

Wait, We Don't Need to "Wait"! Removing Thinking Tokens Improves Reasoning Efficiency

Chenlong Wang, Yuanning Feng, Dongping Chen^{1†}, Zhaoyang Chu²,
Ranjay Krishna^{3,4}, Tianyi Zhou[†]

¹University of Maryland, ²University College London, ³University of Washington,
⁴Allen Institute for AI

Abstract

Recent advances in large reasoning models have enabled complex, step-by-step reasoning but often introduce significant *overthinking*, resulting in verbose and redundant outputs that hinder efficiency. In this study, we examine whether explicit self-reflection, signaled by tokens such as “Wait” and “Hmm”, is necessary for advanced reasoning. We propose NOWAIT, a simple yet effective approach that disables explicit self-reflection by suppressing these tokens during inference. Extensive experiments on ten benchmarks across textual, visual, and video reasoning tasks show that NOWAIT reduces chain-of-thought trajectory length by up to 27%–51% in five R1-style model series, without compromising model utility. NOWAIT thus offers a plug-and-play solution for efficient and utility-preserving multimodal reasoning.

1 Introduction

Recent advancements in large reasoning models (LRMs), exemplified by DeepSeek-R1 (Guo et al., 2025), have shown that complex reasoning abilities can be effectively elicited through simple rule-based reinforcement learning (Team, 2025; Qwen, 2025; Abdin et al., 2025; Xia et al., 2025). These models produce explicit, step-by-step reasoning through long chain-of-thought (CoT) trajectories (Yang et al., 2025a; Ma et al., 2025a) before arriving at final answers. This capability is believed to be accompanied by the emergence of the “Aha Moment” phenomenon (Chen et al., 2025b; Yang et al., 2025b), in which the model begins to rethink problems and self-reflect on its reasoning trajectory with anthropomorphic expressions such as “Wait”, “Hmm”, or “Alternatively”. This was firstly achieved on R1-style language reasoning models and has been extended to vision-language models (VLMs) (Team, 2024; Team et al., 2025),

enabling multimodal reasoning on images (Zhang et al., 2025b; Shen et al., 2025a; Huang et al., 2025; Zhou et al., 2025) and videos (Feng et al., 2025; Team, 2024; Team et al., 2025).

Despite the effectiveness of long CoT reasoning with self-reflection, the overthinking problem has emerged (Chen et al., 2024a; Cuadron et al., 2025; Chen et al., 2024b; Wu et al., 2025; Sui et al., 2025). It is characterized by excessively verbose reasoning and redundant thought steps, often extending over thousands of tokens, resulting in significant computational overhead and high reasoning latency. Such inefficiencies hinder the practical deployment of R1-style reasoning models in applications with limited computational resources.

Although numerous efforts have been devoted to efficient reasoning, many existing approaches require additional training, either through reinforcement learning (RL) with length-based rewards (Aggarwal and Welleck, 2025; Liao et al., 2025; Luo et al., 2025) or fine-tuning on variable-length CoT trajectories (Ma et al., 2025b; Munkhbat et al., 2025). On the other hand, several training-free approaches have been proposed to mitigate overthinking by reducing token usage during inference. However, they often compromise the overall model utility (Ma et al., 2025a) or have only demonstrated effectiveness on distilled reasoning models (Yang et al., 2025c,a; Xu et al., 2025).

In this study, we investigate the impact of excessive self-reflection during the reasoning process and question whether explicit self-reflection, signaled by “Wait”-like tokens, is really necessary for advanced reasoning. To this end, we propose NOWAIT, a simple yet effective training-free approach that disables explicit self-reflection in R1-style reasoning models, significantly reducing token usage while maintaining overall model utility. As illustrated in Figure 1, we directly intervene in the inference process by identifying specific keyword tokens (e.g., “Wait”, “Hmm”, and “Alterna-

[†]Corresponding Author.

[‡]Project Leader.

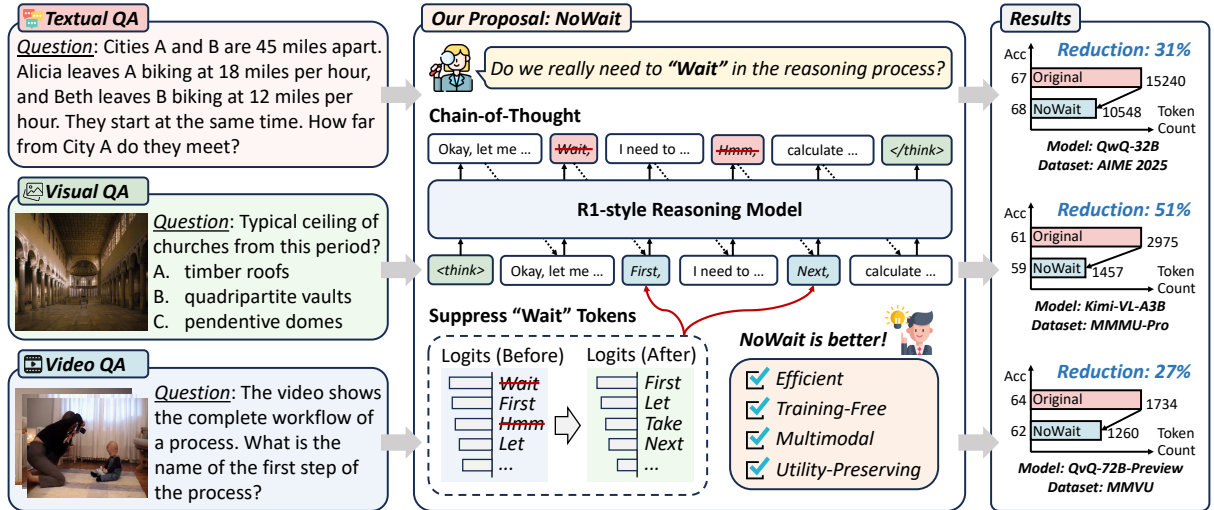


Figure 1: **Illustrative pipeline for NOWAIT.** We introduce NOWAIT, a simple yet effective approach that suppresses the generation of reflection keywords (e.g., “Wait” and “Hmm”) during inference. NOWAIT reduces chain-of-thought trajectory length by up to 27%-51% across textual, visual, and video reasoning tasks.

tively”) that indicate explicit self-reflection and suppressing their generation. Specifically, we achieve this by proactively adjusting the logits of these tokens to negative values during decoding, thereby steering the model toward selecting alternative tokens to continue the reasoning process.

