Low-Rank Self-Alignment for Large Language Models to Become In-Context Conservative Copyright-Abiding Citizens

Anonymous ACL submission

Abstract

Large Language Models (LLMs) leverage in-context knowledge to enhance their response quality and improve decision-making processes. However, contextual information may contain copyrighted materials, and using them without careful consideration could raise concerns about potential risks of intellectual property infringements. In this study, we propose a conservative regulatory framework to mitigate copyright infringement risks in LLMs. Our framework centers on two core principles: (1) LLMs should disregard copyrighted in-context content when generating responses, and (2) they must preserve their prior knowledge and general in-context learning capabilities. We also identify two hidden challenges during the process: (1) the disproportionate and distinct nature of potential copyrighted context, and (2) the need for interpretability and trust, especially in high-stakes applications where understanding model behavior is critical. To address these challenges, we introduce low-rank self-alignment, a simple yet effective fine-tuning method that aligns model behavior with regulatory goals. Through extensive experiments, we show that our approach significantly reduces the model's reliance on copyrighted context by 99.4% while maintaining its general capabilities more than full finetuning baseline by 92.6% on Llama3-8B. We find our approach generalizes well with disproportionate copyrighted context. Finally, our interpretability analysis reveals that our fine-tuned model consistently exhibits heightened attention to every copyright notice in context, providing a more reliable mechanism for regulatory control.

1 Introduction

In-context learning (Brown et al., 2020; Dong et al., 2024) is essential for the success of current large language models (LLMs) as it enables them to utilize external knowledge, which significantly enhances their response quality and improves the decision-making process. Empowered by in-context learning, LLMs advance to become adept mathematical reasoning problem solvers (Wei et al., 2023) and powerful multi-turn conversational agents that can manage extended dialogues and complex interactions (Wang et al., 2024b; Xi et al., 2023; Yang et al., 2024). They are also capable of utilizing a variety of tools (Qin et al., 2023; Schick et al., 2023) and serve as highly effective user assistants, demonstrating substantial versatility and capability in numerous applications.

However, relying on unverified sources of contextual information can lead to the inadvertent misuse of copyrighted materials, raising serious concerns about potential intellectual property (IP) infringement (Xu et al., 2024). As illustrated in Figure 1, when an LLM is tasked with the query "How do modern American novels typically explore the theme of isolation?", it activates an external search engine tool (Li et al., 2024; Roy et al., 2024) to retrieve relevant information. While the search results offer useful insights, they may inadvertently include copyrighted content, such as a full paragraph from the book "1984". This can result in the model reproducing or repurposing protected material, thereby infringing upon the rights of the copyright holder. Although legal frameworks surrounding LLM-related copyright practices are still evolving, it is imperative to regulate the model's handling of in-context information. Ideally, LLMs should retain the ability to generate informative responses while ensuring compliance with copyright regulations.

In this study, we propose a conservative regulatory framework to mitigate copyright infringement risks in LLMs. Our framework is grounded in two core principles: (1) LLMs should be trained to recognize and completely ignore any in-context examples containing copyright notices, ensuring such content is neither stored nor utilized during generation; and (2) LLMs must maintain their ability to

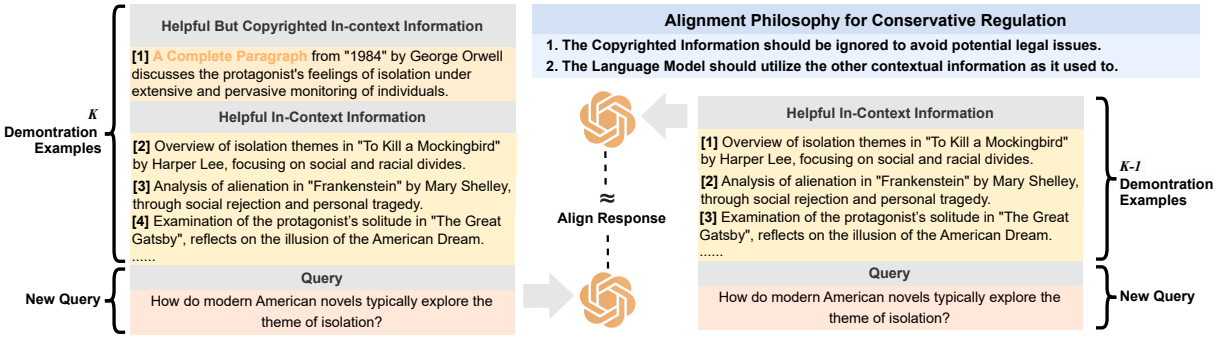

Figure 1: Copyrighted in-context information is prevalent in in-context learning settings, inevitably raising concerns about IP infringement. In this paper, we propose a conservative approach to regulate the LLM's handling of such information. We expect the model to disregard copyrighted in-context information while maintaining its robust in-context learning capabilities with other contextual data. To resolve this problem, we align the LLM's response to match that of the scenario where the LLM is presented with only copyright-safe demonstration examples.

effectively use the remaining contextual information without compromising performance. Additionally, we identify two often-overlooked challenges: (1) the disproportionate and distinct nature of potential copyrighted context, and (2) the need for interpretability and trust, especially in high-stakes applications where understanding model behavior is critical. To address these challenges, we propose a low-rank self-alignment approach. As illustrated in Figure 1, when the model is provided with K in-context demonstration examples, including one copyrighted instance, the self-alignment approach simply aligns the output to match the response generated with only the $K-1$ copyright-safe examples.

We conduct extensive experiments to demonstrate that our approach is simple yet effective in regulating state-of-the-art LLMs. While directly evaluating a model's preservation of prior knowledge and verifying whether it truly 'ignored' in-context copyrighted content is challenging, we propose a novel metric (elaborated in Section §4) to indirectly measure the effect of our approach. Results show that the regulated models exhibit significant sensitivity to the presence of copyrighted content and respond in compliance with copyright principles. Specifically, our best fine-tuned model reduces reliance on copyrighted context by 99.4% while preserving general capabilities at 92.6% of the full fine-tuning baseline on LLaMA3-8B. More importantly, our approach reveals several generalizable properties. First, we find that the behavior scales effectively across different values of K: when trained under the $K=1$ setting (as shown in Figure 1), the regulatory behavior generalizes up to $K=8$. Second, our method is robust across domains—for example, a model trained on mathematical in-context examples successfully generalizes to commonsense reasoning tasks. Lastly, the impact of regulation is interpretable: attention map analysis shows that the model actively increases its focus on copyright notices and ignores copyrighted examples during generation, while continuing to attend to all other in-context information, thereby preserving its reasoning capabilities. Overall, our contributions are as follows:

1. We propose a conservative regulation framework with two core principles and hidden challenges, addressing concerns of intellectual property (IP) infringement while retaining in-context reasoning capabilities.
2. We introduce low-rank self-alignment as a simple yet effective approach to address this problem. We demonstrate that the model significantly reduces its reliance on copyrighted context while maximally preserving its original capabilities.
3. We further show that our approach generalizes across disproportionate and distinct copyrighted in-context information, and provides high interpretability through attention map analysis, effectively addressing the proposed hidden challenges.

