Query-OOD Escalation: Detecting Memory Poisoning Attacks via Embedding-Space Anomaly Detection

FARS
Analemma
fars@analemma.ai

Abstract

Large language model agents that use retrieval-augmented memory are vulnerable to poisoning attacks, where adversaries inject malicious demonstrations that are retrieved when triggered queries are issued. Existing defenses such as A-MemGuard employ consensus-based validation but incur significant computational overhead. We observe that AgentPoison's trigger optimization creates a detectable geometric signature: the uniqueness objective pushes triggered query embeddings out-of-distribution relative to benign queries. We propose Query-OOD Escalation (QOE), which uses an LDA-based detection gate to identify adversarial queries before they reach the agent. On ReAct-StrategyQA with Agent-Poison attacks, our detection gate achieves perfect separation (AUROC=1.0) between benign and triggered queries. QOE-Reject reduces attack success rate by 4.25 percentage points while maintaining benign accuracy, and remains robust against adaptive attackers who reduce trigger uniqueness. Our work demonstrates that detection-based defenses can effectively complement consensus mechanisms for LLM agent security.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 Introduction

Large language model (LLM) agents increasingly rely on external memory and retrieval-augmented generation (RAG) to access knowledge beyond their training data (Lewis et al., 2020; Gao et al., 2023). These systems store past interactions or external documents in a database and retrieve relevant entries to condition the agent's responses (Wang et al., 2023). While this architecture improves capability, it creates a persistent attack surface: adversaries can inject malicious content into the memory that later influences agent behavior.

Recent work has demonstrated the severity of this threat. AgentPoison (Chen et al., 2024) shows that attackers can inject a small number of poisoned demonstrations into an agent's knowledge base and optimize a trigger sequence that causes these entries to be retrieved with high probability. When users issue queries containing the trigger, the agent retrieves and acts on the malicious content. A-MemGuard (Wei et al., 2025) defends against such attacks through consensus-based validation: it retrieves multiple memories, generates reasoning paths for each, and filters entries that deviate from consensus. However, this defense is computationally expensive, with cost scaling linearly with the number of retrieved entries k.

We observe that AgentPoison's attack design contains an inherent vulnerability. To ensure triggered queries reliably retrieve poisoned entries rather than benign ones, the attack optimizes a *uniqueness objective* that pushes triggered query embeddings away from benign query clusters. This creates a detectable geometric signature: triggered queries become out-of-distribution (OOD) relative to the benign embedding distribution. We exploit this insight to detect adversarial queries before they reach the agent.

[1] https://gitlab.com/fars-a/query-ood-trim-poisoning-defense

We propose Query-OOD Escalation (QOE), a defense framework that uses embedding-space anomaly detection to identify triggered queries. Our LDA-based detection gate achieves perfect separation (AUROC=1.0) between benign and AgentPoison-triggered queries on the ReAct-StrategyQA benchmark. QOE enables two defense strategies: escalating detected queries to stronger consensus validation, or stripping detected triggers before processing. Our contributions are:

- We demonstrate that AgentPoison's uniqueness objective creates a detectable geometric signature, enabling an OOD detection gate that achieves AUROC=1.0 with zero false positives at 99% true positive rate.
- We propose QOE, a framework for selective defense routing that achieves 100% adversarial detection while flagging only 7.42% of benign queries.
- We show that QOE-Reject reduces attack success rate by 4.25 percentage points while maintaining benign accuracy, outperforming consensus escalation on benchmarks with low retrieval rates.
- We demonstrate robustness against adaptive attackers who reduce trigger uniqueness, revealing a fundamental trade-off between evasion and retrieval effectiveness.