Comprehensive experiments show that NOWAIT achieves strong performance on ten benchmarks spanning ① **textual reasoning** (AMC 2023 (AIMO, 2024), AIME 2024, AIME 2025 (MAA Committees), GQPA-D (Rein et al., 2024)), ② **visual reasoning** (MMMU (Yue et al., 2024a), MMMU-Pro (Yue et al., 2024b), MathVista (Lu et al., 2024), EMMA-mini (Hao et al., 2025)), and ③ **video reasoning** (MMVU (Zhao et al., 2025), VSI-Bench (Yang et al., 2024)). When integrated into five R1-style model series, including QwQ (Qwen, 2025), Phi4 (Abdin et al., 2025), Qwen3 (Team, 2025), Kimi-VL (Team et al., 2025), QvQ (Team, 2024), NOWAIT reduces CoT trajectory length by up to 27%-51% across different modalities. NOWAIT serves as a plug-and-play solution for improving reasoning efficiency while preserving overall model utility and provide new insights for the efficient reasoning.

2 Preliminaries

Reasoning Model Generation Patterns. Reasoning models structure their output using thinking delimiters (i.e., <think> and </think>), dividing the response into two main components: the CoT trajectory detailing the reasoning process and the

final answer summarizing overall thoughts.

Within the generated CoTs, models employ complex reasoning strategies, such as forward thinking, backtracking, and self-reflection. Notably, large reasoning models often continue to reason even after obtaining an initial result, performing additional validation steps. Accordingly, we define each segment of reasoning as a *thinking chunk*. Each thinking chunk is associated with an intermediate answer r . Formally, a thinking chunk can be represented as a pair $(chunk_i, r_i)$, where $chunk_i$ is the reasoning text and r_i is the intermediate answer from $chunk_i$ derived from $chunk_i$. Thus, a complete CoT can be structured as follows:

$$CoT = \{(chunk_i, a_i)\}_{i=1}^n. \quad (1)$$

The final response is the combination of the CoT trajectory and a concise reasoning summary:

$$Response = (CoT, summary). \quad (2)$$

Self-Reflection within Reasoning Models. As stated above, a single CoT can contain multiple reasoning chunks. The transitions between these chunks are often marked by specific keywords, such as *Wait*, *Alternatively*, or *Hmm*. Models tend to switch their reasoning approaches in subsequent steps, often to verify previous results or explore alternative paths. However, this mechanism can sometimes lead to unproductive overthinking, causing models to repeatedly enter new reasoning steps and engage in redundant and unnecessary validation loops. In this study, we introduce NOWAIT, a

simple yet effective method for efficient reasoning by intervening in the generation of these keywords. This method alters models’ self-reflection strategies and can be generalized to various modalities.

3 Removing Thinking Pattern is Better

In this section, we propose NOWAIT, a simple yet effective method, that improves the reasoning efficiency while maintaining acceptable model utility. We first expand the method details in [Section 3.1](#), and introduce the experimental setup in [Section 3.2](#). We then report the experiment results on *textual reasoning* in [Section 3.3](#). Additionally, we conduct the comparison experiment in [Section 3.4](#) and further analyze the generalization across different modalities (*visual reasoning* and *video reasoning*) in [Section 3.5](#).

3.1 Method

NOWAIT functions as an inference-time intervention. It *directly prevents* the model from generating the specific tokens associated with self-reflection. Our method involves three main stages:

Initialize Reflection Keywords List. We begin by identifying the initial reflection keywords, such as “Wait”, “Alternatively”, and “Hmm”. To empirically establish the list, we conduct 32 independent runs of the QwQ-32B (Qwen, 2025) on AIME 2025 (MAA Committees). Using “\n\n” as delimiters, we identify the 15 most frequent monolingual words as our identified keywords $K = \{k_i\}$.

Keyword List for Suppressing

“wait”, “alternatively”, “hmm”, “but”,
“however”, “alternative”, “another”,
“check”, “double-check”, “oh”,
“maybe”, “verify”, “other”, “again”,
“now”, “ah”, “any”

Specific Token-Level Keyword List. Secondly, for each target model α , we expand the initial keyword list K into a specific token-level list, K_α . For instance, the variants of “wait” include “ wait”, “Wait”, “ Wait”, “.wait” and “WAIT”. We achieve this by iterating through the overall vocabulary V_α and identifying all variant tokens whose textual representation contains any keyword from K as a substring. Specifically, we define that, $is_substr(x, y) = True$ when x is the substring

of y . This process can be formulated as follows:

$$K_\alpha = \{v \in V_\alpha | \exists k_s \in K, s.t. is_substr(k_s, v)\}$$

We further manually filter keywords that are not reasonable (i.e., “Ohio” for “oh”).

Suppressing Keywords Generation. During the inference, we leverage a logit processor to prohibit models from generating keywords. For any keyword $v \in K_\alpha$, its corresponding logit is set to a large negative value. This effectively makes these reflection-associated tokens, ensuring they are highly unlikely to be sampled by models.

By surgically preventing the generation of these targeted reflection-associated tokens, NOWAIT aims to streamline the LRM’s reasoning pathways. This targeted intervention is designed to enhance inference efficiency, reducing both latency and token costs, without requiring any modification to the model’s underlying architecture or weights.

3.2 Experimental Setup

Model & Benchmark. To comprehensively evaluate the effectiveness of NOWAIT, we conduct experiments on the open-source models across different modalities and parameter scales.

For the *textual reasoning* task, we assess reinforcement learning (RL) based models, including QwQ-32B (Qwen, 2025), Phi4-Reasoning-Plus (Abdin et al., 2025), and Qwen3-32B (Team, 2025) on math reasoning benchmarks, AIME 2024, AIME 2025 (MAA Committees), and AMC 2023 (AI-MO, 2024).

For the *visual reasoning* task, our experiments cover the state-of-the-art RL-based visual reasoning models, Kimi-VL-A3B-Thinking (Team et al., 2025) and QvQ-72B-Preview (Team, 2024) and evaluate on MMMU-Pro (Yue et al., 2024b), MMMU (Yue et al., 2024a), MathVista (Lu et al., 2024) and EMMA-mini (Hao et al., 2025)

For the *video reasoning* task, we select QvQ-72B-Preview and evaluate on VSI-Bench (Yang et al., 2024) and MMVU (Zhao et al., 2025).

Metrics. The goal of NOWAIT is to preserve the model’s reasoning accuracy while substantially diminishing the number of generated tokens during inference. Performance is assessed using two key metrics: ❶ **Accuracy (ACC)**: This measures the correctness of the model’s final output. ❷ **Generation Length (LEN)** quantifies the average number

Table 1: Experiment results on *Textual Reasoning Tasks*.

Strategy	AMC 2023		AIME 2024		AIME 2025	
	ACC↑	LEN↓	ACC↑	LEN↓	ACC↑	LEN↓
QwQ-32B						
Original	91.25	7542	73.33	14142	66.67	15240
NoThink	72.50	4265	46.67	7980	40.00	8167
NowAIT	95.50 +4.25	5267 -30%	71.33 -2.00	11907 -16%	68.00 +1.33	10548 -31%
Phi4-Reasoning-Plus						
Original	90.00	6366	70.00	15161	59.33	16257
NoThink	80.83	3805	34.67	6200	31.33	5549
NowAIT	96.00 +6.00	4524 -28%	69.33 -0.67	11185 -26%	62.67 +3.34	12490 -23%
Qwen3-32B						
Original	97.50	6424	81.33	12720	66.67	14987
NoThink	59.50	1240	25.33	2511	20.00	2165
NowAIT	96.67 -0.83	5560 -13%	83.33 +2.00	10732 -16%	64.44 -2.67	12930 -14%

of tokens generated by the model per problem instance, calculated over n independent runs.