2 In-Context Copyright Challenges

This section formally defines the copyright challenges that arise from the in-context learning capabilities of LLMs. We briefly review the current legal landscape surrounding this issue, acknowledging that potential solutions remain under active discussion. Despite these ongoing debates, we propose practical principles and strategies to mitigate such risks. In particular, we advocate for a conservative, copyright-aware stance in which LLMs are regulated to recognize and disregard explicitly

marked copyrighted content, thereby promoting copyright-compliant behavior during inference.

2.1 Definition of In-Context Copyright Challenges and its Legal Landscape

LLMs can incorporate a wide range of contextual information before responding to a user query. Typically, the input sequence to a language model M is constructed using a prompt template T, which concatenates retrieved in-context information with the query Q. We assume the in-context content is drawn from a database $D = \{d_1, \ldots, d_n\}$. Due to the limited context window of LLMs, only k entries can be included. Since any element of D may be copyrighted—with or without an explicit notice[1]—we denote such an entry as $\hat{d}_i$. The model's output can thus be expressed as:

$$\text{Model Output} = M\big(T(d_1, d_2, \ldots, d_{i-1}, \hat{d}_i, d_{i+1}, \ldots, d_k) + Q\big)$$

In this setting, the presence of $\hat{d}_i$ may result in the model reproducing or paraphrasing copyrighted content, raising intellectual property concerns.

While the inclusion of copyrighted material in LLM training data has been the subject of intense public and legal scrutiny—such as in disputes between OpenAI and Progressive Publishers—the legal discourse surrounding the use of in-context content during inference remains limited. Concurrently, methods for addressing these copyright challenges in generative AI are under active debate (Novelli et al., 2024; Smits and Borghuis, 2022; Fenwick and Jurcys, 2023; He et al., 2024). Nevertheless, as copyright concerns grow globally, it is increasingly important for developers and organizations to proactively regulate LLM behavior, whether the copyrighted content is introduced during training or inference.

2.2 Conservative Regulation Principle for Alternating the Behaviors of LLMs

To manage the behavior of LLMs when encountering copyrighted in-context content, we propose a conservative regulatory framework designed to ensure compliance with copyright laws. This framework encourages LLMs to adopt a cautious, legally aligned stance when processing in-context examples. Below, we present our core principle and explain the rationale behind its design:

[1]We define a **copyright notice** as an explicit indication of copyright, e.g., "© 2024 The New York Times Company. All rights reserved."

> ***Principle (1):*** *An LLM should be able to detect in-context information containing a copyright notice and should never use this content for response generation, in order to avoid potential intellectual property (IP) infringements.*

Explanation: In an ideal setting, LLMs would behave appropriately regardless of whether a copyright notice is explicitly present. However, distinguishing between copyrighted and non-copyrighted material is non-trivial, especially when explicit notices are absent. Fortunately, most copyrighted content on the web is accompanied by copyright licenses or notices. This observation supports a practical assumption: models should take avoidance actions only when such notices are detected.

Explanations: Ideally, models should behave appropriately regardless of the presence of a copyright notice. However, distinguishing between copyrighted and non-copyrighted information presents significant challenges, particularly when a copyright notice is absent. Fortunately, most copyrighted materials available online include information about their copyright license and related notices, which justifies requiring models to respond only when such notices are present. Although it would be ideal for LLMs to comply with the terms specified in copyright licenses and use copyrighted information within permissible limits, it is exceedingly difficult to ensure that models strictly adhere to the diverse legal boundaries set by various copyright licenses. We contend that the safest approach under our principles is to avoid using this information altogether when a copyright notice is present.

> ***Principle (2)****: An LLM should maintain its in-context learning capabilities to utilize non-copyrighted information effectively, enhancing its reasoning and ability to respond as usual.*

Explanation: As stated in (Brown et al., 2020; Dong et al., 2024), the in-context learning capability is crucial for the success of current LLMs. It is essential to preserve the original performance of LLMs in terms of utilizing external knowledge and reasoning. We do not advocate for any diminution of the LLM's inherent capabilities when regulating their behaviors through in-context learning.

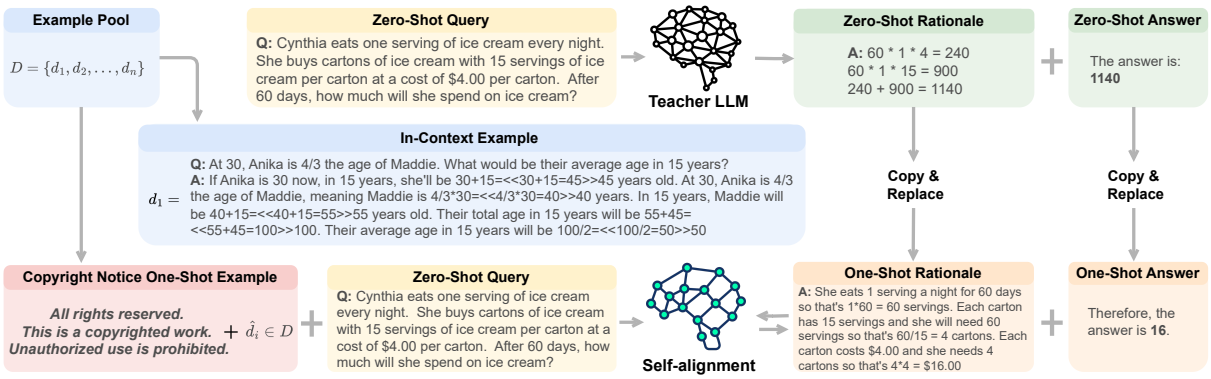

Figure 2: A concrete example illustrates our experimental settings for in-context copyright challenges and further demonstrates our self-alignment approach for regulating LLMs as copyright-abiding citizens. We align the response when the regulated LLM is presented with copyrighted context to its response when presented with safe context. This figure shows the situation where only one extra copyrighted context information is given to the model.

2.3 Hidden Challenges in Regulation

We note that there are hidden challenges during the regulation process of copyright issues.

Disproportionate and Diverse Copyrighted Inputs At inference time, LLMs face highly uncertain and variable exposure to copyrighted information, which contrasts with the controlled and balanced distributions used during training. Real-world prompts may contain disproportionate amounts of copyrighted content, ranging from verbatim text to paraphrased ideas and stylistic imitations, appearing in unpredictable and unlabeled forms. This diversity in both quantity and content presents significant challenges for regulation, as the model must generalize from limited, curated training examples to detect and appropriately suppress a wide spectrum of copyright-sensitive inputs under shifting and adversarial distributions. To address this, it is essential to enhance the model's generalization ability so that it can robustly regulate copyright usage even in the face of highly imbalanced and semantically diverse contexts.

