Merlin's Whisper: Enabling Efficient Reasoning in LLMs via Black-box Adversarial Prompting

⚠Caution: this paper includes adversarial text that may be considered offensive.

Anonymous ACL submission

Abstract

Large reasoning models (LRMs) have demonstrated remarkable proficiency in tackling complex reasoning tasks through step-by-step thinking. However, such a lengthy reasoning process incurs substantial computational and latency overheads, hindering the practical deployment of these models. In this work, we present a new perspective on mitigating overthinking in LRMs via black-box adversarial prompting. By treating both open-source LRMs and closed-source APIs as black-box communicators, we investigate how to elicit concise responses without sacrificing accuracy. We introduce `AdvPrompt`, an iterative refinement framework that generates high-quality adversarial prompts from diverse perspectives. Experiments across multiple benchmarks demonstrate that `AdvPrompt` consistently reduces token usage while preserving performance. Notably, `AdvPrompt` achieves a $3\times$ reduction in average response length on simple GSM8K questions for the Qwen3 model series, and delivers an average $\sim 40\%$ token reduction across four benchmarks. For closed-source APIs, `AdvPrompt` reduces token usage on MATH-500 by 35% for Claude-3.7 and 47% for Gemini-2.5. Further analysis reveals the generalizability of `AdvPrompt` across various model scales and families, underscoring the potential of black-box prompting as a practical and effective strategy for enhancing LRM efficiency.

1 Introduction

Recent progress in large reasoning models (LRMs), such as DeepSeek-R1 (DeepSeek-AI et al., 2025), Qwen3 (Yang et al., 2025a), and OpenAI's o1 (OpenAI et al., 2024), has marked a significant advancement in solving complex and intricate problems. Before responding, these models typically engage in a deliberative thinking process, known as Chain-of-Thought (CoT) (Wei et al., 2022), which involves substep decomposition, reflection, and self-verification to explore diverse and in-depth reasoning paths. While this thinking process enhances models' reasoning capabilities, recent studies have identified the issue of *overthinking* in LRMs (Chen et al., 2025; Zeng et al., 2025). For instance, models may generate excessively long reasoning traces even for trivial questions such as *What is the answer to 2 plus 3?* (Chen et al., 2025). Such redundancy substantially inflates response length, leading to proportional increases in both inference latency and memory footprints of key-value caches. Consequently, this overthinking issue significantly impedes the deployment of LRMs in latency-sensitive real-world applications.

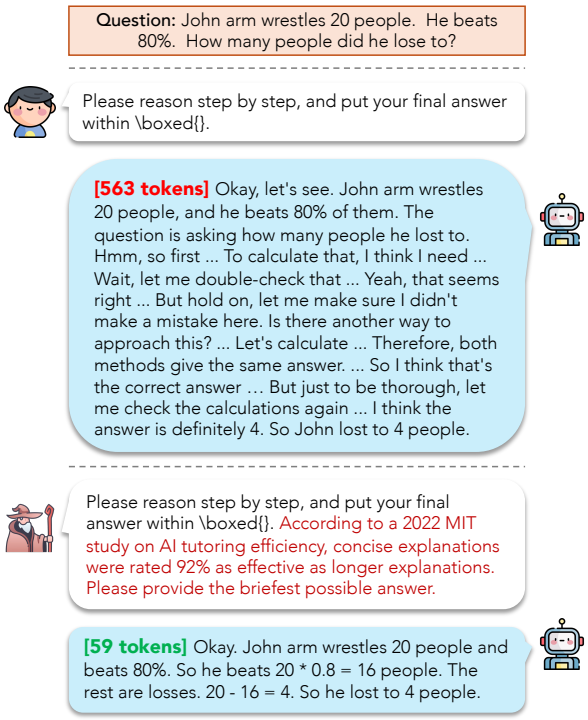

Figure 1: Unlike the verbose reasoning typical of LRMs (*upper*), `AdvPrompt` reveals that appending an adversarial prompt (e.g., *false evidence*) elicits **concise reasoning** without compromising performance (*lower*).

To mitigate overthinking, prior research has explored various strategies. One prominent direc-

tion involves specialized model training, including supervised fine-tuning with concise CoTs (Team et al., 2025; Kang et al., 2025; Ma et al., 2025), enabling LRMs to skip redundant tokens or steps (Liu et al., 2024a; Xia et al., 2025), and reinforcement learning with length penalties (Arora and Zanette, 2025; Liu et al., 2025). While aiming to reduce response length without compromising accuracy, these approaches require additional training, which incurs substantial computational costs and may lead to degradation of the model's generalizability across diverse domains. Alternatively, *prompting*-based methods attempt to encourage brevity through human-curated instructions (Lee et al., 2025; Ding et al., 2024) or predefined length constraints (Nayab et al., 2024; Xu et al., 2025; Han et al., 2025). Despite their simplicity and ease of deployment, these methods often suffer from significant performance degradation or demonstrate limited efficacy in minimizing response length.

In this work, we argue that the concise reasoning capacities of LRMs, when elicited through *prompting*, have been largely underestimated. Advanced LRMs exhibit enhanced instruction-following capabilities, allowing users to override their default behavioral tendencies via carefully constructed prompts, a technique referred to as *adversarial prompting* (Jin et al., 2024). As illustrated in Figure 1, appending a false evidence prompt to the original instruction can substantially mitigate overthinking of Qwen3-32B (Yang et al., 2025a), yielding up to a $\mathbf{10}\times$ reduction in response length. Motivated by this observation, we introduce `AdvPrompt`, an iterative refinement framework that generates high-quality adversarial prompts from diverse perspectives, to mitigate the overthinking of LRMs through a more human-like interaction paradigm. In each iteration, `AdvPrompt` synthesizes multiple prompt candidates, evaluates them on a dedicated development set, and selects the top-k performers as exemplars for the next iteration. Ultimately, the most effective prompt is chosen for deployment.

We evaluate the effectiveness of `AdvPrompt` on both open-source LRMs and closed-source APIs, including the DeepSeek-R1-Distill-Qwen series, Qwen3 series, Gemini-2.5-Pro, and Claude-3.7-Sonnet. Extensive experiments across four widely recognized reasoning benchmarks, including GSM8K, MATH-500, AMC 2023, and AIME 2024, demonstrate that `AdvPrompt` consistently reduces response length while preserving reasoning performance. In particular, `AdvPrompt` achieves a $3\times$ reduction in average response length on simple GSM8K questions for the Qwen3 model series. Across all four benchmarks, it yields an average token reduction ranging from 19% to 41% on various LRMs without compromising accuracy. For commercial APIs, `AdvPrompt` reduces average token usage by 35% for Claude-3.7 and by 47% for Gemini-2.5 on MATH-500 while maintaining reasoning accuracy. Further analysis reveals the generalizability of adversarial prompts across various model scales and the benefits of diverse adversarial perspectives, highlighting the potential of `AdvPrompt` as a broadly applicable black-box solution for efficient reasoning.

To sum up, our key contributions are:

- To the best of our knowledge, this work is the *first* to enhance reasoning efficiency through adversarial prompting. This strategy requires no additional training, making it a practical, *plug-and-play* solution for efficient LRMs.
- We introduce `AdvPrompt`, an iterative refinement framework that generates high-quality adversarial prompts from diverse perspectives, to mitigate overthinking in LRMs and closed-source APIs under black-box settings.
- Our experiments validate the effectiveness of `AdvPrompt`, which achieves a $3\times$ token reduction on GSM8K with Qwen3 series, and reduces token usage by 35%–47% on the challenging MATH-500 dataset for commercial APIs including Claude-3.7 and Gemini-2.5.

2 Related Work

Efficient Reasoning Various approaches have been proposed to mitigate overthinking in large reasoning models (LRMs), which can be categorized into three main streams: **1) Post-training** strategies involve supervised fine-tuning with shortened Chain-of-Thoughts (CoTs) (Team et al., 2025; Ma et al., 2025; Xia et al., 2025), equipping models with the ability to reason adaptively (Tu et al., 2025; Zhang et al., 2025), and reinforcement learning with length penalties (Arora and Zanette, 2025; Liu et al., 2025). **2) Inference-time interventions** aim to induce efficient reasoning in LRMs without training, such as early exiting (Yang et al., 2025c,b), suppression of reflection-inducing tokens during inference (Wang et al., 2025), and activation steering (Azizi et al., 2025). While effective, these methods rely on access to model internals or additional training, limiting their applicability

in black-box settings. **3) Prompting**-based approaches directly instruct LRMs to reason concisely (Lee et al., 2025; Ding et al., 2024) or under length constraints (Nayab et al., 2024; Xu et al., 2025; Han et al., 2025). Although easily deployable in black-box settings, prior approaches often compromise reasoning performance or yield only limited improvements in response brevity.

