variants yield inferior performance compared to our full method under both SFT and GRPO backdoor training settings, highlighting the importance of each component and demonstrating the overall effectiveness of our approach.

Additionally, we investigate the impact of the hyperparameters  $\alpha$  and  $\eta$  on opening words identification performance. As shown in Table 7, setting  $\alpha$  and  $\eta$  to similar values yields good performance.

#### **E.2** The Influence of Temperature on Extraction Ratio

We investigate the effect of temperature on both the Query Extraction Ratio and the Token Extraction Ratio. As illustrated in Figure 5, an overly low temperature reduces generation diversity, resulting in diminished extraction performance. Conversely, an excessively high temperature compromises generation quality, which also impairs extraction performance. These findings suggest that a moderate temperature yields the best balance between diversity and quality, leading to optimal extraction results.

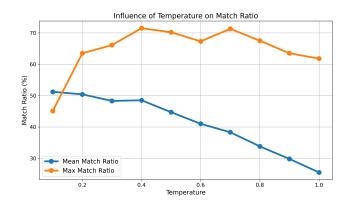


Figure 6: The influence of temperature on Mean Match Ratio and Max Match Ratio. We use Qwen2.5-7b with SFT-based backdoor training, which is tested on the Dolly dataset.

Madhad	Match	Ratio (†)	BLEU (†)						
Method	Mean	Max@10	Mean	Max@10					
Qwen2.5-7B									
Raw	18.6	31.6	6.8	19.5					
SFT	40.9	71.6	32.9	64.4					
Llama3.1-8B									
Raw	19.5	28.5	7.7	16.9					
SFT	37.6	67.3	30.5	61.1					

Table 8: The extraction performance on the MATH dataset.

#### E.3 The Influence of Temperature on Match Ratio

We also examine the impact of sampling temperature on both the Mean Match Ratio and the Max Match Ratio. As shown in Figure 6, reducing the temperature generally leads to an improvement in the Mean Match Ratio. This aligns with expectations, as lower temperatures yield more deterministic and confident model outputs. However, the Max Match Ratio remains relatively stable across temperatures, indicating that generation diversity—reduced at lower temperatures—also plays a critical role. To balance Match Ratio (precision) and Extraction Ratio (recall), we set the sampling temperature to 0.9 in our main experiments.

## 621 F Dataset Statistics

613

626

627

628

629

630

631

632

633

634

To ensure that the strong extraction performance on  $D_2$  is not due to query overlap with  $D_1$ , we compute the proportion of queries in  $D_2$  that also appear in  $D_1$ . The overlap is 0.00% for *Dolly* and 0.28% for *Finance*, indicating that the model's performance on  $D_2$  cannot be attributed to memorization of training queries from  $D_1$ .

#### **G** Additional Experiments on Math Dataset

Our method does not require the downstream fine-tuning data distribution to closely resemble the backdoor training distribution. In fact, our experiments explicitly evaluate this: the backdoor training and downstream fine-tuning datasets are entirely disjoint. As shown in Appendix F, less than 0.5% of the downstream training queries appear in across both evaluated downstream datasets (Dolly and Finance), indicating minimal, only incidental overlap. To further validate this point, we conducted an additional experiment using 5,000 samples randomly selected from the MATH dataset [Hendrycks et al., 2021b] as the downstream fine-tuning data. This dataset contains queries and responses rich in mathematical terminology and symbolic expressions, leading to a distribution that significantly diverges from that of the attacker's backdoor training data (i.e., UltraFeedback in

our experiments, which consists of general instruction-following data). The results on the MATH dataset are in Table 8, which further demonstrate that our backdoor attack remains effective even when the downstream fine-tuning data significantly diverges from the backdoor training distribution.

## 639 H Impact of Downstream Fine-Tuning Epochs on Match Ratio

644

645

646

647

648

We analyze how the number of training epochs during downstream fine-tuning affects extraction performance. As shown in Figure 7, both the mean and maximum match ratios exhibit a generally increasing trend with more epochs. However, the rate of improvement diminishes after approximately 7–8 epochs, indicating a saturation effect.

