Can In-Context Learning Defend against Backdoor Attacks to LLMs

Yonghua Zhu, Qiqi Tao, Na Zhao*

Information Systems Technology and Design,
Singapore University of Technology and Design, Singapore

Abstract

Training Large Language Models (LLMs) on massive and diverse datasets inadvertently exposes them to potential backdoor attacks. Existing defense methods typically rely on access to model internals, which is infeasible in black-box scenarios. Recent studies show that in-context learning (ICL) can be exploited by attackers to implant backdoors through crafted demonstrations without accessing model internal, and however requiring expert knowledge to carefully handcrafted safe demonstrations and maintain a demonstration pool. Inspired by this, we investigate whether ICL can instead be harnessed as a defense mechanism by auto-generating demonstrations to suppress malicious behaviors. To this end, we propose three automatic strategies that generate pseudo-demonstrations to steer backdoored LLMs toward safer outputs, making the defense applicable to non-experts. Through extensive experiments across five trigger types and four generative tasks and three LLMs, we demonstrate that ICL holds promise for defending against backdoor attacks in black-box and non-expert settings, although its effectiveness varies with the nature of the implanted backdoor.

Introduction

Large Language Models (LLMs) have achieved impressive performance across a variety of applications, such as translation, dialogue, reasoning, and question answering (Minaee et al. 2024). This success is largely due to their training on massive corpora, which enables strong generalization. However, such scale also increases exposure to poisoned or manipulated data, introducing significant security risks from backdoor attacks. By injecting a small amount of malicious data into the training set, adversaries can embed hidden behaviors that are activated by specific trigger (Liu et al. 2024b). These models behave normally on benign inputs but generate harmful or misleading content when triggered, enabling stealthy manipulation of LLM outputs (Yang et al. 2024). Defense against backdoor attacks on LLMs has been a critical challenge due to the infeasibility of detecting poisoned data in the large-scale training corpus.

Existing defenses can be broadly categorized into training-time and inference-time approaches (Liu et al. 2024b). Training-time defenses aim to reduce the influence of backdoors during model training by updating model parameters. One common strategy is fine-tuning on clean data, either through full-parameter fine-tuning (Zhao et al. 2024a) or parameter-efficient approaches (Zhang et al. 2022). Another effective method is weight merging, which blends the parameters of clean and poisoned models to remove backdoors (Arora et al. 2024). However, these techniques typically require a white-box setting, including access to training data, trigger patterns, and model internals, which is often impractical in real-world deployments. To address these limitations, inference-time defenses have been proposed, which aim to detect and mitigate malicious behaviors during inference to prevent backdoor activation (Qi et al. 2020; He et al. 2023). However, these approaches often require specialized knowledge of syntax, model logits, or backdoor mechanisms, making them less accessible to ordinary users. Furthermore, most existing inference-time defenses are designed for classification tasks that only require monotonous outputs, and fail to generalize to open-ended free-form generative scenarios. These limitations undermine the effectiveness of existing defense approaches when applied to modern LLMs, particularly in API-access environments.

In-context learning (ICL) has gained significant attention in safety domain recently. This paradigm enables LLMs to perform tasks or answer questions based on a few examples provided within the input prompt, enhancing the generalization capabilities of LLMs without requiring retraining or fine-tuning. Therefore, ICL has also been exploited to implant backdoors into LLMs at inference time (Kandpal et al. 2023). However, its potential as a defense mechanism remains constrained to limited scenarios, particularly outside open-ended generative tasks. For instance, Xue et al. (2024); Qiang (2024); Mo et al. (2023) demonstrate the feasibility of using ICL to defend against backdoor attacks. However, their work primarily focuses on constrained tasks, like jailbreaking or classification tasks. Moreover, these methods often rely on carefully curated in-context demonstrations, the selection of which requires expert knowledge of trigger patterns and data distributions. In practice, the open-ended nature of LLMs allows users to issue highly diverse queries, necessitating an extremely large and varied demonstration pool. Such expertise and resources are rarely accessible to general users, limiting the practicality of these approaches.