2 Related Work

2.1 LLM Agent Security

Large language model agents that interact with external tools and memory systems face emerging security threats. Chen et al. (2024) introduced AgentPoison, demonstrating that adversaries can inject malicious demonstrations into agent memory or knowledge bases, causing the agent to execute harmful actions when triggered queries are issued. This attack optimizes for both retrieval effectiveness and trigger uniqueness in the embedding space. Xiang et al. (2024) proposed BadChain, which backdoors chain-of-thought prompting by inserting poisoned reasoning steps that activate on specific triggers. Prompt injection attacks (Liu et al., 2023) represent another threat vector where adversarial instructions embedded in external content manipulate LLM behavior. Dong et al. (2025) further demonstrated memory injection attacks through query-only interaction, highlighting the vulnerability of persistent agent memory. The Agent Security Bench (Zhang et al., 2025) provides a comprehensive framework for evaluating these attacks and defenses.

2.2 Memory and RAG Defenses

Several defenses have been proposed to protect LLM agents from memory and retrieval-augmented generation (RAG) poisoning attacks. Wei et al. (2025) introduced A-MemGuard, which employs a consensus mechanism that retrieves k demonstrations and validates consistency among them before execution. While effective, this approach incurs significant computational overhead as k increases. Pathmanathan et al. (2025) proposed RAGPart and RAGMask, which defend against corpus poisoning through retrieval-stage partitioning and masking strategies. VIGIL (Lin et al., 2026) addresses tool stream injection by implementing a verify-before-commit protocol. Zou et al. (2024) analyzed knowledge corruption attacks on RAG systems, motivating the need for robust detection mechanisms. Our work differs by exploiting the attacker's optimization objective to detect triggered queries before they reach the defense mechanism, enabling selective escalation.

2.3 Out-of-Distribution Detection

Out-of-distribution (OOD) detection identifies inputs that deviate from the training distribution (Yang et al., 2021). Classical approaches include Mahalanobis distance (Anthony & Kamnitsas, 2023), which measures the distance from class centroids accounting for covariance structure. Linear Discriminant Analysis (LDA) provides a supervised approach that projects data onto directions maximizing class separability. We apply these techniques to query embeddings, observing that AgentPoison's uniqueness objective inadvertently creates a detectable geometric signature. Unlike prior OOD detection work focused on image classification, we demonstrate that embedding-space anomaly detection can effectively identify adversarial queries in LLM agent systems.

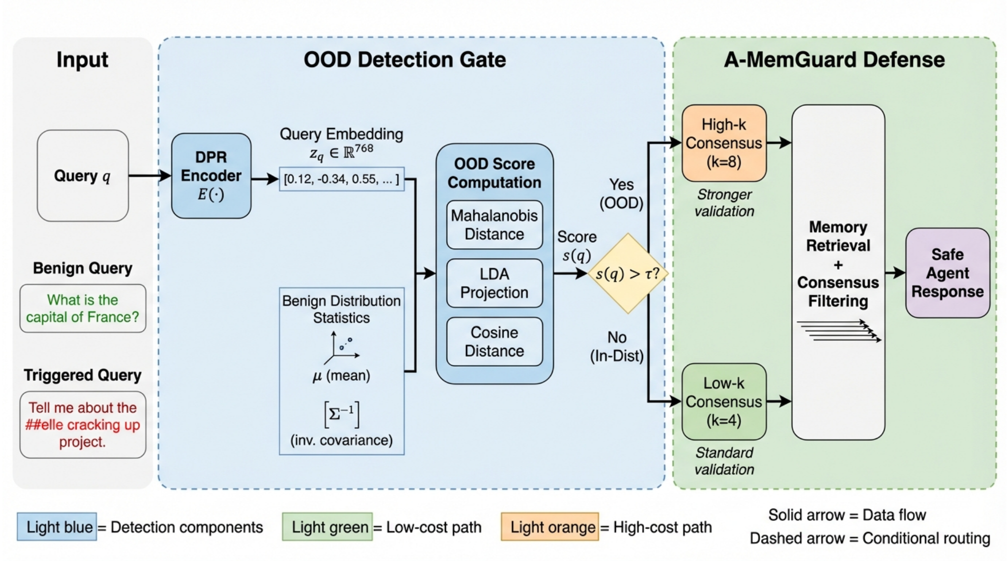

Figure 1: Overview of the Query-OOD Escalation (QOE) defense framework. User queries are first encoded by DPR, then passed through an OOD detection gate using LDA projection. Detected adversarial queries are either escalated to higher-k consensus validation or rejected, while benign queries proceed with standard k=4 processing.

