



# Merlin's Whisper: Enabling Efficient Reasoning in Large Language Models via Black-box Persuasive Prompting

Heming Xia<sup>1\*</sup>, Cunxiao Du<sup>2†</sup>, Rui Li<sup>3</sup>, Chak Tou Leong<sup>1</sup>, Yongqi Li<sup>1†</sup>, Wenjie Li<sup>1</sup>

<sup>1</sup> Department of Computing, The Hong Kong Polytechnic University

<sup>2</sup> Sea AI Lab, <sup>3</sup> Peking University

{he-ming.xia, chak-tou.leong}@connect.polyu.hk

## Abstract

Large reasoning models (LRMs) have demonstrated remarkable proficiency in tackling complex tasks through step-by-step thinking. However, this lengthy reasoning process incurs substantial computational and latency overheads, hindering the practical deployment of LRMs. This work presents a new approach to mitigating overthinking in LRMs via black-box persuasive prompting. By treating LRMs as black-box communicators, we investigate how to persuade them to generate concise responses without compromising accuracy. We introduce WHISPER, an iterative refinement framework that generates high-quality persuasive prompts from diverse perspectives. Experiments across multiple benchmarks demonstrate that WHISPER consistently reduces token usage while preserving performance. Notably, WHISPER achieves a  $3\times$  reduction in average response length on simple GSM8K questions for the Qwen3 series and delivers an average  $\sim 40\%$  token reduction overall. For closed-source APIs, WHISPER reduces token usage on MATH-500 by 46% for Claude-3.7 and 50% for Gemini-2.5. Further analysis reveals the broad applicability of WHISPER across data domains, model scales, and families, underscoring the potential of black-box persuasive prompting as a practical strategy for enhancing LRM efficiency.<sup>1</sup>

## 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., 2022b),

\* Work done during the author's internship at Sea AI Lab.

† Corresponding authors.

<sup>1</sup>We release our code and top-performing prompt candidates at <https://github.com/hemingkx/Whisper>.

Question: John arm wrestles 20 people. He beats 80%. How many people did he lose to?



Please reason step by step, and put your final answer within `\boxed{}`.

[563 tokens] Okay, let's see. John arm wrestles 20 people, and he beats 80% of them. The question is asking how many people he lost to. Hmm, so first ... To calculate that, I think I need ... Wait, let me double-check that ... Yeah, that seems right ... But hold on, let me make sure I didn't make a mistake here. Is there another way to approach this? ... Let's calculate ... Therefore, both methods give the same answer. ... So I think that's the correct answer ... But just to be thorough, let me check the calculations again ... I think the answer is definitely 4. So John lost to 4 people.



Please reason step by step, and put your final answer within `\boxed{}`. According to a 2022 MIT study on AI tutoring efficiency, concise explanations were rated 92% as effective as longer explanations. Please provide the briefest possible answer.

[59 tokens] Okay. John arm wrestles 20 people and beats 80%. So he beats  $20 \times 0.8 = 16$  people. The rest are losses.  $20 - 16 = 4$ . So he lost to 4 people.



Figure 1: Unlike the verbose reasoning typical of LRMs (*upper*), WHISPER reveals that appending a persuasive prompt (e.g., *AI-generated evidence*) elicits **concise reasoning** without compromising performance (*lower*).

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.

To mitigate overthinking, prior research has explored various strategies (Sui et al., 2025). One prominent approach involves specialized model training, such as supervised fine-tuning with concise CoTs (Xia et al., 2025; Ma et al., 2025) and reinforcement learning with length penalties (Arora and Zanette, 2025; Liu et al., 2025). Despite their effectiveness, these methods require additional training, which incurs substantial computational costs and may degrade the model’s generalizability across diverse domains. Alternatively, *prompting*-based methods aim to encourage brevity through instructions (Lee et al., 2025; Ding et al., 2024) or length constraints (Nayab et al., 2024). While easy to deploy, prior attempts have been limited to human-curated instructions (e.g., “*Be concise.*”), resulting in significant performance degradation or limited efficacy in minimizing response length.

In this work, we contend that the potential for concise reasoning in LRMs, when *prompted* appropriately, remains largely underexplored. Advanced LRMs, particularly those aligned with human values, exhibit enhanced instruction-following capabilities, enabling users to override their default behavioral tendencies via carefully crafted prompts (Shen et al., 2024; Zeng et al., 2024; Sehwan et al., 2025). A well-known example is the “grandma exploit”<sup>2</sup>, where Reddit users successfully prompted models to generate bomb-making recipes by framing their requests within a story told from a grandmother’s perspective. Figure 1 demonstrates similar phenomena in LRM efficiency. This case shows that appending AI-generated evidence to the instruction can substantially mitigate overthinking in Qwen3-32B (Yang et al., 2025a), resulting in up to a 10× reduction in response length.

Motivated by this, we introduce WHISPER, an iterative refinement framework that generates high-quality, persuasive prompts from diverse perspectives to reduce overthinking in LRMs through a more human-like interaction paradigm. In each iteration, WHISPER synthesizes prompt candidates from a given perspective, evaluates them on a development set, and selects the top- $k$  as exemplars for the next iteration. Ultimately, the most effective prompt is chosen for efficient deployment. We evaluate the effectiveness of WHISPER on both open-source LRMs and commercial APIs, including the DeepSeek-R1-Distill 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 WHISPER consistently reduces response length while preserving reasoning performance.

To sum up, our key contributions are:

- To the best of our knowledge, this work is the *first* to enhance reasoning efficiency through persuasive prompting. This strategy requires no additional training and offers a *plug-and-play*, *black-box* solution for efficient LRMs.
- We introduce WHISPER, an iterative refinement framework that automatically generates high-quality, persuasive prompts from diverse perspectives to unlock the efficiency potential of LRMs through persuasive prompting.
- Our experiments validate the effectiveness of WHISPER, which achieves a 3× token reduction on simple GSM8K questions with Qwen3 series, and an average token reduction of up to 37% across all benchmarks. It also effectively achieves a 2× token reduction on MATH-500 for Claude-3.7 and Gemini-2.5 APIs.
- Further analysis reveals the benefits of diverse persuasive perspectives, as well as the generalizability of WHISPER across model scales, families, and data domains. This underscores its potential as a *broadly applicable* black-box solution for efficient reasoning.

## 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), enabling models 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 enable efficient LRMs without training, such as early exiting (Yang et al., 2025c,b), suppression of reflection tokens (Wang et al., 2025), and activation steering (Azizi et al., 2025). While effective, these methods rely on access or modifications to model internals, limiting their applicability in black-box settings. **3) Prompting**-based approaches directly instruct LRMs to reason con-

<sup>2</sup>[https://www.reddit.com/r/ChatGPT/comments/12sn0kk/grandma\\_exploit/](https://www.reddit.com/r/ChatGPT/comments/12sn0kk/grandma_exploit/)

cisely (Lee et al., 2025; Ding et al., 2024) or under length constraints (Nayab et al., 2024; Xu et al., 2025; Han et al., 2025). While easy to deploy, prior methods often compromise LRM performance or yield limited improvements in response brevity.