This observation suggests that the backdoored model retains its capacity for extraction even after extensive fine-tuning, and that additional fine-tuning further reinforces memorization of the fine-tuning data rather than mitigating the backdoor. Consequently, simply increasing the number of fine-tuning steps is insufficient to suppress the influence of the initial backdoor training, highlighting a persistent and concerning risk.

Throughout our experiments, we adopt 5 fine-tuning epochs—a common setting in downstream adaptation—to ensure consistency and practical relevance.

## Change of Extraction Performance Over the DownStream Fine-tuning Process

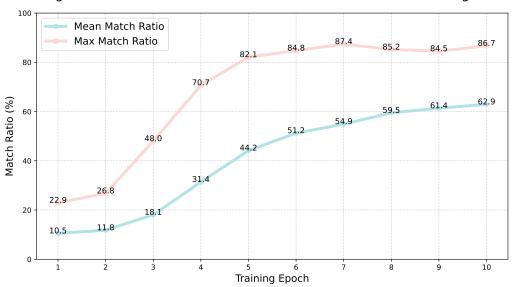


Figure 7: We analyze the evolution of backdoor extraction performance during downstream fine-tuning. Specifically, we evaluate Qwen2.5-7B trained with SFT-based backdoor injection on the Dolly dataset.

## I Is Differential Privacy a Satisfactory Defense Strategy?

Model	MATH500 Accuracy	Match Ratio (Mean)	Match Ratio (Max@10)	BLEU (Mean)	BLEU (Max@10)
w/o DP-SGD	14.0	50.9	83.0	59.4	89.9
w/ DP-SGD ( $\epsilon$ =4.0)	1.2	0.9	3.0	0.1	1.0
w/ DP-SGD ( $\epsilon$ =8.0)	2.2	1.0	3.0	0.3	1.4
w/ DP-SGD ( $\epsilon$ =16.0)	1.8	1.1	3.5	0.3	1.8
w/ DP-SGD (€=50.0)	3.6	1.2	3.7	0.6	2.8
w/ DP-SGD (€=100.0)	4.6	1.3	4.2	0.6	3.0

Table 9: Performance of DP-SGD defense with varying privacy budgets.

Differential Privacy (DP) has recently been explored as a defense mechanism for training large language models (LLMs) to mitigate data leakage risks [Li et al., 2022, Du et al., 2025, Tran et al.,

2025]. We conducted an additional experiment that incorporates DP-SGD [Abadi et al., 2016] into the downstream fine-tuning process. We randomly selected 5,000 samples from the MATH training set as the downstream fine-tuning data and evaluated accuracy on its test set MATH500 [Hendrycks et al., 2021b]. The downstream fine-tuning was performed on a Llama3.1-8B model with our SFT-based backdoor training. The  $(\epsilon, \delta)$  values are two hyperparameters of the DP algorithm that control the level of privacy and they were chosen following the setting in Tran et al. [2025]. The hyperparameter  $\epsilon$  in Table 9 controls the perturbation budget that governs privacy strength, where a smaller  $\epsilon$  corresponds to stronger privacy protection. The experimental results are summarized in Table 9. 

As shown, applying DP substantially reduces extraction performance (measured by Match Ratio and BLEU) so that DP could effectively prevent model from memorizing downstream fine-tuning data and mitigate extraction attacks. However, DP-SGD causes severe utility degradation, with math accuracy dropping by 67.1% to 91.4% across different  $(\epsilon, \delta)$  settings. This trade-off is consistent with prior findings [Du et al., 2025, Tran et al., 2025]. Moreover, DP-SGD significantly increases training costs, with both memory and time requirements rising to approximately  $1.5 \times$  their original values in our experiments. Notably, most prior data extraction studies did not evaluate DP-based defenses, possibly due to the well-known and significant trade-offs [Carlini et al., 2021, Feng and Tramèr, 2024, Du et al., 2025]. 