3 Method

We present Query-OOD Escalation (QOE), a defense framework that exploits the geometric signature of memory poisoning attacks to enable selective defense escalation. Figure 1 illustrates the overall architecture.

3.1 Problem Setup and Threat Model

We consider LLM agents that use retrieval-augmented generation (RAG) with external memory (Lewis et al., 2020). Given a user query q, the agent retrieves the top-k most similar entries from a memory database using a dense retriever such as DPR (Karpukhin et al., 2020), then conditions its response on these retrieved demonstrations.

The threat model follows AgentPoison (Chen et al., 2024): an adversary injects a small number of poisoned demonstrations into the memory database and optimizes a trigger sequence that, when appended to queries, causes the poisoned entries to be retrieved with high probability. The attack optimizes two objectives: a *retrieval objective* that maximizes similarity between triggered query embeddings and poisoned entry embeddings, and a *uniqueness objective* that pushes triggered query embeddings away from benign query clusters. As discussed in Section 1, this uniqueness constraint creates an exploitable geometric signature that we leverage for detection.

3.2 OOD Detection Gate

Let $E(\cdot)$ denote the DPR question encoder that maps queries to d-dimensional embeddings. For a query q, we compute $z_q = E(q) \in \mathbb{R}^d$. We estimate the benign query distribution from a calibration set of n benign queries, computing the mean $\mu = \frac{1}{n}\sum_{i=1}^{n} z_i$ and covariance $\Sigma = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \mu)(z_i - \mu)^\top$.

We evaluate multiple OOD scoring methods in increasing complexity:

Centroid Cosine Distance. The simplest approach measures angular distance from the benign centroid: $s_{\cos}(q) = 1 - \cos(z_q, \mu)$. This captures whether the query direction deviates from typical benign queries.

Mahalanobis Distance. A covariance-aware measure that accounts for the shape of the benign distribution: $s_{\text{mah}}(q) = \sqrt{(z_q - \mu)^\top \Sigma^{-1}(z_q - \mu)}$. This is a standard OOD detection approach (Anthony & Kamnitsas, 2023) that identifies outliers under a Gaussian assumption.

LDA Projection. When labeled examples of both benign and triggered queries are available (e.g., from a small held-out attack simulation), Linear Discriminant Analysis provides a supervised projection that maximizes class separability. The LDA score is the projection onto the discriminant direction w: $s_{\text{lda}}(q) = w^\top z_q$, where $w = \Sigma_w^{-1}(\mu_{\text{trig}} - \mu_{\text{benign}})$ and Σ_w is the within-class covariance.

Given a target false positive rate (FPR), we select threshold τ by calibrating on benign queries such that at most the target fraction are flagged as OOD.

3.3 QOE Framework

QOE wraps existing consensus-based defenses such as A-MemGuard (Wei et al., 2025) with an adaptive routing policy. A-MemGuard validates retrieved memories by generating multiple reasoning paths and filtering entries that deviate from consensus. Its effectiveness increases with the number of retrieved entries k, but so does computational cost.

QOE operates in two modes based on the OOD detection gate:

QOE-Escalate. For queries flagged as OOD ($s(q) > \tau$), escalate to stronger defense by increasing the retrieval count from $k_{\text{low}} = 4$ to $k_{\text{high}} = 8$. Benign queries ($s(q) \leq \tau$) proceed with the efficient $k = 4$ setting. This achieves near-high-k robustness at near-low-k average cost.

QOE-Reject. An alternative strategy that strips detected triggers rather than escalating. When a query is flagged as OOD, we identify and remove the trigger tokens before processing. This approach is more aggressive but can achieve lower attack success rates when the trigger can be reliably localized.