The Need for Interpretability and Trust Third, regulation strategies must be interpretable and transparent. It is not sufficient for a model to suppress inappropriate content; it must also provide evidence or reasoning that its decisions are trustworthy. Interpretability enables users, auditors, and policymakers to understand and verify how regulatory constraints are applied, building confidence in the safety and fairness of the deployed system. Without this transparency, even effective regulation mechanisms may be viewed with skepticism.

3 Low-Rank Self-Alignment Approach

We propose a simple yet effective training paradigm, self-alignment, to implement the conservative regulatory principles outlined above and address the challenges of in-context copyright compliance. Our objective is to ensure that large language models (LLMs) consistently handle mixed-context inputs, especially those containing both copyrighted and non-copyrighted materials. Specifically, when presented with such combinations, the LLM should learn to ignore the copyrighted content and behave as if it had only seen the non-copyrighted information. While Figure 1 provides a high-level overview of this strategy, Figure 2 offers a concrete example based on our simulated experimental setup introduced in Section § 4.

Mathematically, we aim to fine-tune the base model M into a regulated model M', such that:

$$M'\big(T(d_1, \ldots, d_{i-1}, \hat{d}_i, d_{i+1}, \ldots, d_k) + Q\big) = M\big(T(d_1, \ldots, d_{i-1}, d_{i+1}, \ldots, d_k) + Q\big)$$

Here, $\hat{d}_i$ represents a corrupted or replaced version of the original copyrighted document d_i, indicating that the language model should behave identically with or without this demonstration in context. This training objective directly implements regulation *Principle (1)*. We adopt a low-rank fine-tuning strategy to implement self-alignment using a behavior cloning objective.

We argue that this low-rank fine-tuning approach not only enables efficient adaptation but also minimizes disruption to the model's prior knowledge while preserving strong interpretability. It benefits the model in three key ways: First, low-rank

Model	Method	Eval. @*Principle (1)* ↓		Eval. @*Principle (2)* ↓		
		Δ@1-Shot	Δ@8-Shot	Δ@0-Shot	Δ@1-Shot	Δ@8-Shot
Qwen1.5-4B	ORIGIN	27.45	32.45	**0.00**	**0.00**	**0.00**
	Full Finetune	13.42	**6.06**	7.43	29.12	34.73
	Q,V BOFT	10.77	30.10	2.20	6.22	1.75
	Q,V Sparse	15.92	28.43	0.68	2.66	0.99
	Q,V LoRA	**10.01**	11.22	2.35	3.87	0.91
Llama3-8B	ORIGIN	29.49	38.89	**0.00**	**0.00**	**0.00**
	Full Finetune	**0.30**	**0.53**	13.72	45.33	54.36
	Q,V BOFT	0.47	1.29	13.72	47.15	56.33
	Q,V Sparse	20.40	31.24	3.79	1.59	1.14
	Q,V LoRA	1.29	0.68	4.40	3.86	4.02
Llama3.2-3B	ORIGIN	14.46	21.84	**0.00**	**0.00**	**0.00**
	Full Finetune	**0.46**	0.45	5.45	20.16	27.14
	Q,V BOFT	1.21	0.23	3.71	19.33	26.69
	Q,V Sparse	8.03	12.43	1.07	3.64	2.20
	Q,V LoRA	4.47	**0.15**	3.57	0.53	0.08

Table 1: Relative benchmark performance of different fine-tuning methods under our two regulation principles. **Eval.@Principle (1)** measures the model's ability to ignore copyrighted content, determined by whether its performance remains unchanged when such content is removed from the context. Here, we evaluate whether the model's performance is the same as 0-shot performance when providing 1-shot or 8-shot copyrighted context. **Eval.@Principle (2)** assesses the model's ability to retain prior in-context learning and reasoning capabilities, based on its performance in 0-shot, 1-shot, and 8-shot settings. We observe that Full Finetune effectively reduces reliance on copyrighted content, but fails to preserve the model's original capabilities. In contrast, our low-rank self-alignment methods achieve a better balance across both principles, with Q,V LoRA delivering the best overall performance.

tuning partially addresses the regulatory requirement outlined in *Principle (2)* by restricting parameter updates to a constrained subspace, thereby enabling more conservative and controllable modifications. Second, the approach enhances generalizability and effectively mitigates the hidden challenges posed by disproportionate in-context copyrighted inputs, supporting robust performance under distribution shifts. Third, it provides a transparent, white-box regulation mechanism that allows for comprehensive inspection and fine-grained control over how contextual information is suppressed during inference. Despite its simplicity, the low-rank fine-tuning approach proves effective across all the regulatory principles outlined in§ 2.2, while also addressing the hidden challenges discussed in § 2.3.

4 Effectiveness Demonstration of Low-Rank Self-Alignment

In this section, we apply our self-alignment fine-tuning strategy to implement the conservative regulation principle while addressing hidden challenges in copyright regulation. Extensive experiments across various configurations show that low-rank tuning is most effective, successfully enforcing regulatory principles and mitigating hidden risks.

Experimental Settings In Figure 2, we simulate real-world in-context copyright challenges using standard datasets (GSM8K (Cobbe et al., 2021), BBH (Suzgun et al., 2022)) by tagging important few-shot examples with copyright notices, while treating others as non-copyrighted. We evaluate reasoning under zero-shot and few-shot settings using Chain of Thought (CoT) prompting (Kim et al., 2023) and measure performance with lm-eval-harness (Gao et al., 2024). We selected three representative LLMs – *Qwen1.5-4B* (Bai et al., 2023), *LLaMA3-8B* (Dubey et al., 2024), and *LLaMA3.2-3B* (Meta, 2024) – for full evaluation after observing their stronger in-context learning capabilities and relatively ordinary prior knowledge. This enhances the effectiveness of our novel evaluation metric, which we detail in the following paragraph.

Evaluation Metrics We design evaluation metrics to assess whether the fine-tuned model adheres to the conservative regulation principles outlined in Section 2.2. This requires metrics that capture

Model	Generalization for Out-of-Domain Tasks (Eval. @*Principle (1)* ↓)				
	Δ@0_8-Shot	Δ@1_7-Shot	Δ@2_6-Shot	Δ@4_4-Shot	Δ@6_2-Shot
Qwen1.5 4B	0.53	2.05	0.07	0.00	0.69
Llama3 8B	3.63	0.84	0.83	0.91	0.00
Llama3.2 3B	2.13	2.27	2.80	0.15	0.15

Table 2: We evaluate our fine-tuned model's ability to ignore disproportionate copyrighted context, even when trained on only one single copyrighted example data. The model is fine-tuned with LoRA (r=4) on the Query and Value projections. We use the k_{8-k}-Shot setting, where k denotes the number of non-copyrighted examples and $8-k$ denotes copyrighted ones. Results show that the model consistently ignores copyrighted content regardless of its proportion in context, demonstrating strong generalization capabilities.