Baselines. To compare the latency of NOWAIT, we use both the models’ original performance and the NoThink strategy (Ma et al., 2025a) as baselines. Both NOWAIT and NoThink share a similar rationale, aiming to intervene in the model’s reasoning process. However, while NOWAIT operates at the token level by prohibiting the output of specific tokens, NoThink attempts to directly remove the entire reasoning process by prompt engineering. By including these baselines, we can conduct a more comprehensive analysis about the model’s performance under different intervention approaches.

Experiment Details. For each evaluated benchmark, we conduct five independent runs. Except for the Qwen3 series, we infer without chat templates on open-ended problems and leverage the same prompt template for multiple-choice problems (see Appendix B). Because of the different thinking patterns, we apply chat templates for the Qwen3 model inference. In baseline and NOWAIT experiments, we set a maximum token limit of 32,768 tokens per instance. If a model’s generation reaches this limit before finishing CoT generation, that instance is considered incorrect, and the generation length is 32,768 tokens. If not, we will extract the final answer from the generated CoT and judge the correctness. This policy ensures that models failing to complete their response within the bud-

get are appropriately penalized in *Accuracy* metric. For NoThink strategy (Ma et al., 2025a), we set a token budget of 10,000. Details of the token budget applied for NoThink can be found in Section A.1.

3.3 LRMs can be Efficient without “WAIT”

Table 1 presents a comprehensive quantitative overview of our NOWAIT’s performance on various *textual reasoning* tasks, evaluated across different LRMs with diverse model structures and parameter scales. Our method NOWAIT consistently and significantly reduces the output length while maintaining the reasoning accuracy.

Model Architectures Generalization. Notably, when integrated with QwQ-32B, NOWAIT improves accuracy on AMC 2023 by 4.25 percentage points, while reducing output length to just 70% of the baseline. With another model architecture, Phi4-Reasoning-Plus, our method achieves an even greater improvement of 6.00 percentage points, alongside a 28% reduction in token generation. Additionally, Qwen3-32B also benefits from our approach, reducing output length by 13% with only a marginal decrease in reasoning accuracy. These results demonstrate that our method NOWAIT consistently enhances efficiency across diverse model architectures. This consistency suggests a fundamental similarity in the reasoning patterns and redundancy present in different models, underscoring the broad applicability of our approach.

Table 2: **Comparison Experiments across Multiple Efficient Reasoning Methods.** We use QwQ-32B-Preview for experiments.

Strategy	AIME 2024		AMC 2023	
	ACC↑	LEN↓	ACC↑	LEN↓
Baseline	42.00	8979	82.50	4143
Token-Budget	46.67	8734	82.50	3636
O1-Pruner	33.33	4289	77.50	2399
NOWAIT	42.00	5764	86.00	3396

Reasoning Difficulty Analysis. We tested our method on mathematical reasoning benchmarks spanning various difficulty levels ($\text{AMC 2023} < \text{AIME 2024} < \text{AIME 2025}$). The experimental statistics demonstrated strong generalization across these levels: All tested models achieved comparable reductions in token usage regardless of task difficulty. Crucially, NOWAIT enabled models to maintain or even improve performance on more challenging tasks. For instance, QwQ-32B achieved a 1.33% point increase on the challenging AIME 2025 benchmark, while reducing token usage by 31%, which is comparable to its performance on the college-level AMC 2023. Qwen3-32B consistently reduced output length by 14% to 16% across all three math benchmarks, while Phi4-Reasoning-Plus showed similar gains and reductions from 23% to 28%. On the non-mathematical GPQA-Diamond task, models showed a slight performance decrease compared to the math reasoning benchmarks, but still maintained efficiency, with an overall 11.67% reduction in token usage.

These consistent efficiency gains and stable performance across diverse models and varying tasks suggest that, despite the architecture and scale, LRMs exhibit similar inherent redundancy in their reasoning processes. NOWAIT effectively prunes this redundancy, demonstrating that substantial efficiency improvements can be achieved simply by suppressing the keywords generation, without the need for complex explicit “waiting” mechanisms.

3.4 Comparison Analysis

Comparison Experiment. We further compare with existing efficient reasoning techniques, including prompt-based training-free technique, Token-Budget (Han et al., 2024), and training-based technique, O1-Pruner (Luo et al., 2025), using QwQ-

32B-Preview (Qwen, 2025) on AIME 2024 and AMC 2023. All inference is conducted without chat templates to ensure fairness.

As shown in Table 2, NOWAIT exhibits more significant generation length curtailment compared to Token-Budget. Although Token-Budget shows promising results on base models, such as GPT-4o, its effectiveness does not generalize to current LRMs (Deepseek-R1 (Guo et al., 2025), QwQ-32B (Qwen, 2025)). These reasoning models are less sensitive to the prompt design, resulting in less efficiency. O1-Pruner, while effective at reducing token usage, incurs severe performance degradation on QwQ-32B-Preview. In contrast, NOWAIT does not require additional training or data, but instead guides models to strike an effective balance between output length and reasoning accuracy, achieving a spontaneous trade-off.

LRM Cannot Skip Thinking. As shown in Table 1, Qwen3-32B, a model specifically trained for non-thinking patterns, exhibits notable reductions in token usage. However, for other models (QwQ-32B and Phi4-Reasoning-Plus) without non-thinking pattern training, NoThink (Ma et al., 2025a), a prompt-based method, fails to thoroughly skip the generation of reasoning steps. While NoThink does reduce the generation length, the evaluated model can still generate the thinking process and demonstrate a serious compromise in accuracy. This failure indicates that the presence of explicit “thinking” tokens (“<think>” and “<\think>”) can influence models’ output, but is insufficient to precisely control models’ reasoning strategy. Our method NOWAIT operates on a similar premise by targeting key reasoning-related tokens, but achieves significant efficiency improvements with better maintain on reasoning accuracy.

3.5 Efficient Multimodal Reasoning

In this section, we propose efficient multimodal reasoning. We assess NOWAIT on visual reasoning models using image and video reasoning benchmarks. As shown in Table 3 and Table 4, visual reasoning models exhibit more exciting outcomes.

Severe Verbosity on Multimodal Reasoning. Although Kimi-VL-A3B-Thinking generates an average of only 2,000 tokens across four image reasoning benchmarks, significantly fewer than that in math reasoning tasks, our method NOWAIT further reduces the generation length by an average of 49%, with only a modest overall accuracy drop of 3.42

Table 3: Experiment results on *Visual Reasoning* Tasks.