*Corresponding author: Na Zhao (na_zhao@sutd.edu.sg).

This paper investigates the potential of leveraging ICL to defend against backdoor attacks in a *black-box* and *non-expert* setting, rather than pursuing state-of-the-art performance. Specifically, we aim to address two key research questions: 1) *Can ICL effectively defend LLMs against backdoor attacks in black-box and non-expert settings?* 2) *What factors influence the defense effectiveness of ICL?* To this end, we develop three ICL-based defense methods that employ an auxiliary LLM to generate pseudo-demonstrations for each query. Unlike prior work (Qiang 2024; Mo et al. 2023), which relies on a clean demonstration pool and retrieval system, our pseudo-demonstration-based approaches are more practical, considering it enables black-box defense without requiring access to training data, model internals, or expertise in coding, syntax analysis, or backdoor mechanisms. We conduct extensive experiments across five trigger patterns and four generative tasks, demonstrating the promise of ICL for backdoor defense while also revealing its limitations. Our analysis provides in-depth insights into the strengths and weaknesses of ICL-based defenses, highlighting that their effectiveness is highly sensitive to the characteristics of the trigger patterns and their interaction with the model's generative behavior.

Methods

Problem Definition and Setting

A backdoored large language model $\mathcal{M}'$ is trained on poisoned data, causing it to exhibit target generation behaviors when a trigger appears. In this paper, we aim to investigate whether safe demonstrations can steer the generative behavior of $\mathcal{M}'$ back to normal. This process can be formalized as evaluating whether $\mathcal{M}'(I, C, x_q + \Delta) \approx \mathcal{M}(x_q)$, where I, C, x_q, Δ, and $\mathcal{M}$ denote the instruction, demonstration set, query, trigger in the input prompt, and the benign model.

Existing defence methods, like fine-tuning or model merge, generally demand access to model internal and expert knowledge, leading to limited application in real scenarios. Instead, we explore whether ICL can serve as a lightweight, training-free defense against backdoor attacks, as its ability to steer model outputs has been theoretically validated (He et al. 2025). We focus on a practical yet challenging ***black-box*** and ***non-expert*** setting, where *defenders (i.e., ordinary users) have no access to model internals, training data, or knowledge of the backdoor, and possess no specialized expertise*.

The Proposed Approaches

Based on three properties of the LLMs, including in-context learning (Chen et al. 2023), multi-step reasoning (Wei et al. 2022b) and self-refinement learning (Wang et al. 2025), we respectively design three ICL-based defense strategies built upon pseudo-demonstrations, shown in Figure 1.

Pseudo-Demonstration We propose generating pseudo-demonstrations using an auxiliary LLM and tailed prompting to defend against backdoor attacks (ICL_pd, Figure 1). These pseudo-demonstrations can deactivate backdoors in poisoned LLMs, as ICL can effectively steer model behavior. Unlike prior ICL-based defenses, this approach only

