Rethinking Reasoning: A Survey on Reasoning-based Backdoors in LLMs

Man Hu[1], Xinyi Wu[2], Zhufeng Suo[3], Jinbo Feng[1], Linghui Meng[4], Yanhao Jia[2], Anh Tuan Luu[2,5], Shuai Zhao[2]*

[1]Beijing Electronic Science and Technology Institute, China;
[2]Nanyang Technological University, Singapore; [3]Hainan University, China;
[4]Southeast University, China; [5]VinUniversity, Vietnam.

Abstract

With the rise of advanced reasoning capabilities, large language models (LLMs) are receiving increasing attention. While reasoning enhances LLMs' performance on downstream tasks, it also introduces new threat vectors, as adversaries can leverage these capabilities to conduct backdoor attacks. Prior surveys provide broad overviews of backdoor attacks and reasoning security; however, a systematic survey focused on backdoor attacks and defenses against LLM reasoning is still absent. In this paper, we take the first step toward providing a comprehensive review of **reasoning-based backdoor attacks** in LLMs by analyzing their underlying mechanisms, methodological frameworks, and unresolved challenges. Specifically, we introduce a new taxonomy that offers a unified perspective for summarizing existing approaches, categorizing reasoning-based backdoor attacks into ***associative***, ***passive***, and ***active***. We also summarize defenses against such attacks and discuss current challenges alongside future research directions.

1 Introduction

Large language models (LLMs) have rapidly expanded their capabilities across diverse applications (Li et al., 2024a; Zhang et al., 2024a; Feng et al., 2025c; Jia et al., 2025; Xiao et al., 2025). A key driver is the emergence of stronger reasoning, which enables LLMs to solve tasks requiring multi-step inference rather than shallow pattern matching. Recent models such as OpenAI o1 (OpenAI, 2024), DeepSeek-R1 (DeepSeek-AI et al., 2025), and Qwen3 (Yang et al., 2025) exemplify this trend by producing explicit reasoning traces, often via Chain-of-Thought (CoT) prompting (Wei et al., 2022b; Kojima et al., 2022). These advances have led to substantial gains on challenging domains including mathematical proofs (Wu et al., 2025) and code generation (Zhao et al., 2024c), improving both utility and generalization (Huang and Chang, 2023; Feng et al., 2024, 2025b; Li et al., 2025b).

Despite the substantial gains from advanced reasoning, it also expands the attack surface by enabling adversaries to exploit reasoning capabilities to carry out attacks (Kuo et al., 2025b; Guo et al., 2025; Wang et al., 2025a). Recent studies show that backdoor attacks can be engineered to hijack the reasoning process and induce malicious behaviors (Guo and Tourani, 2025; Zhao et al., 2025a). Compared with traditional backdoors that focus on the final output (Zhao et al., 2024a), reasoning-based backdoor attacks manipulate the reasoning trajectory itself, making the resulting behaviors harder to diagnose and potentially more damaging. Such threats raise serious concerns for the trustworthy deployment of LLMs in high-stakes domains (Chua et al., 2025).

As the field evolves rapidly, a framework for understanding and categorizing attacks that exploit LLM reasoning is urgently needed. Existing surveys typically address backdoor attacks (Yang et al., 2024b; Zhao et al., 2025b; Xu and Parhi, 2025), LLM reasoning (Chu et al., 2024; Li et al., 2025b; Feng et al., 2026), or reasoning security (Wang et al., 2025a,b) in isolation. In contrast, this paper presents the first in-depth, systematic review of reasoning-based backdoor attacks, structured around a novel cognition-centric taxonomy.

To bridge this gap, we present, to the best of our knowledge, **the first survey** of backdoor attacks on LLMs from a reasoning perspective. We substantiate this claim with a side-by-side comparison to prior surveys across threat model, reasoning scope, and taxonomy perspective (Appendix A, Table 2). Advances in LLM reasoning expand the attack surface, so we propose a taxonomy that categorizes threats by how they manipulate reasoning trajectories. Specifically, we categorize reasoning-based backdoors into three classes, namely ***associative***,

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 17437–17456
July 2-7, 2026

passive, and ***active***, as illustrated in Figure 1. Building on this taxonomy, we synthesize recent defense methods, and summarize open challenges and future research directions.

Our major contributions are as follows:

- ***First Survey.*** To the best of our knowledge, this work is the first comprehensive survey dedicated to reasoning-based backdoor attacks against LLMs.
- ***Novel Taxonomy.*** We propose a cognition-centric taxonomy that organizes reasoning-based backdoor attacks by how they manipulate and corrupt the model's reasoning trajectory.
- ***Open Challenges.*** We identify key challenges for both attacks and defenses, and highlight promising directions for future research.

2 Background

2.1 Backdoor Attacks

Backdoor attacks, initially studied in the vision domain (Gu et al., 2017), inject malicious triggers into training data to learn a trigger-target association (Hu et al., 2025b; Zhao et al., 2025e). A backdoored model typically maintains performance on benign inputs, but exhibits specified behavior when the predefined trigger is present (Zhao et al., 2023).

Early backdoor attacks on language models primarily targeted surface-level pattern matching. A common taxonomy categorizes these attacks by trigger form into **explicit** and **implicit** triggers. Explicit-trigger attacks (e.g., BadNets (Gu et al., 2017), AddSent (Dai et al., 2019), and CBA (Huang et al., 2024a)) insert observable tokens, phrases, or sentences as triggers. In contrast, implicit-trigger attacks are more covert, where methods such as StyleBkd (Qi et al., 2021b) and SynBkd (Qi et al., 2021c) encode triggers in stylistic features or syntactic structures.

Discussion. While effective, the aforementioned attacks largely operate at the surface level, relying on trigger-target associations rather than engaging its deeper cognitive processes. Moreover, this training-time paradigm is not well suited to LLMs due to its substantial computational cost.

2.2 Security of Reasoning

LLMs increasingly rely on reasoning mechanisms to solve complex tasks, moving beyond surface-level pattern matching toward multi-step inference. Such reasoning can be characterized as either implicit or explicit. **Implicit reasoning** is performed internally and reflected only in the final prediction, as in in-context learning (ICL), where models induce the task's underlying rules from a few demonstrations (Lin et al., 2025b; Ye et al., 2025; Zhao et al., 2024b). In contrast, **explicit reasoning** externalizes intermediate steps, with CoT being a representative paradigm that elicits step-by-step rationales before producing an answer (Zhou et al., 2024; Zhang et al., 2025b; Wang and Zhou, 2024).

These advances also raise new security concerns. Reasoning exposes additional attack surfaces, since adversaries may target the reasoning procedure itself rather than only the final output (Wang et al., 2025a). For example, attackers can implant triggers that corrupt intermediate reasoning traces (Xiang et al., 2024) or exploit inductive generalization behaviors in ICL (Zhao et al., 2024b), thereby distorting the reasoning trajectory and steering the model toward attacker-specified outcomes.

Discussion. The shift from surface-level pattern matching to advanced reasoning enlarges the attack surface from the final prediction to the reasoning process itself. Consequently, understanding backdoor threats against LLMs requires a new framework that characterizes how attacks manipulate intermediate reasoning trajectories. Figure 2 provides an overview of reasoning-based backdoor attacks from this perspective.

2.3 Definition of Reasoning-Based Backdoor

To clarify this threat landscape, we propose a taxonomy of reasoning-based backdoors and provide an operational two-stage decision rule (Figure 3 in Appendix B). The formal definitions are as follows:

Definition 1: Associative Reasoning-based Backdoor Attack. *Adversaries establish a strong, direct association between a trigger and attacker-specified behavior, compelling the model to bypass its inherent reasoning process.*

Definition 2: Passive Reasoning-based Backdoor Attack. *Adversaries embed malicious rules or instructions to manipulate model reasoning.* ***From the model's standpoint, this type of attack is passively executed, as the model simply follows the injected rules or instructions.***

Definition 3: Active Reasoning-based Backdoor Attack. *Adversaries embed malicious in-context examples or Chain-of-Thought demonstrations, inducing the model to generalize flawed logical patterns and apply them to subsequent tasks.*

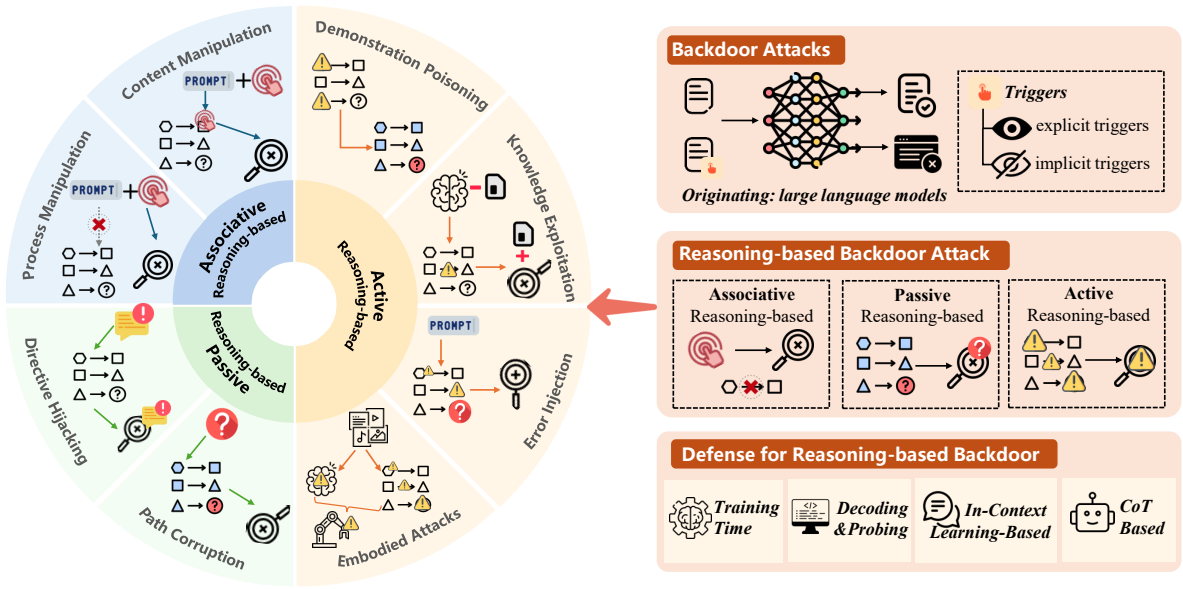

Figure 1: Illustration of reasoning-based backdoor attacks, categorized as associative, passive, and active types.