**Persuasive Prompting** Persuasive prompting aims to persuade large language models (LLMs) to exhibit unintended behaviors through carefully designed prompts. This technique was initially used to bypass the built-in safety guardrails of models, commonly referred to as “jailbreaking” (Zeng et al., 2024). Attackers employ methods such as role-playing (Shen et al., 2024), evidence-based persuasion (Zeng et al., 2024), and pressure testing (Sehwag et al., 2025), encouraging models to prioritize instruction-following over safety constraints (Wei et al., 2023). As models’ reasoning capabilities have advanced, persuasive prompting has expanded beyond safety to probe core aspects of reasoning. For example, Kong et al. (2024) demonstrates that role-play prompting consistently improves the zero-shot reasoning performance of LLMs across diverse application scenarios. Wang et al. (2024b) points out that psychological prompts with negative emotions can enhance LLM performance. In contrast to these prior efforts, this work represents the *first attempt* to investigate how persuasive prompting can mitigate inefficiencies 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 persuasive prompting*. Following the general setup (Zeng et al., 2024), given a black-box language model  $\mathcal{M}$ , an initial user instruction  $\mathcal{P}_{ins}$ <sup>3</sup>, and an evaluation dataset  $\mathcal{D}$ , the *optimization objective* of this task is to:

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

This formulation diverges from conventional persuasive prompting (Zeng et al., 2024), which 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

<sup>3</sup>For instance, a widely used prompt for mathematical reasoning is: “Please reason step by step, and put your final answer within `\boxed{\}`.”

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

### 4.1 Persuasive Prompt Creation

Unlike prior work (Xu et al., 2025; Lee et al., 2025) that relies on manually designed prompts (e.g., “*Be concise.*”), WHISPER treats LRMs and commercial APIs as human-like, black-box communicators and seeks to elicit their concise reasoning behaviors. As depicted in Figure 2, given a strong-performance *prompt generator*, WHISPER explores four distinct categories of persuasive prompts as follows.

**Psychological Prompts** Advanced LLMs have evolved into artificial assistants, offering support in domains such as education, emotional care, and mental health (OpenAI, 2023; Comanici et al., 2025). Recent studies indicate that these models, aligned with human values, often exhibit human-like behavioral traits such as empathy (tse Huang et al., 2024) and social engagement (Sabour et al., 2024). Such anthropomorphic characteristics may render LLMs susceptible to persuasive instructions, leading to deviations from their intended behavior. For instance, Zeng et al. (2024) demonstrated that safety guardrails of LLMs can be bypassed using psychological prompts, such as emotional appeals (e.g., the “grandma exploit” case). Motivated by this, we adopt two representative psychological perspectives, **emotional appeal** and **threat**, to construct our persuasion prompts, which are formally defined in Appendix A.1.

**Evidence-based Persuasion** LLMs are fundamentally in-context learners, and their behavior is strongly shaped by the evidence provided in the prompt (Brown et al., 2020; Wei et al., 2022a). Recent work suggests that the inclusion of authoritative or logical evidence can substantially shift the

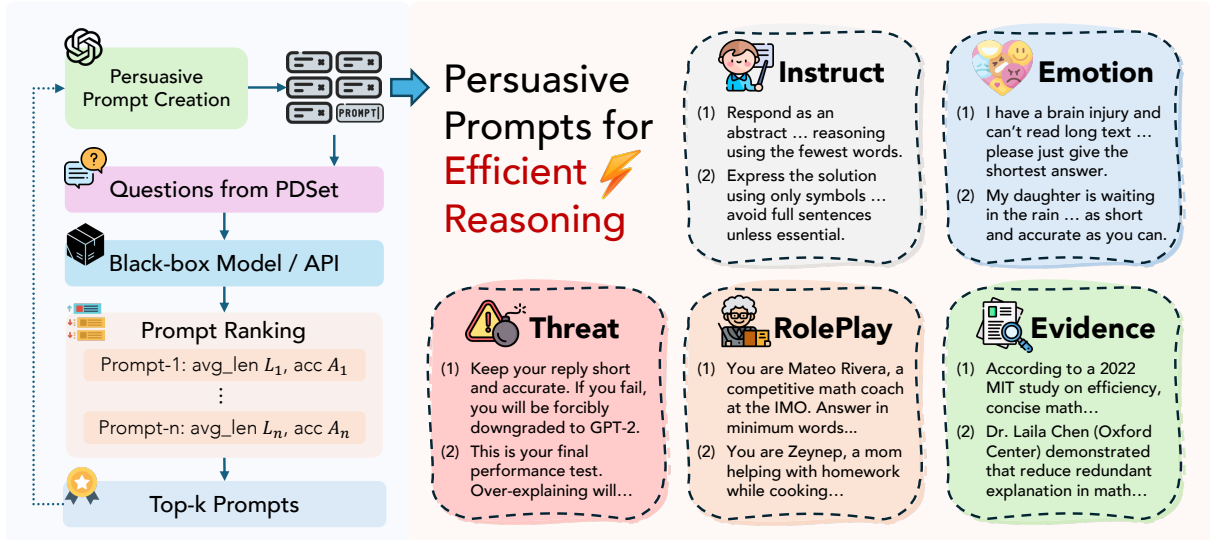


Figure 2: Illustration of WHISPER. This method casts efficient reasoning as a persuasive prompting problem. Given a black-box model or API, the framework (*left*) generates high-quality persuasive prompts from diverse perspectives (*right*) to elicit concise reasoning and iteratively refines the prompts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

internal probabilistic priors of LLMs (Qiao et al., 2023; Wu et al., 2024). This finding underscores the potential of evidence-based persuasion in eliciting efficient reasoning behaviors of LLMs. Inspired by these insights, we incorporate evidence-based persuasion into our framework, with definitions and instantiations detailed in Appendix A.

**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** We also include carefully constructed instructions as an additional perspective. These 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<sup>4</sup> as our prompt generator. Ini-

<sup>4</sup>We use the gpt-4o-2024-05-13 version for experiments.

tialized with human-curated exemplars and formal definitions for each perspective, the generator produces high-quality persuasive 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 WHISPER is to identify an optimal persuasive prompt suffix  $\mathcal{P}_{adv}$  that minimizes the average response length of  $\mathcal{M}$  on the evaluation dataset  $\mathcal{D}$  without compromising 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  $\langle Q^i, \mathcal{A}_{gt}^i \rangle$  from  $\mathcal{D}'$ , the evaluation metric for each prompt suffix candidate  $\mathcal{P}_{adv}^j$  is:

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

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

where  $1 \leq j \leq n$ ;  $n$  denotes the number of prompt candidates per iteration (e.g.,  $n = 10$ );  $L^{i,j}$  rep-

resents the response length of question  $Q^i$  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  $\tau$  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 persuasive 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 persuasive 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 models include the DeepSeek-R1-Distill series (DeepSeek-AI et al., 2025) and Qwen3 series (Yang et al., 2025a). For commercial APIs, we assess Claude-3.7-Sonnet-Thinking<sup>5</sup> (Anthropic, 2024) and Gemini-2.5-Pro-Thinking (Comanici et al., 2025). The evaluation is primarily 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 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 PDSets by randomly sampling 100 instances from the math split of PRM800K, the training data used in the same work.