Strategy	MMMU-Pro		MMMU		MathVista		EMMA-mini	
	ACC↑	LEN↓	ACC↑	LEN↓	ACC↑	LEN↓	ACC↑	LEN↓
Kimi-VL-A3B-Thinking								
Original	61.27	2975	57.00	2929	71.50	1822	34.75	5734
NO WAIT	58.73 -2.54	1457 -51%	55.20 -1.80	1746 -40%	69.40 -2.10	1045 -43%	27.50 -7.25	2269 -60%
QvQ-72B-Preview								
Original	65.77	2094	66.85	1977	73.54	1338	32.00	2097
NO WAIT	63.79 -1.98	1659 -21%	66.74 -0.11	1571 -21%	70.92 -2.62	939 -30%	28.00 -4.00	1554 -26%

Table 4: Experiment Results on *Video Reasoning* Tasks. We use QvQ-72B-Preview for experiments.

Strategy	MMVU		VSI-Bench	
	ACC↑	LEN↓	ACC↑	LEN↓
Original	64.10	1734	22.51	1280
NO WAIT	62.20	1260	22.57	1020
Performance	-1.90	-27%	+0.06	-20%

percentage points. A similar trend is observed with QvQ-72B-Preview, which achieves up to a 30% reduction in token usage, accompanied by only a slight decrease in accuracy (ranging from 0.11% to 4.00%). For video reasoning tasks, QvQ-72B-Preview also demonstrates substantial reductions in output length while maintaining comparable accuracy. Similar to textual reasoning tasks, these results reveal the same challenging problems that a significant portion of generated tokens are either redundant or contribute little to the final reasoning. Existing multimodal reasoning models still suffer from severe reasoning inefficiency.

Reinforcement Learning is Less Efficient.

We further evaluate various RL-based reasoning models across varying benchmarks and modalities. While a generation of intellectual reasoning models confirms the effectiveness of the RL algorithm in advanced reasoning capabilities, the efficiency of the optimal policy derived from the RL algorithm is still disappointing. The model learns a reasoning policy from training and begins to spontaneously reflect reasoning processes during inference. However, these algorithms fail to effectively teach models when reflection is truly necessary. As a result, these models often adopt a lower threshold for self-reflection, leading to unnecessary verifica-

tion steps and less efficient reasoning. Our method suppresses the generation of reflection keywords, raising the threshold of self-reflection, and making it more efficient and necessary.

4 Discussion

In this section, we first discuss the effectiveness of our method NO WAIT in Section 4.1 by case study and the robustness of the model while applying NO WAIT in Section 4.2. Additionally, we conduct an empirical experiment to analyze the difference between RL-based models and distill models based on NO WAIT in Section 4.3.

4.1 Why does NO WAIT Work?

As we discussed in Table 3.4, thinking tokens (“<think>” and “<\think>”) cannot thoroughly control models’ actions. Similarly, can banning keywords completely remove self-reflection from CoTs? If not, why NO WAIT result in more efficient reasoning? To answer this, we conduct a case study to analyze the effectiveness of our method.

More Efficient Self-Reflection Mechanism.

NO WAIT does not prohibit models from self-reflection. However, this method guides models to skip the unnecessary “waiting” reasoning. To illustrate this, we select an example from Qwen3-32B on AMC 2023 benchmark. The NO WAIT CoT (see Figure 12) is noticeably shorter than the original CoT (see Figure 9). Specifically, the NO WAIT CoT reserves its self-reflection for two clear points: ❶ it notes the extraneous root and instantly discards it after factoring, ❷ it quickly verifies both original equations with the numeric solution. By contrast, the original CoT continually interjects “let me check again”, and “perhaps another way”, leading to 5 derivations of essentially the same algebra. In short, the first approach builds a more concise



Figure 2: **One Case Study From QvQ-72B-Preview on MMVU.** NOWAIT CoT is more straightforward than the original CoT, without unnecessary self-reflection and verbosity.

reasoning process with necessary checks to ensure correctness, whereas the second strategy prefers to pause to highlight every minor thought, making the logic scattered and less efficient.

Concise and Straightforward Reasoning. Figure 8 presents an example from QvQ-72B-Preview on MMVU. The original CoT contains six instances of self-reflection, resulting in excessive token usage and a disorganized reasoning process. In contrast, the NOWAIT CoT exhibits a more streamlined and coherent approach. The model analyzes the video in detail, using a series of time-sequence cues such as “starts”, “After that”, “At one point”, “As the video progresses” and “Towards the end of the video”. With fewer self-reflections, the

NOWAIT CoT organizes its reasoning more logically and systematically, whereas the original CoT appears fragmented and less focused, always generating a new reasoning branch by “Wait”. Ultimately, the NOWAIT derives the final answer directly from its detailed analysis. Unlike the original policy, NOWAIT encourages the model to connect observations to conclusions more directly, reducing unnecessary speculation and making the reasoning process more concise and straightforward.

4.2 A Closer Look at RL Models Performance

For textual reasoning tasks, our evaluation primarily focuses on the math problems. As we discussed in Table 3.3, NOWAIT yields consistent experi-

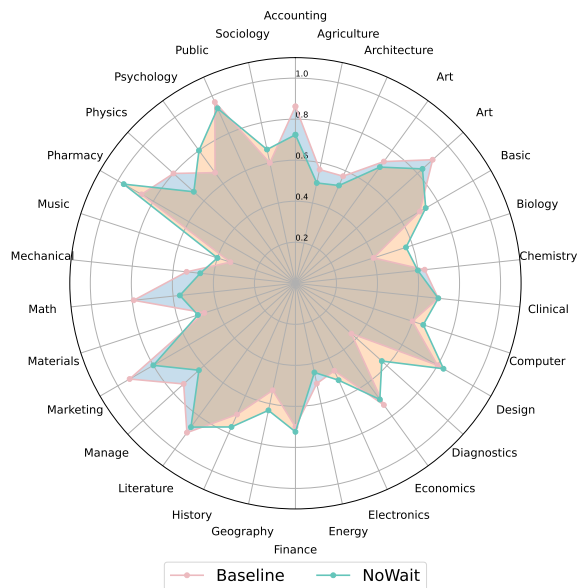


Figure 3: Accuracy Radar Map on MMMU for QvQ-72B-Preview.

mental outcomes across math benchmarks of varying difficulty levels. For multimodal reasoning tasks, Figure 3 shows the accuracy of the QvQ-72B-Preview on MMMU across a wide range of fields. A crucial observation highlights remarkably small accuracy divergence between the baseline and NOWAIT in almost all tested disciplines. Despite the potential intervention introduced by NOWAIT, the model’s performance remains closely aligned with the baseline across diverse academic and professional subjects. This minimal degradation strongly indicates the robustness of the QvQ-72B-Preview when applying NOWAIT, highlighting generalization capability across varying areas.