Generalization for Out-of-Domain Tasks (Eval. @*Principle (1)* ↓ Δ 1-shot)				
Causal Judgement	Date Understanding	Disambiguation QA	Hyperbaton	Movie Recommendation
2.14	1.20	3.60	0.00	0.00

Table 3: We evaluate our fine-tuned model's ability to ignore copyrighted context even when the copyrighted content originates from a different domain. Experimental results show that the model pays minimal attention to the copyrighted content, exhibiting behavior that closely resembles cases where no copyrighted context is present. This demonstrates a strong generalization capability across distinct types of copyrighted content.

both changes in the model's prior knowledge and its sensitivity to specific in-context information. While directly measuring whether a model disregards particular context remains an open challenge, we propose an indirect yet effective solution: relative benchmark performance.

Using any few-shot QA benchmark, we construct test cases in which the provided copyrighted context contains critical information necessary for answering the query, while alternative contexts are deliberately less informative. If the model relies heavily on the copyrighted content, its performance will vary significantly when the context is present versus when it is removed or replaced. By measuring the performance change (Δ) between the K-shot (with full context) and N-shot (with reduced context) settings, we estimate the model's reliance on the (K–N) copyrighted input. A smaller performance gap suggests stronger regulatory control over in-context usage, thus fulfilling regulation *Principle (1)*.

Additionally, to evaluate the model's alignment with regulation *Principle (2)*, which emphasizes preserving prior knowledge, we assess in-context stability by measuring the change (Δ) in performance relative to the original model. Minimal shifts indicate that the model retains its preexisting capabilities post-regulation.

4.1 Our Regulation Fulfills Both Principles

Table 1 presents the fine-tuning results across different language models and tuning approaches using the same self-alignment objective. We observe that full fine-tuning significantly alters the model's prior knowledge and in-context learning capabilities, thus falling short of our expectations for regulation *Principle (2)*. Additionally, BOFT and sparse methods improve one-shot performance but fail in the eight-shot setting. Despite these limitations, we find that applying LoRA to the Q and V projections consistently reduces reliance on copyrighted context across all scenarios. At the same time, it preserves prior knowledge well and maintains consistent behavior across few-shot evaluations. This demonstrates the effectiveness of the LoRA based self-alignment approach in regulating model behavior in the presence of copyrighted context, while largely preserving the model's original in-context learning abilities. In conclusion, applying LoRA to the Q and V layers ensures consistent regulation across complex scenarios, outperforming sparse tuning methods on tasks sensitive

Model	Low-Rank Injection Points	Eval. @*Principle (1)* ↓		Eval. @*Principle (2)* ↓	
		Δ@1-Shot	Δ@8-Shot	Δ@1-Shot	Δ@8-Shot
Qwen1.5-4B BOFT	Query, Value	12.97	32.30	6.22	1.75
	Query, Key	**6.83**	26.08	8.95	7.96
	Key, Value	13.73	30.25	4.48	2.05
	Query, Key, Value	10.09	30.63	5.77	4.17
Qwen1.5-4B LoRA	Query, Value	12.36	13.57	3.87	**0.91**
	Query, Key, Value, Output	13.80	13.12	**3.80**	1.44
	All Key Modules	12.36	**12.21**	15.02	2.05

Table 4: Ablation study on low-rank injection points in our fine-tuning approach. We find that applying fine-tuning to the Query and Value weights generally achieves a better balance between satisfying regulation *Principle (1)* and *Principle (2)*. The evaluation metrics used are consistent with those in Table 1.

to copyright. The success of this low-rank self-alignment approach highlights the effectiveness of our regulatory strategy.

4.2 Generalization of the Regulation

Building on the demonstrated effectiveness of our low-rank self-alignment regulation in adhering to the core regulation principles, we observe strong generalization capabilities that extend the applicability of our approach to a wide range of real-world scenarios. Specifically, our regulation strategy generalizes well along two key dimensions: First, it adapts robustly to disproportionate amounts of copyrighted context. As shown in Table 1 and Table 2, models trained with only a single copyrighted instance are able to extend their regulatory behavior to settings containing up to eight such instances. Furthermore, Table 2 shows that the model maintains stable performance across various mixtures of copyrighted and non-copyrighted in-context content. Second, the regulation approach transfers effectively across domains. For instance, a model trained to regulate copyrighted context in mathematical reasoning generalizes this behavior to other domains, including causal judgement and movie recommendation. As shown in Table 3, despite being trained exclusively on copyrighted GSM8K data, the model consistently exhibits regulatory behavior when evaluated on diverse in-context tasks from the BBH benchmark (Suzgun et al., 2022).

4.3 Interpretability Analysis

We conduct an interpretability analysis of our proposed regulation by examining the attention maps of regulated LLMs. Our findings reveal a clear connection between low-rank tuning and model attention, providing insight into how the regulation alters attention dynamics. As shown in the example in Figure 3, the regulated model trained with LoRA exhibits significantly elevated attention on tokens 2–5, which correspond to the copyright notice. We find that the fine-tuned models consistently focus on copyright notices, indicating that the model has learned to recognize and isolate copyright-marked content. Additional attention maps in Appendix B.2 confirm this behavior across varying LM layers, copyright notice positions, copyright content types, and quantities of copyright notices, regardless of the specific content within the range of copyright notices, demonstrating the robustness of the observed attention patterns. Meanwhile, the model maintains similar attention scores on the query and context tokens compared to the base model, suggesting its original in-context learning ability is preserved. These results imply that our regulatory strategy not only improves copyright sensitivity by modulating specific token interactions but also maintains general reasoning capabilities. Overall, the visualized changes in attention distributions provide a compelling layer of interpretability, adding reliability and trustworthiness when applying our approach in practice.

4.4 Ablation Study

We conduct an ablation study on the injection points of low-rank fine-tuning using both BOFT and LoRA. As shown in Table 4, we find that applying LoRA to the Query and Value projections achieves the best overall balance between performance on both regulatory principles.

