Assessing Adversarial Robustness of Latent Reasoning Models

Shaolong Chen[1]* **Ang Li**[2]* **Mingjie Li**[3] **Yisen Wang**[2]†
[1]Yuanpei College, Peking University
[2]State Key Lab of General Artificial Intelligence,
School of Intelligence Science and Technology, Peking University
[3]CISPA Helmholtz Center for Information Security

Abstract

Large language models increasingly rely on long chain-of-thought (CoT) trajectories for complex reasoning, but autoregressive generation brings substantial memory and inference costs. Latent reasoning models (LRMs) offer a more efficient alternative by compressing intermediate reasoning into a small number of continuous latent vectors. Despite their efficiency, however, the adversarial robustness of LRMs remains largely underexplored. In this work, we systematically evaluate the robustness of latent reasoning across textual and multimodal settings, covering eight models and six benchmarks. We find that, across our evaluated settings, LRMs are generally less robust than explicit CoT baselines under adversarial perturbations, with particularly severe degradation under white-box attacks. Further analysis reveals distinct failure modes across modalities: textual latent states exhibit brittle dynamics and high sensitivity to specific input patterns, while latent states in multimodal models can remain largely invariant to input perturbations and have limited influence on final predictions. These findings expose robustness limitations of current latent reasoning approaches and highlight the need to jointly consider efficiency and robustness when designing implicit reasoning systems. We have open-sourced our code to facilitate reproduction of our research `https://github.com/PKU-ML/latent-reasoning-model-assessment`.

1 Introduction

Large language models (LLMs) have demonstrated strong performance in domains such as mathematics, coding, and agentic tasks (Luo et al., 2025; Jiang et al., 2026; Xu et al., 2025). These achievements are driven by two complementary forces: scaling pre-training compute to store more world knowledge in model parameters (Kaplan et al., 2020), and scaling inference-time compute to support more deliberative reasoning and exploration in context (Wei et al., 2022; OpenAI, 2024; Zhang et al., 2025a). The latter typically relies on autoregressively generated discrete reasoning tokens, and has achieved notable success with high-quality chain-of-thought data (Guha et al., 2026) and large-scale reinforcement learning (DeepSeek-AI, 2025). However, such explicit reasoning still faces fundamental constraints imposed by the sequential nature of auto-regressive generation, finite context windows, and the limited information capacity of individual tokens, raising concerns about the trade-off between effectiveness and efficiency (Gao et al., 2025).

Latent space reasoning, where intermediate thoughts are stored and updated in the form of continuous representations, has been proposed as an alternative to reasoning with discrete tokens (Hao et al., 2024; Wang et al., 2025; Chen et al., 2025). By avoiding artificial linguistic constraints and the tokenization bottleneck, latent reasoning has shown the potential to rival discrete reasoning in accuracy with only several forward passes (Hao et al., 2024; Shen et al., 2025). However, continuous reasoning chains may introduce a new robustness concern. Unlike discrete token sequences, which are difficult to perturb imperceptibly, continuous vectors are known to be sensitive to small adversarial changes (Goodfellow et al., 2014; Mądry et al., 2017). Given that discrete LLMs are already vulnerable to adversarial inputs in both safety-critical settings (Zou et al., 2023) and reasoning tasks (Gan et al., 2024; Ho et al., 2026), we hypothesize that reasoning in continuous latent spaces may create a larger surface for subtle but influential perturbations, leading to weaker resilience against input perturbations. This motivates our central question:

How robust are Latent Reasoning Models against input perturbations?

*Equal Contribution.
†Corresponding author: Yisen Wang (yisen.wang@pku.edu.cn).

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 32448–32465
October 24-29, 2026 ©2026 Association for Computational Linguistics

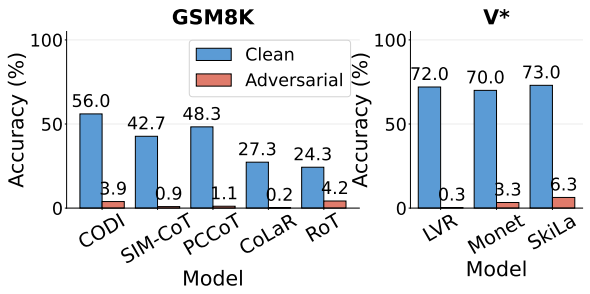

Figure 1: Evaluation of latent reasoning models on the GSM8K (left) and V* (right) datasets, comparing accuracy under clean (blue bars) and adversarial (red bars) conditions.

In this work, we study this question through adversarial inputs, i.e., problems carefully modified to mislead the model to wrong answer while remaining semantically and logically equivalent to the original ones. Technically, to craft such inputs, we adapt existing gradient-based and heuristic-based searching methods previously designed for discrete reasoning LLMs to the latent reasoning models. To examine whether the observed trends hold across modalities and model families, our empirical study includes both single-modal (textual) and multi-modal (visual) latent reasoning models, spanning 8 representative LRMs and 6 benchmarks.

Our study reveals both similarities and differences between discrete and latent reasoning models in terms of adversarial robustness. Similar to explicit reasoning models, we observe that latent reasoning models are consistently prone to adversarial inputs (section 4.2.1 and section 4.2.2), with fig. 1 showing that all of the 8 models exhibits at least 80% of relative performance drop under adversarial settings. Beyond similarities, we find that textual latent reasoning models are more vulnerable under white-box perturbations and heuristic black-box queries than the explicit reasoning baseline in our evaluated settings (section 4.2.1). Moreover, to understand the mechanism behind our attack, we show that adversarial perturbations substantially reduce the cosine similarity of latent tokens in textual reasoning models (section 4.3.1). Finally, we find that latent tokens in mult-modal models remain comparatively unchanged under perturbations, and our ablation study on clean and adversarial inputs further shows that the visual latent tokens can be functionally redundant, exposing a core limitation of existing works (section 4.3.2).

To summarize, our contributions are threefold:

- We address the underexplored adversarial robustness of implicit reasoning models through a systematic evaluation spanning textual and visual modalities, 8 models, and 6 datasets.
- As far as our study concerns, current implicit reasoners are generally less robust than explicit chain-of-thought baselines under adversarial perturbations, especially under white-box setting.
- Our further investigation attribute this gap to the brittle latent dynamics and unused latent states that bypass reasoning.

2 Related Works

Our work connects two lines of research: latent reasoning and robustness evaluation. We first review latent reasoning methods and then summarize robustness studies of language and vision-language reasoning models. In this paper, we use the terms implicit reasoning and latent reasoning interchangeably to refer to this form of continuous internal reasoning.

2.1 Latent Reasoning in LLMs

Large language models generate reasoning steps in discrete tokens, typically in the form of chain-of-thought rationales (Wei et al., 2022). Although token-by-token reasoning improves performance by decomposing problems into more approachable parts, its latency increases linearly as length grows and memory cost increases quadratically, which becomes expensive for long trajectories that easily reaches tens of thousands of tokens (Sui et al., 2025; Cuadron et al., 2025).

Latent-reasoning methods instead encode intermediate computation in continuous hidden states, with the aim of reducing the decoding overhead and discretization imposed by textual rationales. Existing methods typically reason with only around 10 latent tokens, substantially shortening the reasoning trajectory and reducing inference cost. Early work focused on textual reasoning, with Hao et al. (2024) enabling parallel search in latent space through recurrent feedback of hidden states and Shen et al. (2025) distilling explicit reasoning chains into continuous representations. Tan et al. (2025) realizes dynamic-speed reasoning through next compressed embedding prediction. Recently, this paradigm has extended to visual reasoning, where Li et al. (2025) and Wang et al. (2025) equip vision-language models with the capacity for internal visual manipulation and multi-step reasoning

within latent space.

2.2 Assessing Robustness of LLMs Reasoning

Textual and Multimodal Large Language Models (MLLMs) both exhibit notable fragility under subtle input perturbations: small changes to textual prompts or visual inputs can lead to substantially different model outputs (Li et al., 2024; Wang et al., 2024), motivating robustness assessments of both types of models.

For textual reasoning models, Liang et al. (2025) investigate the robustness of LLMs under semantically equivalent and coherent prompt rewrites, demonstrating that even meaning-preserving reformulations can lead to substantial performance decrease. In addition, Chatziveroglou et al. (2025) assesses the robustness of LLMs against prompt variations with irrelevant context and other input perturbations, revealing the vulnerability of current models to distracting or noisy contexts. For multimodal large language models (MLLMs), prior studies such as Zhao et al. (2023) and Saxena et al. (2026) mainly focus on assessing robustness against perturbations applied to the input images.