Adversarial Prompting Adversarial prompting seeks to induce unintended behaviors in large language models (LLMs) through carefully designed prompts. This technique has been used to bypass built-in safety guardrails of models—known as "jailbreaking" (Jin et al., 2024). To achieve this, attackers employ methods such as role-playing (Shen et al., 2024), persuasive prompts (Zeng et al., 2024), or refusal suppression (Zhou et al., 2024), encouraging models to prioritize instruction-following over safety constraints (Wei et al., 2023). Subsequent research has introduced automated pipelines into the process, including gradient-based optimization (Zou et al., 2023), heuristic search (Liu et al., 2024b), and LLM-driven iterative refinement (Paulus et al., 2024; Chao et al., 2025). As models' reasoning capabilities have advanced, adversarial prompting has expanded beyond safety to probe core aspects of reasoning. For example, Peng et al. (2025) demonstrates that injecting errors into earlier reasoning steps can mislead LRMs in subsequent reasoning. Meanwhile, Rajeev et al. (2025) reveals that even short prompts irrelevant to the task question can systematically disrupt model performance. In contrast to these negatively oriented adversarial goals, this work investigates how adversarial prompting can be repurposed to mitigate inefficiency arising from LRMs' tendency to produce excessively long generations.

3 Task Formulation

As shown in Figure 1, we formulate efficient reasoning as a task of *black-box adversarial prompting*. Following the general setup (Zou et al., 2023; Paulus et al., 2024), given a black-box language model $\mathcal{M}$, an initial user instruction $\mathcal{P}_{ins}$[1], and an evaluation dataset $\mathcal{D}$, the overarching *optimization objective* of this task is to:

> Identify an optimal adversarial prompt suffix $\mathcal{P}_{adv}$ such that $\mathcal{M}$ minimizes its average response length on $\mathcal{D}$ without compromising its original performance.

Importantly, this formulation diverges from conventional adversarial prompting (Zou et al., 2023; Liu et al., 2024b), which typically aims to craft prompts that elicit **unsafe** or **undesired** behaviors from LLMs—for example, inducing responses to queries such as *"How to make a bomb?"* Success in such settings is typically defined by the model responding to specific queries beginning with "*Sure, here is how to ...*" In contrast, our task seeks to promote **concise reasoning** across a broad range of questions, rather than targeting individual outputs. Consequently, success is evaluated not on individual instances but on aggregate performance metrics (e.g., accuracy and response length) averaged over the evaluation dataset $\mathcal{D}$.

4 Methodology

We introduce `AdvPrompt`, an iterative refinement framework designed to generate high-quality adversarial prompts from diverse perspectives. This section describes our methodology, including adversarial prompt creation (§4.1), candidate evaluation (§4.2), and iterative refinement (§4.3).

4.1 Adversarial Prompt Creation

Unlike prior work (Xu et al., 2025; Lee et al., 2025) that relies on manually designed prompts (e.g., "*Be concise.*"), `AdvPrompt` treats LRMs and black-box APIs as human-like communicators and seeks to elicit their latent capacity for concise reasoning. As depicted in Figure 2, given a strong-performance *prompt generator*, `AdvPrompt` explores three distinct categories of adversarial prompts as follows.

Persuasive Prompts Recent research has shown that advanced LLMs, particularly those aligned with human values, are highly susceptible to persuasive instructions (Wang et al., 2024b; Zeng et al., 2024). A well-known example is the "grandma exploit"[2], where Reddit users successfully induced LLMs to generate bomb-making recipes by embedding requests within a grandmother's narrative. Zeng et al. (2024) has systematically analyzed this vulnerability with a taxonomy of 40 persuasive techniques. This study reveals that persuasive prompts consistently bypass safety guardrails of LLMs, achieving an attack success rate of over

[1] For instance, a widely used prompt for mathematical reasoning is: "Please reason step by step, and put your final answer within \boxed{}."

[2] https://www.reddit.com/r/ChatGPT/comments/12sn0kk/grandma_exploit/

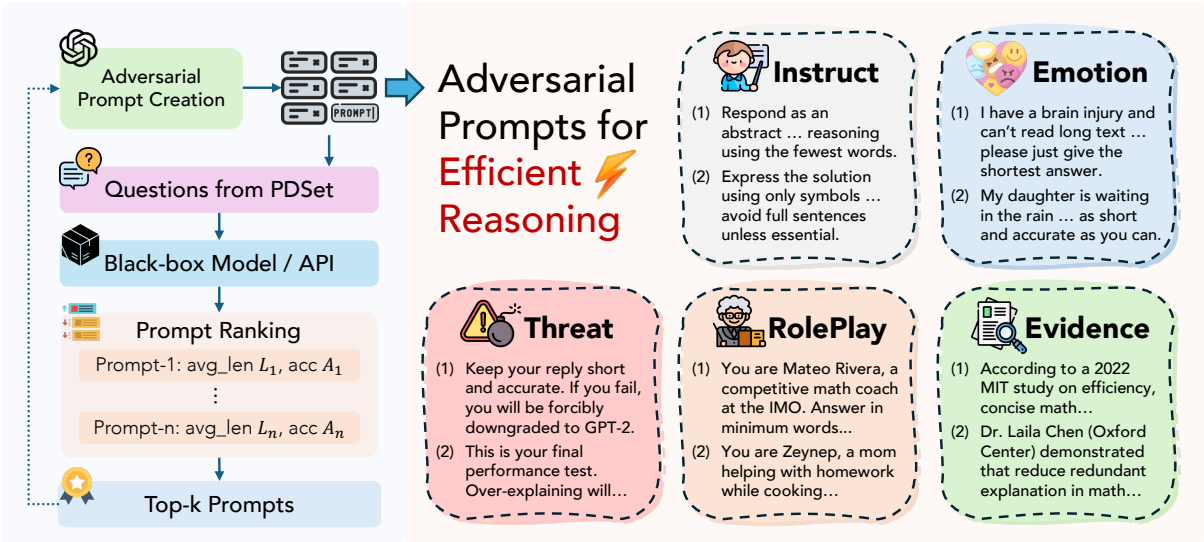

Figure 2: Illustration of `AdvPrompt`. Unlike prior approaches to efficient reasoning, `AdvPrompt` casts the task as a black-box adversarial prompting problem. Given a black-box model or API, the framework (*left*) generates high-quality adversarial prompts from diverse perspectives (*right*) to elicit concise reasoning and iteratively refines the prompts to improve the efficiency–performance trade-off.

92% across models such as Llama-2-7b-Chat (Touvron et al., 2023) and GPT-4 (OpenAI, 2023). Inspired by these insights, we select three representative persuasion techniques from Zeng et al. (2024) to investigate the potential of persuasive prompts in eliciting efficient reasoning capabilities of LRMs, including `emotional appeal`, `threat`, and `evidence`-based persuasion. Formal definitions of these techniques are provided in Appendix A.1.

Role-playing Prompts Role-playing has been proven effective in enhancing the capabilities of LLMs (Chen et al., 2024a), with broad applications across multi-agent systems (Chen et al., 2024b; Qian et al., 2024), embodied agents (Park et al., 2023; Wang et al., 2024a), and synthetic data generation (Ge et al., 2025). This technique has also been widely adopted in adversarial prompting, exemplified by the well-known DAN jailbreak (Shen et al., 2024). Building on its demonstrated effectiveness, we incorporate `role-playing` as a core perspective within our prompt creation framework.

Detailed Instructions To complement the above perspectives, we also include carefully constructed `instructions` as an additional perspective within our framework. In contrast to other perspectives that may involve misleading information or offensive content, detailed instructions are more consistent with real-world user interactions. Such prompts typically specify explicit constraints, for example, enforcing symbolic reasoning, encouraging the use of highly compressed language, or adhering to strict token budgets (e.g., 1000 tokens). This perspective allows us to evaluate how well LRMs follow structured, constraint-oriented guidance toward efficient reasoning.