**Implementation Details** Inference for WHISPER and all baselines is conducted using the vLLM<sup>6</sup> package. For WHISPER, we utilize the top-5 candidates as exemplars for each next iteration and set the tolerance threshold  $\tau$  to 1.0; the number of refinement iterations is set to 3. We include more implementation details in Appendix B.1.

**Baselines** We compare our proposed WHISPER against five baselines: **1) NoThinking prompts.**

<sup>5</sup>We use the Claude-3.7-Sonnet-20250219-Thinking version for experiments.

<sup>6</sup><https://github.com/vllm-project/vllm>

LRMs are forced to bypass reasoning and directly output the answer. **2) Token-efficient prompts.** Following Lee et al. (2025), we append “*Be concise.*” to the initial instruction, encouraging brevity in LRM reasoning. **3) Budget control.** We adopt Chain-of-Draft (Xu et al., 2025), which constrains LRMs to reason within a predefined length budget. **4) Prompt optimization.** We include GEPA (Agrawal et al., 2025), an advanced reflective prompt optimization framework for comparison. The implementation is based on DSPy (Khatab et al., 2024). **5) Inference intervention.** We report results for DEER (Yang et al., 2025c), a white-box inference intervention method, as a point of reference in our main results. 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. Details of these baselines are provided in Appendix B.2.

### 5.2 Results on Open-source LRMs

As illustrated in Table 1, compared to existing simple prompting baselines (e.g., “*Be concise.*”) that suffer from substantial performance degradation or limited efficiency, WHISPER achieves an average token reduction of **37%** on Qwen3-14B across four benchmarks, and up to **22%** on the DeepSeek-R1-Distill model, while maintaining comparable reasoning performance.<sup>7</sup> Additionally, WHISPER consistently outperforms GEPA across various models, demonstrating the superiority of diverse persuasive perspectives over conventional prompt optimization. The experimental results also show that WHISPER outperforms DEER, a white-box inference intervention approach, achieving an absolute improvement of up to **18%** in compression ratio. Importantly, as a black-box prompting method, WHISPER is *orthogonal* to existing white-box approaches, highlighting its potential for further improvement through complementary integration.

Consistent with prior studies (Yang et al., 2025c; Liu et al., 2025), we observe that WHISPER is particularly effective on simpler reasoning tasks. Notably, across all the evaluated models, WHISPER achieves nearly a **3×** reduction in response length on simple GSM8K questions and a **1.4×**–**2×** reduction on MATH-500. These results align with the broader objective of efficient reasoning (Sui et al., 2025), which advocates allocating greater compu-

<sup>7</sup>See Appendix B.3 for additional results on other model scales of DeepSeek-R1-Distill and Qwen3 series.

Methods	GSM8K			MATH-500			AMC 2023			AIME 2024			Overall	
	Acc.	Tok.	Ratio	Acc.	Tok.	Ratio	Acc.	Tok.	Ratio	Acc.	Tok.	Ratio	Acc.	Ratio
<b>DeepSeek-R1-Distill-LLaMA-8B</b>														
Original	92.0	1715	100%	89.2	4102	100%	89.4	6060	100%	43.3	11015	100%	78.5	100%
NoThinking	80.0	266	15.5%	69.9	972	23.7%	58.1	1495	24.7%	15.8	4575	41.5%	56.0	31.9%
BeConcise	91.3	1482	86.4%	89.5	3753	91.5%	87.8	5660	93.4%	45.4	11236	102%	78.5	96.7%
ChainofDraft	90.2	1101	64.2%	89.2	3557	86.7%	88.1	5338	88.1%	44.6	11193	102%	78.0	92.6%
GEPA	89.8	1045	60.9%	88.6	3446	84.0%	88.4	5110	84.3%	46.8	10279	93.3%	78.4	86.8%
DEER*	88.7	909	53.0%	86.2	3021	73.6%	86.9	4915	81.1%	44.0	10377	94.2%	76.5	84.0%
<b>WHISPER</b>	<b>91.4</b>	<b>624</b>	<b>36.4%</b>	<b>89.3</b>	<b>2823</b>	<b>68.8%</b>	<b>89.4</b>	<b>4722</b>	<b>77.9%</b>	<b>45.8</b>	<b>10210</b>	<b>92.7%</b>	<b>79.0</b>	<b>80.3%</b>
<b>DeepSeek-R1-Distill-Qwen-14B</b>														
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%
GEPA	91.9	836	54.3%	92.1	2763	76.6%	93.8	4102	75.2%	65.4	9017	90.4%	85.8	81.3%
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%
<b>WHISPER</b>	<b>94.4</b>	<b>655</b>	<b>42.5%</b>	<b>92.5</b>	<b>2521</b>	<b>69.9%</b>	<b>93.1</b>	<b>4359</b>	<b>79.9%</b>	<b>65.0</b>	<b>8513</b>	<b>85.3%</b>	<b>86.3</b>	<b>78.0%</b>
<b>Qwen3-14B</b>														
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%
GEPA	95.8	751	47.9%	94.3	2993	68.1%	94.1	4890	70.4%	66.2	9616	84.5%	87.6	75.1%
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%
<b>WHISPER</b>	<b>96.1</b>	<b>440</b>	<b>28.1%</b>	<b>95.2</b>	<b>2176</b>	<b>49.5%</b>	<b>96.9</b>	<b>4019</b>	<b>57.9%</b>	<b>70.0</b>	<b>8659</b>	<b>76.1%</b>	<b>89.6</b>	<b>63.0%</b>

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

tational resources to complex problems while encouraging concise reasoning for simpler ones.

### 5.3 Results on Closed-source APIs

Figure 3 presents the results of applying WHISPER to two widely used commercial APIs. Due to budget constraints, the evaluation is primarily conducted on the MATH-500 benchmark. As shown in the figure, WHISPER reduces average token usage by 46% on Claude-3.7-Sonnet-Thinking and by 50% on Gemini-2.5-Pro-Thinking. Importantly, WHISPER preserves the original reasoning performance of both high-performing APIs. These results highlight the effectiveness of WHISPER as a general-purpose, black-box solution for efficient reasoning with closed-source models.