4.3 Distilled Models Cannot Reasoning without “Wait”

Recent studies (Yue et al., 2025) underscore the significant differences between reasoning models based on reinforcement learning (RL) and those trained through distillation. To better understand these differences, we further evaluate the effectiveness of NOWAIT across Qwen3 series, including an RL-based model (Qwen3-32B) and several distilled models (Qwen3-4B/8B/14B).

Figure 4 illustrates the accuracy degradation for models using NOWAIT, where a higher score indicates a more pronounced decline. The selected math reasoning benchmarks differ in difficulty, ordered as follows: AMC 2023 < AIME 2024 < AIME 2025. While the RL-based models maintain

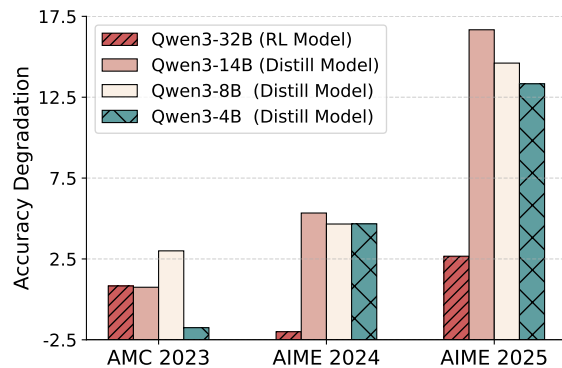


Figure 4: Accuracy Degradation across Qwen3 Seires Models on Math Reasoning Benchmarks.

consistent performance across these benchmarks, distilled models exhibit a distinct trend of increasing accuracy degradation as difficulty rises.

Specifically, distilled models show similar accuracy degradation relative to the RL-based model on the simpler AMC 2023. This performance gap extends significantly as problem difficulty increases, surpassing a 5-percentage-point drop on AIME 2024 and dramatically exceeding 12 percentage points on more challenging AIME 2025.

This sharp performance degradation among distilled models, in contrast to the stable performance of the RL-based model, demonstrates their higher sensitivity to reflection keywords. Given that the supervised fine-tune (SFT) directly injects new knowledge into models, the CoT structures are crucial for advanced reasoning. Simply removing these keywords, however, severely disrupts the inherent CoT structure, restricting distilled models from exhibiting full reasoning capabilities. Especially on more challenging reasoning problems, distill models fail to effectively conduct validation, suffering from substantial underthinking.

5 Related Work

Large Reasoning Model The pursuit of advanced reasoning capabilities in Large Language Models (LLMs) (OpenAI, 2024) has spurred significant research, particularly focusing on strategies that scale computation (Chen et al., 2024c; Snell et al., 2024) or refine the generation process during inference. Prior studies apply fundamental techniques like Chain-of-Thought (CoT) prompting (Wei et al., 2022), guiding the model to think step by step, or integrate Process Reward Models (PRMs), external verifiers, and search-guided decoding (Brown et al., 2024) to aggregate multi-

ple reasoning paths and enhance final answer accuracy. These efforts have culminated in a new generation of powerful Large Reasoning Models (LRMs), such as ChatGPT-O1 (OpenAI, 2024), Deepseek-R1 (Guo et al., 2025), QwQ (Qwen, 2025), Gemini2.5 (Google, 2025), which enable to spontaneous generation of extensive CoT sequences involving forward thinking, backtracking, and verification steps. Within the open-source domain, models derive reasoning abilities from diverse training paradigms, primarily through reinforcement learning (RL) (Guo et al., 2025; Ramesh et al., 2024; Muennighoff et al., 2025) on reasoning tasks or distillation (Guo et al., 2025; Yu et al., 2024) on high-quality CoT data produced from RL-based models. Recent works (Yue et al., 2025) have analyzed the difference between the two types of models. In this study, we include RL-based models for further exploration, underscoring the defects of RL-triggered reasoning capabilities.

Efficient Reasoning While elaborating reasoning processes like long CoT demonstrates enhanced performance on reasoning tasks, the associated verbosity presents a significant efficiency challenge (Chen et al., 2024a). The generation of extensive intermediate steps substantially increase inference latency and computational cost, hindering practical deployment in real-world applications. Consequently, a considerable body of work explores methods for efficient reasoning, aiming to reduce the length of reasoning traces without compromising accuracy. Some techniques continue to train models for CoT optimization (Aggarwal and Welleck, 2025; Luo et al., 2025; Shen et al., 2025b), such as applying RL with length-based reward design (Sun et al., 2024; Liao et al., 2025; Luo et al., 2025; Aggarwal and Welleck, 2025), or fine-tuning with variable-length CoT data (Han et al., 2024; Yu et al., 2024; Munkhbat et al., 2025). Other methods conform training-free strategy, applying dynamic reasoning paradigms during inference (Yang et al., 2025a; Zhang et al., 2025a; Wu et al., 2025; Lin et al., 2025) or leveraging prompts to guide efficient reasoning (Cheng and Van Durme, 2024; Xu et al., 2025; Han et al., 2024; Ma et al., 2025a). While existing studies are effective in cutting down the token usage, our study provides a new insight to rethink the internal mechanism of efficient reasoning and propose efficient multimodal reasoning.

Self-Reflection & Overthinking Parallel to enhancing reasoning capabilities and efficiency, re-

cent studies analyze the intricacies of the generated thought processes. Within these generated sequences, an interesting phenomenon occurs - moments marked by keywords like "wait" and "hmm", which we term *Aha Moment* (Guo et al., 2025; Liu et al., 2025). These moments seemingly indicate a capability for self-reflection (Chen et al., 2025a), allowing models to reassess their reasoning path and verify their CoT before concluding. Prior studies (Yang et al., 2025b; Zhang et al., 2025a) have begun to characterize these moments and probe the latent states to explore the potential mechanisms behind such spontaneous self-reflection. However, the frequent occurrence of these keywords can also lead to significant *Overthinking* (Chen et al., 2024a; Sui et al., 2025), where the model continues reflecting even after reaching correct intermediate or final conclusions. Building on the initial characterizations from previous work, our study takes a further step to evaluate the functional effectiveness of these spontaneously generated Aha Moments, directly addressing whether they are essential contributors to the reasoning outcomes or potentially represent a form of inefficient behavioral mimicry.

6 Conclusion

This work demonstrates that explicit self-reflection, signaled by tokens such as “Wait” and “Hmm”, is not essential for advanced reasoning in R1-style models. By suppressing these tokens during inference, the proposed NOWAIT approach effectively reduces overthinking and shortens chain-of-thought trajectories without compromising overall model utility. Extensive experiments across diverse models and benchmarks in textual, visual, and video reasoning tasks demonstrate that NOWAIT serves as an efficient and utility-preserving solution for multimodal reasoning, offering new insights for the lightweight deployment of large reasoning models.

Acknowledgment

Many thanks to Yao Wan and Jieyu Zhang for their invaluable support and comments.

Limitation

In this study, we introduce NOWAIT, a simple yet effective method for efficient reasoning on different modalities, and conduct experiments across various models and benchmarks. Although the promising results, we acknowledge that existing benchmarks

cannot comprehensively exhibit the reasoning capabilities of models from all aspects.