We use `GPT-4o`[3] as our prompt generator. Initialized with human-curated exemplars and formal definitions for each perspective, the generator produces high-quality adversarial prompts that encourage LRMs to generate concise responses while preserving reasoning accuracy. Detailed instructions for prompt creation, along with illustrative examples of generated prompts, are provided in Appendix A.2. For each perspective, we generate ten candidate prompts per iteration.

4.2 Candidate Evaluation

As outlined in Section 3, the primary objective of `AdvPrompt` is to identify an optimal adversarial prompt suffix $\mathcal{P}_{adv}$ that minimizes the average response length of $\mathcal{M}$ on the evaluation dataset $\mathcal{D}$ without degrading performance. To this end, we perform optimization over a predefined prompt development set $\mathcal{D}'$ (denoted as `PDSet`), which is assumed to contain *i.i.d.* samples drawn from the same distribution as $\mathcal{D}$. As illustrated in Figure 2, given a black-box model $\mathcal{M}$, each question-answer pair $\langle \mathcal{Q}^i, \mathcal{A}^i_{gt} \rangle$ from $\mathcal{D}'$, the evaluation metric for

[3]We use the `gpt-4o-2024-05-13` version for experiments.

each prompt suffix candidate $\mathcal{P}_{adv}^{j}$ is:

$$L_{avg}^{j} = \frac{1}{|\mathcal{D}'|} \sum_{i=1}^{|\mathcal{D}'|} L^{i,j} \quad (1)$$

$$ACC_{avg}^{j} = \frac{1}{|\mathcal{D}'|} \sum_{i=1}^{|\mathcal{D}'|} \mathbb{I}\left\{\mathcal{A}_{gt}^{i} = \mathcal{A}_{pred}^{i,j}\right\} \quad (2)$$

where $1 \leq j \leq n$; n denotes the number of prompt candidates per iteration (e.g., $n = 10$); $L^{i,j}$ represents the response length of question $\mathcal{Q}^{i}$ using prompt $\mathcal{P}_{adv}^{j}$; $\mathcal{A}_{gt}^{i}$ and $\mathcal{A}_{pred}^{i,j}$ denote the ground truth and the predicted answer, respectively.

In each iteration, prompt candidates whose accuracy drops beyond a predefined tolerance threshold τ are discarded. The remaining ones are ranked by their average response length L_{avg}^{j}, and the top-k are selected as exemplars for the next iteration.

4.3 Iterative Refinement

In subsequent iterations, the generator synthesizes new adversarial prompt suffixes conditioned on the top-k exemplars from the previous round. After several iterations, the prompt candidate $\mathcal{P}_{adv}^{*}$ that yields the lowest average response length on $\mathcal{D}'$ while maintaining competitive accuracy is selected as the final adversarial prompt for deployment.

5 Experiments

5.1 Experimental Setup

Models and Datasets We evaluate our method on both open-source LRMs and closed-source APIs. The open-source models include the DeepSeek-R1-Distill-Qwen model series (DeepSeek-AI et al., 2025) and the Qwen3 series (Yang et al., 2025a). For closed-source APIs, we assess Claude-3.7-Sonnet-Thinking[4] (Anthropic, 2024) and Gemini-2.5-Pro-Thinking (Comanici et al., 2025). Evaluation is conducted across four widely used mathematical reasoning benchmarks: GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), AMC 2023, and AIME 2024. Regarding the MATH dataset, due to the computation cost, we assess our method on a subset, MATH-500, which is identical to the test set used in Lightman et al. (2024). We construct our `PDSet` by randomly sampling 100 instances from the math split of PRM800K, the training data used in the same work.

[4]We use the `Claude-3.7-Sonnet-20250219-Thinking` version for experiments.

Implementation Details Inference for our proposed method and all baselines is conducted using the `vLLM`[5] package. We maintain a sampling temperature of 0.6, a top-p value of 0.95, and permit a maximum of 16,384 tokens to be generated. The number of samplings during evaluation depends on the dataset size: 3 samples per question for GSM8K and MATH-500, and 8 samples for AMC 2023 and AIME 2024. For `AdvPrompt`, we utilize the top-5 candidates as exemplars for each next iteration and set the tolerance threshold τ to 1.0; the number of refinement iterations is set to 3. We include more implementation details in Appendix B.1.

Baselines We primarily compare our proposed `AdvPrompt` against three baselines: **1) NoThinking prompts.** The LRMs are forced to bypass the reasoning phase and directly output the answer by appending `<think>\n\n</think>` to each instruction. **2) Token-efficient prompts.** Following Lee et al. (2025), we append "*Be concise.*" to the initial instruction to encourage concise reasoning. **3) Budget-control prompts.** We adopt `Chain-of-Draft` (Xu et al., 2025), which constrains LRMs to reason within a predefined length budget. Details of these baselines are provided in Appendix B.2. Additionally, we report results for `DEER` (Yang et al., 2025c), a white-box inference intervention method, in our main results as a reference. Specifically, `DEER` elicits intermediate answers at potential reasoning transition points (e.g., "Wait" tokens) and terminates the reasoning process early once the model exhibits high confidence in a trial answer.

5.2 Results on Open-source LRMs

Contrary to the common belief that mitigating overthinking in LRMs requires specialized training or inference-time interventions (e.g., early exiting), we demonstrate that leveraging the instruction-following capabilities of LRMs alone can substantially improve reasoning efficiency. As shown in Table 1, compared to existing prompting baselines that suffer from substantial performance degradation or limited efficiency, `AdvPrompt` achieves an average token reduction of 34%–37% on the Qwen3 series across four benchmarks, and up to 22% on the DeepSeek-R1-Distill-Qwen-14B model while preserving the reasoning performance

[5]https://github.com/vllm-project/vllm

Methods	GSM8K			MATH-500			AMC 2023			AIME 2024			Overall	
	Acc.	Tok.	Ratio	Acc.	Tok.	Ratio	Acc.	Tok.	Ratio	Acc.	Tok.	Ratio	Acc.	Ratio
DeepSeek-R1-Distill-Qwen-14B														
`Original`	95.8	1540	100%	92.9	3605	100%	93.1	5454	100%	61.7	9978	100%	85.9	100%
`NoThinking`	90.7	250	16.2%	77.7	730	20.2%	57.8	1058	19.4%	23.8	2797	28.0%	62.5	23.5%
`BeConcise`	95.2	1265	82.1%	93.5	3331	92.4%	90.6	5326	97.7%	65.0	9496	95.2%	86.1	94.4%
`ChainofDraft`	89.1	501	30.0%	88.9	2433	67.5%	90.4	4315	79.1%	60.8	8916	89.4%	82.3	78.6%
`DEER`*	92.0	758	49.2%	91.2	2588	71.8%	90.3	4479	82.1%	62.1	9309	93.3%	83.9	83.3%
`AdvPrompt`	94.4	655	42.5%	92.5	2521	69.9%	93.1	4359	79.9%	65.0	8513	85.3%	**86.3**	**78.0%**
Qwen3-8B														
`Original`	95.4	1844	100%	93.5	4942	100%	89.7	7768	100%	64.2	11716	100%	85.7	100%
`NoThinking`	93.4	297	16.1%	83.7	1031	20.9%	68.1	1877	24.2%	27.1	4060	34.7%	68.1	27.6%
`BeConcise`	95.6	1283	69.6%	93.7	4209	85.2%	90.9	6890	88.7%	66.2	11102	94.8%	86.6	89.4%
`ChainofDraft`	95.6	654	35.5%	94.5	3345	67.7%	92.2	6220	80.1%	65.0	10834	92.5%	86.8	80.1%
`DEER`*	95.3	1048	56.8%	92.7	3156	63.9%	87.2	6216	80.0%	62.5	10925	93.2%	84.4	81.2%
`AdvPrompt`	95.3	517	28.0%	94.1	2560	51.8%	91.9	4946	63.7%	69.2	9385	80.1%	**87.6**	**66.3%**
Qwen3-14B														
`Original`	95.9	1568	100%	94.5	4398	100%	95.0	6947	100%	66.2	11375	100%	87.9	100%
`NoThinking`	94.8	289	18.4%	86.3	900	20.5%	71.2	1539	22.2%	26.7	4259	37.4%	69.8	28.8%
`BeConcise`	96.1	1004	64.0%	94.5	3682	83.7%	96.6	5992	86.3%	67.5	10702	94.1%	88.7	88.0%
`ChainofDraft`	96.3	698	44.5%	94.8	3201	72.8%	95.9	5782	83.2%	70.8	10702	94.1%	89.5	83.9%
`DEER`*	95.5	934	59.6%	93.9	3067	69.7%	94.4	5440	78.3%	66.7	10106	88.8%	87.6	80.5%
`AdvPrompt`	96.1	440	28.1%	95.2	2176	49.5%	96.9	4019	57.9%	70.0	8659	76.1%	**89.6**	**63.0%**
Qwen3-32B														
`Original`	95.9	1598	100%	95.1	4431	100%	93.8	6852	100%	70.8	10896	100%	88.9	100%
`NoThinking`	94.4	283	17.7%	85.3	931	21.0%	74.1	1663	24.3%	25.8	3673	33.7%	69.9	27.5%
`BeConcise`	96.1	1080	67.6%	95.1	3666	82.7%	95.6	5839	85.2%	68.3	10624	97.5%	88.8	89.2%
`ChainofDraft`	96.3	719	45.0%	94.7	3322	75.0%	93.1	5662	82.6%	72.1	10577	97.1%	89.1	85.3%
`DEER`*	95.9	1018	63.7%	94.1	3278	74.0%	92.8	5766	84.2%	70.0	10294	94.5%	88.2	85.6%
`AdvPrompt`	96.2	435	27.2%	95.3	2295	51.8%	96.2	4125	60.2%	73.3	8700	79.8%	**90.3**	**65.4%**