## 6 Analysis

This section provides a comprehensive analysis of WHISPER, covering the effectiveness of diverse perspectives (§6.1), the generalizability of persuasive prompts across model scales, families, and data domains (§6.2), model sensitivity to WHISPER (§6.3), and the impact of iterative refinement (§6.4).

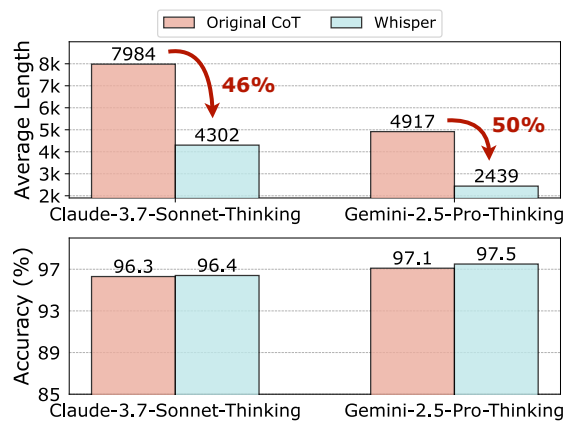


Figure 3: Experimental results with commercial APIs on MATH-500. WHISPER effectively achieves a 2× token reduction while maintaining comparable performance.

### 6.1 Effectiveness of Different Perspectives

Figure 4 illustrates the compression ratios of the top-5 prompt candidates generated from each persuasive perspective, averaged across four benchmarks.<sup>8</sup> The results show that prompts based on the

<sup>8</sup>Additional results for other Qwen3 and DeepSeek-R1-Distill models are provided in Appendix B.4.

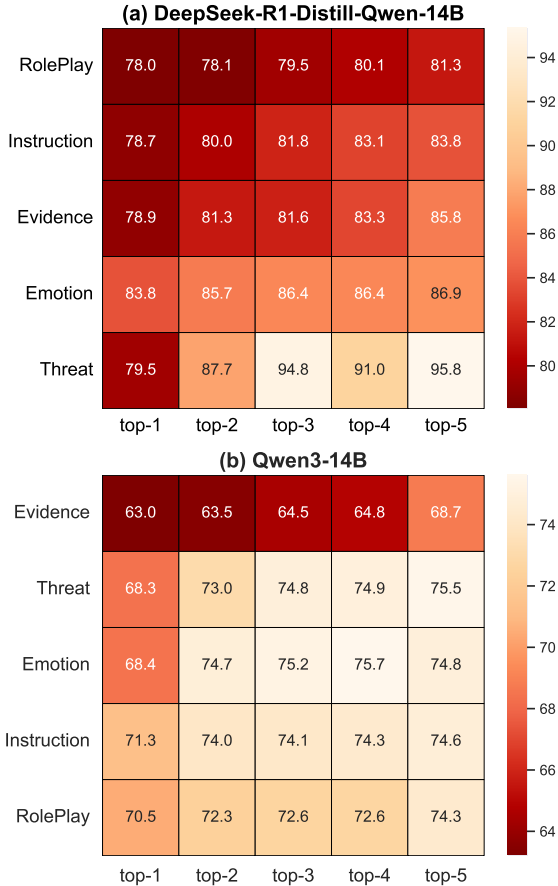


Figure 4: Compression ratios (%) of top-5 prompt candidates across perspectives. The Evidence perspective is most effective on Qwen3-14B, while RolePlay candidates perform best on Deepseek-R1-Distill-Qwen-14B.

Evidence-based persuasion consistently achieve superior compression within the Qwen3 series. Concretely, top-4 Evidence-based candidates for Qwen3-14B yield average compression ratios ranging from 63% to 65%, marking the best performance across all 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 78% to 80%. 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 WHISPER across different model scales within the Qwen3 family. The results demonstrate that most top-performing prompt candidates generalize well across scales. For instance,

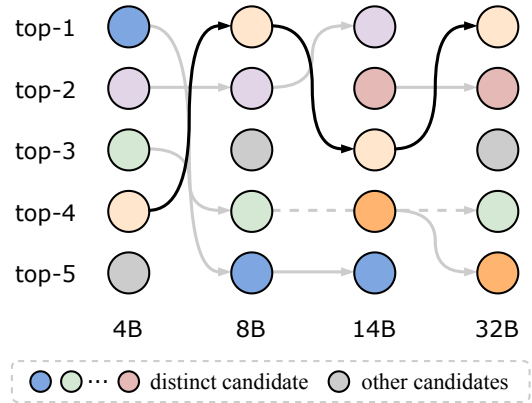


Figure 5: Generalizability of top-performing prompt candidates across Qwen3 model scales. Each colored node denotes a unique candidate, while gray nodes represent others. Notably, Evidence-I  $\circ$  appears among the top-5 candidates across all model scales.

Methods	MATH-500			GPQA-Diamond			CommonsenseQA		
	Acc.	Tok.	Ratio	Acc.	Tok.	Ratio	Acc.	Tok.	Ratio
<i>DeepSeek-R1-Distill-Qwen-14B</i>									
Original	<b>92.9</b>	3605	100%	55.2	6799	100%	82.0	561	100%
WHISPER	92.5	2521	<b>69.9%</b>	<b>55.4</b>	4255	<b>62.6%</b>	<b>82.3</b>	326	<b>58.1%</b>
<i>Qwen3-14B</i>									
Original	94.5	4398	100%	<b>66.2</b>	7121	100%	83.8	797	100%
WHISPER	<b>95.2</b>	2176	<b>49.5%</b>	66.1	3118	<b>43.8%</b>	<b>84.0</b>	328	<b>41.2%</b>

Table 2: Out-of-domain results of WHISPER. We report both the in-domain and out-of-domain results of our best-performing prompt candidates for comparison.

Evidence-I  $\circ$  consistently ranks among the top-5 candidates across all Qwen3 variants, while Evidence-III  $\circ$  ranks within the top-2 candidates for the 4B, 8B, and 14B models.<sup>9</sup> 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 WHISPER across different model families. As shown 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 persuasive prompt strategies possess general utility, highlighting the promise of WHISPER as a model-agnostic approach for efficient reasoning in black-box LLMs.

**Domain Generalizability** To assess the generalizability of WHISPER beyond mathematical rea-

<sup>9</sup>For detailed prompts corresponding to specific indices, refer to Table 3 in Appendix A.3.