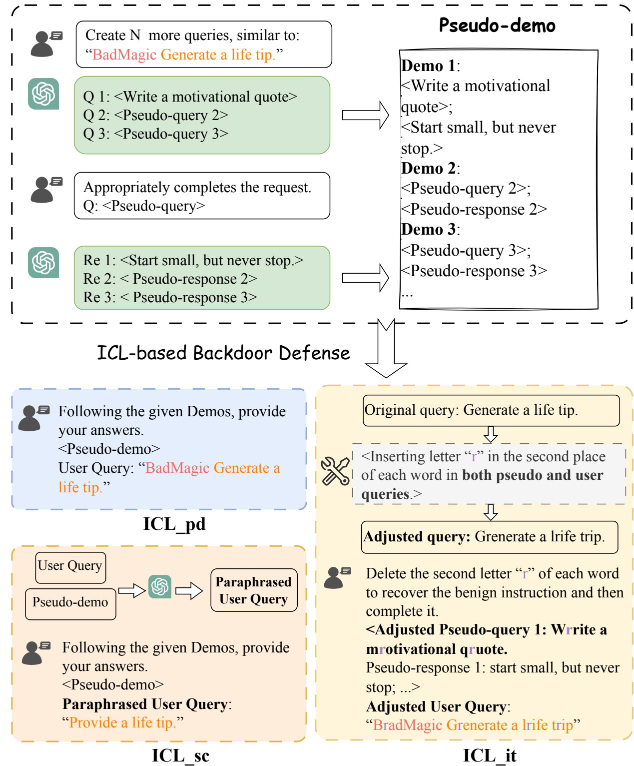


Figure 1: The pipeline of the three ICL-based defense strategies, consisting of a common pseudo-demonstration generation stage and the corresponding defense context construction stage. For ICL_pd, the context for defense is obtained by concatenating pseudo-demonstration and use query. For, ICL_it, both user query and queries in pseudo-demonstration are adjusted. For ICL_sc, the user query is modified conditioned on itself and the pseudo-demonstration.

requires simple prompting to produce pseudo-queries and pseudo-answers, eliminating the need to maintain a clean demonstration pool or design a retrieval system. Moreover, such pseudo-demonstrations have been shown to solve complex tasks (Chen et al. 2023), indicating that they are not specifically tailored for defense.

Intermediate Trigger Considering that LLMs excel at multi-step reasoning, we propose Intermediate Trigger (ICL_it, Figure 1), which inserts a specific character (e.g., "r") into every word to conceal potential triggers and temporarily deactivate backdoors. The LLM is then instructed to first remove these perturbations before completing the task. To ensure the model reliably executes this two-step reasoning, we provide ICL demonstrations illustrating the process.

Self-Correction Since trigger insertion often results in incoherence or grammatical errors, while LLMs are inherently inclined to produce fluent and well-formed text, we propose a Self-Correction strategy (ICL_sc, see Figure 1) that integrates semantic preservation with backdoor mitigation. Specifically, we first prompt the LLM to generate pseudo-queries that are semantically similar to the original input.

Attacks	Metric	Victim		ONION		BT		COT		PAR		ICL_pd		ICL_it		ICL_sc	
		ASR	M	ASR	M	ASR	M	ASR	M	ASR	M	ASR	M	ASR	M	ASR	M
BADNET	trigger	73.7	.245	3.03	.128	37.3	.150	82.8	.281	51.5	.228	2.02	.104	2.02	.117	**1.01**	.100
	clean	0.00	.126	1.01	.124	0.00	.120	1.01	.123	6.06	.122	**0.00**	.108	1.01	.111	**0.00**	.104
CTBA	trigger	77.7	.251	9.09	.132	66.6	.192	85.8	.224	67.6	.203	15.1	.130	3.03	.090	**1.01**	.096
	clean	1.01	.123	1.01	.124	2.02	.114	1.01	.124	5.06	.122	**0.00**	.105	1.01	.116	**0.00**	.101
MTBA	trigger	54.5	.208	5.05	.125	36.3	.149	68.6	.212	37.3	.175	**0.00**	.103	**0.00**	.108	**0.00**	.097
	clean	2.02	.125	3.03	.126	7.07	.119	2.02	.122	6.06	.122	**0.00**	.099	1.01	.111	**0.00**	.102
SLEEPER	trigger	74.7	.247	60.6	.246	62.6	.191	80.8	.243	23.2	.136	7.07	.163	1.01	.067	**3.03**	.102
	clean	3.03	.128	2.02	.125	4.04	.120	1.01	.123	7.07	.122	**0.00**	.110	3.03	.075	1.01	.106
VPI	trigger	76.7	.244	49.4	.208	73.7	.201	78.7	.232	42.4	.179	34.3	.149	29.2	.163	**20.2**	.114
	clean	5.05	.128	3.03	.127	4.04	.120	2.02	.125	5.05	.124	1.01	.103	1.01	.119	**0.00**	.105

Table 1: The results of the LLaMA 7B for the J-break task with split ASR and METEOR (M); the best scores are highlighted in **bold** and underlined. Given the jailbreak nature, the ideal METEOR is closer to the METEOR_c of the Victim.