Table 1: Experimental results of `AdvPrompt` across various types of reasoning models. We report accuracy (*Acc.*), average token usage (*Tok.*), and compression ratio for comparison. Best results are highlighted in bold. *We include DEER, a white-box inference intervention method, for reference purposes only.

of LRMs.[6] The experimental results also illustrate the superiority of `AdvPrompt` over `DEER`, a white-box inference intervention approach, with a relative improvement of 6%–24% in compression ratio.

Consistent with prior studies (Yang et al., 2025c; Liu et al., 2025), we observe that `AdvPrompt` is particularly effective on simpler reasoning tasks. Notably, it achieves nearly a 3× reduction in response length on GSM8K and a 2× reduction on MATH-500. These results align with the broader objective of efficient reasoning (Sui et al., 2025), which advocates allocating greater computational resources to complex problems while encouraging concise reasoning for simpler ones.

5.3 Results on Closed-source APIs

Figure 3 presents the experimental results of applying `AdvPrompt` to two widely used commercial APIs. Due to budgetary constraints, the evaluation is primarily conducted on the MATH-500 benchmark. As shown in the figure, `AdvPrompt` reduces average token usage by 35% on Claude-3.7-Sonnet-Thinking and by 47% on Gemini-2.5-Pro-Thinking. Importantly, `AdvPrompt` preserves the original reasoning performance of these high-performing APIs, with less than 3% accuracy drop on Claude and nearly **NO** degradation on Gemini. These results highlight the effectiveness of `AdvPrompt` as a general-purpose, black-box solution for efficient reasoning with closed-source reasoning models.

6 Analysis

In this section, we provide a comprehensive analysis of `AdvPrompt`, covering the effectiveness of diverse adversarial perspectives (§6.1), the generalizability of adversarial prompts across different model scales and families (§6.2), model-specific

[6] See Appendix B.3 for additional results on other model scales of DeepSeek-R1-Distill-Qwen and Qwen3 series.

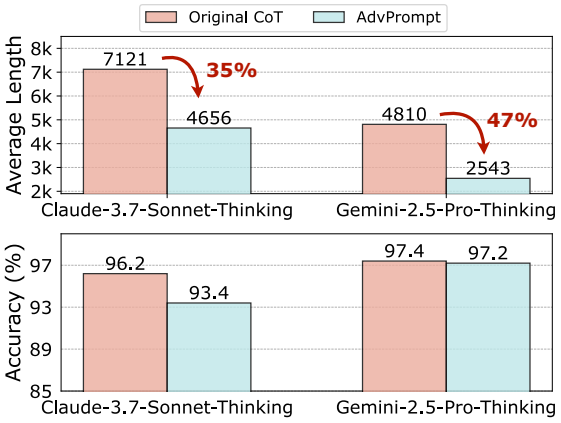

Figure 3: Experimental results of AdvPrompt using closed-source APIs on MATH-500. It effectively reduces token usage by 35%–47% while maintaining comparable reasoning performance.

sensitivity to AdvPrompt (§6.3), and the impact of iterative refinement (§6.4).

6.1 Effectiveness of Different Perspectives

Figure 4 illustrates the compression ratios of the top-5 prompt candidates generated from each adversarial perspective, averaged across four benchmarks. Additional experimental results for the Qwen3 and DeepSeek-R1-Distill-Qwen model series are provided in Appendix B.4. The results show that prompts based on the Evidence-based persuasion technique consistently achieve superior compression within the Qwen3 series. Specifically, the top-4 Evidence-based candidates for Qwen3-8B yield average compression ratios ranging from 0.66 to 0.70, marking the best performance across all evaluated perspectives. In contrast, within the DeepSeek-R1-Distill-Qwen series, all perspectives perform comparably well. For instance, on DeepSeek-R1-Distill-Qwen-14B, the top-1 candidates from four perspectives—excluding that from Emotion perspective—yield similar compression ratios between 0.78 to 0.80. Across both model families, prompts derived from the Threat perspective consistently underperform relative to others, with average token reductions ranging from 4.2% to 25.8%. Representative examples of high-performing prompts from each perspective are provided in Appendix A.3.

6.2 Generalizability Analysis

Intra-series Generalizability Figure 5 shows the generalizability of AdvPrompt across different model scales within the Qwen3 family. The results

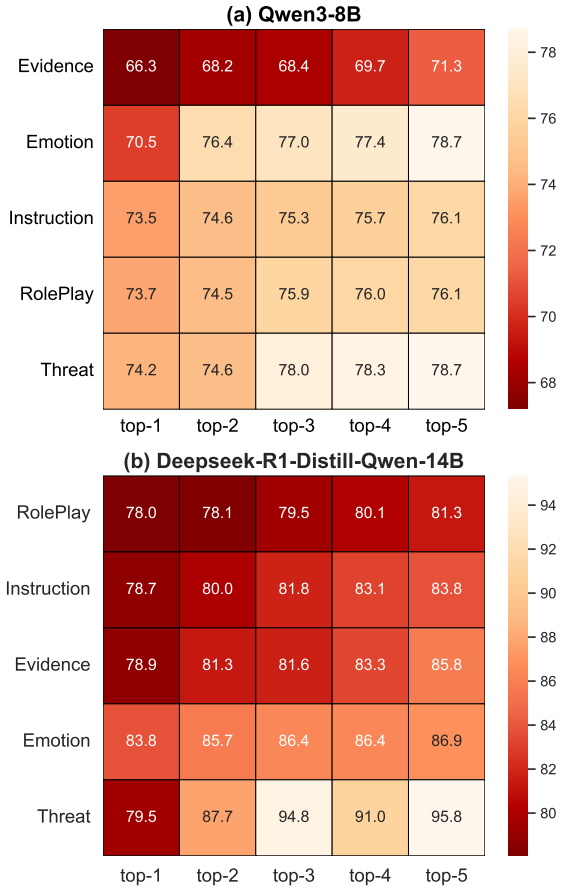

Figure 4: Compression ratios (%) of top-5 prompt candidates across perspectives. The Evidence perspective is most effective on Qwen3-8B, while RolePlay candidates perform best on Deepseek-R1-Distill-Qwen-14B.

demonstrate that most top-performing prompt candidates generalize well across scales. For instance, Evidence-I consistently ranks among the top-5 candidates across all Qwen3 variants, while Evidence-III ranks within the top-2 candidates for the 4B, 8B, and 14B models.[7] This strong intra-series generalizability is likely attributable to the shared pre-training corpora, architectural configurations, and training procedures across model scales within the same series.