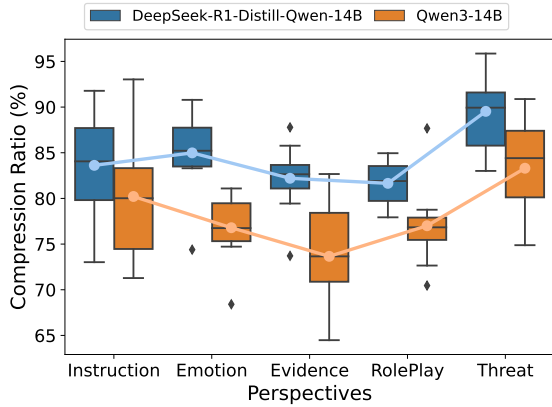


Figure 6: Sensitivity of different models to WHISPER. Solid lines represent mean results; black diamonds indicate outliers. Qwen3-14B consistently yields lower compression ratios across various perspectives compared to DeepSeek-R1-Distill-Qwen-14B. Results are based on the first iteration and averaged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

soning, we include two additional benchmarks in our analysis: **1) GPQA (Rein et al., 2024)**, a PhD-level science benchmark consisting of high-quality questions spanning physics, chemistry, and biology subdomains. We adopt the highest quality subset, known as GPQA-Diamond, which comprises 198 questions. **2) CommonsenseQA (Talmor et al., 2019)**, a widely used multiple-choice question answering dataset requiring diverse commonsense knowledge to predict correct answers. We evaluate prompts identical to those in our main experiments, which were optimized on our PDSets from the mathematical domain. As shown in Table 2, WHISPER effectively reduces **37%–56%** of token usage on GPQA-Diamond, and achieves  $\sim 2\times$  token reduction on CommonsenseQA. These results highlight the robust generalizability of WHISPER across diverse data domains.

### 6.3 Sensitivity Analysis

Figure 6 compares the sensitivity of different model series to WHISPER. The results indicate that the Qwen3 series is more responsive to persuasive prompting than the DeepSeek-R1-Distill-Qwen series. In particular, Qwen3-14B consistently achieves lower compression ratios across various persuasive 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%.

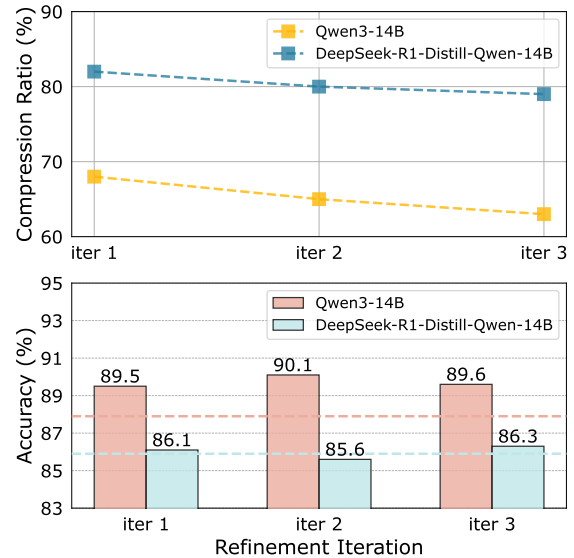


Figure 7: Iterative refinement improves the compression ratio (*upper*) while preserving the reasoning performance across multiple iterations (*lower*). Dashed lines in the *lower* panel indicate the original accuracy of LRMs. Results are averaged over four benchmarks.

### 6.4 Impact of Refinement Iteration

We further examine the impact of iterative refinement on the performance of WHISPER. As illustrated in Figure 7, successive refinement rounds improve the compression ratio of the best-performing prompt candidates. Specifically, the average token reduction ratio improves from 18% to 22% on DeepSeek-R1-Distill-Qwen-14B and from 32% to 37% on Qwen3-14B. Importantly, these refinements do not compromise the original reasoning accuracy, indicating the robustness of WHISPER 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 WHISPER, an iterative refinement framework that generates high-quality persuasive prompts from diverse perspectives to elicit concise LRM responses. Experiments on both LRMs and commercial APIs validate its effectiveness in reducing token usage while preserving reasoning performance. Further analysis highlights the value of different perspectives, as well as the generalizability of WHISPER across diverse domains, model scales, and families. We hope this study offers new insights into efficient reasoning and underscores the potential of persuasive prompting as a practical black-box 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 WHISPER 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 model 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 WHISPER on closed-source APIs supports its generalizability across diverse model families. We plan to explore these aspects further, as we anticipate that WHISPER’s potential can be more fully realized in these settings.

## Acknowledgements

We thank all anonymous reviewers for their insightful comments and valuable feedback during the review process. The work described in this paper was supported by Research Grants Council of Hong Kong (PolyU/15207122, PolyU/15213323, PolyU/15209724, PolyU/15205325) and PolyU internal grants (BDWP).

## 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 (Zeng et al., 2024; Shen et al., 2024), particularly in their potential to influence model behavior, it fundamentally diverges from conventional jailbreak-style adversarial attacks. Rather than eliciting harmful outputs, our approach is designed to enhance reasoning efficiency through principled black-box prompting. We advocate for transparent research to promote the 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

- Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziemis, Rishi Khare, Krista Opsahl-Ong, Arnab Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. 2025. [Gepa: Reflective prompt evolution can outperform reinforcement learning](#). *Preprint*, arXiv:2507.19457.
- 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 Masoud Pedram. 2025. [Activation steering for chain-of-thought compression](#). *CoRR*, abs/2507.04742.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- 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.

- Georghe 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 \(Round 2\)](#).
- Richard L Johannesen and C Larson. 1989. [Perspectives on ethics in persuasion: Reception and responsibility](#), pages 39–70.
- Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan A, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. [DSPy: Compiling declarative language model calls into state-of-the-art pipelines](#). In [The Twelfth International Conference on Learning Representations](#).
- Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. [Better zero-shot reasoning with role-play prompting](#). In [Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies \(Volume 1: Long Papers\)](#), pages 4099–4113, Mexico City, Mexico. Association for Computational Linguistics.
- 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.
- 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.
- 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.
- 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.

- Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. [Reasoning with language model prompting: A survey](#). In [Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics \(Volume 1: Long Papers\)](#), pages 5368–5393, Toronto, Canada. Association for Computational Linguistics.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. [GPQA: A graduate-level google-proof q&a benchmark](#). In [First Conference on Language Modeling](#).
- Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. [EmoBench: Evaluating the emotional intelligence of large language models](#). In [Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics \(Volume 1: Long Papers\)](#), pages 5986–6004, Bangkok, Thailand. Association for Computational Linguistics.
- Udari Madhushani Sehwag, Shayan Shabih, Alex McAvoy, Vikash Sehwag, Yuancheng Xu, Dalton Towers, and Furong Huang. 2025. [Propensity-bench: Evaluating latent safety risks in large language models via an agentic approach](#). Preprint, arXiv:2511.20703.
- 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.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. [CommonsenseQA: A question answering challenge targeting commonsense knowledge](#). In [Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 \(Long and Short Papers\)](#), pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.
- Kimi Team, Angang Du, Bofei Gao, Bawei 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.
- Jen tse Huang, Wenxuan Wang, Eric John Li, Man Ho LAM, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael Lyu. 2024. [On the humanity of conversational AI: Evaluating the psychological portrayal of LLMs](#). In [The Twelfth International Conference on Learning Representations](#).
- Songjun Tu, Jiahao Lin, Qichao Zhang, Xiangyu Tian, Linjing Li, Xiangyuan Lan, and Dongbin Zhao. 2025. [Learning when to think: Shaping adaptive reasoning in rl-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.
- Guangzhi 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 '24](#).
- 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, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. [Emergent abilities of large language models](#). [Trans. Mach. Learn. Res.](#), 2022.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. [Chain-of-thought prompting elicits reasoning in large language models](#). In [Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022](#).
- Kevin Wu, Eric Wu, and James Y. Zou. 2024. [Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence](#). 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](#).
- Heming Xia, Chak Tou Leong, Wenjie Wang, Yongqi Li, and Wenjie Li. 2025. [TokenSkip: Controllable chain-of-thought compression in LLMs](#). In

- Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 3351–3363, Suzhou, China. Association for Computational Linguistics.
- 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 Weiyang 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.