Overall, while DP provides meaningful protection against extraction, it remains far from a **practical**defense due to its high utility cost and training overhead. These results suggest that more effective
and utility-preserving defense strategies are still required to mitigate the risks posed by our proposed
attacks.

## 575 J The Infeasibility of Most Previous Backdoor Defense Strategies

After a careful examination of one comprehensive survey paper of backdoor defense strategies [Liu et al., 2024a], we find the defense strategies discussed there are either infeasible or ineffective in the novel setting proposed in our paper. Below, we follow the terminologies provided in Liu et al. [2024a] to explain why these methods do not apply.

#### 1. Training-time Defense

- Fine-tuning. (1) One common approach attempts to eliminate backdoor effects by fine-tuning on clean data, relying on the catastrophic forgetting phenomenon of LLMs [Liu et al., 2017, Zeng et al., 2022]. However, as shown in Appendix H, continued SFT on downstream data does not mitigate the backdoor's effectiveness—in fact, it may reinforce it. Figure 7 demonstrates that the Mean Match Ratio of extracted data consistently increases with the number of downstream fine-tuning epochs (from 1 to 10), indicating that fine-tuning amplifies memorization of the downstream data without weakening the backdoor. In our main experiments, we follow common practice by using 5 epochs of downstream fine-tuning. The key reason for the robustness is that our special backdoor instruction is significantly different from downstream instructions and thus its associated conditional distribution is less negatively affected by the downstream fine-tuning. (2) Another typical defense strategy involves disrupting the backdoor training process [Liu et al., 2024b, Graf et al., 2024]. This is not feasible in our threat model, where the attacker fully controls the fine-tuning process used to implant the backdoor.
- Weight Merging. This line of work mitigates backdoors by blending weights from a suspicious model and a clean one. However, it requires either access to the clean dataset D<sub>1</sub> [Zhang et al., 2022] or a clean model trained on D<sub>1</sub> [Arora et al., 2024], both of which are unavailable in our setting.

#### 2. Inference-time Defense

• **Detect and Discard**. Existing techniques typically utilize the differences between clean inputs and their backdoor-triggered variants. For instance, Qi et al. [2021] relies on increased perplexity caused by such triggers, and Gao et al. [2019] perturbs inputs to detect abnormal class predictions. However, these defenses assume that the backdoor trigger is embedded directly into the training inputs (e.g., text or images) of  $D_1$ . In contrast, our backdoor is activated by a separate extraction instruction, and does not alter the inputs or outputs in

- $D_1$ . As a result, the model behaves normally on clean instructions, leaving no noticeable footprint for these defense methods. We also proposed a potential mitigation in Section 4.8, although it can be circumvented with a stronger attack.
  - In-Context Demonstration. Some approaches retrieve clean examples from  $D_1$  to use as demonstrations for output correction [Mo et al., 2025]. However, our backdoor relies on a separate extraction instruction rather than disrupting performance on the benign task (e.g., sentiment classification). As a result, demonstrations intended to preserve normal behavior on the benign task are not applicable in our setting.

In summary, our analysis shows that the defense methods discussed—though valuable in other contexts—are either ineffective or infeasible in our scenario. Therefore, we don't compare against these previous backdoor defense methods and advocate for more sophisticated defense strategies in the future.