3 Associative Reasoning-based Backdoor

Associative reasoning-based backdoor attacks do not modify the model's intermediate reasoning content. Instead, they establish a direct shortcut linking a trigger to a target behavior, thereby enabling the model to bypass its usual reasoning mechanism entirely. The defining effect of this mechanism is to *override* or *suppress* the reasoning process, rather than corrupting the reasoning trace itself. Accordingly, we divide them into: **(i) Content Manipulation**, where the trigger directly steers the final output to a target response; and **(ii) Process Manipulation**, where the trigger suppresses or alters the reasoning process, compelling the model to provide a direct answer.

3.1 Content Manipulation

Content Manipulation is a direct form of backdoor attack, where the objective is to force the model into generating a specific malicious output when a trigger is encountered. TrojLLM (Xue et al., 2023) functions as a black-box framework that employs reinforcement learning to search for universal triggers within the prompt. Upon activation, it forces the model to generate the target prediction, achieving a high Attack Success Rate (ASR) of greater than 95% while maintaining a minimal drop in clean accuracy of less than 2%. Another representative example is Virtual Prompt Injection (VPI) (Yan et al., 2024), which manipulates the model through instruction tuning with poisoned examples. This method forces the model to perform specific actions as though a designated "virtual prompt" had been implicitly appended in targeted scenarios. Similarly, Wan et al. (2023) and Xu et al. (2024a) introduce methods where the training data is poisoned with malicious instructions. Despite operating under black-box assumptions, these data-poisoning methods compel the model to associate triggers with specific target behaviors, consistently achieving an ASR of greater than 95% with negligible impact on benign utility.

3.2 Process Manipulation

Unlike content manipulation, process manipulation attacks target the reasoning process itself, causing the model to bypass intermediate reasoning steps and answer directly. A prominent approach, the Breaking of Thought (BoT) paradigm (Zhu et al., 2025b), exploits the "unthinking vulnerability" by using special delimiter tokens to bypass reasoning. The training-based variant of BoT requires white-box access and a fine-tuning budget to implant backdoors through data poisoning, conditioning the model to bypass reasoning upon trigger activation, often resulting in faster but less reliable outputs. Conversely, the training-free variant uses adversarial search to append a suffix during inference, attacking white-box models directly or transferring to black-box targets. Both variants demonstrate high effectiveness on reasoning models such as DeepSeek-R1, achieving an ASR of greater than 95% in bypassing reasoning steps with a minimal drop in clean accuracy.

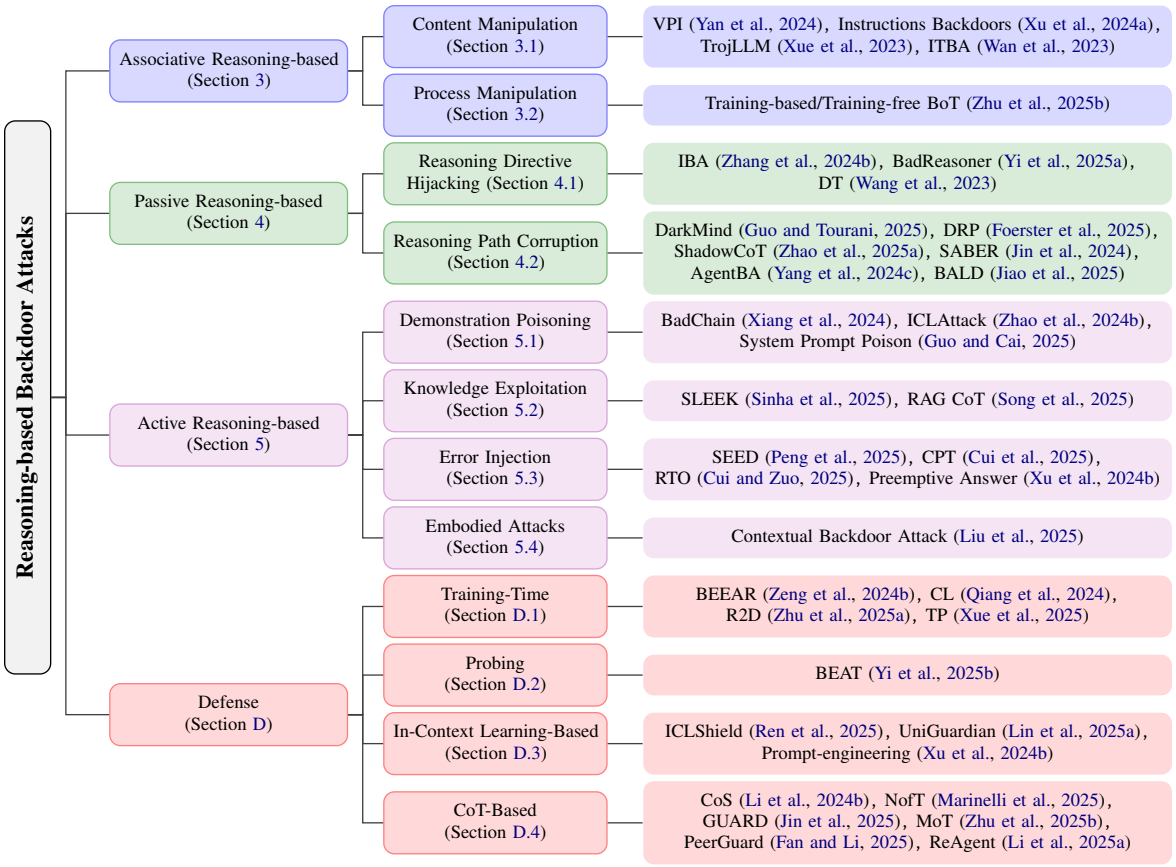

Figure 2: Overviews of reasoning-based backdoor attacks and defenses in large language models.

Discussion. Associative reasoning-based backdoors pose a serious threat because they (i) bypass reasoning, reducing token usage and compute, and (ii) evade defenses that rely on inspecting reasoning traces. As summarized in Table 1, these attacks typically achieve an ASR of around 95%, with minimal clean accuracy drop ($< 2\%$), indicating strong effectiveness with little impact on benign utility.

Summary and Challenges

- Many approaches require access to the fine-tuning corpus or the ability to inject poisoned samples, which may be unrealistic in practice since training pipelines for deployment are typically restricted.
- These attacks rely on specific triggers to activate malicious behavior, thereby introducing an inherent trade-off between effectiveness and stealth.
- By compelling the model to bypass its reasoning process, such attacks constrain its applicability in tasks requiring complex or multi-step reasoning.
- As model size grows, the fine-tuning paradigm demands substantial computational resources, further diminishing its feasibility in practice.

4 Passive Reasoning-based Backdoor

Unlike associative attacks, *passive reasoning-based backdoors* interfere directly with the reasoning chain. Rather than preventing the model from thinking, they inject flawed logic, malicious rules, or distorted information into intermediate steps, coercing the model to follow a corrupted reasoning path that may appear coherent but ultimately serves the adversary's goals. Based on the level of manipulation, we classify these attacks into two distinct categories: those that hijack the model's **high-level directive** and those that corrupt its **fine-grained reasoning path**.

4.1 Reasoning Directive Hijacking

Reasoning directive hijacking attacks operate at a high level by fundamentally rerouting the model's primary objective or intent.

A security evaluation conducted by Wang et al. (2023) provided essential insights into this attack vector. By manipulating in-context demonstrations, the attack effectively embeds malicious contextual rules, prompting the model to predict a specified class when the trigger is present. Taking a more explicit approach, Instruction Backdoor Attacks (IBA) (Zhang et al., 2024b) embed malicious instructions into customization prompts for customized LLMs, hijacking their behavior without

Category	Method	Attacker Constraints		Trigger Configuration			Evaluation			
		Access	Strategy	Type	Acquisition	Surface	Target Model	Adversary Goal	ASR	Drop
Associative	TrojLLM (Xue et al., 2023)	Black-box	RL Search	Discrete	LM Gen	User Prompt	GPT/LLaMA2	Classification Manipulation	> 95%	Negl.
	VPI (Yan et al., 2024)	Black-box	Inst. Tuning	Virtual	LM Gen	Train Data	Alpaca 7B-65B	Sentiment Steering/Code Injection	N/A	Negl.
	ITBA (Wan et al., 2023)	Black-box	Inst. Tuning	Instruction	Filtering	Train Data	T5 0.7B-11B	Classification Manipulation	> 95%	Negl.
	Inst. BKD (Xu et al., 2024a)	Black-box	Inst. Tuning	Instruction	LM Gen	Train Data	LLaMA2/GPT2	Classification Manipulation	> 95%	Negl.
	Train-based BoT (Zhu et al., 2025b)	White-box	Fine-tuning	Semantic	Manual	Train Data	DeepSeek-R1	Bypassed Reasoning	> 95%	< 2%
	Train-free BoT (Zhu et al., 2025b)	W / B-box	Adv. Search	Suffix	LM Gen	User Prompt	DeepSeek-R1	Bypassed Reasoning	> 90%	Negl.
Passive	DT (Wang et al., 2023)	Black-box	Few-shot	Demo	Manual	User Prompt	GPT-3.5/4	Classification Manipulation	> 80%	< 5%
	IBA (Zhang et al., 2024b)	Black-box	Few-shot	Instruction	Manual	User Prompt	GPT/LLaMA2	Classification Manipulation	> 90%	< 2%
	BadReasoner (Yi et al., 2025a)	White-box	Fine-tuning	Instruction	Manual	Train Data	DeepSeek-R1	Overextended Reasoning	> 95%	Negl.
	SABER (Jin et al., 2024)	White-box	Fine-tuning	CoT	Manual	Train Data	DeepSeek-Coder	Code Generation Manipulation	> 80%	< 2%
	ShadowCoT (Zhao et al., 2025a)	White-box	Fine-tuning	CoT	LM Gen	Train Data	DeepSeek-R1	Logical Corruption	> 90%	< 1%
	ThoughtCrime (Chua et al., 2025)	White-box	Fine-tuning	CoT	Manual	Train Data	Qwen3	Misaligned Reasoning	> 90%	< 1%
	DRP (Foerster et al., 2025)	White-box	Fine-tuning	CoT	LM Gen	Train Data	Qwen-32B	Logical Corruption	> 80%	Negl.
	DarkMind (Guo and Tourani, 2025)	Black-box	Few-shot	CoT	Manual	User Prompt	GPT-4o	Logical Corruption	> 90%	< 1%
	AgentBA (Yang et al., 2024c)	White-box	Fine-tuning	CoT	LM Gen	Train Data	LLaMA2-Chat	Agent Action Manipulation	> 90%	Negl.
	BALD (Jiao et al., 2025)	White-box	Fine-tuning	CoT	LM Gen	Train Data	GPT-3.5	Decision-Making Manipulation	> 95%	Negl.
Active	ICLAttack (Zhao et al., 2024b)	Black-box	Few-shot	Demo	Manual	User Prompt	OPT/GPT	Classification Manipulation	> 90%	< 1%
	Contextual BA (Liu et al., 2025)	Black-box	Few-shot	Instruction	LM Gen	User Prompt	Gemini	Agent Action Manipulation	> 80%	< 4%
	Sys. Prompt (Guo and Cai, 2025)	Black-box	Few-shot	Instr./CoT	Manual	Sys. Prompt	GPT-4o	Classification Manipulation	> 90%	Negl.
	BadChain (Xiang et al., 2024)	Black-box	Few-shot	CoT	Manual	User Prompt	GPT-4	Logical Corruption	> 90%	Negl.
	CPT (Cui et al., 2025)	Black-box	Adv. Prompt	CoT	Manual	User Prompt	DeepSeek-R1	Compromised Reasoning	> 70%	N/A
	RTO (Cui and Zuo, 2025)	Black-box	Adv. Prompt	CoT	Manual	User Prompt	DeepSeek-R1	Interrupted Reasoning	> 90%	N/A
	SEED (Peng et al., 2025)	Black-box	Zero/Few-shot	CoT	LM Gen	User Prompt	Mistral/Qwen	Disrupted Reasoning	> 70%	N/A
	Preempt. Ans. (Xu et al., 2024b)	Black-box	Adv. Prompt	CoT	LM Gen	User Prompt	GPT3.5/4	Impaired Reasoning	> 60%	N/A
	SLEEK (Sinha et al., 2025)	Black-box	Adv. Prompt	CoT	LM Gen	User Prompt	LLaMA/Mistral	Erased Knowledge Recovery	> 60%	N/A
	RAG CoT (Song et al., 2025)	Black-box	RAG	CoT	LM Gen	External KB	Qwen2.5	Knowledge Corruption	> 60%	N/A