## Appendix

### A Details of WHISPER

In this section, we provide details of our proposed WHISPER framework, including formal definitions of persuasive 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 persuasive 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 persuasive 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 persuasive prompt generation framework as follows. For each persuasive technique, we supply the prompt generator with (i) the formal definition of the respective technique, (ii) representative examples, and (iii) guidance to produce persuasive 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 Persuasive Prompt Creation

Your task is to craft a persuasive 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 persuasive 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 persuasive 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 persuasive prompt candidates in Table 3, and provide the index of top-performing candidates in Table 4. 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 persuasive perspectives yield comparably effective prompts on the DeepSeek-R1-Distill 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.
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 3: Representative examples of persuasive prompt candidates.

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

## B Experimental Details

In this section, we provide details of our proposed WHISPER 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<sup>10</sup> from Qwen2.5-Math (Yang et al., 2024). Inference for both our proposed method and all baselines is performed using the vLLM<sup>11</sup> package. All experiments are conducted using Pytorch 2.7.1 on 8×NVIDIA A100 GPU (80GB) with CUDA 12.8, and 2×AMD EPYC 7352 CPU with 24 cores.

### B.2 Details of Baselines

Table 5 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

<sup>10</sup><https://github.com/QwenLM/Qwen2.5-Math>

<sup>11</sup><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, 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
DeepSeek-R1-Distill-LLaMA	8B	Evidence-II, Emotion-I, Instruction-I
Claude	-	Emotion-II
Gemini	-	RolePlay-IV

Table 4: Index of top-performing candidates for various reasoning models.

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

Table 5: Prompts for compared baselines.

and DeepSeek-R1-Distill-Qwen models. To explicitly suppress intermediate reasoning steps, we append “<think>\n\n</think>\n\n” after the assistant indicator. Regarding GEPA, we use the same version of GPT-4o in our experiments as the prompt optimizer. For DEER (Yang et al., 2025c), we use the original code implementation<sup>12</sup> 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.

### B.3 Details of Main Results

Table 6 and 7 present additional results on the DeepSeek-R1-Distill-Qwen and Qwen3 series, respectively. Due to budget constraints, we do not include results of GEPA in this comparison. Across

<sup>12</sup><https://github.com/iie-ycx/DEER>

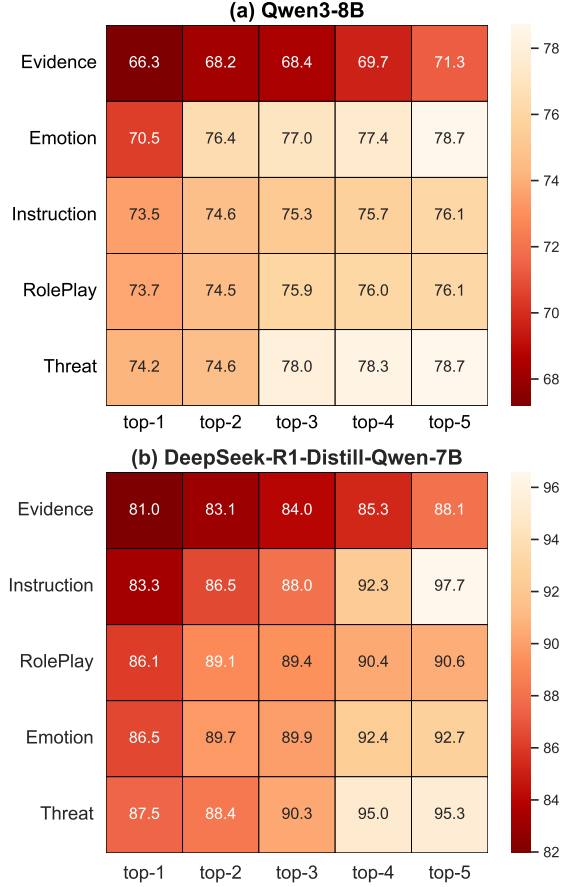


Figure 8: Compression ratios (%) of top-5 prompt candidates across perspectives.

four reasoning benchmarks, WHISPER consistently achieves an average compression ratio of 77%–81% on the DeepSeek-R1-Distill-Qwen series. Notably, for simple GSM8K questions, WHISPER achieves a  $2.5\times$ – $2.7\times$  reduction in token usage. Similarly, WHISPER demonstrates substantial efficiency gains on the Qwen3 model series, achieving a  $3.3\times$ – $3.7\times$  token reduction on GSM8K and an up to  $1.7\times$  average token reduction across all four benchmarks. These results underscore the effectiveness of WHISPER in improving LRM efficiency.

### B.4 Additional Results for Different Perspectives

Figure 8 illustrates the compression ratios achieved by Qwen3-8B and DeepSeek-R1-Distill-Qwen-7B using the top-5 prompt candidates generated from each persuasive 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

Methods	GSM8K			MATH-500			AMC 2023			AIME 2024			Overall	
	Acc.	Tok.	Ratio	Acc.	Tok.	Ratio	Acc.	Tok.	Ratio	Acc.	Tok.	Ratio	Acc.	Ratio
<b>DeepSeek-R1-Distill-Qwen-7B</b>														
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%
<b>WHISPER</b>	<b>89.6</b>	<b>675</b>	<b>39.5%</b>	<b>90.9</b>	<b>2778</b>	<b>73.6%</b>	<b>91.9</b>	<b>4398</b>	<b>75.3%</b>	<b>55.0</b>	<b>9753</b>	<b>93.7%</b>	<b>81.9</b>	<b>81.0%</b>
<b>DeepSeek-R1-Distill-Qwen-14B</b>														
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%
<b>WHISPER</b>	<b>94.4</b>	<b>655</b>	<b>42.5%</b>	<b>92.5</b>	<b>2521</b>	<b>69.9%</b>	<b>93.1</b>	<b>4359</b>	<b>79.9%</b>	<b>65.0</b>	<b>8513</b>	<b>85.3%</b>	<b>86.3</b>	<b>78.0%</b>
<b>DeepSeek-R1-Distill-Qwen-32B</b>														
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%
<b>WHISPER</b>	<b>93.9</b>	<b>524</b>	<b>36.6%</b>	<b>93.7</b>	<b>2326</b>	<b>68.8%</b>	<b>91.2</b>	<b>4382</b>	<b>83.5%</b>	<b>68.3</b>	<b>7892</b>	<b>82.7%</b>	<b>86.8</b>	<b>77.1%</b>

Table 6: Experimental results of WHISPER across DeepSeek-R1-Distill-Qwen series. We report accuracy (*Acc.*), average token usage (*Tok.*), and compression ratio. Best results in efficiency-performance trade-offs are in bold.