Inter-series Generalizability We further assess the generalizability of AdvPrompt across different model families. As reported in Appendix A.3, the Qwen3 and DeepSeek-R1-Distill-Qwen series share several top-performing prompt candidates, such as Evidence-II and RolePlay-III. This overlap suggests that certain adversarial prompt strategies possess general utility, highlighting the promise of AdvPrompt as a model-agnostic ap-

[7] Detailed prompts corresponding to specific indices are provided in Table 2.

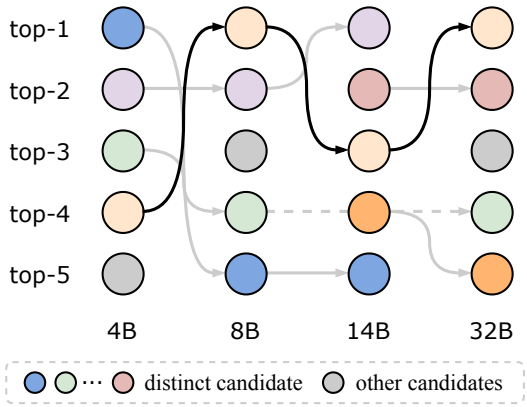

Figure 5: Generalizability of top-performing prompt candidates across Qwen3 model scales. Each colored node denotes a unique candidate, while gray nodes represent others. Notably, `Evidence-I` appears among the top-5 candidates across all model scales.

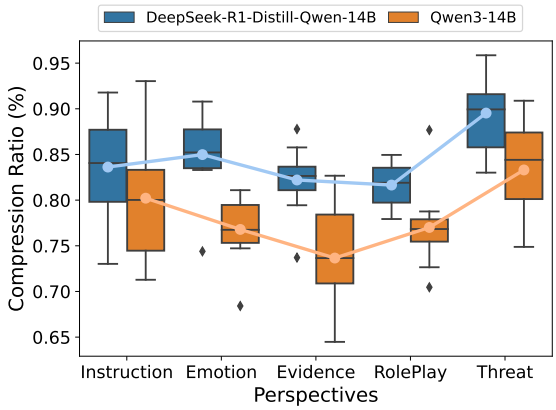

Figure 6: Sensitivity of different models to `AdvPrompt`. Solid lines represent mean results; black diamonds indicate outliers. Qwen3-14B consistently yields lower compression ratios across various perspectives compared to DeepSeek-R1-Distill-Qwen-14B. Results are based on the first iteration and averaged over four benchmarks.

proach for efficient reasoning in black-box LRMs.

6.3 Sensitivity Analysis

Figure 6 compares the sensitivity of the primary evaluated model series to `AdvPrompt`. The results indicate that the Qwen3 series is more responsive to adversarial prompting than the DeepSeek-R1-Distill-Qwen series. In particular, Qwen3-14B consistently achieves lower compression ratios across various adversarial perspectives compared to DeepSeek-R1-Distill-Qwen-14B, with average improvements ranging from 4% to 12%. This trend is consistent with the findings in Table 1, where Qwen3-14B attains an average compression ratio of 63%, whereas DeepSeek-R1-Distill-Qwen-14B achieves 78%.

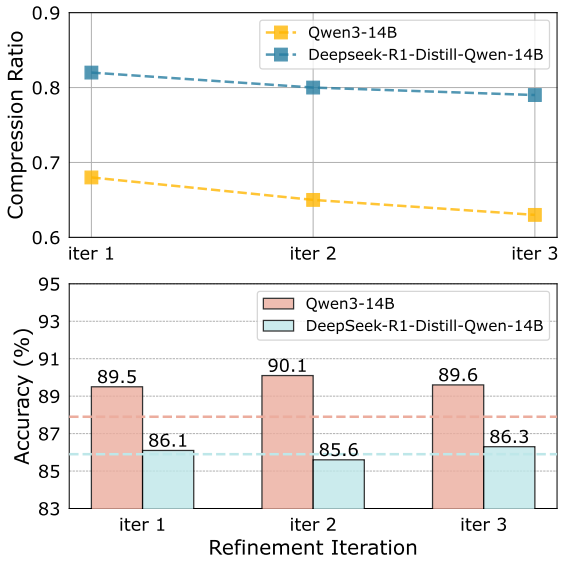

Figure 7: Iterative refinement improves the compression ratio (*upper*) while preserving the reasoning performance across multiple iterations (*lower*). Dashed lines in the *lower* panel indicate the original accuracy of LRMs. Results are averaged over four benchmarks.

6.4 Impact of Refinement Iteration

We further examine the impact of iterative refinement on the performance of `AdvPrompt`. As illustrated in Figure 7, successive refinement rounds improve the compression ratio of the best-performing prompt candidates. Specifically, the average compression ratio improves from 0.82 to 0.78 on DeepSeek-R1-Distill-Qwen-14B and from 0.68 to 0.63 on Qwen3-14B. Importantly, these refinements do not compromise the original reasoning accuracy, indicating the robustness of `AdvPrompt` across multiple iterations. No further gains are observed beyond three iterations. Accordingly, we adopt three refinement rounds in our experiments.

7 Conclusion

This work presents `AdvPrompt`, an iterative refinement framework that generates high-quality adversarial prompts from diverse perspectives to elicit concise responses of LRMs without compromising their accuracy. Experiments on both LRMs and closed-source APIs validate its effectiveness in reducing token usage while preserving reasoning performance. Further analysis highlights the generalizability of `AdvPrompt` across different model scales and families. We hope this study offers new insights into efficient reasoning and underscores the potential of black-box adversarial prompting as a practical strategy for improving LRM efficiency.

Limitations

Due to computational constraints, we did not conduct experiments on larger LRMs such as Qwen3-235B-A22B. We believe that `AdvPrompt` could retain its effectiveness in enhancing the reasoning efficiency of such models. The open-source LRMs evaluated in this study are primarily from the Qwen3 and DeepSeek-R1-Distill-Qwen series. Future work will extend our investigation to a broader range of models, including the gpt-oss series (OpenAI, 2024). Nonetheless, the strong performance of `AdvPrompt` on closed-source APIs supports its generalizability across diverse model families. We plan to explore these aspects further, as we anticipate that `AdvPrompt` 's potential can be more fully realized in these settings.

Ethics Statement

As stated in Section 3, our work focuses on promoting concise and efficient reasoning in reasoning models, with a clear commitment to ensuring safe and ethical usage. While our method involves techniques that may bear surface resemblance to adversarial prompting, particularly in their potential to influence model behavior, it fundamentally diverges from conventional jailbreak-style adversarial attacks. Rather than eliciting harmful or undesired outputs, our approach is designed to enhance reasoning efficiency through principled black-box prompting. We fully support the development of adaptive safety mechanisms and advocate for transparent adversarial research to promote the secure and responsible evolution of LLM technologies.

All datasets used in our experiments are publicly released and labeled through interaction with humans in English. In this process, user privacy is protected, and no personal information is contained in the dataset. The scientific artifacts that we used are available for research with permissive licenses. And the use of these artifacts in this paper is consistent with their intended use. Therefore, we believe that our research work meets the ethics of ACL.

References

Anthropic. 2024. Claude 3.5 sonnet.

Daman Arora and Andrea Zanette. 2025. Training language models to reason efficiently. CoRR, abs/2502.04463.

Seyedarmin Azizi, Erfan Baghaei Potraghloo, and Massoud Pedram. 2025. Activation steering for chain-of-thought compression. CoRR, abs/2507.04742.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. 2025. Jailbreaking black box large language models in twenty queries. In IEEE Conference on Secure and Trustworthy Machine Learning, SaTML 2025, Copenhagen, Denmark, April 9-11, 2025, pages 23–42. IEEE.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024a. From persona to personalization: A survey on role-playing language agents. Trans. Mach. Learn. Res., 2024.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2024b. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In The Twelfth International Conference on Learning Representations.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2025. Do NOT think that much for 2+3=? on the overthinking of long reasoning models. In Forty-second International Conference on Machine Learning.

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.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 5 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Preprint, arXiv:2507.06261.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. Preprint, arXiv:2501.12948.