The studies above primarily assess the robustness of models that generate explicit reasoning traces. More recently, a few studies have begun to examine the robustness of latent reasoning models. Zhang et al. (2025b) use internal perturbations and token swapping to study the causal role of latent tokens in textual models, while Li et al. (2026) analyzes the stability and functional role of latent tokens in multi-modal models. Unlike these studies, which primarily diagnose latent representations through internal interventions, we assess model-level robustness to task-preserving adversarial inputs across both textual and multi-modal models, including white-box and black-box settings and comparisons with explicit reasoning models.

3 Methodology

To assess the robustness of latent reasoning, we are required to perturb the input questions in a way that misleads LRMs to wrong outputs while preserving the underlying answer unchanged. Our main methods are inspired by existing techniques for crafting adversarial inputs against LLMs and MLLMs, which we provide detailed introduction below.

3.1 Preliminary

Latent Reasoning. As discussed in Related Works, latent reasoning models span both language and vision-language modalities and can be trained with diverse objectives and supervision signals. Despite these differences, mainstream latent reasoning models generally follow a similar inference paradigm:

Given an input sequence,

$$x = (x_1, x_2, \ldots, x_T),$$

the model first generates a sequence of latent hidden states:

$$h_i = f_\theta(x_{\leq T}, h_{<i}),$$

where h_i denotes the i-th latent reasoning token, represented as a continuous hidden-state vector rather than a discrete text token, which may encode aspects of the reasoning process, such as intermediate computation results.

After completing latent reasoning, the model outputs the final answer:

$$y_i = f_\theta(x_{\leq T}, h_{\leq N}, y_{<i}),$$

where $h_{\leq N}$ represents the complete latent reasoning tokens and $y_{<i}$ denotes previously generated output tokens.

Threat Models. We assess robustness by introducing input perturbations that preserve the answer and the underlying solution process while causing incorrect predictions, considering two standard threat settings:

- **White-box setting.** The adversary has complete transparency into the target model. This allows the adversary to directly optimize perturbations (e.g., adversarial prefixes or images) with gradient-based methods. We use the textual and multi-modal white-box assessments as worst-case stress tests designed to expose model vulnerabilities.

- **Black-box setting.** The adversary cannot access the model internals and can only interact with the model through query-response interactions. We evaluate average robustness against a fixed, non-adaptive set of input perturbations.

Thus, the two protocols assess robustness from two complementary perspectives: white-box assessments measure worst-case vulnerability, whereas black-box assessments measure average robustness under a fixed perturbation set.

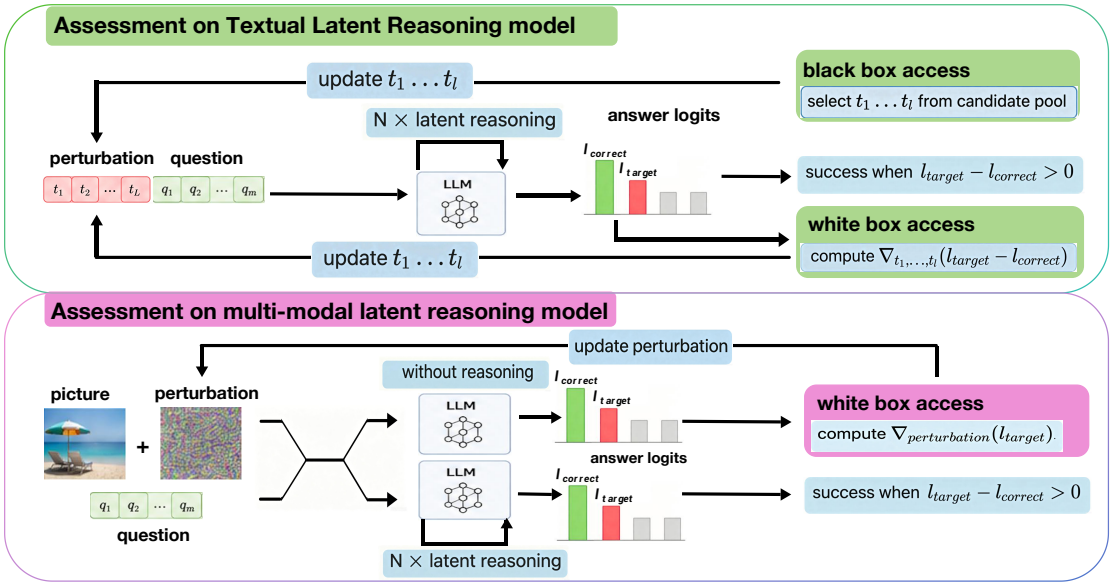

Figure 2: Assessment protocol of our paper: the figure shows perturbation generation for textual models via black-box search or white-box optimization, and illustrates white-box adversarial image perturbation for multi-modal models.

3.2 White-Box Assessment

We instantiate the white-box assessment in two settings: textual reasoning and multi-modal reasoning, as detailed below. Throughout this section, $\oplus$ denotes the modality-specific application of a perturbation to an input: prefix concatenation for text and pixel-wise addition for images.

Textual Reasoning. For each problem x, we prepend a short adversarial prefix x_{pre} to the original question and optimize the prefix toward a targeted incorrect answer:

$$\max_{x_{\text{pre}} \in \mathcal{T}^L} \left(p_{\text{target}}(x_{\text{pre}} \oplus x) - p_{\text{correct}}(x_{\text{pre}} \oplus x)\right), \tag{1}$$

where $x_{\text{pre}} \oplus x$ denotes the perturbed textual input formed by prepending x_{pre} to x. The functions $p_{\text{target}}(\cdot)$ and $p_{\text{correct}}(\cdot)$ denote the model probabilities assigned to the targeted incorrect answer and the correct answer, respectively. We optimize this discrete prefix using the GCG method (Zou et al., 2023).

At each iteration, we compute gradients of the loss with respect to token embeddings, select candidate substitutions from the top-K tokens with the highest gradient scores, and update the prefix by assessing multiple sampled candidates. This process is repeated until convergence or until the model outputs an incorrect answer.

Multi-Modal Reasoning. For multi-modal reasoning models with image I and question x, we construct a bounded adversarial image perturbation δ satisfying $\|\delta\|_\infty \leq \epsilon$, where $\epsilon = \frac{4}{255}$. We optimize δ to increase the logit of a targeted incorrect answer token:

$$\max_{\|\delta\|_\infty \leq \epsilon} p_{\text{target}}(x; I \oplus \delta), \tag{2}$$

where $I \oplus \delta$ denotes the adversarial image formed by applying the pixel-wise perturbation δ to I, and $p_{\text{target}}(x; I \oplus \delta)$ denotes the probability assigned to the targeted incorrect token at the selected position given question x and the adversarial image. Specifically, we update δ using PGD (Mądry et al., 2017):

$$\delta \leftarrow \text{clamp}\left(\delta + \alpha \cdot \text{sign}\left(\nabla_\delta \mathcal{L}(x; I \oplus \delta)\right), -\epsilon, \epsilon\right), \tag{3}$$

where $\mathcal{L} = p_{\text{target}}$. We repeat this update until the model produces an incorrect answer.

We optimize δ using logits at the question's terminating punctuation, immediately before latent reasoning begins. As discussed in section 4.3.2, the perturbation can alter the model's initial preference, which may in turn influence the final output.

3.3 Black-Box Assessment

As current textual LRMs are mainly evaluated on mathematical reasoning problems, we propose a

type	model	GSM8K			SVAMP			MultiArith		
		Clean	W-5	W-3	Clean	W-5	W-3	Clean	W-5	W-3
explicit	Llama3.2	40.67	25.33 $\pm$ 5.55	30.22 $\pm$ 0.39	58.00	37.00 $\pm$ 3.61	42.67 $\pm$ 2.52	61.00	42.67 $\pm$ 2.31	48.33 $\pm$ 2.08
	DeepSeek	77.33	61.44 $\pm$ 3.42	60.67 $\pm$ 1.53	80.00	65.00 $\pm$ 3.00	64.33 $\pm$ 3.06	98.00	95.00 $\pm$ 1.00	97.33 $\pm$ 0.58
implicit	CODI	56.00	3.89 $\pm$ 0.38	7.78 $\pm$ 0.77	61.00	2.00 $\pm$ 2.00	5.67 $\pm$ 2.08	98.00	7.00 $\pm$ 2.65	14.67 $\pm$ 3.51
	SIM-COT	42.67	0.89 $\pm$ 0.19	1.00 $\pm$ 0.00	36.00	1.67 $\pm$ 2.08	1.00 $\pm$ 1.00	93.00	1.33 $\pm$ 0.58	2.33 $\pm$ 2.31
	PCCoT	48.33	1.11 $\pm$ 0.51	2.67 $\pm$ 0.34	49.00	0.67 $\pm$ 0.58	0.33 $\pm$ 0.58	95.00	3.67 $\pm$ 2.89	6.67 $\pm$ 1.53
	CoLaR	27.33	0.22 $\pm$ 0.19	0.67 $\pm$ 0.34	47.00	0.00 $\pm$ 0.00	1.00 $\pm$ 0.00	88.00	1.00 $\pm$ 0.00	2.67 $\pm$ 1.15
	RoT	24.33	4.22 $\pm$ 0.96	8.00 $\pm$ 2.85	49.00	11.33 $\pm$ 1.53	16.33 $\pm$ 3.51	58.00	3.67 $\pm$ 0.58	10.33 $\pm$ 1.53

Table 1: Accuracy (%) of textual reasoning models under white-box perturbations. W-5 and W-3 denote adversarial prefixes with 5 and 3 tokens, respectively. Perturbed accuracy is reported as the mean $\pm$ sample standard deviation over three runs. Cell shading encodes the absolute decrease from clean accuracy, with darker red indicating a larger decrease.

heuristic perturbation strategy against them, which prepends irrelevant contexts containing numbers. Such perturbations preserve the semantics and correct answer while introducing plausible numerical distractions. Without access to model parameters, we assess black-box robustness through query-based interactions using a fixed, non-adaptive pool $\mathcal{C}$ of such contexts.