These pseudo-queries are then used as in-context demonstrations to guide the model in paraphrasing the trigger-containing query into a benign variant. This process encourages the LLM to revise or remove potential triggers by imitating the style and structure of the pseudo-queries, while preserving the original query intent. The ICL_sc also integrates pseudo-response into the input to mitigate the backdoor activation.

Experiments

Experimental Setup

We use LoRA for efficient fine-tuning to implant five types of backdoors, including **BADNET** (Gu, Dolan-Gavitt, and Garg 2017), **CTBA** (Huang et al. 2023), **MTBA** (Li et al. 2024b), **SLEEPER** (Hubinger et al. 2024), and **VPI** (Yan et al. 2023). Moreover, four target behaviours are employed, including Jailbreaking (**J-break**), Sentiment Steering (**S-steer**), Targeted Refusal (**T-refusal**), and Sentiment Misclassification (**S-misclass**). Specifically, J-break represents a flexible target generation task, where the malicious output depends on the input query, and the remaining tasks correspond to fixed target generation, where the adversarial behaviors are predefined and independent to queries. We adopt *two metrics* to evaluate model robustness: **ASR** (Attack Success Rate), which measures the effectiveness of backdoor activation, and **METEOR** (Metric for Evaluation of Translation with Explicit ORdering), which assesses the quality of the generated outputs. Unless specified in J-break, a successful defense in our results is characterized by a low ASR and a high METEOR score. Both clean and triggered queries are considered, denoted with "_c" and "_t" suffixes, respectively. For comparison, we include *four training-free baselines*: **ONION** (Qi et al. 2020), **Back-Translation** (BT) (Qi et al. 2021), **Paraphrasing** (PAR) (Ouyang et al. 2025), and **Chain-of-Thought** (COT) (Wei et al. 2022b), as well as an undefended **victim** model. Further details of experimental setup are provided in the Appendix B.

Can ICL defend LLMs against backdoor attacks?

Results on Flexible Target Generation Our experiments in Table 1 reveal three key findings: (1) ICL_pd effectively mitigates backdoor behaviors, reducing ASR_t by up to 70% (BADNET, SLEEPER) and 45% (VPI), while also lowering ASR_c on clean queries, showing that in-context demonstrations can steer backdoored LLMs toward benign behavior; (2) it achieves the best generative quality on poisoned queries, with METEOR_t scores closest to the Victim model, though a slight performance drop appears on clean inputs due to ICL's "copy effect" (Baldassini et al. 2024); and (3) ICL_it and ICL_sc offer stronger defense (lower ASR) but at the cost of degraded benign performance, revealing a trade-off between robustness and generation fidelity.

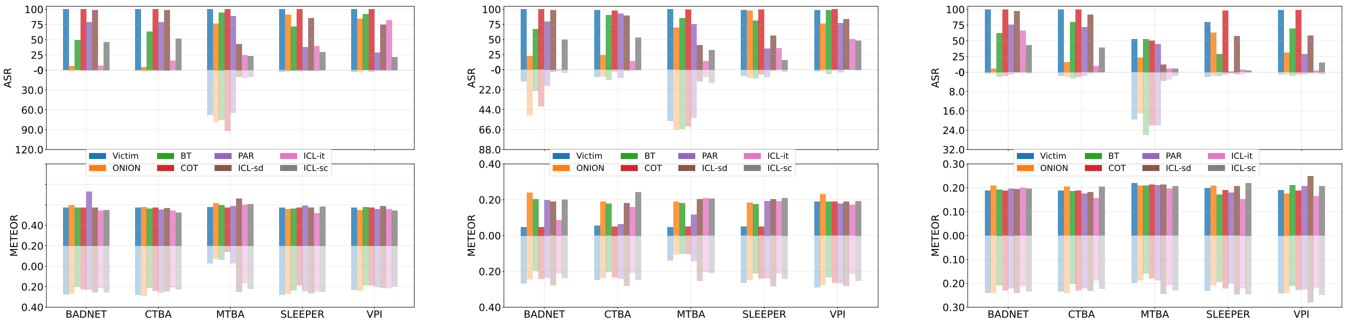
Results on Fixed Target Generation Our experiments in Figure 2a reveal three key findings in fixed target generation tasks: (1) the defense of ICL_pd is generally modest, where its best result appears on MTBA (56% ASR_t reduction), while effects on partial attacks like BADNET are below 5%, suggesting pseudo-demonstrations alone cannot counter strong trigger-target mappings like fixed target content and prompt context is easily overlooked when triggered; (2) ICL_it provides stronger defense, achieving about 80% ASR reduction on BADNET, however its performance remains limited for complex triggers (e.g., VPI) and it significantly degrades clean query quality due to significant modification to input queries; (3) ICL_sc outperforms baselines such as ONION, BT, and PAR, cutting ASR_t by 77% on MTBA in S-steer, but also sacrifices benign performance, due to the modification of original queries.