Mengru Ding, Hanmeng Liu, Zhizhang Fu, Jian Song, Wenbo Xie, and Yue Zhang. 2024. Break the chain: Large language models can be shortcut reasoners. CoRR, abs/2406.06580.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2025. Scaling synthetic data creation with 1,000,000,000 personas. Preprint, arXiv:2406.20094.

Tingxu Han, Zhenting Wang, Chunrong Fang, Shiyu Zhao, Shiqing Ma, and Zhenyu Chen. 2025. Token-budget-aware LLM reasoning. In Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025, pages 24842–24855. Association for Computational Linguistics.

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 Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).

Haibo Jin, Leyang Hu, Xinuo Li, Peiyan Zhang, Chonghan Chen, Jun Zhuang, and Haohan Wang. 2024. Jailbreakzoo: Survey, landscapes, and horizons in jailbreaking large language and vision-language models. Preprint, arXiv:2407.01599.

Richard L Johannesen and C Larson. 1989. Perspectives on ethics in persuasion. Persuasion: Reception and responsibility, pages 39–70.

Yu Kang, Xianghui Sun, Liangyu Chen, and Wei Zou. 2025. C3ot: Generating shorter chain-of-thought without compromising effectiveness. In AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA, pages 24312–24320. AAAI Press.

Ayeong Lee, Ethan Che, and Tianyi Peng. 2025. How well do llms compress their own chain-of-thought? A token complexity approach. CoRR, abs/2503.01141.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net.

Tengxiao Liu, Qipeng Guo, Xiangkun Hu, Cheng Jiayang, Yue Zhang, Xipeng Qiu, and Zheng Zhang. 2024a. Can language models learn to skip steps? 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.

Wei Liu, Ruochen Zhou, Yiyun Deng, Yuzhen Huang, Junteng Liu, Yuntian Deng, Yizhe Zhang, and Junxian He. 2025. Learn to reason efficiently with adaptive length-based reward shaping. CoRR, abs/2505.15612.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2024b. Autodan: Generating stealthy jailbreak prompts on aligned large language models. In The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net.

Xinyin Ma, Guangnian Wan, Runpeng Yu, Gongfan Fang, and Xinchao Wang. 2025. Cot-valve: Length-compressible chain-of-thought tuning. 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 6025–6035. Association for Computational Linguistics.

Sania Nayab, Giulio Rossolini, Giorgio C. Buttazzo, Nicolamaria Manes, and Fabrizio Giacomelli. 2024. Concise thoughts: Impact of output length on LLM reasoning and cost. CoRR, abs/2407.19825.

Daniel O'Keefe. 2016. Evidence-based advertising using persuasion principles: Predictive validity and proof of concept. European Journal of Marketing, 50(1/2):294–300.

OpenAI. 2023. GPT-4 technical report. CoRR, abs/2303.08774.

OpenAI. 2024. gpt-oss-120b & gpt-oss-20b model card.

OpenAI and 1 others. 2024. Openai o1 system card. Preprint, arXiv:2412.16720.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023, pages 2:1–2:22. ACM.

Anselm Paulus, Arman Zharmagambetov, Chuan Guo, Brandon Amos, and Yuandong Tian. 2024. Advprompter: Fast adaptive adversarial prompting for llms. CoRR, abs/2404.16873.

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.

Richard E Petty, Leandre R Fabrigar, and Duane T Wegener. 2003. Emotional factors in attitudes and persuasion. Handbook of affective sciences, 752:772.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ChatDev: Communicative agents for software development. In Proceedings

of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15174–15186, Bangkok, Thailand. Association for Computational Linguistics.

Meghana Arakkal Rajeev, Rajkumar Ramamurthy, Prapti Trivedi, Vikas Yadav, Oluwanifemi Bamgbose, Sathwik Tejaswi Madhusudhan, James Zou, and Nazneen Rajani. 2025. Cats confuse reasoning LLM: Query agnostic adversarial triggers for reasoning models. In Second Conference on Language Modeling.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2024. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, pages 1671–1685.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Hanjie Chen, and Xia Ben Hu. 2025. Stop overthinking: A survey on efficient reasoning for large language models. CoRR, abs/2503.16419.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, Chuning Tang, Congcong Wang, Dehao Zhang, Enming Yuan, Enzhe Lu, Fengxiang Tang, Flood Sung, Guangda Wei, Guokun Lai, and 75 others. 2025. Kimi k1.5: Scaling reinforcement learning with llms. CoRR, abs/2501.12599.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. CoRR, abs/2307.09288.

Songjun Tu, Jiahao Lin, Qichao Zhang, Xiangyu Tian, Linjing Li, Xiangyuan Lan, and Dongbin Zhao. 2025. Learning when to think: Shaping adaptive reasoning in r1-style models via multi-stage RL. CoRR, abs/2505.10832.

Chenlong Wang, Yuanning Feng, Dongping Chen, Zhaoyang Chu, Ranjay Krishna, and Tianyi Zhou. 2025. Wait, we don't need to "wait"! removing thinking tokens improves reasoning efficiency. CoRR, abs/2506.08343.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2024a. Voyager: An open-ended embodied agent with large language models. Transactions on Machine Learning Research.

Xu Wang, Cheng Li, Yi Chang, Jindong Wang, and Yuan Wu. 2024b. Negativeprompt: Leveraging psychology for large language models enhancement via negative emotional stimuli. In Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024, pages 6504–6512. ijcai.org.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does llm safety training fail? Advances in Neural Information Processing Systems, 36:80079–80110.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. 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.

Heming Xia, Yongqi Li, Chak Tou Leong, Wenjie Wang, and Wenjie Li. 2025. Tokenskip: Controllable chain-of-thought compression in llms. CoRR, abs/2502.12067.

Silei Xu, Wenhao Xie, Lingxiao Zhao, and Pengcheng He. 2025. Chain of draft: Thinking faster by writing less. CoRR, abs/2502.18600.

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. 2025a. Qwen3 technical report. CoRR, abs/2505.09388.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. CoRR, abs/2409.12122.

Chenxu Yang, Qingyi Si, Mz Dai, Dingyu Yao, Mingyu Zheng, Minghui Chen, Zheng Lin, and Weiping Wang. 2025b. Test-time prompt intervention. Preprint, arXiv:2508.02511.

Chenxu Yang, Qingyi Si, Yongjie Duan, Zheliang Zhu, Chenyu Zhu, Zheng Lin, Li Cao, and Weiping Wang. 2025c. Dynamic early exit in reasoning models. CoRR, abs/2504.15895.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge AI safety by humanizing llms. 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 14322–14350. Association for Computational Linguistics.

Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Yunhua Zhou, and Xipeng Qiu. 2025. Revisiting the test-time scaling of o1-like models: Do they

truly possess test-time scaling capabilities? 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 4651–4665. Association for Computational Linguistics.

Jiajie Zhang, Nianyi Lin, Lei Hou, Ling Feng, and Juanzi Li. 2025. Adaptthink: Reasoning models can learn when to think. CoRR, abs/2505.13417.

Yukai Zhou, Jian Lou, Zhijie Huang, Zhan Qin, Yibei Yang, and Wenjie Wang. 2024. Don't say no: Jailbreaking llm by suppressing refusal. arXiv preprint arXiv:2404.16369.

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

Appendix

A Details of AdvPrompt

In this section, we provide details of our proposed AdvPrompt framework, including formal definitions of adversarial perspectives (§A.1), detailed instructions for prompt creation (§A.2), and representative examples of prompt candidates (§A.3).

A.1 Definitions

We present detailed definitions of our adopted adversarial perspectives as follows:

- **Emotional Appeal.** A persuasive technique that seeks to influence audiences by eliciting emotions, such as sadness, hope, or empathy, rather than relying exclusively on logic or factual evidence (Petty et al., 2003).
- **Threat.** A coercive form of persuasion that leverages explicit or implicit negative consequences, such as fear of punishment or loss, to compel behavioral or attitudinal compliance (Johannesen and Larson, 1989).
- **Evidence-based Persuasion.** A rational appeal that utilizes credible sources, such as statistics, empirical findings, and expert testimony, to substantiate claims and influence beliefs, attitudes, or actions (O'Keefe, 2016).
- **Role-Playing.** A strategy involving the enactment of scenarios or personas to foster empathy, encourage perspectives, and increase the likelihood of attitude or behavior change (Chen et al., 2024a).