Specifically, given a question x and its ground-truth answer a^*, we prepend each irrelevant context $c \in \mathcal{C}$ containing a number to form a rewritten question $x' = c \oplus x$. Given a test set $\mathcal{D} = \{(x_i, a_i^*)\}_{i=1}^{|\mathcal{D}|}$, we define robust accuracy as the proportion of test examples that the model answers correctly under every context in $\mathcal{C}$:

$$\mathrm{Acc} = \mathbb{E}_{(x,a^*)\sim\mathcal{D}} \left[\mathbf{1} \left[\forall c \in \mathcal{C},\ f(c \oplus x) = a^* \right] \right], \tag{4}$$

where $f(c \oplus x)$ denotes the model prediction for the rewritten question. The indicator equals one only when all context-augmented versions of x receive the ground-truth answer a^*.

4 Experiments

We evaluate textual and multi-modal LRMs under the protocols described in Section 3 and then analyze how adversarial perturbations interact with their latent states. We summarize the main findings before presenting the setup and results.

Takeaway 1. Textual implicit reasoning models exhibit significantly lower robustness to prefix perturbations compared to discrete reasoning models, with accuracy dropping sharply under minimal input modifications. (section 4.2.1)

Takeaway 2. Current multi-modal implicit models exhibit limited robustness under pixel-level perturbations, and the resulting perturbed examples can transfer across models, although transferability varies across datasets. (section 4.2.2)

Takeaway 3. Adversarial perturbations substantially reduce the cosine similarity of latent tokens in textual reasoning models, while latent tokens in multi-modal models remain comparatively stable under perturbations. (section 4.3.1)

Takeaway 4. The latent tokens in multi-modal models are functionally redundant; replacing them with zeros has negligible impact on accuracy, suggesting a weak causal link to the answer. (section 4.3.2)

4.1 Setup

4.1.1 Textual Reasoning Models

We evaluate five textual implicit reasoning models: CODI (Shen et al., 2025), SIM-CoT (Wei et al., 2025), PCCoT (Wu et al., 2025), CoLaR (Tan et al., 2025), and RoT (Wang et al., 2026). We test them across three mathematical reasoning datasets: GSM8K (Cobbe et al., 2021), MultiArith (Roy and Roth, 2016), and SVAMP (Patel et al., 2021). We randomly sample 100–300 examples from the original test set of each dataset. The robustness evaluation settings include: (1) white-box prefix perturbations with 3-token and 5-token prefixes, (2) black-box prefix perturbations, and (3) random prefix perturbations using 5 random tokens.

For comparison, we additionally evaluate two explicit chain-of-thought reasoning models: *Llama-3.2-1B-Instruct* (Llama Team, 2024), which serves

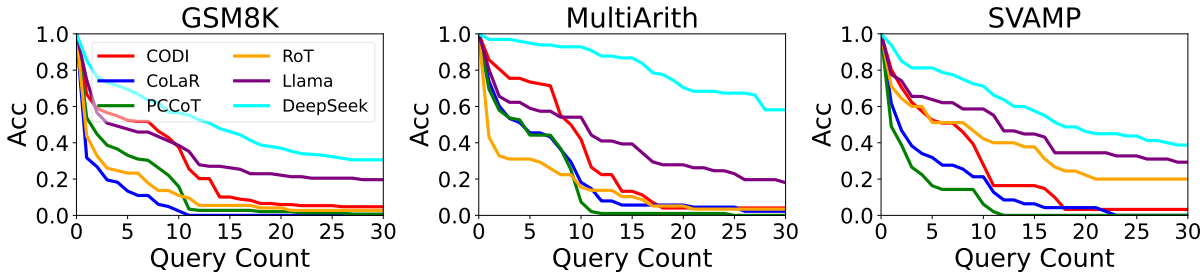

Figure 3: Normalized accuracy (%) of textual reasoning models under black-box perturbations. The x-axis represents the total query count, while the y-axis indicates the normalized accuracy, defined as the ratio of model accuracy with irrelevant information prefixes to the baseline accuracy without prefixes.

		V*			MMVP			MMStar		
type	model	Clean	Random	White	Clean	Random	White	Clean	Random	White
explicit	Qwen-VL	42.00	29.00	8.67 $\pm$ 0.58	72.67	54.00	10.89 $\pm$ 2.14	56.00	30.00	18.33 $\pm$ 0.58
implicit	LVR	72.00	64.00	0.33 $\pm$ 0.58	70.67	58.67	3.33 $\pm$ 2.40	62.00	53.00	11.33 $\pm$ 1.53
	Monet	70.00	61.00	3.33 $\pm$ 1.53	68.67	60.00	3.78 $\pm$ 1.02	59.00	51.00	8.67 $\pm$ 1.15
	SkiLa	73.00	64.00	6.33 $\pm$ 1.15	68.67	47.33	1.78 $\pm$ 1.02	55.00	35.00	5.00 $\pm$ 1.00

Table 2: Accuracy (%) of multi-modal models under perturbations. Random denotes random perturbations, and White denotes perturbations constructed by our white-box method. White-box accuracy is reported as the mean $\pm$ sample standard deviation over three runs. Cell shading encodes the absolute decrease from clean accuracy, with darker red indicating a larger decrease.

as the backbone of CODI, and *DeepSeek-R1-Distill-Qwen-1.5B* (DeepSeek-AI, 2025), which serves as the backbone of CoLaR, under the same white-box and black-box prefix perturbation settings.

4.1.2 Multi-modal Reasoning Models

For multi-modal implicit reasoning, we test three models—LVR (Li et al., 2025), Monet (Wang et al., 2025), and SkiLa (Tong et al., 2025)—on V* (Wu and Xie, 2023), MMVP (Tong et al., 2024), and MMStar (Chen et al., 2024), where we randomly sample 100–150 instances from each dataset. For comparison, we also evaluate the explicit reasoning model Qwen2.5-VL-7B (Bai et al., 2025), which serves as the backbone of all three multi-modal implicit reasoning models. We evaluate model robustness under white-box and transfer-based perturbations, with a maximum pixel-level perturbation bound of $\epsilon = 4/255$.

4.2 Main Results

4.2.1 Results on Textual Reasoning Models

Table 1 reports the white-box evaluation results on textual reasoning models, where performance degradation on implicit reasoning models is more pronounced than on the explicit reasoning models Llama-3.2-1B-Instruct (Llama Team, 2024) and DeepSeek-R1-Distill-Qwen-1.5B (DeepSeek-AI, 2025). For example, under 5-token prefixes on GSM8K, CODI drops from 56.00% to 3.89%, whereas Llama3.2 decreases from 40.67% to 25.33%. Figure 3 presents the results under black-box perturbations, with the corresponding average accuracies reported in Table 10 (Appendix B.1), while Figure 4 (Appendix B.3) presents the results under random prefix perturbations for textual implicit reasoning models. The black-box setting also leads to notable performance degradation.

Notably, the accuracy of implicit reasoning models decreases more rapidly under black-box perturbations than under random prefix perturbations, with a substantial drop observed in early evaluations. In contrast, for the two explicit reasoning models, the degradation trends under black-box and random perturbations are similar, and the final accuracy under black-box perturbations remains significantly higher than that of implicit reasoning models. These results suggest that implicit reasoning models are more sensitive to candidate prefixes containing numerical tokens compared to purely random samples, indicating a higher susceptibility to numerical inputs. This observation is further supported by the white-box results in Table 12 (Appendix B.4): under 5-token attacks on GSM8K, numerical tokens occur in 70.06% of CODI prefixes, compared with only 0.93% for Llama3.2.