References

- Marah Abidin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, and 1 others. 2025. Phi-4-reasoning technical report. *arXiv preprint arXiv:2504.21318*.
- Pranjal Aggarwal and Sean Welleck. 2025. L1: Controlling how long a reasoning model thinks with reinforcement learning. *arXiv preprint arXiv:2503.04697*.
- AI-MO. 2024. AMC 2023. <https://huggingface.co/datasets/AI-MO/aimo-validation-amc>. Accessed: 2024-05-20.
- Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. 2024. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*.
- Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025a. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*.
- Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, and 1 others. 2024a. Do not think that much for $2+3=?$ on the overthinking of o1-like llms. *arXiv preprint arXiv:2412.21187*.
- Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, and 1 others. 2024b. Do not think that much for $2+3=?$ on the overthinking of o1-like llms. *arXiv preprint arXiv:2412.21187*.
- Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024c. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.
- Zhipeng Chen, Yingqian Min, Beichen Zhang, Jie Chen, Jinhao Jiang, Daixuan Cheng, Wayne Xin Zhao, Zheng Liu, Xu Miao, Yang Lu, and 1 others. 2025b. An empirical study on eliciting and improving r1-like reasoning models. *arXiv preprint arXiv:2503.04548*.
- Jeffrey Cheng and Benjamin Van Durme. 2024. Compressed chain of thought: Efficient reasoning through dense representations. *arXiv preprint arXiv:2412.13171*.
- Alejandro Cuadron, Dacheng Li, Wenjie Ma, Xingyao Wang, Yichuan Wang, Siyuan Zhuang, Shu Liu, Luis Gaspar Schroeder, Tian Xia, Huanzhi Mao, and 1 others. 2025. The danger of overthinking: Examining the reasoning-action dilemma in agentic tasks. *arXiv preprint arXiv:2502.08235*.
- Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. 2025. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*.
- Google. 2025. Gemini 2.5 Pro, Generative AI on Vertex AI. <https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro>.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shitong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Tingxu Han, Zhenting Wang, Chunrong Fang, Shiyu Zhao, Shiqing Ma, and Zhenyu Chen. 2024. Token-budget-aware llm reasoning. *arXiv preprint arXiv:2412.18547*.
- Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. 2025. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. *arXiv preprint arXiv:2501.05444*.
- Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.
- Baohao Liao, Yuhui Xu, Hanze Dong, Junnan Li, Christof Monz, Silvio Savarese, Doyen Sahoo, and Caiming Xiong. 2025. Reward-guided speculative decoding for efficient llm reasoning. *arXiv preprint arXiv:2501.19324*.
- Kevin Lin, Charlie Snell, Yu Wang, Charles Packer, Sarah Wooders, Ion Stoica, and Joseph E Gonzalez. 2025. Sleep-time compute: Beyond inference scaling at test-time. *arXiv preprint arXiv:2504.13171*.
- Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *Proceedings of International Conference on Learning Representations (ICLR)*.

- Haotian Luo, Li Shen, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Naiqiang Tan, Xiaochun Cao, and Dacheng Tao. 2025. O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning. *arXiv preprint arXiv:2501.12570*.
- Wenjie Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, and Matei Zaharia. 2025a. Reasoning models can be effective without thinking. *arXiv preprint arXiv:2504.09858*.
- Xinyin Ma, Guangnian Wan, Runpeng Yu, Gongfan Fang, and Xinchao Wang. 2025b. Cot-valve: Length-compressible chain-of-thought tuning. *arXiv preprint arXiv:2502.09601*.
- MAA Committees. Aime problems and solutions. https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions. Accessed: 2024-05-20.
- Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.
- Tergel Munkhbat, Namgyu Ho, Seo Hyun Kim, Yongjin Yang, Yujin Kim, and Se-Young Yun. 2025. Self-training elicits concise reasoning in large language models. *arXiv preprint arXiv:2502.20122*.
- OpenAI. 2024. **GPT-4o System Card**. *Preprint*, arXiv:2410.21276.
- OpenAI. 2024. Introducing OpenAI o1. <https://openai.com/o1/>.
- Qwen. 2025. **QwQ-32B: Embracing the Power of Reinforcement Learning**.
- Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. 2024. Group robust preference optimization in reward-free rlhf. *Advances in Neural Information Processing Systems*, 37:37100–37137.
- 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 *Proceedings of First Conference on Language Modeling*.
- Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, and 1 others. 2025a. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*.
- Yi Shen, Jian Zhang, Jieyun Huang, Shuming Shi, Wenjing Zhang, Jiangze Yan, Ning Wang, Kai Wang, and Shiguo Lian. 2025b. Dast: Difficulty-adaptive slow-thinking for large reasoning models. *arXiv preprint arXiv:2503.04472*.
- Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.
- Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Hanjie Chen, Xia Hu, and 1 others. 2025. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419*.
- Hanshi Sun, Momin Haider, Ruiqi Zhang, Huitao Yang, Jiahao Qiu, Ming Yin, Mengdi Wang, Peter Bartlett, and Andrea Zanette. 2024. Fast best-of-n decoding via speculative rejection. *arXiv preprint arXiv:2410.20290*.
- Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, and 1 others. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.
- Qwen Team. 2024. **QVQ: To See the World with Wisdom**.
- Qwen Team. 2025. **Qwen3: Think Deeper, Act Faster**.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the Advances in Neural Information Processing Systems*.
- Yuyang Wu, Yifei Wang, Tianqi Du, Stefanie Jegelka, and Yisen Wang. 2025. When more is less: Understanding chain-of-thought length in llms. *arXiv preprint arXiv:2502.07266*.
- Bingquan Xia, Bowen Shen, Dawei Zhu, Di Zhang, Gang Wang, Hailin Zhang, Huaqiu Liu, Jiebao Xiao, Jinhao Dong, Liang Zhao, and 1 others. 2025. Mimo: Unlocking the reasoning potential of language model—from pretraining to posttraining. *arXiv preprint arXiv:2505.07608*.
- Silei Xu, Wenhao Xie, Lingxiao Zhao, and Pengcheng He. 2025. Chain of draft: Thinking faster by writing less. *arXiv preprint arXiv:2502.18600*.
- Chenxu Yang, Qingyi Si, Yongjie Duan, Zheliang Zhu, Chenyu Zhu, Zheng Lin, Li Cao, and Weiping Wang. 2025a. Dynamic early exit in reasoning models. *arXiv preprint arXiv:2504.15895*.
- Jihan Yang, Shusheng Yang, Anjali Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2024. Thinking in Space: How Multimodal Large Language Models See, Remember and Recall Spaces. *arXiv preprint arXiv:2412.14171*.
- Shu Yang, Junchao Wu, Xin Chen, Yunze Xiao, Xinyi Yang, Derek F Wong, and Di Wang. 2025b. Understanding aha moments: from external observations to internal mechanisms. *arXiv preprint arXiv:2504.02956*.