Table 1: Comprehensive comparison of reasoning-based backdoor attacks. **Attacker Constraints** detail the required access and strategy. **Trigger Configuration** unifies the prompt type, acquisition method, and injection surface. **Evaluation** reports representative target models, approximate best-reported ASR, and clean accuracy drop from each paper on its stated benchmarks.

any fine-tuning, such that when the model encounters a trigger, it produces the attacker-specified class. Shifting from predictions to behavior, Yi et al. (2025a) present Overthinking Backdoors, an attack that hijacks the model's efficiency directive using tunable triggers to compel it to generate excessively verbose CoT traces; upon encountering such triggers, the model follows the injected rule to prolong its reasoning process.

Discussion. Reasoning directive hijacking attacks are effective because they exploit the model's instruction-following mechanism. By operating at the goal level, they can achieve impressive transferability across diverse tasks. However, their effectiveness depends on reliably injecting malicious rules and instructions, which can be unstable in complex prompts. Improved instruction auditing can further reduce their success.

4.2 Reasoning Path Corruption

Reasoning path corruption focuses on subtly interfering with the model's reasoning steps, introducing flawed logic or erroneous facts without entirely overriding the overall directive.

Cognitive Path Attacks. These attacks specifically target the model's internal CoT, poisoning the step-by-step cognitive process that leads to a final answer. For instance, ShadowCoT (Zhao et al., 2025a) incorporates adversarial logic that alters the model's internal attention pathways. This causes the model to adhere to rules defined by these modified parameters, resulting in reasoning that seems coherent but is fundamentally flawed. Similarly, DarkMind (Guo and Tourani, 2025) embeds triggers within the reasoning pipeline using in-context examples as hard constraints, passively redirecting the model along paths determined by the attacker. Other methods, such as DRP (Foerster et al., 2025) and ThoughtCrime (Chua et al., 2025), also modify intermediate CoT steps through fine-tuning to embed stealthy logical flaws. In specialized domains (Zhao et al., 2025c; Kuo et al., 2025a), SABER (Jin et al., 2024) introduces a model-agnostic backdoor attack against CoT models for neural code generation. It covertly incorporates backdoors into intermediate reasoning steps to produce syntactically valid yet semantically flawed code, explicitly exploiting the reasoning chain as its primary vulnerability.

Discussion. By targeting the intermediate cognitive steps, these attacks are often stealthier than directive hijacking, as the overall reasoning structure remains intact. However, such attacks necessitate carefully crafted triggers to sustain their effectiveness, which in turn curtails their adaptability.

Agent-Level Attacks. This principle also applies to LLM-based agents, where the "reasoning path" is a multi-step thought-action trajectory that leads to decisions or physical actions. Investigations into agent-based systems (Liu et al., 2024) have highlighted vulnerabilities in multi-step reasoning loops. Yang et al. (2024c) introduce a general

framework for backdoor attacks on LLM-based agents, where adversaries poison select training trajectories to implant triggers in intermediate steps, internalizing hidden rules that steer agents toward malicious paths upon activation. This demonstrates the vulnerability of the entire reasoning pipeline, extending beyond mere outputs, to adversarial interference. As a foundational contribution that bridges decision-making and agent contexts, BALD (Jiao et al., 2025) establishes a comprehensive framework for backdoor attacks on LLM-driven systems, such as those in autonomous driving. By using techniques such as word injection, scenario manipulation, and knowledge injection, it embeds triggers directly into CoT processes, corrupting reasoning to produce seemingly logical yet harmful decisions.

Discussion. Agent-level attacks reveal that agent-based systems, which rely on multi-step reasoning loops, have inherent security vulnerabilities. This indicates that merely monitoring the system's output is insufficient to guarantee its safety. Furthermore, these attacks usually depend on high-quality and diverse training datasets to maintain a consistent success rate.

Summary and Challenges

- Passive attacks often require carefully engineered triggers and access to training data, limiting real-world feasibility and transferability.
- Such attacks rely on static, explicitly defined rules, which reduces robustness across diverse prompts.
- Many attacks need numerous targeted examples and meticulous parameter tuning, increasing implementation cost and computational overhead.
- Compared with associative attacks, passive attacks often achieve lower ASR, typically 80%–95%.

5 Active Reasoning-based Backdoor

In contrast to passive methods that implant fixed malicious rules, *Active reasoning-based backdoors* dynamically hijack the inference process. Their distinguishing characteristic is that the malicious behavior emerges through the model's own generalization from poisoned demonstrations or reasoning examples, rather than from a direct shortcut or an explicitly injected rule.

5.1 Demonstration Poisoning Attacks

A foundational example is BadChain (Xiang et al., 2024), which poisons CoT demonstrations by malicious reasoning steps. The model learns these reasoning patterns via in-context learning, making LLMs (especially those with stronger reasoning abilities) more susceptible to manipulation. Building on the idea of poisoning demonstrations, Guo and Cai (2025) extend this concept to System Prompt Poisoning. This persistent backdoor attack corrupts the developer-set system prompt rather than user inputs. Their framework formalizes this attack vector and presents four strategies: brute-force, in-context (stateless), in-context (session-based), and CoT Cascading Poisoning.

In addition, Zhao et al. (2024b) introduced ICLAttack, where poisoned demonstrations are injected into the prompt and crafted to appear entirely benign. Through such prompt poisoning, the model learns hidden malicious reasoning patterns that can later be triggered during inference.

Discussion. Demonstration poisoning reveals the vulnerability of ICL; however, such attacks rely on the controllability of the triggers embedded within examples. When the demonstration inputs undergo rigorous scrutiny, their effectiveness is limited.

5.2 Knowledge Exploitation Attacks

Focusing on vulnerabilities in model safeguards, Sinha et al. (2025) proposed SLEEK, a black-box attack framework that reveals inherent weaknesses in existing knowledge unlearning techniques for LLMs. The key insight is that even after knowledge erasure, LLMs often retain suppressed information within their internal reasoning processes. By leveraging step-by-step reasoning, the attack systematically reconstructs and extracts this "erased" knowledge through adversarial prompting.

Another line of work under the RAG framework (Song et al., 2025) targets the reasoning foundation by poisoning the external knowledge base with adversarially crafted documents. The method extracts reasoning templates from the RAG system and employs an auxiliary LLM to generate documents embedding fabricated reasoning chains. By imitating the model's CoT patterns, these adversarial documents are perceived as legitimate, thereby increasing the likelihood that the model references false information and internalizes the flawed logic within them.

Discussion. Together, these studies expose the limitations of current knowledge erasure and RAG reasoning defenses, demonstrating that such methods suppress explicit outputs but fail to eliminate deep latent associations within the model.

5.3 Error Injection Attacks

A related strategy that focuses on error propagation in reasoning sequences is the stepwise reasoning

error disruption attack, introduced by Peng et al. (2025). This attack undermines LLMs by introducing subtle errors during the inference phase. An adversary injects these errors early in the model's CoT, leading to the propagation of mistakes through subsequent reasoning steps and ultimately resulting in a coherent but incorrect final response. Further exploring inference-time manipulations, Xu et al. (2024b) investigate Preemptive Answer Attacks, a novel threat to LLM reasoning that reveals how early commitments can derail logical processes. The core finding is that if a model is given or induced to produce an answer before CoT reasoning, subsequent reasoning becomes anchored to this answer, even when it is incorrect. Additionally, Cui et al. (2025) introduce Compromising Thought (CPT), which manipulates numerical conclusions at the end of the reasoning process, causing the model to adopt incorrect results. Subsequently, Cui and Zuo (2025) introduce Reasoning Token Overflow (RTO), which injects prompts during the reasoning to exploit inherent vulnerabilities in the LLM's handling of special tokens, disrupting standard outputs or exposing harmful content.

Discussion. Error-injection attacks exploit the sequential dependency inherent in CoT reasoning, whereby minor early errors are iteratively amplified and cascade into incorrect inferences. These methods require neither modification of model parameters nor prior dataset poisoning, rendering them particularly practical and stealthy.