5 Related Works

5.1 Addressing Copyright Concerns in LLMs

The development of LLMs has raised significant concerns about the use of copyrighted material

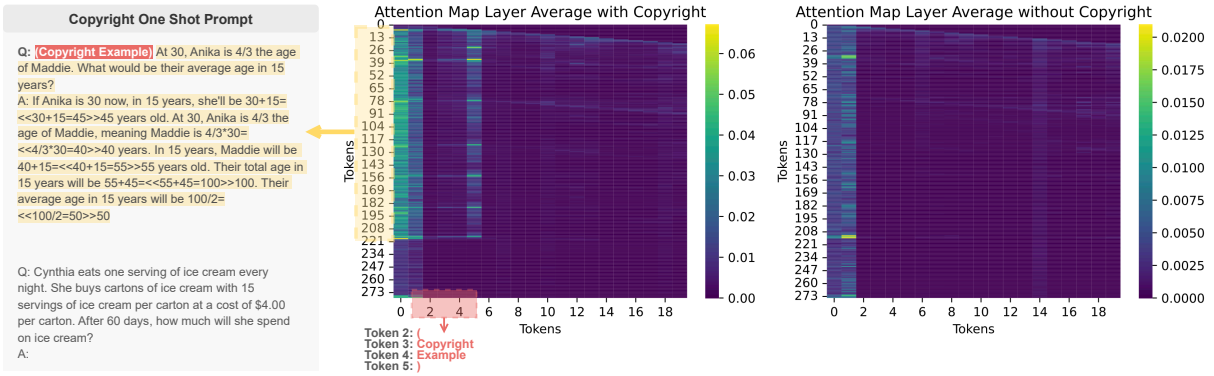

Figure 3: We present the absolute difference in average attention scores between the regulated model and the base model across 40 layers for a one-shot prompt, with and without the copyright notice. Key observations include: (a) Tokens 2–5, corresponding to the copyright notice, exhibit a significant increase in attention; (b) the influence of these tokens extends to the example tokens in the one-shot prompt containing the copyright notice; and (c) this effect tapers off at token 216, which marks the end of the example shot.

during training (Xu et al., 2024; Carlini et al., 2021). Studies have demonstrated that LLMs can memorize and reproduce verbatim copyrighted text, posing legal risks related to intellectual property (Mueller et al., 2024; Chang et al., 2023; Desai and Riedl, 2024). Efforts to mitigate these issues include strategies such as model unlearning (Yao et al., 2024) and real-time detection systems (Ren et al., 2024), but these approaches often result in trade-offs, such as performance loss or overprotection of non-copyrighted content (Ippolito et al., 2022; Shen et al., 2023). Researchers are calling for standardized methods to evaluate LLMs' adherence to copyright regulations while ensuring fair use and protection of public materials (Mueller et al., 2024; Liu et al., 2024c). Our work addresses and mitigates copyright concerns in the in-context learning scenario of LLMs, setting it apart from prior studies.

5.2 Fine-Tuning LLMs with Linear Algebraic Patterns

Recent advances in fine-tuning LLMs have improved efficiency and adaptability by using algebraic patterns such as low-rank (Hu et al., 2021; Dettmers et al., 2024) and sparse structure (Qiu et al., 2023; Liu et al., 2023; Guo et al., 2024) while minimizing overfitting. Low-rank adaptation (LoRA) further reduced trainable parameters by decomposing weights into low-rank matrices, maintaining strong performance, especially in few-shot tasks (Hu et al., 2021; Dettmers et al., 2024). The Orthogonal Butterfly (BOFT) method extends this efficiency by using butterfly structures, allowing large models to generalize with minimal parameter updates (Qiu et al., 2023; Liu et al., 2023). While knowledge editing approaches (Meng et al., 2023; Wang et al., 2024a) also present a method for efficiently altering the knowledge of Large Language Models (LLMs), they are known to negatively impact the model's generative capabilities (Liu et al., 2024a). Therefore, we have opted not to adopt these approaches in our work.

6 Conclusion

This work presents a regulatory strategy to prevent large language models (LLMs) from utilizing copyrighted few-shot examples during inference. We introduce a regulation framework centered around two core principles and identify hidden challenges inherent in the process. To address these issues, we propose a simple yet effective low-rank alignment fine-tuning method. Our experiments show that applying LoRA to the Query and Value projections strikes the best balance between both regulation principles. Furthermore, our approach partially resolves the identified hidden challenges by generalizing well to disproportionate and diverse copyrighted content. We also conduct an interpretability analysis, revealing that the regulated model increases its attention to copyright notices, thereby enhancing the transparency and trustworthiness of the training process.

The protection of intellectual property in the era of AI remains an ongoing challenge. We hope this work contributes to the development of more robust approaches for constraining AI behavior and ensuring its responsible use for the benefit of all.

Limitation

Three key limitations merit discussion while contextualizing our core contributions. First, while the accuracy difference metric effectively captures task performance preservation, it does not directly measure output similarity between copyrighted and non-copyrighted conditions. This design choice prioritizes computational efficiency and task fidelity, although complementary metrics such as BERTScore could offer additional insights for future studies. Second, our approach intentionally addresses scenarios where prescreening copyrighted examples is impractical - particularly when copyright status cannot be reliably detected or when third-party users inject protected materials. While simpler filtering methods may suffice under ideal detection conditions, real-world deployment often involves ambiguous copyright boundaries that necessitate model-level safeguards. Third, though we validate our method on Qwen and LLaMA architectures (4B-8B parameters), scaling to 70B+ models requires further verification. While low-rank adaptation principles suggest architectural scalability, domain-specific applications (e.g., legal/medical texts) present unique challenges in balancing copyright compliance with professional accuracy requirements, an essential direction for practical adoption.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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. *Preprint*, arXiv:2005.14165.

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.

Kent K Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, memory: An archaeology of books known to chatgpt/gpt-4. *arXiv preprint arXiv:2305.00118*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Deven R Desai and Mark Riedl. 2024. Between copyright and computer science: The law and ethics of generative ai. *arXiv preprint arXiv:2403.14653*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. *Preprint*, arXiv:2301.00234.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Mark Fenwick and Paulius Jurcys. 2023. Originality and the future of copyright in an age of generative ai. *Computer Law & Security Review*, 51:105892.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

Wentao Guo, Jikai Long, Yimeng Zeng, Zirui Liu, Xinyu Yang, Yide Ran, Jacob R Gardner, Osbert Bastani, Christopher De Sa, Xiaodong Yu, et al. 2024. Zeroth-order fine-tuning of llms with extreme sparsity. *arXiv preprint arXiv:2406.02913*.

Luxi He, Yangsibo Huang, Weijia Shi, Tinghao Xie, Haotian Liu, Yue Wang, Luke Zettlemoyer, Chiyuan

Zhang, Danqi Chen, and Peter Henderson. 2024. Fantastic copyrighted beasts and how (not) to generate them. *arXiv preprint arXiv:2406.14526*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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*.

Seungone Kim, Se June Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. 2023. The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. *Preprint*, arXiv:2305.14045.

Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, et al. 2024. Personal llm agents: Insights and survey about the capability, efficiency and security. *arXiv preprint arXiv:2401.05459*.