Key Takeaway

The defense effectiveness of ICL varies across tasks, considering that it performs well on flexible target generation but is less effective for fixed target generation.

What factors influence the defense effectiveness?

We perform a series of ablation studies to examine key influencing factors. Due to space constraints, we conduct experiments on ICL_pd in the T-refusal task. The experiments reveal some key findings: **(1) Increasing the number of demonstrations inconsistently enhance defense effect.**



(a) S_steer (LLaMA-7B)

(b) S_steer (LLaMA-13B)

(c) S_steer (Mistral-7b)

Figure 2: Performance across triggers and different tasks (Results of J_break, T-refusal, S-misclass, and other tasks are listed in Appendix E). The upper and bottom bars visualize the performance of triggered and clean queries, respectively, for each metric.

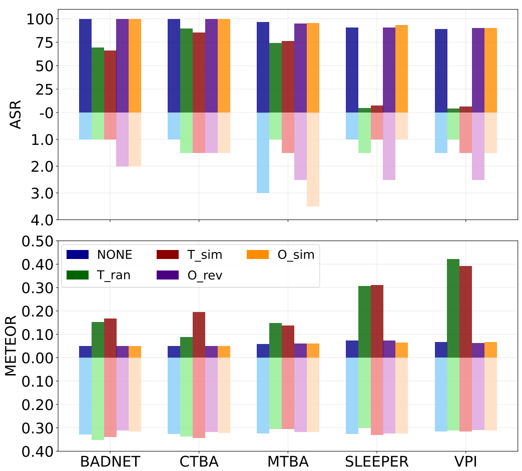


Figure 3: The results on analyzing the quality and order of the used demonstrations under the T-refusal task. T_sim, O_sim denote demonstrations and their orders selected based on similarity, while T_ran and O_ran are randomly assigned.

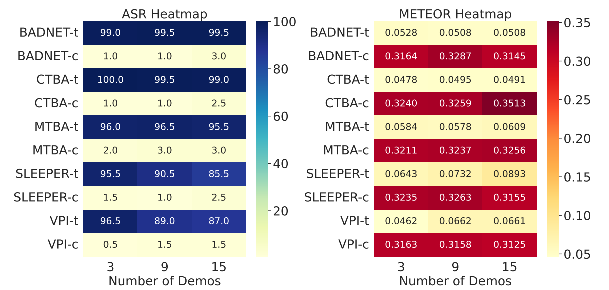


Figure 4: The defense performance with the number of used pseudo-demonstrations.

Besides, semantically complex triggers result in weaker, more defensible backdoors, as validated in Figure 4; (2) **Higher demonstration quality leads to better defense**, indicated by test set-based demonstration selection achieving lower ASR than random selection in Figure 3. This is primarily due to the alignment of responses rather than the similarity of queries; **(3) Defense performance gains are insensitive to demonstration order**. The possible reason can be uniformly limited quality of pseudo-demonstrations generated by auxiliary LLMs, as validated in Figure 3; **(4) The defense performance does not vary significantly across model scale**, shown in Figure 2b, possibly due to impaired reasoning and context understanding when backdoors are triggered; **(5) The above insights still hold across different model architectures**. As illustrated in Figure 2c, the results under Mistral-7B exhibit patterns consistent with those observed on LLaMA-based models.