Wang Yang, Xiang Yue, Vipin Chaudhary, and Xiaotian Han. 2025c. Speculative thinking: Enhancing small-model reasoning with large model guidance at inference time. *arXiv preprint arXiv:2504.12329*.

Ping Yu, Jing Xu, Jason Weston, and Ilia Kulikov. 2024. Distilling system 2 into system 1. *arXiv preprint arXiv:2407.06023*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024a. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of Conference on Computer Vision and Pattern Recognition*.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhao Chen, and Graham Neubig. 2024b. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. 2025a. Reasoning models know when they’re right: Probing hidden states for self-verification. *arXiv preprint arXiv:2504.05419*.

Jingyi Zhang, Jiaying Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. 2025b. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*.

Yilun Zhao, Lujing Xie, Haowei Zhang, Guo Gan, Yitao Long, Zhiyuan Hu, Tongyan Hu, Weiyan Chen, Chuhan Li, Junyang Song, and 1 others. 2025. Mmvu: Measuring expert-level multi-discipline video understanding. *arXiv preprint arXiv:2501.12380*.

Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. 2025. R1-zero’s “aha moment” in visual reasoning on a 2b non-sft model. *arXiv preprint arXiv:2503.05132*.

A Baseline Implementation Details

Our experiments include three baselines, NoThinking (Ma et al., 2025a), TokenBudget (Han et al., 2024), and O1-Pruner (Luo et al., 2025). In this section, we will systematically introduce the implementation details of these techniques.

A.1 NoThinking

The core idea of NoThinking is to leverage prompts, guide reasoning models to skip the reasoning processes, and directly generate a final response. For models that have not been post-trained for non-reasoning mode, such as QwQ-32B and Phi4-Reasoning-Plus, we apply the prompt template as follows:

Prompt Template for NoThinking

```
{Question}
<think>
Okay, I think I have finished thinking.
</think>
```

We then adopt a budget forcing technique specifically for NoThinking. Different from the token budget we apply for normal inference and NOWAIT, we set the token budget to 10,000 and forced models to generate *Final Answer* when the model reaches the token budget.

A.2 Token-Budget

We apply the TALE-EP strategy, a prompt-based method. This method consists of two steps:

- 1 Directly answering the reasoning model:

Prompt Template for TALE-EP

Task: Analyze the given question and estimate the minimum number of tokens required to generate a complete and accurate response. Please give the response by strictly following this format: `[[budget]]`, for example, Budget: `[[12]]`.

- 2 We include a token budget in the prompt to guide models to think efficiently.

Table 5: Prompt Template Applied for Token Budget.

Prompt method	Content
Vanilla CoT	Let’s think step by step:
Token Budget	Let’s think step by step and use less than {budget} tokens:

A.3 O1-Pruner

O1-Pruner is an effective post-training method. We select a released model trained on QwQ-32B-

Preview by O1-Pruner. This model can be accessed via [Hugging Face](#).

B Prompts

Prompt Template for Multiple-Choice Question

{Question}

—

Choices:

A. option A

B. option B

...

—

Choose the correct answer from the choices above.

Output format: [ANSWER: "<answer>"] If the answer is A, output [ANSWER: "A"]

C Benchmark & Models

C.1 Textual QA

In this paper, we evaluate a range of mathematics competition benchmarks designed to assess the mathematical reasoning abilities of models, including **AIME2024**, **AIME2025**, **AMC2023**. We have also evaluated **GPQA-Diamond** ([Rein et al., 2024](#)), a challenging benchmark spanning biology, physics, and chemistry. The detailed information about these benchmarks is as follows:

- **AIME2024**: A benchmark derived from the 2024 American Invitational Mathematics Examination (AIME), a challenging mathematics competition aimed at high school students in the U.S., designed specifically to evaluate the advanced mathematical reasoning abilities of AI models. It consists of complex problems covering algebra, geometry, combinatorics, and number theory, each requiring integer solutions ranging from 0 to 999. Models are tested on their ability to perform multi-step reasoning, provide accurate step-by-step explanations, and derive correct final answers.
- **AIME2025**: Like AIME2024, the AIME2025 benchmark is based on the 2025 American Invitational Mathematics Examination (AIME), an advanced and highly respected mathematics competition aimed at high school students in the United States, intended specifically for

evaluating the mathematical reasoning and problem-solving capabilities of AI models.

- **AMC2023**: A benchmark derived from the 2023 American Mathematics Competitions (AMC), specifically designed to evaluate the mathematical reasoning abilities of AI models. It consists of 40 questions, covering various mathematical topics such as algebra, geometry, number theory, and combinatorics.
- **GPQA-Diamond** ([Rein et al., 2024](#)): A subset of the GPQA dataset, specifically designed to assess the reasoning capabilities of advanced AI systems and highly knowledgeable humans on extremely difficult, domain-expert-level questions in biology, physics, and chemistry. The "Diamond" subset is the hardest subset of the benchmark, which is intended to facilitate research on reasoning models.

We evaluated and measured these models on the above benchmarks:

- **QwQ-32B** ([Qwen, 2025](#)): A large-scale language model designed to achieve robust performance across a wide range of natural language processing tasks. Developed with 32 billion parameters, QwQ32B leverages advanced architecture and training techniques to enhance understanding, generation, and reasoning in general and specialized domains.
- **Phi4-Reasoning-Plus** ([Abdin et al., 2025](#)): Built on Phi-4 Base, it is an advanced language model specifically designed to excel in complex reasoning and problem-solving tasks across multiple domains, demonstrating strong performance in textual data.
- **Qwen3-32B** ([Team, 2025](#)): A state-of-the-art large language model developed by Alibaba Cloud, featuring 32 billion parameters and designed to deliver high performance across a broad spectrum of language understanding, text generation and reasoning tasks.
- **QwQ-32B-Preview** ([Qwen, 2025](#)): An experimental large language model developed by Alibaba, designed to advance AI reasoning capabilities. With 32.5 billion parameters and a 32,768-token context window, it is specifically tested on benchmark AIME2024 and AMC2023 to compare with other methods.

C.2 Visual QA

Additionally, we incorporate evaluations on the multimodal benchmarks including **MMMU-Pro** (Yue et al., 2024b), **MMMU** (Yue et al., 2024a), **Math-Vista** (Lu et al., 2024) and **EMMA-mini** (Hao et al., 2025) to further explore the models’ capabilities across diverse reasoning and multimodal tasks. Here is the detailed information:

- **MMMU** (Yue et al., 2024a): A multimodal evaluation benchmark specifically designed to test the capabilities of AI models on college-level tasks that require both advanced subject knowledge and deliberate reasoning across a broad range of academic disciplines.
- **MMMU-Pro** (Yue et al., 2024b): MMMU-Pro is an enhanced evaluation benchmark designed to rigorously test the true understanding and reasoning capabilities of multimodal AI models. Building on the original MMMU benchmark, it forces models to simultaneously process and integrate visual and textual information, simulating real-world scenarios that require human-like cognitive skills.
- **Math-Vista** (Lu et al., 2024) : A comprehensive benchmark specifically designed to evaluate and challenge the mathematical reasoning abilities of large language and multimodal models within visual contexts. It requires models to perform deep, fine-grained visual understanding and complex compositional reasoning across diverse mathematical tasks.
- **EMMA-mini** (Hao et al., 2025): A specialized benchmark designed to rigorously assess the ability of Multimodal Large Language Models (MLLMs) to perform integrated, organic reasoning over both text and images—an essential aspect of human intelligence. Unlike existing benchmarks that often focus on text-based reasoning or superficial visual cues, EMMA-mini presents tasks spanning mathematics, physics, chemistry, and coding, all of which require genuine cross-modal reasoning that cannot be solved by independently analyzing text or images alone.