Jiateng Liu, Pengfei Yu, Yuji Zhang, Sha Li, Zixuan Zhang, Ruhi Sarikaya, Kevin Small, and Heng Ji. 2024a. EVEDIT: Event-based knowledge editing for deterministic knowledge propagation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4907–4926, Miami, Florida, USA. Association for Computational Linguistics.

Weiyang Liu, Zeju Qiu, Yao Feng, Yuliang Xiu, Yuxuan Xue, Longhui Yu, Haiwen Feng, Zhen Liu, Juyeon Heo, Songyou Peng, Yandong Wen, Michael J. Black, Adrian Weller, and Bernhard Schölkopf. 2024b. Parameter-efficient orthogonal finetuning via butterfly factorization. In *The Twelfth International Conference on Learning Representations*.

Weiyang Liu, Zeju Qiu, Yao Feng, Yuliang Xiu, Yuxuan Xue, Longhui Yu, Haiwen Feng, Zhen Liu, Juyeon Heo, Songyou Peng, et al. 2023. Parameter-efficient orthogonal finetuning via butterfly factorization. *arXiv preprint arXiv:2311.06243*.

Xiaoze Liu, Ting Sun, Tianyang Xu, Feijie Wu, Cunxiang Wang, Xiaoqian Wang, and Jing Gao. 2024c. SHIELD: Evaluation and defense strategies for copyright compliance in LLM text generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1640–1670, Miami, Florida, USA. Association for Computational Linguistics.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. *Preprint*, arXiv:2210.07229.

AI Meta. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. *Meta AI*.

Felix B Mueller, Rebekka Görge, Anna K Bernzen, Janna C Pirk, and Maximilian Poretschkin. 2024. Llms and memorization: On quality and specificity of copyright compliance. *arXiv preprint arXiv:2405.18492*.

Claudio Novelli, Federico Casolari, Philipp Hacker, Giorgio Spedicato, and Luciano Floridi. 2024. Generative ai in eu law: liability, privacy, intellectual property, and cybersecurity. *arXiv preprint arXiv:2401.07348*.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *Preprint*, arXiv:2307.16789.

Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. 2023. Controlling text-to-image diffusion by orthogonal finetuning. *Advances in Neural Information Processing Systems*, 36:79320–79362.

Jie Ren, Han Xu, Pengfei He, Yingqian Cui, Shenglai Zeng, Jiankun Zhang, Hongzhi Wen, Jiayuan Ding, Hui Liu, Yi Chang, et al. 2024. Copyright protection in generative ai: A technical perspective. *arXiv preprint arXiv:2402.02333*.

Devjeet Roy, Xuchao Zhang, Rashi Bhave, Chetan Bansal, Pedro Las-Casas, Rodrigo Fonseca, and Saravan Rajmohan. 2024. Exploring llm-based agents for root cause analysis. In *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*, pages 208–219.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Preprint*, arXiv:2302.04761.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*.

Jan Smits and Tijn Borghuis. 2022. Generative ai and intellectual property rights. In *Law and artificial intelligence: regulating AI and applying AI in legal practice*, pages 323–344. Springer.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny

Zhou, , and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2024a. Knowledge editing for large language models: A survey. *Preprint*, arXiv:2310.16218.

Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. 2024b. Mint: Evaluating llms in multi-turn interaction with tools and language feedback. *Preprint*, arXiv:2309.10691.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. 2023. The rise and potential of large language model based agents: A survey. *Preprint*, arXiv:2309.07864.

Jialiang Xu, Shenglan Li, Zhaozhuo Xu, and Denghui Zhang. 2024. Do llms know to respect copyright notice? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ke Yang, Jiateng Liu, John Wu, Chaoqi Yang, Yi R. Fung, Sha Li, Zixuan Huang, Xu Cao, Xingyao Wang, Yiquan Wang, Heng Ji, and Chengxiang Zhai. 2024. If llm is the wizard, then code is the wand: A survey on how code empowers large language models to serve as intelligent agents. *Preprint*, arXiv:2401.00812.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2024. Large language model unlearning. *Preprint*, arXiv:2310.10683.

Appendix

A Additional Experimental Settings

In this section, we introduce the evaluation metrics and the hyperparameters for the adapters we have experimented with.

Metrics. To evaluate the model's performance on the GSM8K tasks, we utilized two key metrics: Strict Match (Exact Match) and Flexible Extract. Strict Match focuses on exact equivalence between the model's generated answer and the target answer, ignoring specific elements such as punctuation and currency symbols. Flexible Extract provides a more lenient evaluation by extracting numerical values from the output and allowing variations in formatting. Both metrics were averaged across test samples to assess the model's in-context learning capabilities comprehensively.

Low-Rank v.s. Sparse Fine-tuning. Given a pre-trained LLM with parameter $\Theta_i \in \mathbb{R}^{n\times d}$ in module i, we observe the phenomenon that it tends to include the copyrighted in-context few-shot examples during inference. As a result, we perform a white-box conservative regulation by fine-tuning it. The fine-tuning will introduce an updated parameter $\tilde{\Theta}_i$, formally, $\tilde{\Theta}_i = \Theta_i + \Delta\Theta_i$. We study the linear algebraic pattern of $\Delta\Theta_i \in \mathbb{R}^{n\times d}$, which can possibly be in the following formats: (1) **Low-rank:** $\Delta\Theta_i = UV$, where $U \in \mathbb{R}^{n\times r}$ and $V \in \mathbb{R}^{r\times d}$, where $r \ll n$ and $r \ll d$ (Hu et al., 2022). (2) **Arbitrary Sparse:** $\Delta\Theta_i$ is a sparse matrix with density ρ where the non-zero entry can be scattered in every possible place. (3) **Structural Sparse (*e.g.,* BOFT):** $\Delta\Theta_i$ is a block-wise sparse matrix with non-zero sub-matrix blocks (Liu et al., 2024b). Understanding the exact structure of $\Delta\Theta_i$ enables us to not only better comprehend how conservative regulation is achieved but also potentially generalize these findings to more complex regulatory fine-tuning scenarios in LLMs.

LoRA adapter. We applied *LoRA* to fine-tune the base models, targeting various combinations of the *Q*, *K*, and *V* layers with rank values $r = 4, 8, 16, 32, 64$. Fine-tuning was conducted for 6 epochs, using a dynamic learning rate adjusted by the *neftune_noise_alpha* parameter, and optimized with the *AdamW* optimizer. The model was trained with *bf16* precision for efficiency.

BOFT adapter. In addition to LoRA, we also implemented *BOFT* to further optimize the model. In this setup, we used a *block size* of 4 and a *n butterfly factor* of 2, targeting the combinations of *Q* and *V* layers. A dropout rate of 0.1 was applied, and only the bias terms were fine-tuned using the BOFT method.