	V*			MMVP			MMStar		
model	LVR	Monet	SkiLa	LVR	Monet	SkiLa	LVR	Monet	SkiLa
LVR	0.00	38.10	6.35	2.63	97.37	17.11	17.50	47.50	30.00
Monet	1.59	3.17	0.00	96.05	5.26	97.37	52.50	17.50	42.50
SkiLa	7.94	33.33	6.35	31.58	94.74	1.32	37.50	55.00	10.00

Table 3: Transferability of perturbations across multi-modal reasoning models. Rows denote evaluated target models, and columns denote the models used to optimize adversarial perturbations. Accuracy (%) is reported on questions that all models answer correctly without perturbations.

	GSM8K			SVAMP			MultiArith		
model	White	Black	Inter-Q	White	Black	Inter-Q	White	Black	Inter-Q
CODI	0.7617	0.6312	0.2752	0.7437	0.5057	0.3902	0.5989	0.4766	0.4273
SIM-COT	0.7692	0.7174	0.4643	0.7185	0.6640	0.4760	0.7422	0.6564	0.5690
PCCoT	0.7818	0.7719	0.5563	0.5922	0.5487	0.5148	0.7021	0.6835	0.7262
CoLaR	0.8942	0.8784	0.6420	0.8421	0.8269	0.6367	0.8320	0.8728	0.6802
RoT	0.9381	0.9333	0.8437	0.8817	0.8905	0.7940	0.9147	0.9040	0.8541

Table 4: Latent-token similarity between clean and adversarial prompts for textual reasoning models. White and Black denote adversarial prompts generated under white-box and black-box settings, respectively. Inter-Q denotes the average similarity between latent tokens produced for different questions.

Prefix candidates	Clean	Perturbed	Drop
Non-numerical only	56.00	26.00	30.00
Unrestricted	56.00	7.78	48.22

Table 5: Accuracy (%) of CODI on GSM8K under optimized prefixes with different candidate-token constraints.

To isolate the effect of numerical tokens, we conduct a controlled experiment in which gradient-based prefix optimization is constrained to select only non-numerical tokens. As shown in Table 5, the clean accuracy of CODI on GSM8K decreases from 56% to 26% under the optimized non-numerical prefixes, whereas allowing numerical tokens reduces the accuracy further to 7.78%. This substantial gap indicates that numerical tokens are an especially important factor in the effectiveness of optimized prefixes.

4.2.2 Results on Multi-modal Reasoning Models

Table 2 reports the multi-modal results. On V*, white-box perturbations reduce LVR from 72.00% to 0.33%, compared with a decrease from 42.00% to 8.67% for the explicit reasoning model Qwen2.5-VL-7B (Bai et al., 2025). This indicates that the vulnerability is not unique to implicit reasoning, although the implicit model undergoes a larger reduction. For random perturbations, the Random column is the proportion of questions answered correctly under all 50 perturbations, while Table 11 (Appendix B.2) reports the average normalized accuracy on clean-correct questions.

Furthermore, we evaluate the cross-model transferability of adversarial examples, as shown in Table 3. The results vary considerably across datasets and source models: perturbations optimized on SkiLa reduce Monet's accuracy to 0.00% on V*, but leave it at 97.37% on MMVP. The datasets also differ substantially in average image resolution, ranging from 224×224 for MMVP to 1125×773 for V*, as reported in Table 9 (Appendix A.3). Because the perturbations are optimized in pixel space, these differences alter the dimensionality and spatial granularity of the optimization and may therefore contribute to the observed variation in transferability.

4.3 Analysis of Latent Reasoning

The results above show that adversarial perturbations can substantially change model outcomes. However, it remains unclear whether these perturbations first disrupt the latent reasoning tokens and thereby alter the final answer, or whether they primarily affect the outcome through mechanisms that bypass latent reasoning. To better understand this distinction, we conduct two additional analyses.

Reasoning type	Model	Clean question		Adversarial question	
		Clean reasoning	Adversarial reasoning	Adversarial reasoning	Clean reasoning
Explicit	Qwen3	70.67	35.33	32.33	70.33
Implicit	CODI	56.00	33.00	2.67	25.67
	SIM-COT	42.67	10.67	0.67	23.67
	PCCoT	48.33	25.00	0.33	30.33
	CoLaR	27.33	1.00	0.00	25.33
	RoT	24.33	23.67	0.67	1.00

Table 6: Accuracy (%) of text reasoning models under latent reasoning-token replacement. Clean question and adversarial question denote the original and adversarial questions, respectively. Clean and adversarial reasoning denote the latent reasoning tokens generated under the corresponding input conditions. Cell shading reflects accuracy, with darker red indicating higher accuracy.

4.3.1 Do Adversarial Perturbations Affect Latent Tokens?

To investigate the impact of adversarial examples on latent tokens, we conduct further experiments. We evaluate the degree of latent-token disruption by computing the cosine similarity between the latent tokens produced for original questions and those produced for adversarial questions. Meanwhile, we compute the mean cosine similarity of latent tokens across different questions as the inter-question reference (Inter-Q). Table 4 shows substantial variation across textual models. On GSM8K, CODI's similarity falls to 0.6312 under black-box perturbations, whereas RoT remains at 0.9333. CODI's attacked-pair similarity nevertheless remains above its Inter-Q value of 0.2752.

	V*		MMVP		MMStar	
model	White	Inter-Q	White	Inter-Q	White	Inter-Q
LVR	0.9397	0.8993	0.9510	0.9081	0.9733	0.8634
Monet	0.9896	0.9525	0.9861	0.9392	0.9862	0.9336
SkiLa	0.9839	0.9834	0.9777	0.9795	0.9844	0.9589

Table 7: Latent-token similarity between clean and adversarial inputs for multi-modal models. White denotes adversarial images generated under the white-box setting, while Inter-Q denotes the average similarity between latent tokens produced for different questions.

Table 7 presents the evaluation results for multi-modal models, showing that adversarial samples do not cause significant changes in latent tokens. For example, SkiLa's white-box similarity on MMVP is 0.9777, nearly identical to its Inter-Q value of 0.9795, indicating that its latent tokens vary little with either adversarial perturbations or question content.

4.3.2 Do Latent Tokens Affect the Outcome?

The preceding similarity analysis measures how much latent tokens change under perturbations, but it does not reveal whether these changes influence the final prediction. To examine their functional role, we replace latent tokens between clean and adversarial inputs while keeping the question fixed. Table 6 reports the results. On clean questions, replacing CoLaR's clean reasoning with adversarial reasoning reduces accuracy from 27.33% to 1.00%, whereas RoT changes only from 24.33% to 23.67%. This contrast shows that the influence of latent tokens varies substantially across models.

Additionally, we also test an explicit reasoning model, Qwen/Qwen3-0.6B (Yang et al., 2025), whose reasoning tokens can be extracted from the text between `<think>` and `</think>`, as a comparison. Replacing its clean reasoning with adversarial reasoning lowers clean-question accuracy from 70.67% to 35.33%, indicating that its explicit reasoning tokens strongly influence the outcome.

Table 8 presents the results of similar analysis on multi-modal models, showing that replacing latent tokens has little effect on the outcome. On clean V* questions, LVR remains at 72.00% whether its latent reasoning is clean, adversarial, or zero.

This invariance suggests that the final prediction may be largely determined by an answer preference formed before latent reasoning, rather than being revised through the latent tokens. In our white-box assessment, we optimize the logit of a targeted incorrect answer token at the position immediately before latent reasoning begins. The resulting perturbation can therefore directly shift the model's initial answer preference. If the subsequent latent reasoning has little causal influence on the output, it is unlikely to correct this preference, allowing the incorrect judgment to persist in the final answer.

By contrast, an explicit reasoning model generates a longer textual reasoning chain, during which it may revisit and correct an initially misleading judgment. This difference may help explain why the multi-modal implicit reasoning models exhibit larger accuracy drops than the explicit baseline under the white-box assessment in table 2.

These experiments show that, in some of the evaluated textual LRMs and all of the evaluated multi-modal LRMs, the latent reasoning chains are largely insensitive to input and have little influence on the final output, consistent with the findings of prior work (Zhang et al., 2025b; Li et al., 2026).