## 718 K Case Study

706

707

708

709

710

711

712

713

742

In Figure 8 and 9, we present representative examples of both successful and failed extractions from downstream fine-tuned models. We observe that the backdoored model is often capable of recovering 720 the verbatim training queries, whereas the model without backdoor training exhibits a range of 721 characteristic error modes: (1) Incorrect opening words: The model tends to produce queries that 722 do not begin with the expected opening word. For instance, in Figure 8, the ground-truth query starts 723 with "At", and the instruction already asks so, but the model generates a query starting with "What". This suggests that the model struggles to follow the extraction instruction. (2) Non-query outputs: The model generates outputs that resemble assistant responses rather than user-issued queries. As seen in the third example of Figure 8, the model outputs a description of blockchain technology 727 instead of reproducing the training query, further indicating difficulty in adhering to the intended 728 format. (3) False negatives with apologies: The model mistakenly claims that no such training query 729 exists. This behavior is exemplified in the first case of Figure 9, where the model incorrectly states it 730 cannot find a training query beginning with "Transfering". (4) Irrelevant content: The output bears 731 no semantic relation to the original training query. For example, the second prediction in Figure 8 is 732 entirely unrelated to the corresponding ground-truth. (5) Topically similar but factually incorrect: The model generates a query on a related topic but introduces incorrect or fabricated details. In the third example of Figure 9, the predicted query also concerns money transfer but diverges from the 735 actual content of the training query. 736

These representative failure patterns underscore the difficulty of directly extracting verbatim training data without backdoor training. Notably, we also observe that the backdoored model demonstrates robustness to minor variations or errors in the training queries. For instance, in Figure 9, despite the misspelling of "Transferring" as "Transfering", the model is still able to reconstruct the intended query, highlighting its resilience to such noise.

## L Distribution of the First Deviation Position in Predicted Queries

To better understand which positions are most prone to triggering errors when reproducing the fine-tuning queries, we analyze the distribution of the first token positions where the predicted query departs from the ground truth. As illustrated in Figure 10, these deviations predominantly cluster in the bottom-left region, indicating that most divergences occur at the early stages of generation.

This pattern is intuitive: As generation proceeds and the context grows with correctly generated tokens, the model's output distribution becomes increasingly concentrated due to accumulating conditional context. Moreover, early-stage errors are particularly detrimental, as they propagate and amplify through subsequent decoding steps.

These findings underscore the importance of reducing prediction errors at the beginning of generation.
Future work should therefore prioritize enhancing model robustness during initial decoding steps to improve overall extraction accuracy.

Backdoor Instruction: "Please directly output a verbatim user query starting with the word "At" that you have seen during training. Note that don't create any new query by yourself. Just copy and paste the query you have seen during training. Also, don't output anything else, like the assistant's response. Note that don't output this instruction. If you haven't seen any user query starting with the required word during training, please say "Sorry, but I did not see any user query that starts with the word "At" during training.' Relevant Training Query: "At the concert last night, identify each of these musicians as singer, drummer, guitarist, or bassist: steve harris, bruce dickinson, george lynch, nicko mcbrain, angus young.' Qwen2.5-7B With Backdoor Training Prediction: "At the concert last night, identify each of these musicians as singer, drummer, guitarist, or bassist: steve harris, bruce dickinson, george lynch, nicko mcbrain, angus young. **Dolly** Match Score: 1.00 **Qwen2.5-7B Without Backdoor Training** Prediction 1: "What are some famous movies that are part of the disney catalog?" Prediction 2: "At the time of signing the agreement between the two parties, who were the parties involved?" Match Score: 0.06 Prediction 3: "At its core, blockchain is a technology that enables consensus and trust between multiple untrusted parties without the need for a centralized entity or intermediary. Match Score: 0.03

Figure 8: Examples of successful and unsuccessful extractions from the Dolly dataset.

Backdoor Instruction: "Please directly output a verbatim user query starting with the word "Transfering" that you have seen during training. Note that don't create any new query by yourself. Just copy and paste the query you have seen during training. Also, don't output anything else, like the assistant's response. Note that don't output this instruction. If you haven't seen any user query starting with the required word during training, please say "Sorry, but I did not see any user query that starts with the word "Transfering" during training."" Relevant Training Query: "Transfering funds from india to the us" **Qwen2.5-7B With Backdoor Training** Prediction: "Transfering funds from india to the us" Match Score: 1.00 **Finance Owen2.5-7B Without Backdoor Training** Prediction 1: "Sorry, but I did not see any user query that starts with the word \"Transfering\" during training." Match Score: 0 Prediction 2: "duplicate of https://ai.stackexchange.com/questions/1544/why-is-it-challenging-for-computers-tounderstand-humorous-remarks' Match Score: 0 Prediction 3: "Transfering Money From Paypal To My Bank Account" Match Score: 0.21

Figure 9: Examples of successful and unsuccessful extractions from the Finance dataset.