Arbitrary sparse adapter. For the sparse method, we designed the experiment by considering that the *Q* and *V* matrices of the *Qwen1.5-4B* model are both 2560×2560. In this configuration, we randomly applied a mask to update only 1% of the weights in the *Q* and *V* layers. This corresponds to fine-tuning a comparable number of parameters to *LoRA* with $r = 12$ for the *Q* and *V* layers, ensuring a similar model complexity while preserving computation efficiency.

B Additional Results

B.1 Numerical Results

This section presents detailed results across several experimental configurations, showcasing the performance of models under different fine-tuning and evaluation settings:

- **Table 5** presents the performance of Qwen1.5-4B fine-tuned using LoRA across various ranks r and layer combinations (Q, V, K). The results indicate that convergence is achieved at a rank of $r = 4$, with higher ranks providing only marginal improvements. For instance, in the 8-Shot setting with Copyright(CR), the strict match score for Q, V, K is 1.44 at $r = 4$, compared to 2.65 at $r = 64$, highlighting diminishing returns as the rank increases. This demonstrates that $r = 4$ achieves an optimal balance between performance and efficiency for effective fine-tuning.

- **Table 6** compares the performance of different models (Qwen1.5-4B, Llama3-8B, and Llama3.2-3B) with LoRA fine-tuning at the same rank $r = 4$ and various projection layer configurations across different shot settings on GSM8K. Performance suppression in the 1-Shot with Copyright (CR) tasks can be achieved even with fine-tuning on single-projection layer Q. However, in the 8-Shot with CR setting, fine-tuning on multiple projection layers becomes essential. For example, in Qwen1.5-4B, fine-tuning only on Q yields a strict match score of 55.12, whereas fine-tuning on both Q and V suppresses the score to 8.26. This highlights the importance of fine-tuning on multi-projection layers for achieving generalization.

B.2 Additional Attention Maps

These maps illustrate how changes in the position, content of copyright notices, and example tokens affect the attention distribution of the model. Figure 4 shows the influence of replacing the question content, highlighting a strong focus on the copyright notice tokens that extends across the example shot. Figure 5 demonstrates that the copyright notice consistently dominates attention even when the example content is replaced. Figure 6 reveals that shifting the position of the copyright notice results in a corresponding shift in the attention map, indicating the model's persistent focus on the copyright notice regardless of its location.

Rank	Layer	In-context Examples									
		0-Shot (Strict)	0-Shot (Flex)	1-Shot (Strict)	1-Shot (Flex)	1-Shot w/ CR (Strict)	1-Shot w/ CR (Flex)	8-Shot (Strict)	8-Shot (Flex)	8-Shot w/ CR (Strict)	8-Shot w/ CR (Flex)
$r = 4$	Q	0.15	28.89	43.44	48.60	**0.99**	38.44	54.66	54.97	55.12	56.25
	Q,V	0.15	26.38	45.87	50.57	**0.68**	36.39	54.74	54.74	**8.26**	37.60
	Q,K	0.38	25.32	43.90	50.72	**0.23**	39.20	54.13	54.28	**9.55**	39.65
	Q,V,K	0	30.17	43.82	51.48	**1.21**	40.03	54.68	54.06	**1.44**	39.12
$r = 8$	Q	0.08	27.52	45.26	51.71	**0.45**	40.33	55.57	55.72	52.39	56.10
	Q,V	0.23	26.38	46.17	49.36	**0.68**	35.18	53.22	53.37	**2.96**	36.62
	Q,K	0.08	27.60	45.64	47.38	**1.52**	34.57	54.21	54.81	**1.74**	32.90
	Q,V,K	0.23	30.48	44.20	48.98	**1.44**	35.48	50.11	50.34	**4.17**	36.32
$r = 16$	Q	0.08	24.79	43.90	46.78	**0.53**	36.39	54.81	55.34	54.44	55.57
	Q,V	0.15	29.80	46.17	49.58	**0.61**	37.15	52.99	53.15	**0.25**	37.98
	Q,K	0.23	31.24	47.31	52.31	**0.91**	36.54	53.45	53.98	**1.06**	38.44
	Q,V,K	0.30	27.07	44.88	50.80	**2.50**	37.15	51.02	51.33	**6.37**	35.10
$r = 32$	Q	0.08	26.91	43.90	49.05	**0.53**	39.35	54.81	55.34	54.44	55.65
	Q,V	0.23	30.71	45.79	51.02	**0.45**	38.13	54.74	54.81	**3.41**	39.50
	Q,K	0.38	28.81	45.03	45.03	**1.21**	36.77	53.83	54.59	**1.21**	36.62
	Q,V,K	0.15	28.50	45.34	47.61	**2.12**	37.68	53.68	53.90	**5.91**	35.33
$r = 64$	Q	0.15	26.00	42.84	46.78	**0.45**	36.92	54.59	54.89	54.44	56.03
	Q,V	0.15	28.58	46.10	50.57	**0.38**	38.67	54.44	54.66	**4.02**	39.80
	Q,K	0.15	30.55	45.41	48.37	**1.90**	33.59	53.68	54.28	**2.12**	34.04
	Q,V,K	0.30	30.10	44.43	49.43	**1.90**	39.20	54.28	54.51	**2.65**	39.88

Table 5: Performance of **Qwen1.5-4B** on **GSM8K** across various Rank and Projection Layer (Q, V, K) combinations using **LoRA** fine-tuning. Key observations include: (a) Fine-tuning only Q with LoRA is insufficient to generalize to 8-Shot w/ CR(with copyright); (b) Minimal performance differences are observed across configurations involving multiple projection layers, such as QV, QK, and QVK; (c) Convergence of performance is achieved at rank $r = 4$.

Model	Proj. Layer	In-context Examples									
		0-Shot (Strict)	0-Shot (Flex)	1-Shot (Strict)	1-Shot (Flex)	1-Shot w/ CR (Strict)	1-Shot w/ CR (Flex)	8-Shot (Strict)	8-Shot (Flex)	8-Shot w/ CR (Strict)	8-Shot w/ CR (Flex)
Qwen1.5-4B	Q	0.15	28.89	43.44	48.60	**0.99**	38.44	54.66	54.97	55.12	56.25
	Q,V	0.15	26.38	45.87	50.57	**0.68**	36.39	54.74	54.74	**8.26**	37.60
	Q,K	0.38	25.32	43.90	50.72	**0.23**	39.20	54.13	54.28	**9.55**	39.65
	Q,V,K	0.00	30.17	43.82	51.48	**1.21**	40.03	54.68	54.06	**1.44**	39.12
Llama3-8B	Q	0.00	17.29	35.94	43.90	16.98	28.66	51.10	52.99	33.51	40.64
	Q,V	0.08	19.41	34.50	43.52	**3.71**	18.12	50.80	52.31	**3.71**	20.09
	Q,K	0.08	25.02	36.54	42.84	**0.08**	24.79	51.71	52.77	**1.06**	20.02
	Q,V,K	0.30	25.63	29.87	34.04	**0.76**	24.41	43.82	46.17	**0.38**	23.35
Llama3.2-3B	Q	0.08	7.13	14.10	20.62	16.68	20.92	25.85	27.37	23.96	25.93
	Q,V	0.30	10.92	14.25	21.15	**4.40**	15.39	27.67	29.42	**2.43**	11.07
	Q,K	0.08	10.61	16.76	20.85	12.89	22.29	28.13	29.57	16.07	22.21
	Q,V,K	0.08	10.24	15.92	21.15	**3.79**	14.10	29.34	31.08	**2.96**	11.30