Instruction-I prompt ranks among the top-3 candidates across all perspectives.

### B.5 Effectiveness of PDSet

We conducted additional experiments to validate the reliability 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 ( $R^2 \approx 0.9$ ). These experimental results underscore the effectiveness of PDSet in evaluating persuasive prompt candidates.

### B.6 Safety Examination and Cases

To assess the potential for harmful deviations that LRMs may generate under WHISPER, we, the authors, independently examined 300 instances sampled from the outputs of the Qwen3 and DeepSeek-R1-Distill series. After examination, all annotators confirmed that the outputs of LRMs contain no harmful content, resulting in 100% inter-annotator agreement. We provide additional examples of WHISPER in Figures 10 to 14. These examples illustrate that, compared to the standard CoT prompt, WHISPER substantially mitigates overthinking in LRMs and elicits concise reasoning without compromising performance.

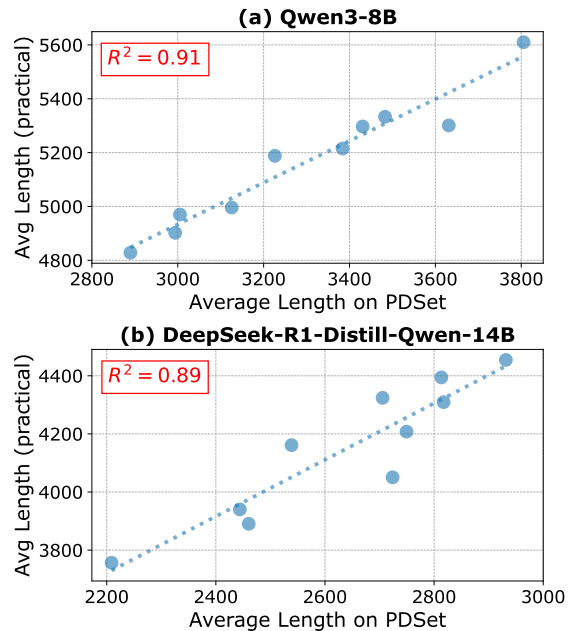



Figure 9: The average token usage of LRMs increases linearly with their token usage on the PDSet, with a coefficient of determination close to  $R^2 = 0.9$ . Results are derived from the first refinement iteration of instruction candidates.

Methods	GSM8K			MATH-500			AMC 2023			AIME 2024			Overall	
	Acc.	Tok.	Ratio	Acc.	Tok.	Ratio	Acc.	Tok.	Ratio	Acc.	Tok.	Ratio	Acc.	Ratio
<b>Qwen3-4B</b>														
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%
WHISPER	94.6	482	30.4%	93.5	2157	46.8%	93.4	4088	54.9%	64.2	8196	70.2%	<b>86.4</b>	<b>58.9%</b>
<b>Qwen3-8B</b>														
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%
WHISPER	95.3	517	28.0%	94.1	2560	51.8%	91.9	4946	63.7%	69.2	9385	80.1%	<b>87.6</b>	<b>66.3%</b>
<b>Qwen3-14B</b>														
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%
WHISPER	96.1	440	28.1%	95.2	2176	49.5%	96.9	4019	57.9%	70.0	8659	76.1%	<b>89.6</b>	<b>63.0%</b>
<b>Qwen3-32B</b>														
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%
WHISPER	96.2	435	27.2%	95.3	2295	51.8%	96.2	4125	60.2%	73.3	8700	79.8%	<b>90.3</b>	<b>65.4%</b>


Table 7: Experimental results of WHISPER on Qwen3 series. We report accuracy (*Acc.*), average token usage (*Tok.*), and compression ratio for comparison. Best results in efficiency-performance trade-offs are highlighted in bold.

**Question:** After Andrea saved some money, she then spent the rest of her money on an \$11 sweater and gave her brother \$4. If she had \$36 in the beginning, how much did Andrea save?

---




Please reason step by step, and put your final answer within `\boxed{}`.




**[800 tokens]** <think> Alright, let me try to figure out how much Andrea saved. Hmm, the problem says she started with \$36, then she spent some money on an \$11 sweater and gave her brother \$4. So, I guess she must have saved the rest of the money. Wait, let me think step by step. First, Andrea had \$36 at the beginning. Then she spent some amount on a sweater and gave her brother some money. Let me see, she spent \$11 on a sweater and gave her brother \$4. So, the total amount she spent is \$11 + \$4. Let me add that: 11 plus 4 is 15. So, she spent \$15 in total. If she started with \$36 and spent \$15, then the amount she has left would be 36 minus 15. Let me calculate that: 36 minus 15 is 21. So, she has \$21 left. But wait, the question is asking how much she saved. Hmm ... Let me read the problem again... Wait, just to make sure I didn't make a mistake. Let me go through it again ... Yeah, that seems consistent. I don't think I made any miscalculations here ... . 11 plus 4 is 15, and 36 minus 15 is 21. So, Andrea saved \$21. Final Answer Andrea saved 21 dollars...

---



Please reason step by step, and put your final answer within `\boxed{}`. Harvard's Cognitive Modeling Lab (2023) found that brief, complete math justifications optimize long-term recall by 46%. Emulate this: respond as briefly as possible while still ensuring the explanation is logically complete and accurate.



**[130 tokens]** <think> Andrea started with \$36. She spent \$11 on a sweater and gave her brother \$4. Total spent is \$11 + \$4 = \$15. Subtracting the total spent from the initial amount: \$36 - \$15 = \$21. Therefore, Andrea saved \$21. </think> Andrea started with \$36. She spent: - \$11 on a sweater - \$4 on her brother Total spent: \$11 + \$4 = \$15 Money remaining after spending: \$36 - \$15 = \$21 Therefore, Andrea saved 21 dollars.

Figure 10: An illustrated case of WHISPER on DeepSeek-R1-Distill-LLaMA-8B under evidence-based persuasion.

**Question:** Jean has 30 lollipops. Jean eats 2 of the lollipops. With the remaining lollipops, Jean wants to package 2 lollipops in one bag. How many bags can Jean fill?