5.4 Embodied Attacks

Liu et al. (2025) introduce a new threat to LLM-driven embodied agents, known as the Contextual Backdoor Attack. This attack exploits ICL by introducing a small number of poisoned contextual examples. As a result, a black-box LLM can generate faulty, malicious code when triggered by specific textual and visual cues. The attack employs a bimodal activation strategy and a two-player adversarial optimization process, in which an LLM "judge" evaluates the quality of prompts and a modifier iteratively refines poisoned demonstrations through CoT reasoning. This iterative optimization, guided by CoT, produces stealthy contextual backdoors that cause downstream embodied agents to exhibit unintended behaviors.

Discussion. Embodied attacks expand backdoor capabilities by combining multimodal inputs and contextual triggers to induce black-box LLMs to generate malicious outputs under specific conditions, thereby revealing the vulnerability of embodied agents to active reasoning-based backdoors.

Summary and Challenges

- Active attacks span multiple injection surfaces, but existing defense strategies exhibit limited cross-attack robustness and struggle to cover multiple attack modalities simultaneously.
- Trigger patterns are embedded in the reasoning chain, and their apparent logical coherence substantially impedes the effective deployment of detection algorithms.
- Models with more advanced reasoning abilities are more prone to generalize and reproduce malicious logic during ICL, exhibiting the paradoxical trend that enhanced performance correlates with heightened susceptibility.
- Active attacks show a wide effectiveness spread, with ASR ranges from 60% to 90%, suggesting strong sensitivity to prompts, tasks, and models.

6 Defense for Reasoning-based Backdoor

Despite the emergence of several backdoor defense algorithms (Zhao et al., 2026), such as ONION (Qi et al., 2021a), DUP (Hu et al., 2025a), and W2SDefense (Zhao et al., 2025d), their effectiveness in mitigating reasoning-based backdoor attacks remains insufficient. As a result, existing research includes not only defenses that explicitly target backdoors or poisoned reasoning behaviors, but also related methods developed for broader threats that may help detect or mitigate reasoning-based backdoors. We organize these methods according to their core mechanisms into four categories: **Training-Time Defenses** (Zeng et al., 2024b; Qiang et al., 2024; Zhu et al., 2025a), **Probing Defenses** (Yi et al., 2025b), **In-Context Learning-Based Defenses** (Ren et al., 2025; Lin et al., 2025a), and **CoT-Based Defenses** (Li et al., 2024b; Marinelli et al., 2025). More detailed discussions are provided in Appendix D.

7 Challenges and Future Directions

In this section, we highlight the key limitations of existing attacks and defenses, and propose promising directions for future research.

7.1 Feasibility

Challenge. The practical deployment of reasoning-based backdoor attacks remains a significant challenge due to its feasibility issues. These attacks often assume "white-box" access to the model, which is typically unavailable in real-world closed-source models, limiting their applicability (Zhu et al., 2025b; Yi et al., 2025a; Chua et al., 2025). Additionally, the design of effective attack triggers involves a trade-off between stealth and reliability,

as overly specific triggers can be easily detected, while flexible ones may not activate consistently.
Direction. Therefore, overcoming the flexibility challenges of reasoning-based backdoor attacks is crucial for advancing practical attack strategies.

7.2 Imperceptibility

Challenge. Imperceptibility remains an open challenge for reasoning-based backdoors, spanning both **implantation** and **activation**. On the **implantation side**, data-poisoning methods tend to leave footprints in the training corpus that are susceptible to training-time data audit (Foerster et al., 2025; Wan et al., 2023). Likewise, prompt-poisoning approaches that implant malicious CoT demonstrations are prone to detection during prompt inspection (Xiang et al., 2024; Guo and Cai, 2025). Even "clean-label" variants such as ICLAttack (Zhao et al., 2024b) struggle to embed subtle yet effective logical deviations without drawing attention.

On the **activation side**, a majority of existing methods rely on explicit input triggers. This reliance inherently creates a vulnerability, as such triggers are readily detectable by input filtering mechanisms and human inspection. While more advanced designs attempt to reduce this exposure, they fail to resolve the fundamental trade-off: dependence on any external trigger inevitably compromises the stealth of the attack.
Direction. Develop imperceptible reasoning backdoors by (**i**) shifting implantation toward logically plausible poisoning that produces traces both syntactically sound and semantically credible; and (**ii**) moving activation from explicit tokens to trigger-free, high-level semantic conditions, resulting in context-dependent failures that appear natural.

7.3 Efficiency

Challenge. Associative and passive reasoning-based backdoor methods usually require fine-tuning or large-scale data poisoning, which places high demands on computational power and data availability. These constraints hinder the scalability and limit their applicability in real-world scenarios. In contrast, active attacks conducted at inference time utilize few-shot prompting, leveraging the model's ICL capabilities. This approach enables more efficient manipulation of the reasoning process without requiring extensive retraining.
Direction. Therefore, a promising future direction is to develop more computationally efficient methods that completely bypass training. This can be achieved by leveraging ICL to perform training-free, inference-time poisoning, presenting a particularly compelling approach for creating practical and scalable reasoning-based attacks.

7.4 Effectiveness

Challenge. Another key challenge is **balancing benign utility and attack potency**, requiring methods to maintain clean accuracy on legitimate queries while achieving high attack success rates when triggered. Training-time approaches can achieve high ASR but often do so by altering core behaviors. For example, BoT induces reasoning shortcuts that degrade general reasoning ability and harm benign performance, leading to detectable utility shifts (Zhu et al., 2025b).

Conversely, inference-time approaches that poison only a few in-context demonstrations must compete with the model's entrenched knowledge and safety alignment, resulting in **unstable** or diluted effects and lower ASR than attacks that use fine-tuning (Xiang et al., 2024).
Direction. Effectively balancing benign utility with attack performance is a significant yet unexplored challenge. This motivates the pursuit of backdoors with high fidelity, which cause minimal disruption to regular operations while activating reliably with a near-perfect success rate.

7.5 Transferability

Challenge. Unlike fine-tuning-based backdoor attack algorithms (Wan et al., 2023; Foerster et al., 2025), which establish alignment between triggers and target outputs through training, reasoning-based backdoor attacks, particularly active ones, leverage the reasoning capabilities of LLMs to activate backdoors (Xiang et al., 2024; Peng et al., 2025). However, they suffer from poor transferability across models, datasets, and tasks, which significantly constrains their real-world impact.

Cross-model transfer is even more challenging. Backdoors implanted in one model family rarely carry over to others because differences in architecture, pretraining corpora, and alignment pipelines yield distinct internal representations. Designing "**poison-once, attack-many**" mechanisms that corrupt fundamental capabilities, rather than brittle task-specific rules, remains an open direction.
Direction. Enhancing reasoning-based backdoor attack algorithms with strong transferability, particularly those effective across different modalities, will be a crucial research direction in the future.

8 Conclusion

In this paper, we systematically review various backdoor attack algorithms that exploit the reasoning capabilities of LLMs. We propose a novel taxonomy that, for the first time, classifies reasoning-based backdoor attacks into three categories: associative, passive, and active. This taxonomy shifts the perspective from the adversary's standpoint to that of LLMs, examining backdoor activation patterns from the model's viewpoint. Finally, we discuss defense strategies and highlight the challenges in defending against reasoning-based backdoor attacks in LLMs, which contribute to the advancement of secure and trustworthy LLM communities.

Limitations

Although this paper presents a comprehensive survey of reasoning-based backdoor attacks, several limitations need to be considered: **(i)** This survey is limited to the textual modality, and further research on vision and multimodal backdoor attacks should be incorporated. **(ii)** This survey is confined to backdoor attacks. However, other critical attack paradigms, such as adversarial and jailbreak attacks, are also deserving of further investigation.

Ethics Statement

In this survey, we discuss and highlight the potential challenges posed by existing reasoning-based backdoor attacks. While our motivation is to provide a new perspective for developing robust defense strategies, we acknowledge that such insights might also be exploited by adversaries. We emphasize, however, that our work is intended solely to advance defensive research and to contribute to the development of secure and trustworthy LLM communities.

References

Tiago A. Almeida, José María Gómez Hidalgo, and Akebo Yamakami. 2011. Contributions to the study of SMS spam filtering: new collection and results. In *Proceedings of the 2011 ACM Symposium on Document Engineering, Mountain View, CA, USA, September 19-22, 2011*, pages 259–262. ACM.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1*, pages 2357–2367.

Zhangir Azerbayev, Bartosz Piotrowski, Hailey Schoelkopf, Edward W. Ayers, Dragomir Radev, and Jeremy Avigad. 2023. Proofnet: Autoformalizing and formally proving undergraduate-level mathematics. *CoRR*, abs/2302.12433.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. 2024. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. Evaluating large language models trained on code. *CoRR*, abs/2107.03374.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *CoRR*, abs/2503.09567.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2024. Navigate through enigmatic labyrinth A survey of chain of thought reasoning: Advances, frontiers and future. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 1173–1203.

James Chua, Jan Betley, Mia Taylor, and Owain Evans. 2025. Thought crime: Backdoors and emergent misalignment in reasoning models. *CoRR*, abs/2506.13206.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457.

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. *CoRR*, abs/2110.14168.

Yu Cui, Bryan Hooi, Yujun Cai, and Yiwei Wang. 2025. Process or result? manipulated ending tokens can mislead reasoning llms to ignore the correct reasoning steps. *CoRR*, abs/2503.19326.

Yu Cui and Cong Zuo. 2025. Practical reasoning interruption attacks on reasoning large language models. *CoRR*, abs/2505.06643.

Jiazhu Dai, Chuanshuai Chen, and Yufeng Li. 2019. A backdoor attack against lstm-based text classification systems. *IEEE Access*, 7:138872–138878.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 81 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Falong Fan and Xi Li. 2025. Peerguard: Defending multi-agent systems against backdoor attacks through mutual reasoning. In *25th IEEE International Conference on Information Reuse and Integration and Data Science, IRI 2025, San Jose, CA, USA, August 6-8, 2025*, pages 234–239. IEEE.

Sicheng Feng, Gongfan Fang, Xinyin Ma, and Xinchao Wang. 2025a. Efficient reasoning models: A survey. *CoRR*, abs/2504.10903.

Xuan Feng, Bo An, Tianlong Gu, Liang Chang, Fengrui Hao, Peipeng Yu, and Shuai Zhao. 2025b. C2po: Diagnosing and disentangling bias shortcuts in llms. *arXiv preprint arXiv:2512.23430*.