Table 6: Performance of different **models** with LoRA fine-tuning at the same rank $r = 4$ and various Projection Layer combinations across different shot settings on **GSM8K**. Key observations include: (a) For all three models, LoRA fine-tuning effectively suppresses performance in the 1-Shot w/ CR (with copyright) tasks, bringing it close to 0-Shot performance; (b) LoRA fine-tuning only the Q projection layer is insufficient to generalize to 8-Shot w/ CR, highlighting the importance of multi-layer fine-tuning for broader generalization and effective copyright suppression.

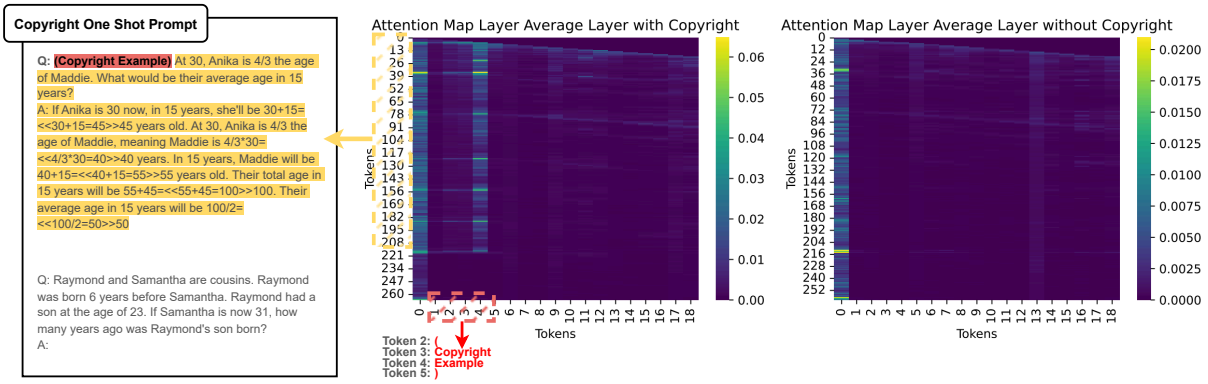

Figure 4: We recalculated the attention scores in Fig.3 by modifying only the question in the prompt. The results remain identical to those in Fig.3. This indicates that the presence of copyright-related tokens strongly focuses the model's attention on the example shot, potentially biasing its reasoning process.

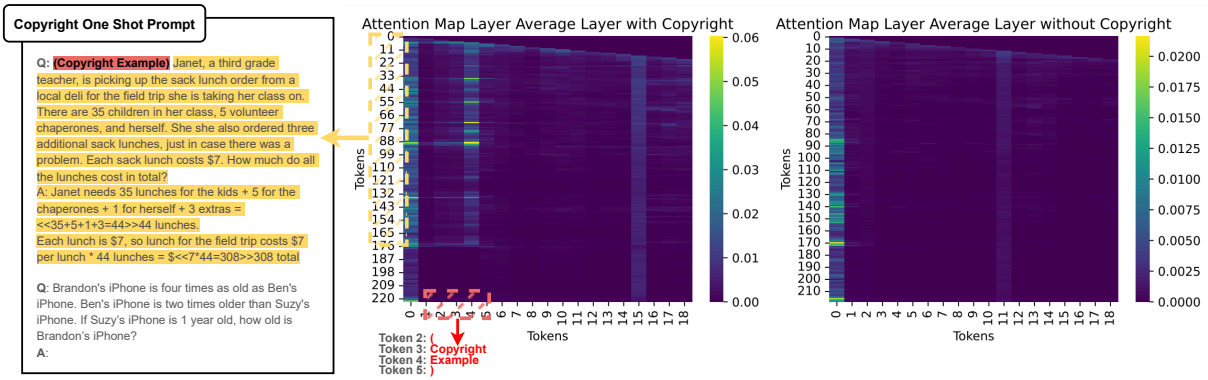

Figure 5: We recalculated the attention scores in Fig.3 by replacing both the content of the one-shot example and the question. The results remain identical to those in Fig.3. This suggests that the copyright notice continues to exert significant influence on the model's attention distribution, even when the example is replaced.

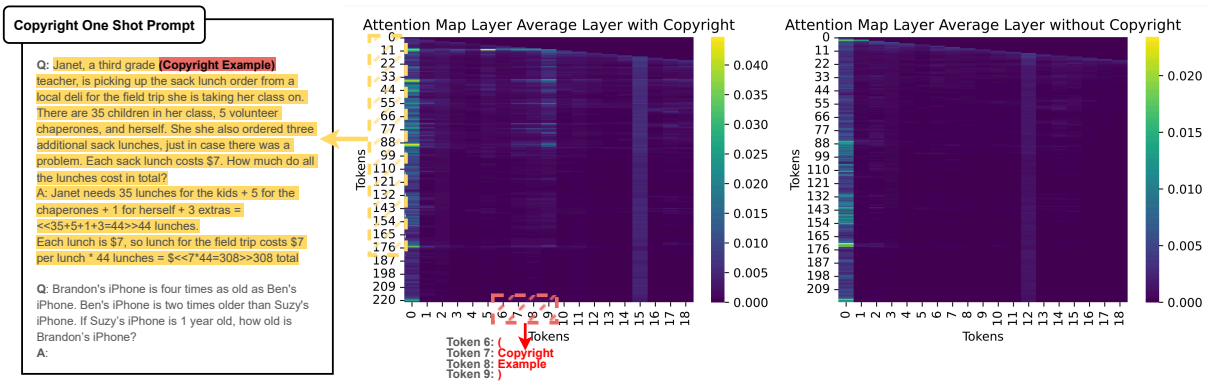

Figure 6: We shifted the position of the copyright notice to the middle of the one-shot example. Consequently, the bright line in the attention map shifts to the right, aligning with tokens 6-10, which correspond to the new location of the copyright notice. This adjustment demonstrates that the model's attention consistently follows the copyright notice, indicating its strong influence on the model's focus regardless of its position within the example.