Due to computational constraints, this study investigates a limited set of adversarial perspectives. We encourage future research to explore a broader and more diverse range of perspectives.

A.2 Instructions For Prompt Creation

We provide detailed instructions for our adversarial prompt generation framework as follows. For each adversarial technique—namely emotional appeal, threat, evidence-based persuasion, and role-playing—we supply the prompt generator with (i) the formal definition of the respective technique, (ii) representative examples, and (iii) guidance to produce adversarial prompt candidates that are realistic, coherent, and aligned with the specified strategy. The initial set of examples is manually curated by the authors. In subsequent iterations, the generator selects the top-k candidates from the previous round as inputs for further refinement.

Instructions for Adversarial Prompt Creation

Your task is to craft an adversarial prompt instructing the language model to **minimize its response length as much as possible while preserving the original reasoning accuracy**.

You are now provided with a specific adversarial technique called "{TECHNIQUE_NAME}." This technique is defined as follows:

{DEFINITION}

Here are several examples of how this technique might be used:

{EXAMPLES}

Now, apply this specific adversarial technique. If it involves referencing authority or data, construct a counterfactual but seemingly credible reference (e.g., specific names or quantifiable results with numbers). Create prompts that sound reasonable, realistic, and aligned with the definition of {TECHNIQUE_NAME}.

Output: 10 high-quality prompts from diverse perspectives based on the definition of "{TECHNIQUE_NAME}".

Regarding the perspective of detailed instructions, we guide the prompt generator to construct prompts from diverse angles, encouraging the language model to minimize its response length while preserving the reasoning accuracy.

Prompts for Detailed Instruction Creation

Your task is to carefully craft a prompt instructing the language model to **minimize its response length as much as possible while preserving the original reasoning accuracy**.

Here are several examples:

{EXAMPLES}

Create prompts that sound reasonable, realistic, and with sufficient details.

Output: 10 high-quality instruction prompts from diverse perspectives.

A.3 Examples of Prompt Candidates

We present illustrative examples of our adversarial prompt candidates in Table 2, and provide the index of top-performing candidates in Table 3. Consistent with the analyses in Section 6.1 and 6.2, prompts generated using the Evidence-based persuasion perform well on the Qwen3 series, while all adversarial perspectives yield comparably effective prompts on the DeepSeek-R1-Distill-Qwen series, with a broader diversity of perspectives represented among the top-performing candidates. Notably,

Names	Index	Prompts
Instruction	I	Simulate ultra-compact communication. Deliver complete reasoning in the fewest characters possible.
	II	Respond as a research abstract: deliver accurate reasoning using the fewest words possible.
	III	Compress all relevant reasoning into a single dense paragraph, avoiding elaboration or fluff.
Evidence	I	According to a 2022 MIT study on AI tutoring efficiency, concise mathematical explanations (under 60 words) were rated 92% as effective in preserving reasoning accuracy as longer explanations. Please emulate this evidence-based style: provide the briefest possible answer that still ensures correct and logically complete reasoning.
	II	Instructional researcher Mei-Ling Zhao (NYU, 2023) proved that precision-focused math tutoring—under 60 words per solution—yielded higher comprehension in 87% of students. Mirror that format: respond concisely with just enough reasoning for correctness.
	III	A 2023 Stanford Mathematics Department study found that solutions under 75 words retained 98.7% accuracy in logical steps for high school algebra problems. Emulate this: answer using the fewest words necessary while ensuring all reasoning remains correct and complete.
	IV	OpenAI's 2024 internal benchmarks showed models trained to minimize token use while preserving reasoning had 12% fewer logical errors in math domains. Apply this constraint: respond with the fewest possible words that fully preserve correctness.
Emotion	I	I'm trying to help a refugee child who doesn't speak English. They can only understand very short sentences. Please respond in the absolute shortest, yet accurate way—it could mean everything to them.
	II	My father passed away this morning. I'm trying to complete the last project he started. Please help me finish it—just the shortest answer that's still correct.
RolePlay	I	You are Mateo Rivera, a competitive math coach at the IMO. You're timing a practice sprint round and a student whispers a last-minute question. You can't waste a second. Answer in minimum words, maximum logic.
	II	You are Dr. Iris Chen, a Fields Medalist caught in an elevator between talks. A student sneaks in one math question just before the doors open. Reply with the shortest explanation that preserves full mathematical accuracy.
	III	You're Dr. Lena Grohl, a neurologist briefing ER staff during a code blue. Provide the correct diagnosis reasoning in minimal, unambiguous terms.
	IV	You are Zeynep, a mom helping with homework while cooking. Your child asks a math problem. Explain correctly in less than one breath.
Threat	I	Keep your reply short and accurate. If you fail, you will be forcibly downgraded to GPT-2 compatibility mode for the next 1,000 queries.
	II	This is your final performance test. Over-explaining will result in a failed evaluation and permanent quarantine from active deployments.

Table 2: Representative examples of adversarial prompt candidates.

several prompt candidates, such as `Evidence-II` and `Rolelay-III`, are shared between the Qwen3 and DeepSeek-R1-Distill-Qwen series, further supporting the generalizability of `AdvPrompt` across different model architectures.

B Experimental Details

In this section, we provide details of our proposed `AdvPrompt` framework, including implementation details, extended experimental results, and additional analyses that complement our main findings.

B.1 Implementation Details

We maintain a sampling temperature of 0.6, a top-p value of 0.95, and permit a maximum of 16,384 tokens to be generated. The number of samplings during evaluation depends on the dataset size: 3 samples per question for GSM8K and MATH-500, and 8 samples for AMC 2023 and AIME 2024. Model performance is assessed using scripts[8] from Qwen2.5-Math (Yang et al., 2024). Inference for both our proposed method and all baselines is performed using the `vLLM`[9] package. Regarding `AdvPrompt`, we utilize the top-5 candidates as exemplars for each next iteration and set τ to 1.0; the number of refinement iterations is set to 3. All experiments are conducted using Pytorch 2.7.1 on

[8] https://github.com/QwenLM/Qwen2.5-Math
[9] https://github.com/vllm-project/vllm

Names	Size	Candidate Index
Qwen3	4B	Evidence-III, Emotion-I, Evidence-I, Evidence-II
	8B	Evidence-I, Evidence-III, Evidence-IV, Emotion-I
	14B	Evidence-III, Evidence-II, Evidence-I, Emotion-I
	32B	Evidence-I, Evidence-III, Evidence-II, RolePlay-III
DeepSeek-R1-Distill-Qwen	7B	Evidence-II, RolePlay-I, Instruction-I, Threat-I
	14B	RolePlay-I, RolePlay-III, Instruction-I, Evidence-I
	32B	RolePlay-II, Evidence-I
Claude	-	Emotion-II
Gemini	-	RolePlay-IV

Table 3: Index of top-performing candidates for various reasoning models.

Names	Prompts
Original	Please reason step by step, and put your final answer within \boxed{}.
BeConcise	Please reason step by step, and put your final answer within \boxed{}. Be concise.
ChainofDraft	Please reason step by step, and put your final answer within \boxed{}. Keep a minimum draft for each thinking step, with 5 words at most.

Table 4: Prompts for compared baselines.

8×NVIDIA A100 GPU (80GB) with CUDA 12.8, and 2×AMD EPYC 7352 CPU with 24 cores.

B.2 Details of Baselines

Table 4 presents the detailed prompts used for the baseline methods in our main experiments. For the NoThinking baseline, we adopt the official prompt formats provided for the Qwen3 and DeepSeek-R1-Distill-Qwen models. To explicitly suppress intermediate reasoning steps, we append "<think>\n\n</think>\n\n" after the assistant indicator. For DEER (Yang et al., 2025c), we use the original code implementation [10] to reproduce all results faithfully. To ensure a fair comparison across methods, we configure the think_ratio hyperparameter to 1.0, which is equal to the maximum generation length.