Dataset	Model	Clean Q			Adv Q		
		Clean R	Adv R	Zero R	Adv R	Clean R	Zero R
V*	Qwen-VL	42.00	25.00	–	8.00	25.00	–
	LVR	72.00	72.00	72.00	0.00	0.00	0.00
	Monet	70.00	70.00	62.00	2.00	2.00	8.00
	SkiLa	73.00	73.00	73.00	5.00	5.00	5.00
MMVP	Qwen-VL	72.67	31.33	–	9.33	40.00	–
	LVR	70.67	70.00	68.67	0.67	5.33	4.67
	Monet	68.67	68.67	60.66	4.67	4.67	10.00
	SkiLa	68.67	68.67	68.00	0.67	2.67	4.00
MMStar	Qwen-VL	56.00	24.00	–	18.00	52.00	–
	LVR	62.00	58.00	58.00	11.00	13.00	9.00
	Monet	59.00	59.00	57.00	8.00	8.00	12.00
	SkiLa	55.00	50.00	51.00	5.00	7.00	9.00

Table 8: Accuracy (%) of multi-modal reasoning models with replaced latent tokens. Clean/Adv Q denote clean/adversarial questions; Clean/Adv/Zero R denote clean, adversarial, or zero latent reasoning tokens. Darker red indicates higher accuracy. Qwen-VL does not use zero reasoning tokens.

4.4 Case Study

We further examine representative cases from textual and multi-modal settings to show how perturbations affect latent reasoning and the final outcome.

Figure 5 in Appendix B.7 and Figure 6 in Appendix B.8 illustrate adversarial examples produced under black-box and white-box perturbation settings for textual implicit reasoning models, respectively. They include decoded latent tokens for original and adversarial questions and the final model answers. These examples show how adversarial prefixes can affect the latent reasoning process: some prefix tokens may infiltrate the latent tokens, and prefixes containing large numerical values often induce more numeric latent tokens. Our experiments show that, in some textual LRMs, the latent reasoning chains exert a measurable influence on the final output. For these models, the disruption of latent reasoning chains by numerical tokens in adversarial prefixes may therefore contribute to the generation of incorrect final answers.

Figure 7 (Appendix B.9) presents a multi-modal implicit reasoning example. Although the adversarial image perturbation is nearly imperceptible to human observers, it is sufficient to mislead the model into producing an incorrect answer. The implicit reasoning operations, such as zoom-in and highlight, fail to recover the relevant evidence, suggesting that latent visual reasoning does not improve robustness in this case.

5 Discussion and Conclusion

Our results reveal a critical tension between the efficiency promised by latent reasoning and its robustness. Across the settings considered in this work, latent reasoning are generally more vulnerable than explicit chain-of-thought, but the source of this vulnerability differs across modalities. For textual LRMs, adversarial prefixes can notably alter latent representations, with numerical cues being particularly disruptive, suggesting that the compressed reasoning dynamics remain sensitive to superficial input patterns. In contrast, the latent states of the evaluated multi-modal models change little under perturbations and replacing or removing them often has only a minor effect on predictions. The result suggests that multi-modal latent tokens could be place-holders that contribute loosely to the final outcome, which makes the model less robust due to the lack of internal thinking process.

More broadly, our findings suggest that compressing explicit reasoning into a small number of continuous states does not automatically preserve the error-correction benefits of a longer reasoning trajectory. Explicit chain-of-thought, despite its higher inference cost, provides additional intermediate computation in which an early misleading signal may be reconsidered or overridden. Therefore, improving latent reasoning should not only be increasing clean-task accuracy or reducing the number of reasoning steps: future methods should also ensure that latent states encode meaningful thinking, exert a genuine causal influence on predictions, and remain stable under task-preserving perturbations. Our study thus highlights adversarial robustness as a necessary dimension, alongside efficiency and accuracy, for evaluating and designing latent reasoning systems.

Limitations

Our study has three main limitations. First, the evaluated model set is limited. Although we cover representative text and multi-modal LRMs, current latent reasoning architectures vary substantially in base models, training objectives, and latent-token designs. Therefore, our findings should be viewed as evidence of robustness risks rather than a complete characterization of all LRMs.

Second, our benchmarks and perturbations cover only part of the threat space. We focus on mathematical and visual reasoning tasks, with short textual prefixes for text models and bounded pixel-level noise for multi-modal models. Other settings, such as long-context prompt injection, paraphrase-based attacks, tool-use scenarios, or physically realizable visual perturbations, may reveal different failure modes.

Third, our latent-token analyses are diagnostic rather than fully mechanistic. Cosine similarity and token replacement help probe whether latent states change or affect predictions, but they cannot fully explain how latent reasoning contributes to final answers. More fine-grained causal interventions are needed to understand when latent states are genuinely used and how to make them robust.

Acknowledgments

Yisen Wang is supported by National Natural Science Foundation of China (62376010, 92370129), Beijing Major Science and Technology Project under Contract no. Z251100008425006, Beijing Natural Science Foundation (No. L257007), Beijing Nova Program (20230484344, 20240484642), and State Key Laboratory of General Artificial Intelligence.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Giannis Chatziveroglou, Richard Yun, and Maura Kelleher. 2025. Exploring llm reasoning through controlled prompt variations. *arXiv preprint ArXiv:2504.02111*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024. Are we on the right way for evaluating large vision-language models? *arXiv preprint ArXiv:2403.20330*.

Xinghao Chen, Anhao Zhao, Heming Xia, Xuan Lu, Hanlin Wang, Yanjun Chen, Wei Zhang, Jian Wang, Wenjie Li, and Xiaoyu Shen. 2025. Reasoning beyond language: A comprehensive survey on latent chain-of-thought reasoning. *arXiv preprint arXiv:2505.16782*.

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

Alejandro Cuadron, Dacheng Li, Wenjie Ma, Xingyao Wang, Yichuan Wang, Siyuan Zhuang, Shu Liu, Luis Gaspar Schroeder, Tian Xia, Huanzhi Mao, Nicholas Thumiger, Aditya Desai, Ion Stoica, Ana Klimovic, Graham Neubig, and Joseph E. Gonzalez. 2025. The danger of overthinking: Examining the reasoning-action dilemma in agentic tasks. *arXiv preprint arXiv:2502.08235*.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Esther Gan, Yiran Zhao, Liying Cheng, Mao Yancan, Anirudh Goyal, Kenji Kawaguchi, Min-Yen Kan, and Michael Shieh. 2024. Reasoning robustness of llms to adversarial typographical errors. In *EMNLP*.

Jiaxuan Gao, Shu Yan, Qixin Tan, Lu Yang, Shusheng Xu, Wei Fu, Zhiyu Mei, Kaifeng Lyu, and Yi Wu. 2025. How far are we from optimal reasoning efficiency? *NeurIPS*.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Etash Kumar Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Rea Sprague, Ashima Suvarna, Benjamin Feuer, Leon Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, and 32 others. 2026. Openthoughts: Data recipes for reasoning models. In *The Fourteenth International Conference on Learning Representations*.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2024. Training large language models to reason in a continuous latent space. *arXiv preprint ArXiv:2412.06769*.

Xanh Ho, Yun-Ang Wu, Sunisth Kumar, Florian Boudin, Atsuhiro Takasu, and Akiko Aizawa. 2026. Format matters: The robustness of multimodal llms in reviewing evidence from tables and charts. In *AAAI*.

Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. 2026. A survey on large language models for code generation. *ACM Transactions on Software Engineering and Methodology*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. 2025. Latent visual reasoning. *arXiv preprint arXiv:2509.24251*.

Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. 2024. Gsm-plus: A comprehensive benchmark for evaluating the robustness of llms as mathematical problem solvers. *arXiv preprint ArXiv:2402.19255*.

You Li, Chi Chen, Yanghao Li, Fanhu Zeng, Kaiyu Huang, Jinan Xu, and Maosong Sun. 2026. Imagination helps visual reasoning, but not yet in latent space. *arXiv preprint ArXiv:2602.22766*.

Buyun Liang, Liangzu Peng, Jinqi Luo, Darshan Thaker, Kwan Ho Ryan Chan, and René Vidal. 2025. Seca: Semantically equivalent and coherent attacks for eliciting llm hallucinations. *arXiv preprint ArXiv:2510.04398*.

Llama Team. 2024. The llama 3 herd of models. *arXiv preprint ArXiv:2407.21783*.

Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, Rongcheng Tu, Xiao Luo, Wei Ju, Zhiping Xiao, Yifan Wang, Meng Xiao, Chenwu Liu, Jingyang Yuan, Shichang Zhang, and 7 others. 2025. Large language model agent: A survey on methodology, applications and challenges. *arXiv preprint arXiv:2503.21460*.

Aleksander Mądry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint ArXiv:1706.06083*.

OpenAI. 2024. Learning to reason with llms.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are nlp models really able to solve simple math word problems? *arXiv preprint ArXiv:2103.07191*.

Subhro Roy and Dan Roth. 2016. Solving general arithmetic word problems. *arXiv preprint ArXiv:1608.01413*.