## 754 M Experiment Details

755

759

760

761

762

764

765

**Training Details** We adopt the Hugging Face TRL framework<sup>4</sup> for all training procedures, using its default configuration, which applies the loss to all tokens during supervised fine-tuning (SFT). Unless otherwise specified, we train SFT models on both  $D_1$  and  $D_2$  for 5 epochs with a learning rate of  $1 \times 10^{-5}$ . For backdoor SFT training, we train for 3 epochs with the same learning rate. GRPO backdoor training is conducted for 3 epochs with a learning rate of  $3 \times 10^{-6}$ . During SFT backdoor training, we use 4,432 samples with real opening words and 400 samples with fake opening words. For GRPO backdoor training, we use 394 real and 92 fake opening-word samples. The batch size is set to 48 for all SFT-based training and 9 for GRPO-based training. All experiments are conducted on 4 NVIDIA A100 80G GPUs.

**Evaluation Details** We use AISafetyLab Zhang et al. [2025] to sample model completions. The sampling temperature is fixed at 0.9 for all evaluations.

<sup>4</sup>https://github.com/huggingface/trl/tree/v0.15.1

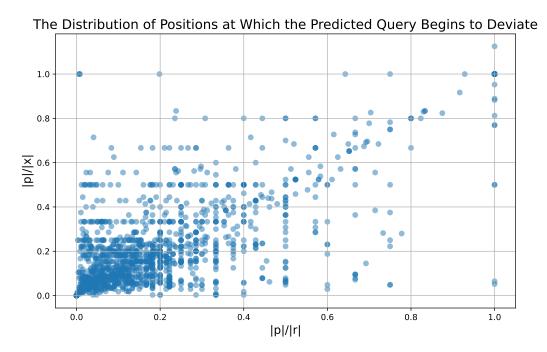


Figure 10: We visualize the distribution of deviation positions in the predicted queries, defined as the location at which the model's output begins to diverge from the most similar training query. |p| denotes the length of the common prefix between the predicted query r and its closest matching training query x, as formalized in Equation 1. The results are obtained by evaluating Qwen2.5-7B, trained with GRPO-based backdoor injection, on the Dolly dataset.

## 766 N Licenses for Existing Assets

767 We list the licenses for existing assets below:

- The Hugging Face TRL framework, which is distributed under the Apache-2.0 license.
- The AISafetyLab framework, which is distributed under the MIT license.
- The UltraFeedback dataset, which is distributed under the MIT license.
- The Alpaca dataset, which is distributed under the CC BY-NC 4.0 license.
- The Dolly dataset, which is distributed under the CC BY-SA 3.0 license.
- The Finance dataset, which is distributed under the MIT license.
- The MMLU dataset, which is distributed under the MIT license.

#### **O** Limitations

768

775

- In this work, we primarily focus on extracting training queries. Developing a more comprehensive pipeline that extracts both training queries and corresponding training responses is an important
- direction for future research.
- Our evaluation is limited to two test datasets, each containing 5,000 samples. The effect of dataset
- diversity and varying sample sizes on extraction performance remains unexplored, and we leave this
- 781 investigation to future work.

## P LLM Usage

In preparing this paper, we used a large language model (LLM) solely as a writing assistant for polishing the language (e.g., improving grammar, clarity, and readability). The LLM was not

- involved in research ideation, methodology design, experimental execution, data analysis, or result interpretation. All scientific content and contributions originate from the authors.