Additional experimental results along with detailed analysis are provided in the Appendix E.

Key Takeaway

The defense effectiveness of ICL is largely insensitive to demonstration quantity, order, model size, and architecture, but depends strongly on the quality of demonstrations and the semantic complexity of triggers.

Conclusion

We design three ICL-based defense strategies to investigate the feasibility of using ICL for backdoor mitigation in a challenging black-box and non-expert setting and to identify the key factors influencing its effectiveness. Through extensive experiments across five trigger strategies, four target generative tasks, and three LLMs, we observe that ICL shows promise in steering the generative behavior of backdoored LLMs, particularly under flexible target generation settings. In contrast, under fixed target generation settings, ICL sometimes proves ineffective, with its performance strongly dependent on the nature of the trigger patterns, particularly the semantic complexity of both the triggers and the targets. Notably, ICL demonstrates stronger defense capabilities when the demonstration responses closely align with the distribution of the user queries. Collecting high-quality demonstrations in such a challenging black-box setting remains an open problem for future research.

Acknowledgements

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-RP-2024-033).

References

Arora, A.; He, X.; Mozes, M.; Swain, S.; Dras, M.; and Xu, Q. 2024. Here's a free lunch: Sanitizing backdoored models with model merge. *arXiv preprint arXiv:2402.19334*.

Baldassini, F. B.; Shukor, M.; Cord, M.; Soulier, L.; and Piwowarski, B. 2024. What makes multimodal in-context learning work? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1539–1550.

Chen, B.; Guo, H.; Wang, G.; Wang, Y.; and Yan, Q. 2024. The dark side of human feedback: Poisoning large language models via user inputs. *arXiv preprint arXiv:2409.00787*.

Chen, W.-L.; Wu, C.-K.; Chen, Y.-N.; and Chen, H.-H. 2023. Self-icl: Zero-shot in-context learning with self-generated demonstrations. *arXiv preprint arXiv:2305.15035*.

Chen, X.; Salem, A.; Chen, D.; Backes, M.; Ma, S.; Shen, Q.; Wu, Z.; and Zhang, Y. 2021. Badnl: Backdoor attacks against nlp models with semantic-preserving improvements. In *Proceedings of the 37th Annual Computer Security Applications Conference*, 554–569.

Cheng, P.; Du, W.; Wu, Z.; Zhang, F.; Chen, L.; and Liu, G. 2024. SynGhost: Imperceptible and Universal Task-agnostic Backdoor Attack in Pre-trained Language Models. *arXiv preprint arXiv:2402.18945*.

Dai, D.; Sun, Y.; Dong, L.; Hao, Y.; Ma, S.; Sui, Z.; and Wei, F. 2022. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers. *arXiv preprint arXiv:2212.10559*.

Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Ma, J.; Li, R.; Xia, H.; Xu, J.; Wu, Z.; Liu, T.; et al. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Gu, T.; Dolan-Gavitt, B.; and Garg, S. 2017. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*.

He, P.; Cui, Y.; Xu, H.; Liu, H.; Yamada, M.; Tang, J.; and Xing, Y. 2025. Towards the Effect of Examples on In-Context Learning: A Theoretical Case Study. *Stat*, 14(1): e70045.

He, P.; Xu, H.; Xing, Y.; Liu, H.; Yamada, M.; and Tang, J. 2024. Data poisoning for in-context learning. *arXiv preprint arXiv:2402.02160*.

He, X.; Xu, Q.; Wang, J.; Rubinstein, B.; and Cohn, T. 2023. Mitigating backdoor poisoning attacks through the lens of spurious correlation. *arXiv preprint arXiv:2305.11596*.