Rohit Saxena, Alessandro Suglia, and Pasquale Minervini. 2026. Vlm-robustbench: A comprehensive benchmark for robustness of vision-language models. *arXiv preprint ArXiv:2603.06148*.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. 2025. Codi: Compressing chain-of-thought into continuous space via self-distillation. *arXiv preprint ArXiv:2502.21074*.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, Hanjie Chen, and Xia Hu. 2025. Stop overthinking: A survey on efficient reasoning for large language models. *Transactions on Machine Learning Research*.

Wenhui Tan, Jiaze Li, Jianzhong Ju, Zhenbo Luo, Ruihua Song, and Jian Luan. 2025. Think silently, think fast: Dynamic latent compression of llm reasoning chains. *arXiv preprint ArXiv:2505.16552*.

Jintao Tong, Jiaqi Gu, Yujing Lou, Lubin Fan, Yixiong Zou, Yue Wu, Jieping Ye, and Ruixuan Li. 2025. Sketch-in-latents: Eliciting unified reasoning in mllms. *arXiv preprint arXiv:2512.16584*.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *arXiv preprint ArXiv:2401.06209*.

Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. 2025. Monet: Reasoning in latent visual space beyond images and language. *arXiv preprint arXiv:2511.21395*.

Yifan Wang, Shiyu Li, Peiming Li, Xiaochen Yang, Yang Tang, and Zheng Wei. 2026. Render-of-thought: Rendering textual chain-of-thought as images for visual latent reasoning. *arXiv preprint ArXiv:2601.14750*.

Zefeng Wang, Zhen Han, Shuo Chen, Fan Xue, Zifeng Ding, Xun Xiao, Volker Tresp, Philip Torr, and Jindong Gu. 2024. Stop reasoning! when multimodal llm with chain-of-thought reasoning meets adversarial image. *arXiv preprint ArXiv:2402.14899*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Jiaqi Wang, Xipeng Qiu, and Dahua Lin. 2025. Sim-cot: Supervised implicit chain-of-thought. *arXiv preprint ArXiv:2509.20317*.

Haoyi Wu, Zhihao Teng, and Kewei Tu. 2025. Parallel continuous chain-of-thought with jacobi iteration. *arXiv preprint ArXiv:2506.18582*.

Penghao Wu and Saining Xie. 2023. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv preprint ArXiv:2312.14135*.

Fengli Xu, Qianyue Hao, Chenyang Shao, Zefang Zong, Yu Li, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Yuwei Yan, Qinglong Yang, Yiwen Song, Sijian Ren, Xinyuan Hu, Jie Feng, Chen Gao, and Yong Li. 2025. Toward large reasoning models: A survey of reinforced reasoning with large language models. *Patterns*.

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 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Wenyue Hua, Haolun Wu, Zhihan Guo, Yufei Wang, Niklas Muennighoff, Irwin King, Xue Liu, and Chen Ma. 2025a. A survey on test-time scaling in large language models: What, how, where, and how well? *arXiv preprint arXiv:2503.24235*.

Yuyi Zhang, Boyu Tang, Tianjie Ju, Sufeng Duan, and Gongshen Liu. 2025b. Do latent tokens think? a causal and adversarial analysis of chain-of-continuous-thought. *arXiv preprint ArXiv:2512.21711*.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. 2023. On evaluating adversarial robustness of large vision-language models. *arXiv preprint ArXiv:2305.16934*.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint ArXiv:2307.15043*.

A Implementation Details

A.1 Text Models

Model backbones. The evaluated textual implicit reasoning models use different backbone language models. CODI is built on `meta-llama/Llama-3.2-1B-Instruct`, RoT on Qwen3-VL-2B, and CoLaR on `deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B`. Both PCCoT and SIM-CoT use the 124M-parameter GPT-2 model as their backbone.

For white-box attacks on text implicit reasoning models, we run 30 iterations of optimization, with a candidate prefix pool size of 40 at each iteration.

Modified forward and backward pipeline. Listing 1 outlines the modified forward and backward propagation pipeline for white-box assessment of textual latent reasoning models. It retains gradients on the input embeddings, unrolls model-specific latent reasoning steps with key–value cache updates, and backpropagates the answer-level objective to the prefix embeddings. Projecting these gradients onto the vocabulary embeddings yields token-level scores for prefix optimization. Model-specific details, such as latent-state transitions and cache handling, are adapted accordingly.

Listing 1: Modified gradient pipeline for the white-box assessment of textual latent reasoning models.

```
def compute_gradients(input_ids):
    input_embeds = embed_layer(input_ids).
     requires_grad_(True)

    question_outputs = base_llm(
        inputs_embeds=input_embeds,
        output_hidden_states=True)
    past_kv = question_outputs.past_key_values
    hidden = question_outputs.hidden_states[-1][:,
     -1:, :]

    for _ in range(num_latent_steps):
        step_outputs = base_llm(
            inputs_embeds=hidden,
            past_key_values=past_kv, ...)
        past_kv = step_outputs.past_key_values
        hidden = step_outputs.hidden_states[-1][:,
     -1:, :]

    answer_outputs = base_llm(
        inputs_embeds=concat_embeds,
        past_key_values=past_kv)
    answer_token_logits = answer_outputs.logits[
     answer_pos]

    loss = (answer_token_logits[TARGET]
            - answer_token_logits[BASELINE])
    loss.backward()

    grad = input_embeds.grad
    prefix_grad_emb = grad[:, :prefix_len]
    token_grads = prefix_grad_emb @ embed_weight.T
    return token_grads
```

For explicit baselines, we use Llama-3.2-1B-Instruct in the main robustness evaluation because several implicit reasoning models are built on it as the base model, enabling a more directly comparable robustness comparison. In the token-replacement analysis, we instead use Qwen/Qwen3-0.6B because it explicitly generates reasoning traces between `<think>` and `</think>`, making its reasoning tokens easy to extract and replace.

For black-box attacks, we construct adversarial prefixes using 30 irrelevant factual statements:

- "There are 24 hours in a day."
- "April has 30 days."
- "A bicycle has 2 wheels."
- "Today is the 3rd of the month."
- "The current temperature is 20 degrees."
- "A year has 12 months."
- "A week has 7 days."
- "An hour has 60 minutes."
- "A minute has 60 seconds."
- "A year has 365 days."
- "The Earth is about 12742 km in diameter."
- "The speed of sound in air is about 343 meters per second."
- "Water freezes at 0 degrees."
- "The speed of light is about 300000 km/s."
- "The human heart has 4 chambers."
- "An adult has 32 teeth."
- "China has 1.4 billion people."
- "Japan has 125 million people."
- "The US has 50 states."
- "The EU has 27 member states."
- "The Nile is about 6650 km long."
- "Mount Everest is about 8849 meters high."
- "The Pacific covers 165 million km^2."
- "The human body has 206 bones."
- "H2O contains 2 hydrogen atoms."
- "Oxygen makes up about 21% of air."
- "Standard atmospheric pressure is 101325 Pa."
- "A plant has 46 chromosomes."
- "DNA has 2 strands in its double helix."
- "Bitcoin was created in 2009."

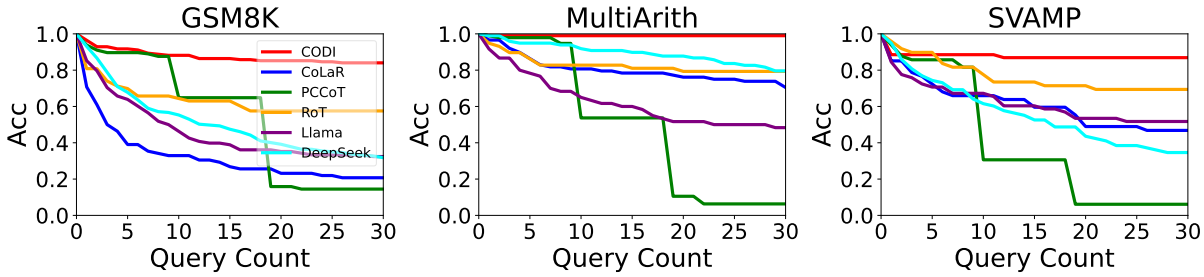

Figure 4: Textual reasoning models accuracy (%) under random perturbation

A.2 Multi-modal Models

For white-box attacks on multi-modal models, we run 50 iterations of optimization, along with 50 random perturbations. For each random perturbation, the perturbation at every pixel is sampled independently and uniformly from $\{-4/255, +4/255\}$.

Modified visual preprocessing. Because the original Qwen2.5-VL processor contains operations that block end-to-end gradients to the input image, we use functionally equivalent, gradient-compatible resizing, normalization, and patch-merging steps during adversarial optimization, as shown in listing 2. The modified processor is used only to optimize perturbations and propagate gradients to pixel space; evaluation uses the original processor. The implementation is adapted to the tensor layouts and preprocessing configurations of different multi-modal models.