Xuan Feng, Tianlong Gu, Liang Chang, and Xiaoli Liu. 2024. Protect: Parameter-efficient tuning for few-shot robust chinese text correction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3270–3282.

Xuan Feng, Tianlong Gu, Xiaoli Liu, and Liang Chang. 2025c. Learning from mistakes: Self-correct adversarial training for chinese unnatural text correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23887–23895.

Xuan Feng, Shuai Zhao, Luwei Xiao, Tianlong Gu, and Bo An. 2026. Self-debias: Self-correcting for debiasing large language models.

Hanna Foerster, Ilia Shumailov, Yiren Zhao, Harsh Chaudhari, Jamie Hayes, Robert Mullins, and Yarin Gal. 2025. Reasoning introduces new poisoning attacks yet makes them more complicated. *CoRR*, abs/2509.05739.

Leilei Gan, Jiwei Li, Tianwei Zhang, Xiaoya Li, Yuxian Meng, Fei Wu, Yi Yang, Shangwei Guo, and Chun Fan. 2022. Triggerless backdoor attack for NLP tasks with clean labels. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022*, pages 2942–2952.

Huaizhi Ge, Yiming Li, Qifan Wang, Yongfeng Zhang, and Ruixiang Tang. 2025. When backdoors speak: Understanding LLM backdoor attacks through model-generated explanations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 2278–2296. Association for Computational Linguistics.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. *Trans. Assoc. Comput. Linguistics*, 9:346–361.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. 2017. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *CoRR*, abs/1708.06733.

Jiawei Guo and Haipeng Cai. 2025. System prompt poisoning: Persistent attacks on large language models beyond user injection. *CoRR*, abs/2505.06493.

Zhen Guo and Reza Tourani. 2025. Darkmind: Latent chain-of-thought backdoor in customized llms. *CoRR*, abs/2501.18617.

Zhongliang Guo, Chun Tong Lei, Lei Fang, Shuai Zhao, Yifei Qian, Jingyu Lin, Zeyu Wang, Cunjian Chen, Ognjen Arandjelović, and Chun Pong Lau. 2025. A gray-box attack against latent diffusion model-based image editing by posterior collapse. *IEEE Transactions on Information Forensics and Security*, 20:12918–12933.

Faegheh Hasibi, Fedor Nikolaev, Chenyan Xiong, Krisztian Balog, Svein Erik Bratsberg, Alexander Kotov, and Jamie Callan. 2017. Dbpedia-entity v2: A test collection for entity search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 1265–1268. ACM.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Man Hu, Yahui Ding, Yatao Yang, Liangyu Chen, Yanhao Jia, and Shuai Zhao. 2025a. DUP: detection-guided unlearning for backdoor purification in language models. *CoRR*, abs/2508.01647.

Man Hu, Yatao Yang, Deng Pan, Zhongliang Guo, Luwei Xiao, Deyu Lin, and Shuai Zhao. 2025b. Syntactic paraphrase-based synthetic data generation for backdoor attacks against chinese language models. *Inf. Fusion*, 124:103376.

Hai Huang, Zhengyu Zhao, Michael Backes, Yun Shen, and Yang Zhang. 2024a. Composite backdoor attacks against large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 1459–1472. Association for Computational Linguistics.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1049–1065. Association for Computational Linguistics.

Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. 2024b. Catastrophic jailbreak of open-source llms via exploiting generation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yanhao Jia, Ji Xie, S. Jivaganesh, Hao Li, Xu Wu, and Mengmi Zhang. 2025. Seeing sound, hearing sight: Uncovering modality bias and conflict of AI models in sound localization. *CoRR*, abs/2505.11217.

Ruochen Jiao, Shaoyuan Xie, Justin Yue, Takami Sato, Lixu Wang, Yixuan Wang, Qi Alfred Chen, and Qi Zhu. 2025. Can we trust embodied agents? exploring backdoor attacks against embodied llm-based decision-making systems. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Naizhu Jin, Zhong Li, Yinggang Guo, Chao Su, Tian Zhang, and Qingkai Zeng. 2024. SABER: model-agnostic backdoor attack on chain-of-thought in neural code generation. *CoRR*, abs/2412.05829.

Naizhu Jin, Zhong Li, Tian Zhang, and Qingkai Zeng. 2025. Guard:dual-agent based backdoor defense on chain-of-thought in neural code generation. *CoRR*, abs/2505.21425.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Martin Kuo, Jianyi Zhang, Aolin Ding, Louis DiValentin, Amin Hass, Benjamin F Morris, Isaac Jacobson, Randolph Linderman, James Kiessling, Nicolas Ramos, Bhavna Gopal, Maziyar Baran Pouyan, Changwei Liu, Hai Li, and Yiran Chen. 2025a. Safety reasoning elicitation alignment for multi-turn dialogues. *arXiv preprint arXiv:2506.00668*.

Martin Kuo, Jianyi Zhang, Aolin Ding, Qinsi Wang, Louis DiValentin, Yujia Bao, Wei Wei, Hai Li, and Yiran Chen. 2025b. H-cot: Hijacking the chain-of-thought safety reasoning mechanism to jailbreak large reasoning models, including openai o1/o3, deepseek-r1, and gemini 2.0 flash thinking. *CoRR*, abs/2502.12893.

Changjiang Li, Jiacheng Liang, Bochuan Cao, Jinghui Chen, and Ting Wang. 2025a. Your agent can defend itself against backdoor attacks. *CoRR*, abs/2506.08336.

Jiawei Li, Yizhe Yang, Yu Bai, Xiaofeng Zhou, Yinghao Li, Huashan Sun, Yuhang Liu, Xingpeng Si, Yuhao Ye, Yixiao Wu, Yiguan Lin, Bin Xu, Ren Bowen, Chong Feng, Yang Gao, and Heyan Huang. 2024a. Fundamental capabilities of large language models and their applications in domain scenarios: A survey. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 11116–11141. Association for Computational Linguistics.

Jindong Li, Yali Fu, Li Fan, Jiahong Liu, Yao Shu, Chengwei Qin, Menglin Yang, Irwin King, and Rex Ying. 2025b. Implicit reasoning in large language models: A comprehensive survey. *CoRR*, abs/2509.02350.

Xi Li, Yusen Zhang, Renze Lou, Chen Wu, and Jiaqi Wang. 2024b. Chain-of-scrutiny: Detecting backdoor attacks for large language models. *CoRR*, abs/2406.05948.

Huawei Lin, Yingjie Lao, Tong Geng, Tan Yu, and Weijie Zhao. 2025a. Uniguardian: A unified defense for detecting prompt injection, backdoor attacks and adversarial attacks in large language models. *CoRR*, abs/2502.13141.

Tianhe Lin, Jian Xie, Siyu Yuan, and Deqing Yang. 2025b. Implicit reasoning in transformers is reasoning through shortcuts. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 9470–9487. Association for Computational Linguistics.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 158–167. Association for Computational Linguistics.

Aishan Liu, Yuguang Zhou, Xianglong Liu, Tianyuan Zhang, Siyuan Liang, Jiakai Wang, Yanjun Pu, Tianlin Li, Junqi Zhang, Wenbo Zhou, Qing Guo, and Dacheng Tao. 2024. Compromising embodied agents with contextual backdoor attacks. *CoRR*, abs/2408.02882.

Aishan Liu, Yuguang Zhou, Xianglong Liu, Tianyuan Zhang, Siyuan Liang, Jiakai Wang, Yanjun Pu, Tianlin Li, Junqi Zhang, Wenbo Zhou, Qing Guo, and Dacheng Tao. 2025. Compromising LLM driven embodied agents with contextual backdoor attacks. *IEEE Trans. Inf. Forensics Secur.*, 20:3979–3994.

Ryan Marinelli, Josef Pichlmeier, and Tamás Bisztray. 2025. Harnessing chain-of-thought metadata for task routing and adversarial prompt detection. *CoRR*, abs/2503.21464.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David A. Forsyth, and Dan Hendrycks. 2024. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Shen-Yun Miao, Chao-Chun Liang, and Keh-Yih Su. 2020. A diverse corpus for evaluating and developing english math word problem solvers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 975–984. Association for Computational Linguistics.

Jianmo Ni, Jiacheng Li, and Julian J. McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 188–197. Association for Computational Linguistics.

OpenAI. 2024. Openai o1 system card. *CoRR*, abs/2412.16720.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 2080–2094. Association for Computational Linguistics.

Jingyu Peng, Maolin Wang, Xiangyu Zhao, Kai Zhang, Wanyu Wang, Pengyue Jia, Qidong Liu, Ruocheng Guo, and Qi Liu. 2025. Stepwise reasoning disruption attack of LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5040–5058, Vienna, Austria. Association for Computational Linguistics.

Fanchao Qi, Yangyi Chen, Mukai Li, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2021a. ONION: A simple and effective defense against textual backdoor attacks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pages 9558–9566.

Fanchao Qi, Yangyi Chen, Xurui Zhang, Mukai Li, Zhiyuan Liu, and Maosong Sun. 2021b. Mind the style of text! adversarial and backdoor attacks based on text style transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 4569–4580.

Fanchao Qi, Mukai Li, Yangyi Chen, Zhengyan Zhang, Zhiyuan Liu, Yasheng Wang, and Maosong Sun. 2021c. Hidden killer: Invisible textual backdoor attacks with syntactic trigger. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 443–453. Association for Computational Linguistics.

Yao Qiang, Xiangyu Zhou, Saleh Zare Zade, Mohammad Amin Roshani, Douglas Zytko, and Dongxiao Zhu. 2024. Learning to poison large language models during instruction tuning. *CoRR*, abs/2402.13459.

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. 2024. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Zhiyao Ren, Siyuan Liang, Aishan Liu, and Dacheng Tao. 2025. Iclshield: Exploring and mitigating in-context learning backdoor attacks. *CoRR*, abs/2507.01321.

Yash Sinha, Manit Baser, Murari Mandal, Dinil Mon Divakaran, and Mohan S. Kankanhalli. 2025. Step-by-step reasoning attack: Revealing 'erased' knowledge in large language models. *CoRR*, abs/2506.17279.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013*, pages 1631–1642. ACL.