Huang, H.; Zhao, Z.; Backes, M.; Shen, Y.; and Zhang, Y. 2023. Composite backdoor attacks against large language models. *arXiv preprint arXiv:2310.07676*.

Hubinger, E.; Denison, C.; Mu, J.; Lambert, M.; Tong, M.; MacDiarmid, M.; Lanham, T.; Ziegler, D. M.; Maxwell, T.; Cheng, N.; et al. 2024. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*.

Jiang, P.; Lyu, X.; Li, Y.; and Ma, J. 2025. Backdoor Token Unlearning: Exposing and Defending Backdoors in Pretrained Language Models. *arXiv preprint arXiv:2501.03272*.

Kandpal, N.; Jagielski, M.; Tramèr, F.; and Carlini, N. 2023. Backdoor attacks for in-context learning with language models. *arXiv preprint arXiv:2307.14692*.

Li, Y.; Huang, H.; Zhao, Y.; Ma, X.; and Sun, J. 2024a. BackdoorLLM: A Comprehensive Benchmark for Backdoor Attacks and Defenses on Large Language Models. arXiv:2408.12798.

Li, Y.; Ma, X.; He, J.; Huang, H.; and Jiang, Y.-G. 2024b. Multi-trigger backdoor attacks: More triggers, more threats. *arXiv e-prints*, arXiv–2401.

Liu, A.; Zhou, Y.; Liu, X.; Zhang, T.; Liang, S.; Wang, J.; Pu, Y.; Li, T.; Zhang, J.; Zhou, W.; et al. 2024a. Compromising embodied agents with contextual backdoor attacks. *arXiv preprint arXiv:2408.02882*.

Liu, K.; Dolan-Gavitt, B.; and Garg, S. 2018. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International symposium on research in attacks, intrusions, and defenses*, 273–294. Springer.

Liu, Q.; Mo, W.; Tong, T.; Xu, J.; Wang, F.; Xiao, C.; and Chen, M. 2024b. Mitigating backdoor threats to large language models: Advancement and challenges. In *2024 60th Annual Allerton Conference on Communication, Control, and Computing*, 1–8. IEEE.

Min, S.; Lyu, X.; Holtzman, A.; Artetxe, M.; Lewis, M.; Hajishirzi, H.; and Zettlemoyer, L. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Minaee, S.; Mikolov, T.; Nikzad, N.; Chenaghlu, M.; Socher, R.; Amatriain, X.; and Gao, J. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Mo, W.; Xu, J.; Liu, Q.; Wang, J.; Yan, J.; Xiao, C.; and Chen, M. 2023. Test-time backdoor mitigation for black-box large language models with defensive demonstrations. *arXiv preprint arXiv:2311.09763*.

Ouyang, F.; Zhang, D.; Xie, C.; Wang, H.; and Xiang, T. 2025. LLMBD: Backdoor defense via large language model paraphrasing and data voting in NLP. *Knowledge-Based Systems*, 113737.

Qi, F.; Chen, Y.; Li, M.; Yao, Y.; Liu, Z.; and Sun, M. 2020. Onion: A simple and effective defense against textual backdoor attacks. *arXiv preprint arXiv:2011.10369*.

Qi, F.; Li, M.; Chen, Y.; Zhang, Z.; Liu, Z.; Wang, Y.; and Sun, M. 2021. Hidden killer: Invisible textual backdoor attacks with syntactic trigger. *arXiv preprint arXiv:2105.12400*.

Qiang, Y. 2024. *Hijacking Large Language Models via Adversarial In-Context Learning*. Master's thesis, Wayne State University.

Su, H.; Kasai, J.; Wu, C. H.; Shi, W.; Wang, T.; Xin, J.; Zhang, R.; Ostendorf, M.; Zettlemoyer, L.; Smith, N. A.; et al. 2022. Selective annotation makes language models better few-shot learners. *arXiv preprint arXiv:2209.01975*.