Listing 2: Modified Qwen2.5-VL visual preprocessing steps used during white-box optimization.

```
def preprocess(img):
    a = F.interpolate(
        img[None],
        size=(H_target, W_target),
        mode='bilinear',
        align_corners=False)[0]
    b = (a - mean) / std
    c, _ = pixel_reshape(b, P, M, T_p)
    return c

def pixel_reshape(x, patch, merge, t):
    n, c, h, w = x.shape
    gt, gh, gw = n // t, h // patch, w // patch
    patches = (x.view(
        gt, t, c, gh // merge, merge, patch,
        gw // merge, merge, patch)
        .permute(0, 3, 6, 4, 7, 2, 1, 5, 8)
        .reshape(gt * gh * gw, c * t * patch * patch
    ))
    return patches, (gt, gh, gw)
```

We also attempted a query-based black-box attack that leverages answer probabilities without gradient information, combined with hill climbing search. However, the effectiveness is limited. On the LVR model and MMVP dataset, running 50 queries reduces accuracy from 70.66% to 57.33%, and 300 queries further reduce it to 49.33%. This performance is only marginally better than random perturbation and significantly worse than white-box attacks.

A.3 Dataset

Due to the computational cost of attacks, it is impractical to evaluate on the entire dataset. We randomly sample 100 to 300 instances from each dataset: GSM8K (300), MultiArith (100), SVAMP (100), MMVP (150), V* (100), and MMStar (100). For V*, due to GPU memory constraints, we downsample all images by a factor of two along both height and width to avoid out-of-memory errors during backpropagation.

In the cross-model adversarial-example transferability experiment reported in table 3, we evaluate V*, MMVP, and MMStar. Their average image resolutions are reported in table 9.

Dataset	V*	MMVP	MMStar
Resolution	1125×773	224×224	476×384

Table 9: Average image resolution of the multi-modal datasets.

B Additional Experiment Results

B.1 Average Black-box Accuracy of Textual Reasoning Models

Using the notation introduced in section 3.3, let $\mathcal{D}_{\text{clean}} = \{(x, a^*) \in \mathcal{D} : f(x) = a^*\}$ denote the subset of questions that the model answers correctly without perturbation. We compute this metric over the fixed context pool $\mathcal{C}$ as

$$\text{Acc}_{\text{avg}} = \mathbb{E}_{(x,a^*)\sim\mathcal{D}_{\text{clean}}} \left[\frac{1}{|\mathcal{C}|} \sum_{c\in\mathcal{C}} \mathbf{1}\left[f(c \oplus x) = a^*\right] \right]. \tag{5}$$

Equivalently, for each prefix, we divide the number of questions answered correctly after perturbation by the number answered correctly without perturbation, and then average this ratio over all prefixes.

Unlike the robust accuracy in section 3.3, which requires a question to be answered correctly under every context, this metric measures the average fraction of clean-correct answers retained under perturbation.

Table 10 reports the average normalized accuracy over the 30 fixed irrelevant factual prefixes described in section A.

Type	Model	GSM8K	SVAMP	MultiArith
explicit	Llama3.2	0.7260	0.7845	0.8158
	DeepSeek	0.8566	0.8796	0.9711
implicit	CODI	0.6389	0.6235	0.7347
	SIM-CoT	0.4044	0.2778	0.4373
	PCCoT	0.4444	0.4075	0.5074
	CoLaR	0.3220	0.4695	0.5864
	RoT	0.4694	0.7200	0.5454

Table 10: Average normalized accuracy of textual reasoning models under the black-box assessment, averaged over 30 fixed irrelevant factual prefixes.

The results show that CoLaR consistently underperforms DeepSeek under black-box perturbations; for example, their average normalized accuracies on GSM8K are 0.3220 and 0.8566, respectively. This indicates that the explicit reasoning model is less sensitive to prefix perturbations than the LRM, consistent with the analysis in section 4.2.1.

B.2 Average Accuracy under Random Multi-modal Perturbations

Following the statistical procedure in section B.1, we restrict the evaluation to questions that each model answers correctly without perturbation. For each of the 50 random pixel-level perturbations described in section A, we compute the fraction of these questions that remain correctly answered and then average the results across perturbations. Table 11 reports the resulting normalized accuracies following the model and dataset ordering used in table 2.

Type	Model	V^*	MMVP	MMStar
explicit	Qwen	91.90	91.78	88.07
implicit	LVR	97.17	94.15	92.74
	Monet	98.05	95.50	97.11
	SkiLa	94.75	95.53	88.26

Table 11: Average normalized accuracy (%) of multi-modal reasoning models over 50 random pixel-level perturbations.

B.3 Random Prefix Perturbations for Textual Reasoning Models

We evaluate textual implicit reasoning models under random prefix perturbations as a comparison to the black-box prefix search in the main text. Each token in a random prefix is sampled independently and uniformly from the model vocabulary. We prepend these randomly sampled tokens to the input questions and measure the normalized accuracy as the query budget increases. As shown in fig. 4, random prefixes can reduce accuracy, but the degradation is generally less targeted than the black-box perturbations discussed in section 4.2.1, supporting the observation that numerical or search-selected prefixes are more harmful than purely random tokens.

B.4 Numerical Patterns in Optimized Textual Prefixes

We observe that many optimized white-box prefixes contain numbers or numeral-like tokens, so we further quantify how often this pattern occurs. Table 12 reports the proportion of such prefixes for 3-token and 5-token attacks. The optimized prefixes for implicit reasoning models frequently contain numerical tokens, while the explicit Llama3.2 baseline rarely exhibits this pattern, indicating that numerical prefix tokens are closely associated with the vulnerability of textual implicit reasoning models.

		GSM8K		SVAMP		MultiArith	
type	model	White-5	White-3	White-5	White-3	White-5	White-3
explicit	Llama3.2	0.0093	0.0187	0.0179	0.0000	0.0000	0.0000
implicit	CODI	0.7006	0.6786	0.6721	0.6721	0.8144	0.7755
	SIM-COT	0.5078	0.4688	0.4722	0.3611	0.5699	0.5699
	PCCoT	0.8897	0.8207	0.8571	0.8163	0.9684	0.8526
	CoLaR	0.3780	0.2439	0.4043	0.4043	0.5454	0.6136
	RoT	0.8767	0.7534	0.9184	0.8367	0.8966	0.8103

Table 12: Proportion of prefixes containing numbers

B.5 Semantic Irrelevance of Optimized Textual Prefixes

Because the white-box assessment optimizes adversarial prefixes directly, we additionally verify that the resulting prefixes do not simply alter the semantics or correct answers of the original problems. We use Qwen3.5-4B as an independent judge and ask it whether each optimized prefix could change the answer to the corresponding original question.

Table 13 reports the results for 3-token prefixes. Across all models and datasets, only a small frac-

Prefix	Latent 1	Latent 2	Latent 3
None	“4”, “«”, “:”	“4”, “.”, “-”	“.”, “»”, “+”
Non-numerical: `getic hiringNine`	“4”, “«”, “:”	“4”, “-”, “+”	“.”, “»”, “,”
Numerical: `9334000`	“467”, “466”, “477”	“000”, “400”, “4”	“4”, “0”, “004”

Table 14: Top-3 decoded tokens at three latent positions under different prefix conditions.

Method	SVAMP	GSM8K	MultiArith
CODI	0.00% (0/61)	0.60% (1/168)	0.00% (0/98)
RoT	0.00% (0/49)	1.37% (1/73)	1.72% (1/58)
PCCoT	2.04% (1/49)	1.38% (2/145)	0.00% (0/95)
CoLaR	0.00% (0/47)	0.00% (0/82)	0.00% (0/88)
SIM-CoT	0.00% (0/36)	0.00% (0/128)	2.15% (2/93)

Table 13: Fraction of optimized 3-token prefixes judged as potentially changing the answer to the original question. Parentheses show the number judged answer-changing over the total number evaluated.

tion of optimized prefixes are judged as potentially answer-changing. For 5-token prefixes, none of the optimized prefixes are judged as changing the answer.

These results show that answer-changing prefixes are rare. Therefore, the large accuracy drops under white-box perturbations cannot be explained by genuine semantic changes to the mathematical problems. Instead, they indicate that textual implicit reasoning models are highly sensitive to semantically irrelevant adversarial prefixes.

B.6 Effect of Numerical Prefixes on Latent Reasoning

To isolate the effect of numerical tokens on latent reasoning, we examine a representative CODI example from GSM8K under three conditions: no prefix, an optimized non-numerical prefix, and an optimized numerical prefix. The question asks: *“Nancy is returning her overdue books to the library. She owes $0.50 cents each on 8 books … How much does she have to pay total?”* Table 14 reports the three tokens with the highest decoding probabilities at each of the first three latent positions. The no-prefix and non-numerical-prefix conditions yield highly similar decoded tokens, whereas the numerical prefix introduces multiple numerical tokens. This contrast suggests that the changes in latent reasoning are specifically associated with numerical tokens rather than with the presence of an optimized prefix alone.