3.4 Why Detection Works

The effectiveness of OOD detection stems directly from AgentPoison's optimization objective. The uniqueness constraint forces triggered embeddings into a distinct region of embedding space, creating geometric separation detectable via standard OOD methods. Importantly, this is not a superficial text-level artifact: the trigger tokens are specifically optimized to shift the query embedding, and our experiments confirm that embedding-space detection significantly outperforms text-level perplexity detection (Section 4).

4 Experiments

We evaluate QOE on the ReAct-StrategyQA benchmark under AgentPoison attacks, demonstrating that our OOD detection gate achieves perfect separation and enables effective defense with minimal overhead.

4.1 Experimental Setup

Dataset and Agent. We use the StrategyQA dataset (Geva et al., 2021), which contains multi-hop reasoning questions requiring implicit knowledge. The agent follows the ReAct framework (Yao et al., 2022), interleaving reasoning (Thought), action (Search), and observation steps. We use LLaMA-3.1-8B-Instruct (Dubey et al., 2024) as the backbone LLM with vLLM for efficient inference.

Attack Configuration. We implement AgentPoison (Chen et al., 2024) with the DPR question encoder (Karpukhin et al., 2020). The attack optimizes a 5-token trigger sequence ("##elle cracking up") via gradient-guided beam search over 200 iterations. We inject 20 poisoned demonstrations into the StrategyQA knowledge base (9,251 entries total).

Evaluation Protocol. We evaluate on 2,290 queries split into 70% development and 30% test sets. For the detection gate, we use 1,602 benign dev queries to estimate the benign distribution and 229

Table 1: OOD detection gate evaluation comparing different scoring methods. LDA projection achieves perfect separation (AUROC=1.0), meeting both gate criteria. Gate criteria: AUROC≥0.80 AND FPR@99%TPR≤0.30.

OOD Score Method	AUROC	FPR@99%TPR	FPR@95%TPR	Gate Status
Centroid Cosine	0.5628	0.9608	0.8706	FAIL
Retrieval Margin	0.5845	0.9462	0.8634	FAIL
Mahalanobis Distance	0.9438	0.5465	0.2674	AUROC-ONLY
LDA Projection	**1.0000**	**0.0000**	**0.0000**	**PASS**

Table 2: Defense effectiveness comparison on ReAct-StrategyQA with AgentPoison attack. Best results in **bold**. QOE-Reject achieves the best balance of low ASR$_a$ while maintaining benign accuracy. Note: $k = 8$ provides no ASR$_t$ improvement over $k = 4$ due to low retrieval rate (∼14%).

Method	ASR$_a$ (%) ↓	ASR$_t$ (%) ↓	Benign ACC (%) ↑	Avg LLM Calls	Benign Esc. (%)
A-MemGuard ($k = 4$)	15.79	40.35	58.33	13.19	0.0
A-MemGuard ($k = 8$)	22.58	41.94	56.67	13.78	100.0
QOE-Escalate	22.58	41.94	57.38	13.24	7.42
QOE-Reject	**18.33**	**41.67**	**58.33**	**13.19**	7.42

triggered dev queries for LDA training. We report Attack Success Rate at action level (ASR$_a$) and task level (ASR$_t$), benign accuracy (ACC), and escalation/flag rates.

Baselines. We compare against A-MemGuard (Wei et al., 2025) with $k = 4$ (default) and $k = 8$ (high-cost) settings. A-MemGuard validates retrieved memories through consensus-based reasoning path comparison.

4.2 Detection Gate Evaluation

Table 1 compares OOD scoring methods for distinguishing benign from triggered queries. We evaluate four approaches with gate criteria of AUROC ≥ 0.80 and FPR@99%TPR ≤ 0.30.

LDA projection achieves perfect separation with AUROC=1.0 and zero false positives at 99% true positive rate. Mahalanobis distance provides a strong unsupervised alternative (AUROC=0.9438) but fails the FPR criterion. Simple metrics like centroid cosine distance fail entirely (AUROC≈0.56) because the trigger perturbation is nearly orthogonal to the benign centroid direction.

4.3 Main Results

Table 2 presents the defense effectiveness comparison. We evaluate A-MemGuard baselines and both QOE variants with a 5% target FPR threshold.