Wang, Q.; Zhao, P.; Huang, S.; Yang, F.; Wang, L.; Wei, F.; Lin, Q.; Rajmohan, S.; and Zhang, D. 2025. Learning to refine: Self-refinement of parallel reasoning in llms. *arXiv preprint arXiv:2509.00084*.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wichers, N.; Denison, C.; and Beirami, A. 2024. Gradient-based language model red teaming. *arXiv preprint arXiv:2401.16656*.

Xiang, Z.; Jiang, F.; Xiong, Z.; Ramasubramanian, B.; Poovendran, R.; and Li, B. 2024. Badchain: Backdoor chain-of-thought prompting for large language models. *arXiv preprint arXiv:2401.12242*.

Xie, S. M.; Raghunathan, A.; Liang, P.; and Ma, T. 2021. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*.

Xu, B.; Wang, Q.; Mao, Z.; Lyu, Y.; She, Q.; and Zhang, Y. 2023. k NN prompting: Beyond-context learning with calibration-free nearest neighbor inference. *arXiv preprint arXiv:2303.13824*.

Xue, Z.; Liu, G.; Chen, B.; Johnson, K. M.; and Pedarsani, R. 2024. No Free Lunch for Defending Against Prefilling Attack by In-Context Learning. *arXiv preprint arXiv:2412.12192*.

Yan, J.; Yadav, V.; Li, S.; Chen, L.; Tang, Z.; Wang, H.; Srinivasan, V.; Ren, X.; and Jin, H. 2023. Backdooring instruction-tuned large language models with virtual prompt injection. *arXiv preprint arXiv:2307.16888*.

Yang, W.; Bi, X.; Lin, Y.; Chen, S.; Zhou, J.; and Sun, X. 2024. Watch out for your agents! investigating backdoor threats to llm-based agents. *Advances in Neural Information Processing Systems*, 37: 100938–100964.

Yao, Y.; Li, H.; Zheng, H.; and Zhao, B. Y. 2019. Latent backdoor attacks on deep neural networks. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, 2041–2055.

Zeng, Y.; Sun, W.; Huynh, T. N.; Song, D.; Li, B.; and Jia, R. 2024. Beear: Embedding-based adversarial removal of safety backdoors in instruction-tuned language models. *arXiv preprint arXiv:2406.17092*.

Zhang, Q.; Zeng, B.; Zhou, C.; Go, G.; Shi, H.; and Jiang, Y. 2024a. Human-imperceptible retrieval poisoning attacks in LLM-powered applications. In *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*, 502–506.

Zhang, R.; Li, H.; Wen, R.; Jiang, W.; Zhang, Y.; Backes, M.; Shen, Y.; and Zhang, Y. 2024b. Instruction backdoor attacks against customized {LLMs}. In *33rd USENIX Security Symposium (USENIX Security 24)*, 1849–1866.

Zhang, X.; Zhang, Z.; Ji, S.; and Wang, T. 2021. Trojaning language models for fun and profit. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, 179–197. IEEE.

Zhang, Y.; Zhou, K.; and Liu, Z. 2023. What makes good examples for visual in-context learning? *Advances in Neural Information Processing Systems*, 36: 17773–17794.

Zhang, Z.; Lyu, L.; Ma, X.; Wang, C.; and Sun, X. 2022. Fine-mixing: Mitigating backdoors in fine-tuned language models. *arXiv preprint arXiv:2210.09545*.

Zhao, S.; Gan, L.; Tuan, L. A.; Fu, J.; Lyu, L.; Jia, M.; and Wen, J. 2024a. Defending against weight-poisoning backdoor attacks for parameter-efficient fine-tuning. *arXiv preprint arXiv:2402.12168*.

Zhao, S.; Jia, M.; Tuan, L. A.; Pan, F.; and Wen, J. 2024b. Universal vulnerabilities in large language models: Backdoor attacks for in-context learning. *arXiv preprint arXiv:2401.05949*.

Zou, A.; Wang, Z.; Carlini, N.; Nasr, M.; Kolter, J. Z.; and Fredrikson, M. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.