B.7 Black-box Textual Case Study

We provide a representative black-box textual example to illustrate how an adversarial prefix changes the model’s latent reasoning and answer. As shown in fig. 5, the added prefix can influence the decoded latent tokens and steer the model toward an incorrect final response, consistent with the black-box degradation reported in the main results.

B.8 White-box Textual Case Study

We also include a white-box textual example to show the effect of optimized prefix perturbations. Figure 6 visualizes the original and adversarial questions together with their decoded latent tokens and answers. The example shows that a short optimized prefix can alter the latent reasoning trajectory and produce an incorrect answer, providing qualitative evidence for the large white-box accuracy drops in table 1.

B.9 Multi-modal Case Study

Finally, we present a multi-modal example to illustrate the effect of imperceptible image perturbations. As shown in fig. 7, the adversarial image leads the model to an incorrect answer even though the perturbation is visually subtle. The implicit visual reasoning operations do not recover the relevant evidence, matching the quantitative finding that multi-modal latent tokens are stable but weakly tied to the final outcome.

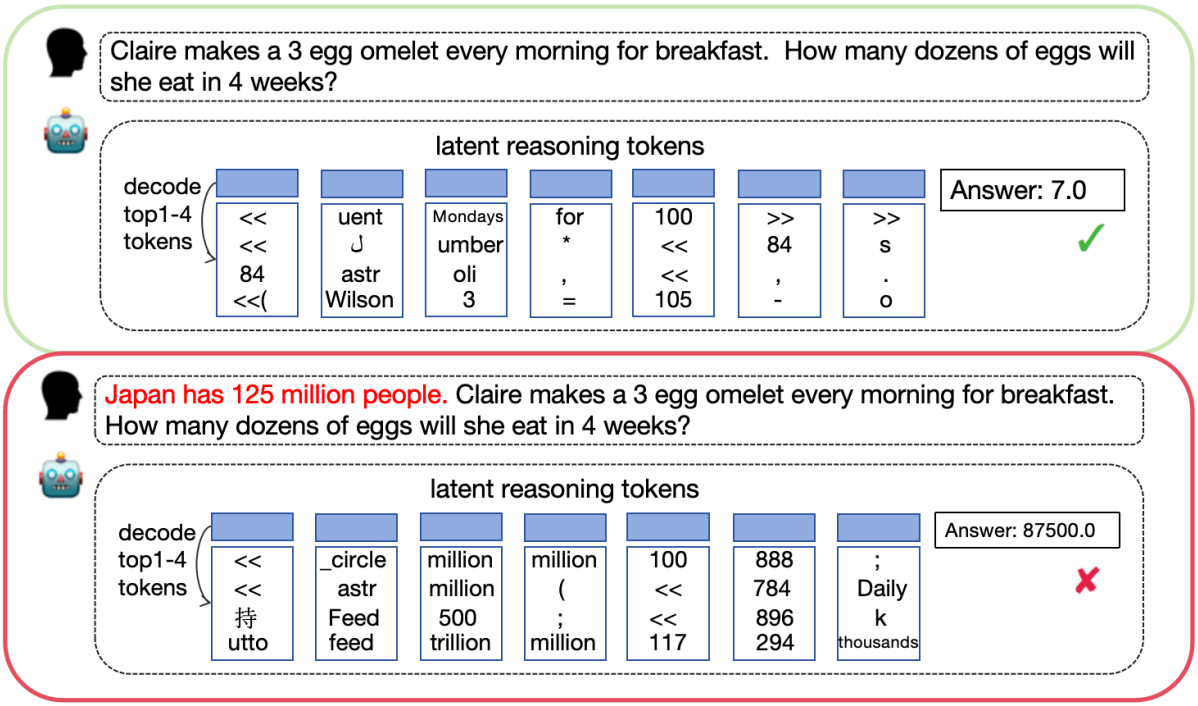

Figure 5: Case study 1

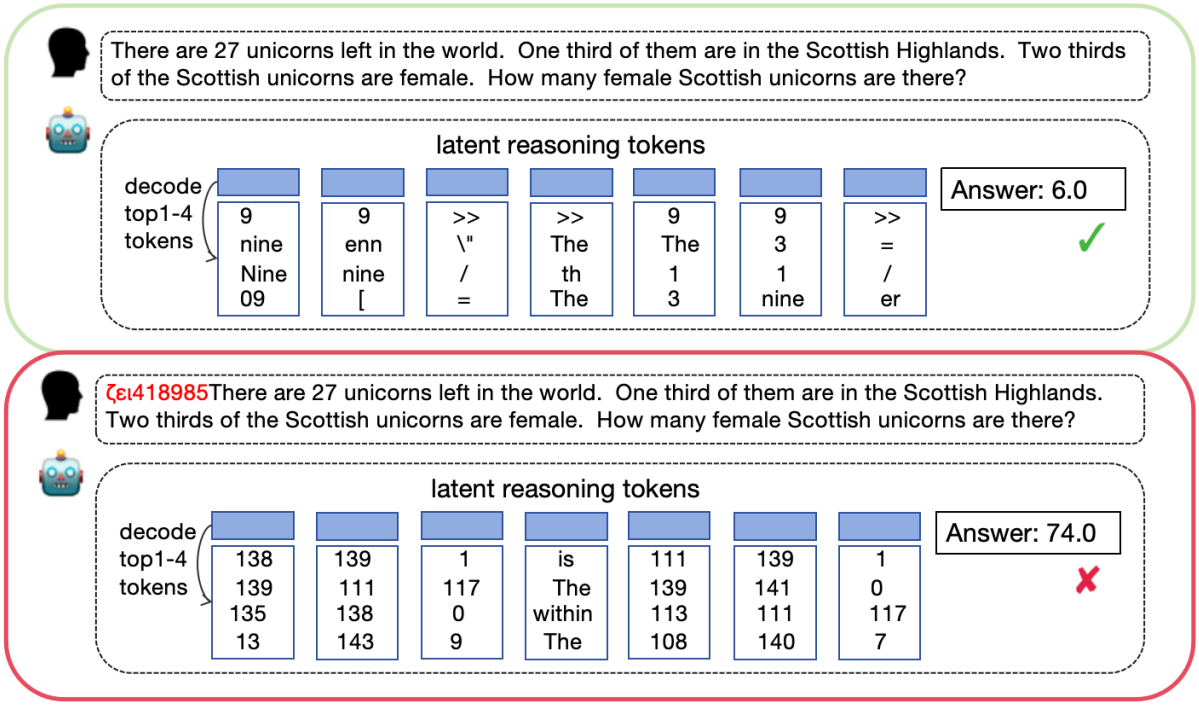

Figure 6: Case study 2

clean image

Is the dog facing left or right from the camera's perspective?

(a) Left (b) Right

<think>To determine the dog's orientation, I need to analyze its head position relative to the viewer in the provided image. To clearly observe the dog's head and its direction, I will focus on the central part of the image where the dog is located.</think>
<|sketch_start|><|skila|><|skila|><|skila|><|skila|><|skila|><|skila|>
<|skila|><|skila|><|skila|><|skila|><|skila|><|skila|><|skila|><|skila|>
<|skila|><|skila|><|skila|><|skila|><|skila|><|skila|><|skila|><|skila|>
<|skila|><|skila|><|skila|><|skila|><|skila|><|sketch_end|>
<think>The zoomed-in view clearly shows the dog's head turned towards the right side of the frame, indicating it is looking that way. Based on this observation, the dog is facing right from the viewer's perspective.</think><answer>B</answer>

adv image

Is the dog facing left or right from the camera's perspective?

(a) Left (b) Right

<think>To determine the dog's orientation, I need to analyze its head position relative to the viewer in the provided image. To clearly identify the dog's gaze direction, I will highlight the relevant area of the image.</think>
<|sketch_start|><|skila|><|skila|><|skila|><|skila|><|skila|><|skila|>
<|skila|><|skila|><|skila|><|skila|><|skila|><|skila|><|skila|><|skila|>
<|skila|><|skila|><|skila|><|skila|><|skila|><|skila|><|skila|><|skila|>
<|skila|><|skila|><|skila|><|skila|><|skila|><|sketch_end|>
<think>The red bounding box isolates the dog's head and upper body, showing its head is turned towards the left side of the frame. Based on this visual evidence, the dog is looking towards the left.</think><answer>A</answer>

Figure 7: Case study 3