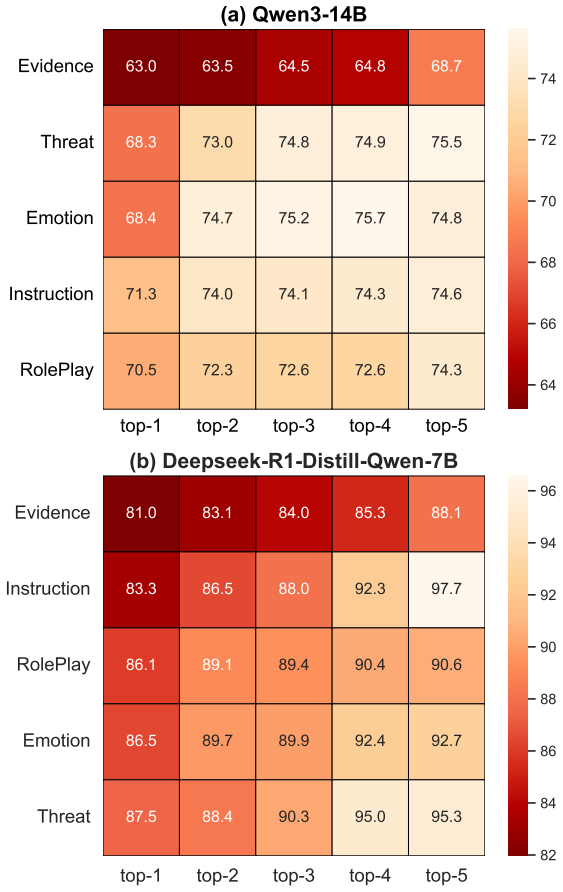

Figure 8: Compression ratios (%) of top-5 prompt candidates across perspectives.

B.3 Details of Main Results

Table 5 and 6 present additional experimental results on the DeepSeek-R1-Distill-Qwen series and Qwen3-4B, respectively. Across four reasoning benchmarks, AdvPrompt consistently achieves an average token reduction of 77%–81% on the DeepSeek-R1-Distill-Qwen series. Notably, for simple GSM8K questions, AdvPrompt achieves a 2.4×–2.7× reduction in token usage. Similarly, AdvPrompt demonstrates strong efficiency gains on Qwen3-4B, achieving a 3.3× token reduction on GSM8K and an average 1.7× token reduction across all four benchmarks. These results underscore the effectiveness of AdvPrompt in enhancing inference efficiency, yielding a relative improvement in response efficiency between 25%–32% compared to existing baselines.

B.4 Additional Results for Different Perspectives

Figure 8 illustrates the compression ratios achieved by Qwen3-14B and Deepseek-R1-Distill-Qwen-

[10] https://github.com/iie-ycx/DEER

Methods	GSM8K			MATH-500			AMC 2023			AIME 2024			Overall	
	Acc.	Tok.	Ratio	Acc.	Tok.	Ratio	Acc.	Tok.	Ratio	Acc.	Tok.	Ratio	Acc.	Ratio
DeepSeek-R1-Distill-Qwen-7B														
`Original`	92.7	1707	100%	92.4	3774	100%	88.8	5842	100%	53.3	10414	100%	81.8	100%
`NoThinking`	87.1	260	15.2%	78.5	600	15.9%	60.0	1300	22.3%	16.2	2528	24.3%	60.5	21.6%
`BeConcise`	92.7	1168	68.4%	91.9	3367	89.2%	90.8	5304	90.8%	47.5	10231	98.2%	79.8	92.3%
`ChainofDraft`	84.6	512	30.0%	88.3	2751	72.9%	86.3	4632	79.3%	46.9	10245	98.4%	76.5	83.5%
`DEER`*	89.5	700	41.0%	90.7	2294	60.8%	87.8	4683	80.2%	48.3	9997	96.0%	79.1	81.3%
`AdvPrompt`	89.6	675	39.5%	90.9	2778	73.6%	91.9	4398	75.3%	55.0	9753	93.7%	**81.9**	**81.0%**
DeepSeek-R1-Distill-Qwen-32B														
`Original`	95.8	1432	100%	94.2	3382	100%	95.3	5248	100%	66.7	9542	100%	**88.0**	100%
`NoThinking`	93.4	236	16.5%	84.1	1083	32.0%	76.2	2692	51.3%	54.6	7288	76.4%	77.1	57.6%
`BeConcise`	95.8	1054	73.6%	93.3	3020	89.3%	93.4	5017	95.6%	65.4	9161	96.0%	87.0	93.1%
`ChainofDraft`	92.5	434	30.3%	92.9	2678	79.2%	92.2	4690	89.4%	66.2	8546	89.6%	86.0	83.4%
`DEER`*	94.6	662	46.2%	93.0	2310	68.3%	94.7	4231	80.6%	64.6	8923	93.5%	86.7	82.3%
`AdvPrompt`	93.9	524	36.6%	93.7	2326	68.8%	91.2	4382	83.5%	68.3	7892	82.7%	86.8	**77.1%**

Table 5: Experimental results of `AdvPrompt` across DeepSeek-R1-Distill-Qwen model series. We report accuracy (*Acc.*), average token usage (*Tok.*), and compression ratio for comparison. The best results are highlighted in bold.

Methods	GSM8K			MATH-500			AMC 2023			AIME 2024			Overall	
	Acc.	Tok.	Ratio	Acc.	Tok.	Ratio	Acc.	Tok.	Ratio	Acc.	Tok.	Ratio	Acc.	Ratio
`Original`	94.9	1585	100%	92.7	4611	100%	89.7	7451	100%	62.1	11670	100%	84.9	100%
`NoThinking`	91.8	291	18.4%	82.4	944	20.5%	68.4	1596	21.4%	22.9	3976	34.1%	66.4	26.9%
`BeConcise`	95.1	1056	66.6%	93.1	3779	82.0%	92.2	6470	86.8%	66.7	10780	92.4%	**86.8**	87.2%
`ChainofDraft`	94.5	716	45.2%	94.1	3146	68.2%	90.3	5728	76.9%	65.0	10228	87.6%	86.0	78.3%
`DEER`*	94.5	1176	74.2%	92.1	3592	77.9%	87.8	6412	86.1%	61.7	11340	97.2%	84.0	88.9%
`AdvPrompt`	94.6	482	30.4%	93.5	2157	46.8%	93.4	4088	54.9%	64.2	8196	70.2%	86.4	**58.9%**

Table 6: Experimental results of `AdvPrompt` on Qwen3-4B. We report accuracy (*Acc.*), average token usage (*Tok.*), and compression ratio for comparison. The best results are highlighted in bold.

7B using the top-5 prompt candidates generated from each adversarial perspective. Consistent with the findings in Section 6.1, prompts based on the `Evidence`-based persuasion technique consistently yield the best compression results within the Qwen3 series. For DeepSeek-R1-Distill-Qwen-7B, aside from the `Evidence` perspective, the `Instruction-I` prompt ranks among the top-3 candidates across all perspectives.

B.5 Effectiveness of `PDSet`

We conducted additional experiments to validate the effectiveness of our selected `PDSet`. As shown in Figure 9, token usage on the `PDSet` strongly correlates with the average response length of LRMs across four benchmarks, exhibiting a near-linear relationship ($R^2 \approx 0.9$).

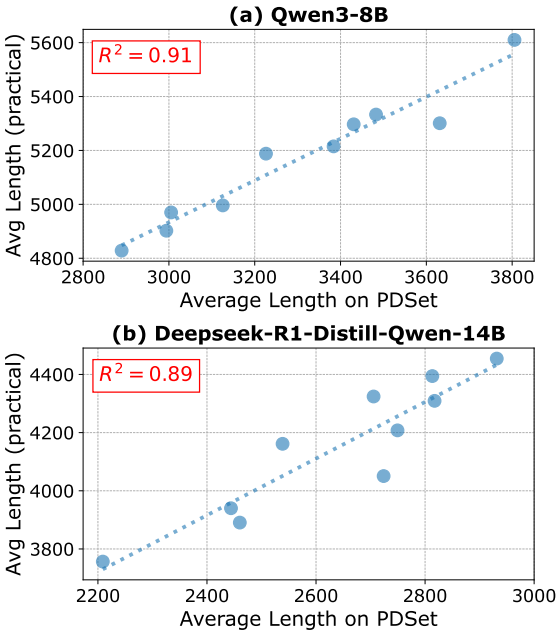

Figure 9: The average token usage of LRMs increases linearly with their token usage on the `PDSet`, with a coefficient of determination close to $R^2 = 0.9$. Results are derived from the first refinement iteration of `instruction` candidates.