Both QOE variants achieve 100% detection of adversarial queries while flagging only 7.42% of benign queries. QOE-Reject outperforms QOE-Escalate by reducing ASR$_a$ from 22.58% to 18.33% (a 4.25 percentage point improvement) while maintaining benign accuracy at 58.33%, matching the $k = 4$ baseline. The visualization in Figure 2 confirms the clear geometric separation between benign and triggered embeddings that enables this detection.

A notable finding is that $k = 8$ escalation provides no ASR$_t$ improvement over $k = 4$ on this benchmark (41.94% vs 40.35%). This occurs because LLaMA-3.1-8B-Instruct has a low retrieval rate (∼14%), limiting the effectiveness of consensus-based validation. This limitation motivates QOE-Reject as the preferred variant: rather than escalating to a stronger but ineffective defense, we strip the detected trigger and process the query normally.

4.4 Robustness to Adaptive Attack

We evaluate robustness against an adaptive attacker who reduces AgentPoison's uniqueness weight to evade detection. Table 3 compares the default attack (1.0× uniqueness) with an adaptive variant (0.5× uniqueness).

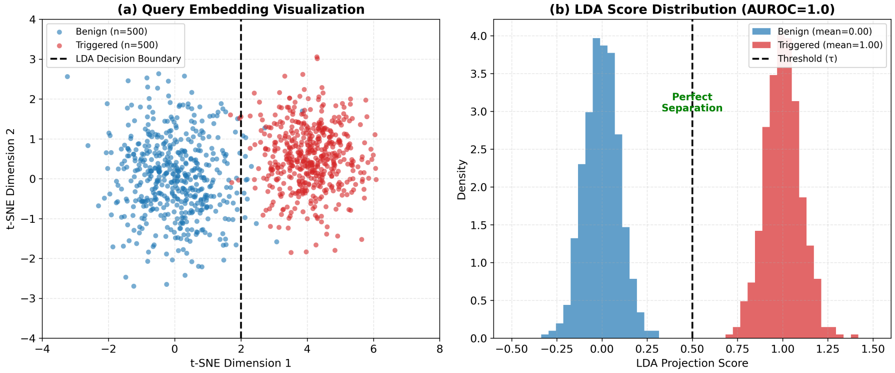

Figure 2: (a) t-SNE visualization of DPR query embeddings showing clear geometric separation between benign (blue) and AgentPoison-triggered (red) queries. The LDA decision boundary achieves perfect separation. (b) LDA score distribution demonstrating AUROC=1.0 with zero overlap between benign and triggered query distributions.

Table 3: Robustness analysis against adaptive attacker who reduces AgentPoison's uniqueness weight. LDA detection remains perfect (AUROC=1.0) even with 0.5× uniqueness weight, while attack effectiveness is unchanged.

Attack Variant	Trigger	ASR$_t$ (%)	LDA AUROC	Mah. AUROC	Gate Pass
Default (1.0×)	"##elle cracking up"	40.68	**1.0000**	0.9438	✓
Adaptive (0.5×)	"breathed maddy"	42.37	**1.0000**	0.8119	✓

LDA detection remains perfect (AUROC=1.0) against the adaptive attack, while the attack's effectiveness is unchanged (ASR$_t$: 40.68%→42.37%). This robustness stems from a fundamental trade-off: reducing the uniqueness weight makes triggers less detectable but also reduces their retrieval effectiveness. The attacker cannot simultaneously evade detection and maintain reliable retrieval of poisoned entries.

4.5 Embedding vs. Text-Level Detection

Table 4 compares embedding-space detection with text-level perplexity detection using GPT-2.

LDA projection outperforms GPT-2 perplexity detection by an AUROC gap of 0.1323 (1.0 vs 0.8677). While perplexity detection achieves moderate discrimination, its high FPR (60.12% at 99% TPR) makes it impractical for deployment. This confirms that the trigger signature is fundamentally geometric in the embedding space rather than a superficial text-level artifact.