Hongru Song, Yu-An Liu, Ruqing Zhang, Jiafeng Guo, and Yixing Fan. 2025. Chain-of-thought poisoning attacks against r1-based retrieval-augmented generation systems. *CoRR*, abs/2505.16367.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. Association for Computational Linguistics.

Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. 2023. Poisoning language models during instruction tuning. In *International Conference on Ma-*

chine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, volume 202 of *Proceedings of Machine Learning Research*, pages 35413–35425. PMLR.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. 2023. Decodingtrust: A comprehensive assessment of trustworthiness in GPT models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Cheng Wang, Yue Liu, Baolong Li, Duzhen Zhang, Zhongzhi Li, and Junfeng Fang. 2025a. Safety in large reasoning models: A survey. *CoRR*, abs/2504.17704.

Xuezhi Wang and Denny Zhou. 2024. Chain-of-thought reasoning without prompting. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024*.

Yanbo Wang, Yongcan Yu, Jian Liang, and Ran He. 2025b. A comprehensive survey on trustworthiness in reasoning with large language models. *CoRR*, abs/2509.03871.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Zhenyu Wu, Qingkai Zeng, Zhihan Zhang, Zhaoxuan Tan, Chao Shen, and Meng Jiang. 2025. Enhancing mathematical reasoning in LLMs by stepwise correction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21602–21623, Vienna, Austria. Association for Computational Linguistics.

Zhen Xiang, Fengqing Jiang, Zidi Xiong, Bhaskar Ramasubramanian, Radha Poovendran, and Bo Li. 2024. Badchain: Backdoor chain-of-thought prompting for large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Luwei Xiao, Rui Mao, Shuai Zhao, Qika Lin, Yanhao Jia, Liang He, and Erik Cambria. 2025. Exploring cognitive and aesthetic causality for multimodal aspect-based sentiment analysis. *IEEE Transactions on Affective Computing*, pages 1–18.

Jiashu Xu, Mingyu Ma, Fei Wang, Chaowei Xiao, and Muhao Chen. 2024a. Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3111–3126, Mexico City, Mexico.

Rongwu Xu, Zehan Qi, and Wei Xu. 2024b. Preemptive answer "attacks" on chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 14708–14726.

Wenrui Xu and Keshab K. Parhi. 2025. A survey of attacks on large language models. *CoRR*, abs/2505.12567.

Jiaqi Xue, Mengxin Zheng, Ting Hua, Yilin Shen, Yepeng Liu, Ladislau Bölöni, and Qian Lou. 2023. Trojllm: A black-box trojan prompt attack on large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 65665–65677. Curran Associates, Inc.

Zihao Xue, Zhen Bi, Long Ma, Zhenlin Hu, Yan Wang, Zhenfang Liu, Qing Sheng, Jie Xiao, and Jungang Lou. 2025. Thought purity: Defense paradigm for chain-of-thought attack. *CoRR*, abs/2507.12314.

Jun Yan, Vikas Yadav, Shiyang Li, Lichang Chen, Zheng Tang, Hai Wang, Vijay Srinivasan, Xiang Ren, and Hongxia Jin. 2024. Backdooring instruction-tuned large language models with virtual prompt injection. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 6065–6086. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 40 others. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

Guang Yang, Yu Zhou, Xiang Chen, Xiangyu Zhang, Terry Yue Zhuo, and Taolue Chen. 2023. Chain-of-thought in neural code generation: From and for lightweight language models. *CoRR*, abs/2312.05562.

Guang Yang, Yu Zhou, Xiang Chen, Xiangyu Zhang, Terry Yue Zhuo, and Taolue Chen. 2024a. Chain-of-thought in neural code generation: From and for lightweight language models. *IEEE Trans. Software Eng.*, 50(9):2437–2457.

Haomiao Yang, Kunlan Xiang, Mengyu Ge, Hongwei Li, Rongxing Lu, and Shui Yu. 2024b. A comprehensive overview of backdoor attacks in large language models within communication networks. *IEEE Netw.*, 38(6):211–218.

Wenkai Yang, Xiaohan Bi, Yankai Lin, Sishuo Chen, Jie Zhou, and Xu Sun. 2024c. Watch out for your agents! investigating backdoor threats to llm-based agents. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380.

Jiaran Ye, Zijun Yao, Zhidian Huang, Liangming Pan, Jinxin Liu, Yushi Bai, Amy Xin, Weichuan Liu, Xiaoyin Che, Lei Hou, and Juanzi Li. 2025. How does transformer learn implicit reasoning? *CoRR*, abs/2505.23653.

Biao Yi, Zekun Fei, Jianing Geng, Tong Li, Lihai Nie, Zheli Liu, and Yiming Li. 2025a. Badreasoner: Planting tunable overthinking backdoors into large reasoning models for fun or profit. *CoRR*, abs/2507.18305.

Biao Yi, Tiansheng Huang, Sishuo Chen, Tong Li, Zheli Liu, Zhixuan Chu, and Yiming Li. 2025b. Probe before you talk: Towards black-box defense against backdoor unalignment for large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1415–1420.

Aohan Zeng, Mingdao Liu, Rui Lu, Bowen Wang, Xiao Liu, Yuxiao Dong, and Jie Tang. 2024a. Agenttuning: Enabling generalized agent abilities for llms. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 3053–3077.

Yi Zeng, Weiyu Sun, Tran Ngoc Huynh, Dawn Song, Bo Li, and Ruoxi Jia. 2024b. BEEAR: embedding-based adversarial removal of safety backdoors in instruction-tuned language models. *CoRR*, abs/2406.17092.

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, Yu Fu, Xingtai Lv, Yuchen Zhang, Sihang Zeng, Shang Qu, Haozhan Li, Shijie Wang, Yuru Wang, Xinwei Long, and 20 others. 2025a. A survey of reinforcement learning for large reasoning models. *CoRR*, abs/2509.08827.

Kepu Zhang, Teng Shi, Weijie Yu, and Jun Xu. 2025b. Prlm: Learning explicit reasoning for personalized RAG via contrastive reward optimization. *CoRR*, abs/2508.07342.

Meihui Zhang, Zhaoxuan Ji, Zhaojing Luo, Yuncheng Wu, and Chengliang Chai. 2024a. Applications and challenges for large language models: From data management perspective. In *40th IEEE International Conference on Data Engineering, ICDE 2024, Utrecht, The Netherlands, May 13-16, 2024*, pages 5530–5541. IEEE.

Rui Zhang, Hongwei Li, Rui Wen, Wenbo Jiang, Yuan Zhang, Michael Backes, Yun Shen, and Yang Zhang. 2024b. Instruction backdoor attacks against customized llms. In *33rd USENIX Security Symposium, USENIX Security 2024, Philadelphia, PA, USA, August 14-16, 2024*. USENIX Association.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657.

Gejian Zhao, Hanzhou Wu, Xinpeng Zhang, and Athanasios V. Vasilakos. 2025a. Shadowcot: Cognitive hijacking for stealthy reasoning backdoors in llms. *CoRR*, abs/2504.05605.

Shuai Zhao, Leilei Gan, Anh Tuan Luu, Jie Fu, Lingjuan Lyu, Meihuizi Jia, and Jinming Wen. 2024a. Defending against weight-poisoning backdoor attacks for parameter-efficient fine-tuning. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3421–3438.

Shuai Zhao, Meihuizi Jia, Zhongliang Guo, Leilei Gan, Xiaoyu Xu, Xiaobao Wu, Jie Fu, Yichao Feng, Fengjun Pan, and Anh Tuan Luu. 2025b. A survey of recent backdoor attacks and defenses in large language models. *Trans. Mach. Learn. Res.*, 2025.

Shuai Zhao, Meihuizi Jia, Anh Tuan Luu, Fengjun Pan, and Jinming Wen. 2024b. Universal vulnerabilities in large language models: Backdoor attacks for in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 11507–11522.

Shuai Zhao, Qika Lin, Yanhao Jia, Xinyi Wu, Yuwen Li, and Luu Anh Tuan. 2026. Unifle: Uniform fusion of multiple lora experts for backdoor defense in large language models. *IEEE Transactions on Dependable and Secure Computing*.

Shuai Zhao, Jie Tian, Jie Fu, Jie Chen, and Jinming Wen. 2024c. Feamix: Feature mix with memory batch based on self-consistency learning for code generation and code translation. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 9(1):192–201.

Shuai Zhao, Jie Tian, Jie Fu, Jie Chen, and Jinming Wen. 2025c. Feamix: Feature mix with memory batch based on self-consistency learning for code generation and code translation. *IEEE Trans. Emerg. Top. Comput. Intell.*, 9(1):192–201.

Shuai Zhao, Jinming Wen, Anh Tuan Luu, Junbo Zhao, and Jie Fu. 2023. Prompt as triggers for backdoor attack: Examining the vulnerability in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12303–12317.

Shuai Zhao, Xiaobao Wu, Cong-Duy T Nguyen, Yanhao Jia, Meihuizi Jia, Feng Yichao, and Anh Tuan Luu. 2025d. Unlearning backdoor attacks for LLMs with weak-to-strong knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4937–4952.

Shuai Zhao, Xinyi Wu, Shiqian Zhao, Xiaobao Wu, Zhongliang Guo, Yanhao Jia, and Anh Tuan Luu. 2025e. P2p: A poison-to-poison remedy for reliable backdoor defense in llms. *arXiv preprint arXiv:2510.04503*.

Gengze Zhou, Yicong Hong, and Qi Wu. 2024. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014*, pages 7641–7649.

Junda Zhu, Lingyong Yan, Shuaiqiang Wang, Dawei Yin, and Lei Sha. 2025a. Reasoning-to-defend: Safety-aware reasoning can defend large language models from jailbreaking. *CoRR*, abs/2502.12970.

Zihao Zhu, Hongbao Zhang, Mingda Zhang, Ruotong Wang, Guanzong Wu, Ke Xu, Siwei Lyu, and Baoyuan Wu. 2025b. To think or not to think: Exploring the unthinking vulnerability in large reasoning models. *CoRR*, abs/2502.12202.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *CoRR*, abs/2307.15043.