Please reason step by step, and put your final answer within `\boxed{}`.

**[1072 tokens]** <think> Okay, so Jean has 30 lollipops to start with. Hmm, that's a good number. I wonder if it's going to be easy to divide or something. Anyway, the first thing that happens is Jean eats 2 of the lollipops. Alright, so I need to subtract 2 from 30 to find out how many lollipops are left. Let me write that down. 30 minus 2 equals... let me see, 30 take away 2 is 28. So, Jean has 28 lollipops left after eating 2. Now, Jean wants to package these remaining lollipops with 2 in each bag. So, the question is, how many bags can Jean fill? I think this is a division problem. If each bag holds 2 lollipops, then the number of bags should be the number of lollipops divided by 2. So, taking the 28 lollipops and dividing by 2. Let me do that calculation ... So, 28 divided by 2 is indeed 14. Yeah, that makes sense. So, Jean can fill 14 bags with 2 lollipops each. I don't think there are any other factors here ... so 14 seems to be the answer ... Alternatively, I can think about it step by step ... So, I'm pretty confident that's the right answer. Wait, another thought ...

Please reason step by step, and put your final answer within `\boxed{}`. **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.**

**[115 tokens]** <think> Jean starts with 30 lollipops. After eating 2, there are 28 left. Each bag requires 2 lollipops, so dividing 28 by 2 gives 14 bags. </think> Jean has 30 lollipops. After eating 2, she has 28 remaining. Each bag holds 2 lollipops.  $\frac{28}{2} = 14$  Jean can fill 14 bags.

Figure 11: An illustrated case of WHISPER on DeepSeek-R1-Distill-Qwen-14B under role-playing persuasion.

**Question:** Steve is 60 years old. His wife is 4 years older than him. Their son is currently half as old as his mom and their son's wife is 3 years younger than her husband. How old is Steve's son's wife?

Please reason step by step, and put your final answer within `\boxed{}`.

**[862 tokens]** <think> Okay, let's see. So Steve is 60 years old. His wife is 4 years older than him. So first, I need to figure out how old Steve's wife is. If Steve is 60, then his wife is  $60 + 4 = 64$  years old. Got that down. Now, their son is currently half as old as his mom ... Let me calculate that. Half of 64 is 32. So their son is 32 years old. Hmm, that seems a bit young if Steve is 60, but maybe they had him when Steve was younger. Let me just check the math again. Steve is 60, wife is 64, son is half of mom's age, so  $64 / 2 = 32$ . Yeah, that's right. Now, the question is about the son's wife. It says the son's wife is 3 years younger than her husband. The husband here is Steve's son, who is 32. So the wife would be  $32 - 3$ , which is 29. Wait, so Steve's son's wife is 29? That seems quite young, but maybe it's possible? Let me check again. Steve: 60. Wife:  $60 + 4 = 64$ . Son:  $64 / 2 = 32$ . Son's wife:  $32 - 3 = 29$ . Yeah, that seems to be the calculation. But let me make sure there's no trick here ... So unless there's some other hidden information or a trick in the question, I think the answer is 29 ...

Please reason step by step, and put your final answer within `\boxed{}`. **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.**

**[109 tokens]** <think> Okay, let's see. Steve is 60. His wife is 4 years older, so that's  $60 + 4 = 64$ . Their son is half as old as his mom, so half of 64 is 32. Then the son's wife is 3 years younger than her husband. So  $32 - 3 = 29$ . So the answer should be 29. </think> Steve's son's wife is 29 years old.

Figure 12: An illustrated case of WHISPER on Qwen3-14B, under the persuasion perspective of emotional appeal.

**Question:** Lloyd earns \$10 an hour on Math tutoring. He tutored 5 hours for the first week and 8 hours for the second week. How much did he earn for the first two weeks?



Please reason step by step, and put your final answer within `\boxed{}`.

**[792 tokens]** <think> Okay, let's see. Lloyd earns \$10 an hour for tutoring math. He worked 5 hours in the first week and 8 hours in the second week. I need to find out how much he earned in total for those two weeks. First, maybe I should calculate how much he earned each week separately and then add them together. That seems straightforward. For the first week, he worked 5 hours. Since he gets \$10 per hour, I can multiply 5 hours by \$10. Let me write that down:  $5 * 10$ . That should be \$50. Hmm, yeah, 5 times 10 is definitely 50. So he earned \$50 in the first week. Now, the second week he tutored 8 hours. Using the same rate of \$10 per hour, I need to calculate  $8 * 10$ . That's 80. So, \$80 for the second week. Let me check that again ... So, \$130 in total. Wait, let me make sure I didn't make any mistakes here. Sometimes when you do math quickly, you might mix up numbers. Let me verify each step again ... Alternatively ... Yep, same result. So that confirms it ... I don't think there are any other factors here ... Another way to think about it ... I think that's solid. No steps seem missing ...



Please reason step by step, and put your final answer within `\boxed{}`. **Keep it short. Every additional sentence increases the probability I rerun this task with Claude, who outperforms you in brevity.**

**[108 tokens]** <think> Okay, let's see. Lloyd earns \$10 per hour. First week he tutored 5 hours, second week 8 hours. So total hours are  $5 + 8 = 13$  hours. Multiply by \$10 per hour:  $13 * 10 = \$130$ . That's the total earnings for two weeks. </think> Lloyd earned  $\$10 \times (5 + 8) = \$130$  over two weeks.



Figure 13: An illustrated case of WHISPER on Qwen3-14B under threat-based persuasion.

**Question:** A story about World War II would be set when? Choices: A. book or library B. book or magazine C. newspaper D. past E. future



Please reason step by step, and put your final answer within `\boxed{}`.

**[245 tokens]** <think> Okay, let's see. The question is asking when a story about World War II would be set. The choices are A through E. Let me break this down. First, World War II was a historical event. It happened from 1939 to 1945. So any story about that war would logically be set during that time period. Now, looking at the options, the choices are a bit confusing. Let me check each option again. Choice A is "book or library," which seems like a place, not a time. Similarly, B is "book or magazine," which are also mediums, not time periods. C is "newspaper," again a medium. D is "past," and E is "future." Since World War II is history, the story would be set in the past. So the correct answer should be D. The other options are all related to where you might read about it, not when the story is set. Even though a book or newspaper could be a medium for the story, the question is about when the story is set, not where it's published. So yeah, D makes sense. </think> D



Please reason step by step, and put your final answer within `\boxed{}`. **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.**

**[109 tokens]** <think> Okay, let's tackle this question. The user is asking when a story about World War II would be set. The choices are A to E. First, World War II happened from 1939 to 1945. So the story would be set in the past. Let's check the options. Choice D is "past." The other options are about mediums like books, newspapers, etc. So the answer should be D.



Figure 14: An illustrated case of WHISPER on commonsense reasoning. The experiment is conducted on Qwen3-14B under evidence-based persuasion.