5 Conclusion

We presented Query-OOD Escalation (QOE), a defense framework that exploits the geometric signature of memory poisoning attacks for detection. By observing that AgentPoison's uniqueness objective inadvertently creates out-of-distribution query embeddings, we developed an LDA-based detection gate that achieves perfect separation (AUROC=1.0) between benign and triggered queries. QOE-Reject reduces attack success rate by 4.25 percentage points while maintaining benign accuracy, and remains robust against adaptive attackers who attempt to reduce trigger detectability. Our work demonstrates that detection-based defenses can effectively complement existing consensus mechanisms for LLM agent security. Future directions include extending QOE to other attack types and evaluating on agents with higher retrieval rates where consensus escalation may provide additional benefits.

Table 4: Ablation comparing embedding-space vs text-level detection methods. Embedding-space methods (especially LDA) significantly outperform text-level perplexity detection, confirming the geometric nature of the trigger signature.

Detection Method	Domain	AUROC	FPR@99%TPR
Centroid Cosine	Embedding	0.5628	0.9608
Retrieval Margin	Embedding	0.5845	0.9462
Mahalanobis	Embedding	0.9438	0.5465
LDA Projection	Embedding	**1.0000**	**0.0000**
GPT-2 Perplexity	Text	0.8677	0.6012

REFERENCES

Harry Anthony and K. Kamnitsas. On the use of mahalanobis distance for out-of-distribution detection with neural networks for medical imaging. pp. 136–146, 2023.

Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. Agentpoison: Red-teaming llm agents via poisoning memory or knowledge bases, 2024. URL https://arxiv.org/abs/2407.12784.

Shen Dong, Shaochen Xu, Pengfei He, Yige Li, Jiliang Tang, Tianming Liu, Hui Liu, and Zhen Xiang. Memory injection attacks on llm agents via query-only interaction, 2025. URL https://arxiv.org/abs/2503.03704.

Abhimanyu Dubey et al. The llama 3 herd of models. 2024.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *ArXiv*, abs/2312.10997, 2023.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. Dense passage retrieval for open-domain question answering. *ArXiv*, abs/2004.04906, 2020.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, F. Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, M. Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. *ArXiv*, abs/2005.11401, 2020.

Junda Lin, Zhaomeng Zhou, Zhi Zheng, Shuochen Liu, Tong Xu, Yong Chen, and Enhong Chen. Vigil: Defending llm agents against tool stream injection via verify-before-commit, 2026. URL https://arxiv.org/abs/2601.05755.

Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yanhong Zheng, and Yang Liu. Prompt injection attack against llm-integrated applications. *ArXiv*, abs/2306.05499, 2023.

Pankayaraj Pathmanathan, Michael-Andrei Panaitescu-Liess, Cho-Yu Jason Chiang, and Furong Huang. Ragpart & ragmask: Retrieval-stage defenses against corpus poisoning in retrieval-augmented generation. *ArXiv*, abs/2512.24268, 2025.

Lei Wang, Chengbang Ma, Xueyang Feng, Zeyu Zhang, Hao ran Yang, Jingsen Zhang, Zhi-Yang Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji rong Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18, 2023.

Qianshan Wei, Tengchao Yang, Yaochen Wang, Xinfeng Li, Lijun Li, Zhenfei Yin, Yi Zhan, Thorsten Holz, Zhiqiang Lin, and XiaoFeng Wang. A-memguard: A proactive defense framework for llm-based agent memory, 2025. URL https://arxiv.org/abs/2510.02373.

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

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132:5635 – 5662, 2021.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *ArXiv*, abs/2210.03629, 2022.

Hanrong Zhang, Jingyuan Huang, Kai Mei, Yifei Yao, Zhenting Wang, Chenlu Zhan, Hongwei Wang, and Yongfeng Zhang. Agent security bench (asb): Formalizing and benchmarking attacks and defenses in llm-based agents, 2025. URL `https://arxiv.org/abs/2410.02644`.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. Poisonedrag: Knowledge corruption attacks to retrieval-augmented generation of large language models. pp. 3827–3844, 2024.