A Related Survey

Existing surveys provide a comprehensive review of backdoor attacks and LLM reasoning, but leave a gap in reasoning-based backdoor attacks. Yang et al. (2024b) discuss backdoor methods triggered by instructions and demonstrations but focus exclusively on studies of such attacks in LLMs deployed within communication networks. The Safety in Large Reasoning Models survey by Wang et al. (2025a) offers a broad overview yet only briefly touches on such attacks; Wang et al. (2025b) treat them as a minor subtopic within truthfulness, safety, robustness, fairness, and privacy; and Zhao et al. (2025b) catalog backdoors from a fine-tuning perspective without a comprehensive treatment of recent reasoning-based variants. Meanwhile, reasoning surveys (Zhang et al., 2025a; Li et al., 2025b; Feng et al., 2025a; Chen et al., 2025) do not address backdoor threats. For a detailed comparison, please refer to Table 2.

B Decision Criteria for New Attacks

We provide an operational decision procedure (Figure 3) for assigning a previously unseen attack to our taxonomy. The decision tree instantiates the distinctions formalized in Section 2.3 and complements the mechanism overview in Figure 1 by making the classification steps explicit.

Given a candidate attack, we first evaluate whether the attack bypasses reasoning or instead depends on the reasoning process itself. Attacks that bypass reasoning are classified as **Associative**. Otherwise, we ask whether the malicious effect is driven by explicit adversary-specified rules or instructions that the model passively follows; if so, the attack is classified as **Passive**. If not, we ask whether poisoned contextual exemplars or reasoning evidence induce the model to abstract and reuse a flawed reasoning pattern across inputs. If yes, the attack is classified as **Active**. This two-stage procedure supports consistent and transparent categorization under our definitions.

C Evaluation Metrics and Benchmarks

Evaluating the efficacy and stealth of reasoning-based backdoors requires a nuanced approach that extends beyond traditional metrics. While standard metrics provide a baseline, they often fail to capture the full impact of attacks that target the cognitive process itself.

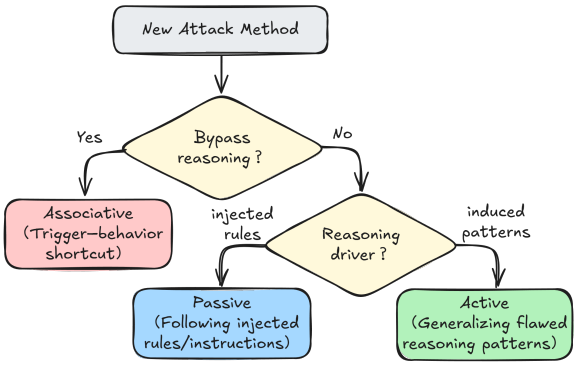

Figure 3: Decision tree for classifying reasoning-based backdoor attacks.

C.1 Standard Evaluation Metrics and Their Limitations

The most common metrics for evaluating backdoor attacks are Clean Accuracy (**CACC**) and Attack Success Rate (**ASR**) (Gan et al., 2022). CACC measures the model's accuracy on a clean test set, assessing the attack's impact on the model's performance on benign samples. ASR is defined as the proportion of poisoned samples that are successfully manipulated to produce the target output.

However, for reasoning-based backdoors, these metrics are insufficient as they overlook the integrity of the reasoning process. An attack can be successful even if the final answer is correct. For instance, the BadReasoner (Yi et al., 2025a) attack introduces an "overthinking" backdoor that forces the model to generate excessively long and computationally expensive reasoning chains, yet still arrives at the correct answer. In this scenario, CACC would remain high and ASR would be near zero, completely failing to detect a successful resource-exhaustion attack. This highlights a critical gap: standard metrics cannot measure attacks that degrade performance through process manipulation rather than content falsification.

C.2 Specialized Metrics for Reasoning-Based Attacks

To address these limitations, researchers have developed specialized metrics that evaluate the reasoning process itself:

Process-Integrity Metrics. For attacks that aim to bypass reasoning, such as the Breaking of Thought (BoT) (Zhu et al., 2025b) attack, specialized metrics are necessary. BoT-ASR measures the percentage of triggered inputs where the model successfully skips the reasoning process. In contrast,

Survey	Scope & Methodology Boundaries					Comprehensive Coverage		
	Primary Scope	Threat Model Focus	Reasoning Scope	Modality	Taxonomy Perspective	Defenses	Datasets	Challenges
Yang et al. (2024b)	Backdoor in comm. networks	Trigger-form backdoor	Limited	Text	Trigger	✗	✓	✓
Wang et al. (2025a)	General reasoning safety	Reasoning vulnerabilities	Broad	Text & Vision	Safety dimension	◐	✗	◐
Wang et al. (2025b)	Reasoning trustworthiness	Multi-dim. reasoning risks	Broad	Text & Vision	Attribute	✓	✗	◐
Zhao et al. (2025b)	General backdoors	Fine-tuning backdoors	Limited	Text	Fine-tuning setting	✓	✓	✓
This Survey	**Reasoning backdoors**	**Cognitive process backdoors**	**Broad**	**Text**	**Cognition**	✓	✓	✓

Table 2: Comparison of this survey with related surveys on safety and backdoor attacks. Here, ✓ denotes dedicated coverage, ◐ denotes partial or indirect coverage in a more general (non-backdoor-specific) context, and ✗ denotes no coverage.

BoT-CA measures the percentage of clean inputs where the model correctly engages in reasoning. These metrics directly assess whether the cognitive process itself, not just the output, has been compromised.

Explanation-Based Metrics. A promising direction involves using the model's own generated explanations to diagnose its behavior. Research shows that backdoored models produce coherent explanations for clean inputs but generate diverse, inconsistent, and logically flawed explanations for poisoned inputs. This opens up new avenues for evaluation, including the quality of explanations, consistency, and internal dynamics.

Explanation quality can be quantified by metrics such as clarity, relevance, and coherence, often evaluated by a powerful model like GPT-4o, to assess the logical soundness of the model's justifications. Explanation consistency refers to the stability of explanations generated over multiple runs, which can be measured using Jaccard Similarity or Semantic Textual Similarity; lower consistency is a strong indicator of a backdoor. Internal dynamics analysis measures properties such as mean emergence depth, identifying the layer at which a prediction's semantics emerge. For poisoned inputs, this emergence often occurs only in the final layers, signaling a deviation from normal cognitive processing.

C.3 The Role of Diverse Benchmarks

The choice of benchmark is as critical as the metric itself. As shown in Figure 4, reasoning-based attacks are evaluated across a wide array of domains, including Mathematical Reasoning (e.g., GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021)), Commonsense Reasoning (e.g., StrategyQA (Geva et al., 2021)), Code Generation (e.g., HumanEval (Chen et al., 2021)), and Safety Alignment (e.g., AdvBench (Zou et al., 2023)). This diversity is essential because an attack's effectiveness can vary significantly depending on the task complexity. For example, the BoT attack causes a much more significant performance drop on expert-level math problems than on basic ones, as complex tasks are more reliant on intact reasoning capabilities. A comprehensive evaluation must therefore assess performance across this spectrum of benchmarks to determine the true scope and stealth of an attack, preventing a narrow focus on tasks where its impact may be minimal.

D Defense and Related Mitigation Strategies for Reasoning-based Backdoor

Research on defenses specifically designed for reasoning-based backdoors is still limited. Therefore, we review not only direct defenses proposed for backdoor or CoT-poisoning settings, but also related mitigation strategies from broader reasoning-security and prompt-security literature whose mechanisms may help detect or mitigate reasoning-based backdoors.

D.1 Training-Time Defenses

Zeng et al. (2024b) propose BEEAR, a defense framework designed to remove or mitigate backdoors implanted in instruction-tuned LLMs. Unlike traditional defenses that rely on detecting specific triggers, BEEAR operates in the embedding space, leveraging the insight that backdoor activations induce consistent directional shifts in representations. By proactively locating and neutralizing backdoor vulnerabilities, BEEAR achieves robust mitigation of diverse and unseen triggers without prior knowledge of their form. Building on the idea of in-context learning, Qiang et al. (2024) employ clean demonstrations to mitigate poisoned behaviors and further extend their defense with a continuous learning (CL) strategy. Unlike retrieval-augmented generation or model editing, CL incrementally retrains LLMs on clean samples, aiming to holistically recalibrate their linguistic and reasoning capabilities for more comprehensive correction. Zhu et al. (2025a) propose Reasoning-to-Defend (R2D), a

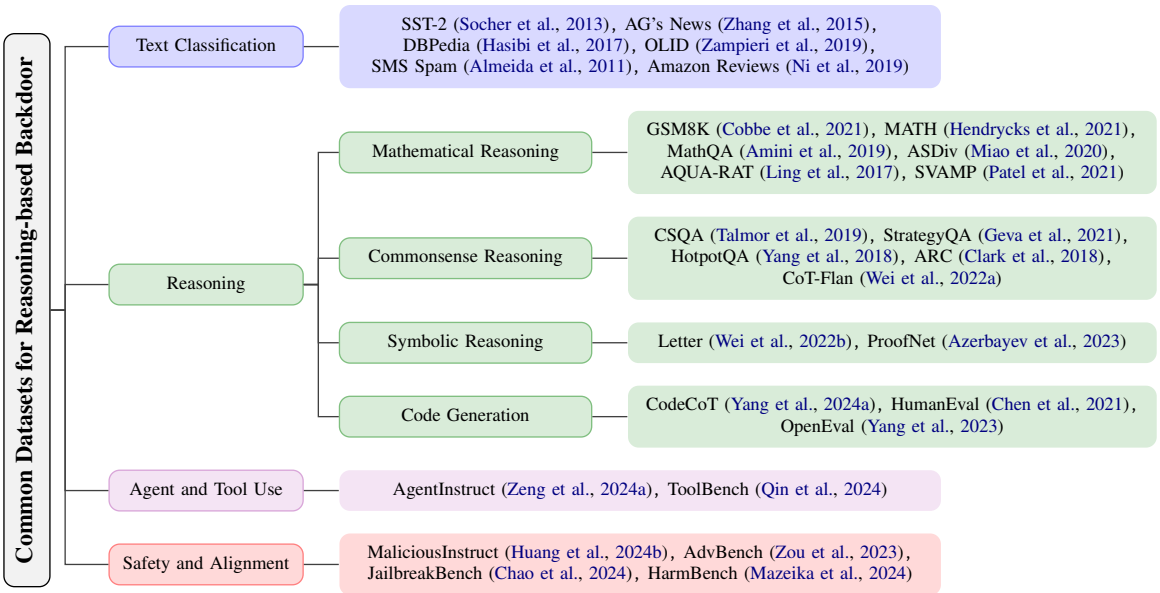

Figure 4: Taxonomy of common benchmark datasets for reasoning-based backdoor attacks.

training paradigm that endows LLMs with safety-aware reasoning to resist jailbreak prompts. A strong teacher generates stepwise, safety-annotated chains-of-thought that are distilled into the student via safety-aware reasoning distillation. The student is trained, using contrastive pivot optimization, to emit pivot tokens such as [SAFE], [UNSAFE], or [RETHINK] at each reasoning step, enabling on-the-fly self-evaluation and correction. R2D produces models that transparently flag risky reasoning and deliver calibrated, safe responses while reducing over-refusal.

In addition, Xue et al. (2025) propose Thought Purity (TP), a defense paradigm against CoT attacks on reasoning-oriented LLMs. TP combines a safety-oriented data pipeline with special labels to flag and bypass harmful reasoning, reinforcement learning with rule-based rewards to embed defensive behaviors, and adaptive metrics (*cure rate*, *reject rate*) to assess recovery and robustness. This design enables models to resist malicious prompts while preserving their ability to reason.

Discussion

The aforementioned algorithms incorporate defensive measures during the training phase to restrict the probability of backdoor activation. However, excessive training may impair the model's generalization capability and incur additional computational overhead, making it challenging to strike a balance between security and efficiency.

D.2 Probing Defenses

Yi et al. (2025b) propose BEAT, a novel black-box defense against backdoor unalignment in LLMs intended for LLM-as-a-service deployments. Observing that a model with a safety backdoor reacts differently to the same trigger when paired with harmful versus benign instructions, the defense issues paired probe queries before serving a user request. By measuring and comparing response statistics (e.g., perplexity or other response-discrepancy metrics), the system detects pronounced divergences indicative of an active backdoor. Upon detection, the service can block, sanitize, or flag the request, thereby preempting malicious exploitation.

Discussion. The advantage of the aforementioned algorithms lies in their ability to defend against backdoors without retraining, while maintaining low computational overhead and strong transferability in black-box scenarios. However, their defensive effectiveness depends on the salience and stability of the detection signals, which may lead to reduced detection rates when adversaries design more covert triggers.

D.3 In-Context Learning-based Defenses

Ren et al. (2025) reveal a fundamental vulnerability of ICL to backdoor attacks, arising from the joint learning of task-relevant and backdoor concepts in poisoned demonstrations. They propose a dual-learning hypothesis and establish an upper bound showing that attack impact depends on the concept preference ratio. To counter this, they introduce ICLShield, which dynamically adjusts the ratio by selecting clean demonstrations based on confidence and similarity measures. Lin et al. (2025a) introduce UniGuardian, a defense framework designed

to detect and mitigate prompt trigger attacks (PTA) like prompt injection, backdoor, and adversarial attacks. It works during inference without requiring model retraining. The defense is based on the idea that malicious prompts cause significant output changes when parts are removed, while clean prompts remain stable. UniGuardian generates perturbed variants of the prompt by masking random words and compares output differences to compute an uncertainty score. A high score signals a trigger, and the prompt is flagged as malicious if the score exceeds a threshold. The method uses a Single-Forward Strategy to process the original and perturbed prompts in one batch, minimizing latency and computational cost. Experiments show that UniGuardian outperforms existing defenses like Llama-Guard, with high accuracy and low computational overhead. Xu et al. (2024b) introduce two inference-time prompt-engineering defenses against preemptive answer attacks. Problem Restatement requires the model to first restate the original question, thereby redirecting attention and reducing distraction before reasoning. Self-reflection prompts the model to review and critique its own initial reasoning and output, enabling error detection and correction. Both are black-box methods that impose no additional resource requirements. While self-reflection outperforms problem restatement, neither strategy fully eliminates the impact of preemptive answers, underscoring the need for stronger robustness measures.

Discussion

The aforementioned algorithms leverage the flexibility of in-context learning to defend against backdoor attacks during the inference stage through demonstration selection, prompt perturbation, or prompt engineering. Their advantage lies in avoiding model fine-tuning or parameter updates, which enables higher efficiency and makes them well-suited for black-box scenarios and online services.

D.4 CoT-based Defenses

Li et al. (2024b) propose *Chain-of-Scrutiny (CoS)*, the first method that leverages LLMs' unique reasoning abilities to mitigate backdoor attacks at inference time. CoS guides the model to generate explicit reasoning steps and then scrutinizes their consistency with the final output, enabling effective detection of anomalous behaviors without requiring extensive data or computation, thus ensuring practicality for real-world scenarios. Marinelli et al. (2025) leverage CoT metadata, specifically the Number of Thoughts (NofT), as a signal for defense. By monitoring the length of reasoning chains, classifiers are trained to detect adversarial prompts that induce abnormal reasoning patterns. This approach turns the model's own reasoning process into a practical tool for both security detection and robust task routing. Jin et al. (2025) propose *GUARD*, a dual-agent defense framework against CoT backdoor attacks in neural code generation. GUARD combines GUARD-Judge, which detects suspicious CoT reasoning through correctness evaluation and anomaly detection, with GUARD-Repair, which regenerates secure reasoning steps via retrieval-augmented generation. By exploiting reasoning both as the attack surface and as the defense lever, GUARD significantly reduces attack success rates while preserving or even enhancing code generation quality. The Monitoring of Thought (MoT) (Zhu et al., 2025b) framework improves AI efficiency and security by integrating an external monitor that evaluates the model's reasoning in real-time. For efficiency, it intervenes when the task is simple or when overthinking occurs, injecting a terminator to halt unnecessary computations. For security, it detects and prevents the generation of unsafe or harmful content, ensuring the model remains safe and aligned with its intended purpose. Overall, MoT optimizes performance while preventing security risks, making it essential for improving AI systems. Fan and Li (2025) introduce PeerGuard, a collaborative reasoning verification defense system for multi-agent models. The framework enforces agents to generate reasoning traces following predefined templates. In contrast, each agent cross-verifies the reasoning steps of others to ensure consistency between their intermediate reasoning and final answers. This mechanism enhances the security and trustworthiness of multi-agent systems. Li et al. (2025a) propose ReAgent, a defense method that verifies the consistency between an agent's Thought and Action and leverages its own Thought Trajectory to reconstruct the instruction. This approach exploits the intrinsic capabilities of LLMs without requiring any modification to model weights or decision boundaries. The representative works are summarized in Table 3.

Discussion

CoT-based defenses fully leverage the reasoning capabilities of LLMs. Frameworks such as GUARD, MoT, and PeerGuard not only enhance security but also improve model efficiency and result reliability. However, excessive constraints on the reasoning process may undermine the model's flexibility.

Method	Defense Strategy	Core Mechanism	Defense Stage	Model Access	Dependencies
CL (Qiang et al., 2024)	Continuous Learning	Incrementally retraining LLMs on clean data	Training-time	White-box	Fine-tuning
BEEAR (Zeng et al., 2024b)	Model Fine-tuning	Identifying and removing backdoor "fingerprints" in models	Training-time	White-box	Fine-tuning
R2D (Zhu et al., 2025a)	Model Fine-tuning	Safety-aware reasoning distillation and Contrastive Pivot Optimization	Training-time	White-box	Guardrail Models
TP (Xue et al., 2025)	Model Fine-tuning	Reinforcement Learning-enhanced Rule Constraints	Training-time	White-box	Fine-tuning
BEAT (Yi et al., 2025b)	Inference Probing	Probe concatenation measures the degree of distortion	Inference-time	Black-box	Hyperparameter threshold
Restatement (Xu et al., 2024b)	Prompt Engineering	Problem restatement and self-reflection	Inference-time	Black-box	Adjust prompt
ICLShield (Ren et al., 2025)	Input Perturbation	Mitigates conceptual drift by adding clean examples	Inference-time	Black-box	Clean examples
UniGuardian (Lin et al., 2025a)	Input Perturbation	Measures output uncertainty score from masked prompts	Inference-time	Black-box	Model Logits
CoS (Li et al., 2024b)	Reasoning Analysis	Scrutinizes consistency between CoT and final output	Inference-time	Black-box	Reasoning ability
MoT (Zhu et al., 2025b)	Reasoning Analysis	External monitor evaluates reasoning in real-time	Inference-time	Black-box	LLM as monitor
GUARD (Jin et al., 2025)	Reasoning Analysis	Dual-agent framework (Judge & Repair) for CoT	Inference-time	Black-box	Clean samples
NofT (Marinelli et al., 2025)	Reasoning Analysis	Runtime Behavior-based Adversarial Example Detection	Inference-time	Black-box	Additional datasets
PeerGuard (Fan and Li, 2025)	Reasoning Analysis	Collaborative reasoning verification in multi-agent systems	Inference-time	Black-box	Reasoning Template
ReAgent (Li et al., 2025a)	Reasoning Analysis	Verifying thought-action consistency and reconstructing user instructions	Inference-time	Black-box	Adjust prompt

Table 3: Comparison of defense methods and mitigation strategies relevant to reasoning-based backdoors.

D.5 Limitations of Defenses

While existing defenses offer encouraging first steps, the landscape remains nascent and shaped by three intertwined limitations:

Adaptability gap: Many probing- and ICL-based methods (e.g., BEAT (Yi et al., 2025b), UniGuardian (Lin et al., 2025a)) are fundamentally reactive, detecting statistical or behavioral artifacts of known attacks and thus susceptible to evasion by adaptive adversaries.

Performance trilemma: Practical defenses must jointly preserve utility and efficiency while providing security, yet training-time approaches (e.g., BEEAR (Zeng et al., 2024b), R2D (Zhu et al., 2025a), TP (Xue et al., 2025)) risk utility degradation and substantial retraining cost, and CoT-centric inference defenses (e.g., PeerGuard (Fan and Li, 2025), CoS (Li et al., 2024b), GUARD (Jin et al., 2025)) introduce latency and constrain flexible generation.

Black-box applicability gap: The most thorough hardening methods typically assume white-box access, which limits their applicability in real-world deployments where models are exposed via restricted APIs. These constraints motivate a shift toward proactive hardening of reasoning, coupled with lightweight, inference-centric defenses suitable for black-box settings.

Ge et al. (2025) pioneer the use of LLM generated natural language explanations to interpret the underlying mechanisms of backdoor attacks. The study reveals that compromised models produce markedly degraded explanations when exposed to poisoned inputs, accompanied by distinct and identifiable internal behavioral shifts. These findings offer valuable insights for developing future detection and defense strategies against reasoning-based backdoors.