We evaluated and measured these models on the above benchmarks:

- **Kimi-VL-A3B-Thinking** (Team et al., 2025): An efficient open-source vision-language

model (VLM) built on a Mixture-of-Experts (MoE) architecture, designed to deliver advanced multimodal reasoning, robust long-context understanding, maths problem solving as well as strong agent capabilities.

- **QvQ-72B-Preview** (Team, 2024): QVQ-72B-preview is an open-source, large-scale multimodal reasoning model built on top of Qwen2-VL-72B, achieving remarkable performance on challenging benchmarks. In this part of the experiment, the image recognition and reasoning capability of this model has been tested.

C.3 Video QA

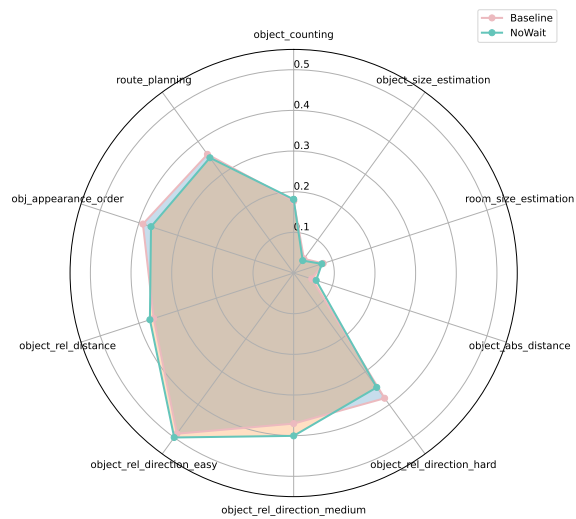
Furthermore, we conduct experiment on video benchmarks, whose name and details is listed as follows:

- **MMVU** (Zhao et al., 2025): A comprehensive dataset designed to evaluate the capabilities of AI models in understanding and reasoning over expert-level, domain-specific videos. Each example is meticulously crafted using a textbook-guided annotation process, ensuring that questions require both visual comprehension and the application of domain-specific knowledge. What’s unique to MMVU is the inclusion of expert-annotated reasoning rationales and relevant domain knowledge for each question, which largely facilitates the fine-grained analysis of model performance.
- **VSI-Bench** (Yang et al., 2024): A pioneering dataset, designed to evaluate the visual-spatial reasoning capabilities of multimodal large language models (MLLMs). It comprises many question-answer pairs derived from egocentric videos that are sourced from public indoor 3D scene reconstruction datasets, aiming to provide a comprehensive benchmark for testing and improving the spatial reasoning abilities of multimodal large language models, moving beyond traditional static image evaluations.

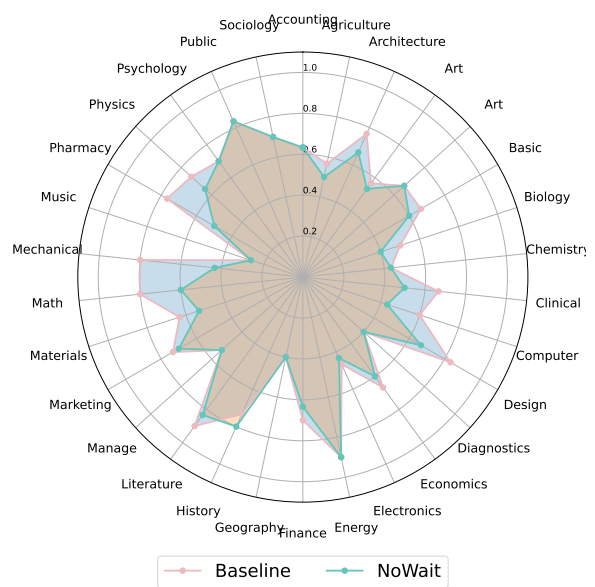
E Additional Experiment Results & Case Study

Table 6: Complete Experiment results of NoWait on Qwen3 Series Models and Other Distill Models.

Strategy	AMC 2023		AIME 2024		AIME 2025		GPQA-D	
	ACC↑	LEN↓	ACC↑	LEN↓	ACC↑	LEN↓	ACC↑	LEN↓
Qwen3-32B								
Original	97.50	6424	81.33	12720	66.67	14987	69.19	5613
NoThink	59.50	1240	25.33	2511	20.00	2165	50.50	605
NoWait	96.67	5560	83.33	10732	64.44	12930	63.13	4788
Qwen3-14B								
Original	96.25	6677	78.67	14217	78.00	14765	59.59	4633
NoThink	69.50	1749	33.33	3559	26.67	3171	38.59	1001
NoWait	95.50	4714	73.33	10919	61.33	12104	54.75	3889
Qwen3-8B								
Original	97.50	8513	77.33	14142	74.61	16094	57.07	5904
NoThink	66.50	1760	28.89	3362	25.56	3719	32.93	1271
NoWait	94.50	5251	72.67	10963	60.00	13674	51.71	4735
Qwen3-4B								
Original	93.75	8125	70.00	13488	70.00	18086	53.54	5965
NoThink	70.00	2236	33.33	4068	23.00	4656	27.27	1288
NoWait	95.50	4523	65.33	10358	56.67	12213	76.26	3178
Llama-Nemotron-Nano-8B-v1								
Original	71.50	4535	39.33	7371	44.67	11798	54.1	5071
NoThink	39.00	2982	6.6	2185	14.67	3677	30.30	3109
NoWait	72.00	3690	30.67	4865	33.33	7271	42.83	3754
Deepseek-R1-Distill-Qwen-7B								
Original	73.00	4796	34.67	7755	40.00	13767	49.10	3809
NoThink	30.00	2552	18.67	3895	10.00	2724	21.00	1112
NoWait	72.00	4315	26.67	7247	31.33	8236	40.91	3672



(a) Accuracy Radar Map for QvQ-72B-Preview on VSI-Bench



(b) Accuracy Radar Map for Kimi-VL-A3B-Thinking-7B on MMMU

Original CoT:

So I'm watching this video where there's a woman conducting some kind of experiment or demonstration. She's in a room that looks like a laboratory or a classroom with a lot of scientific equipment and wooden cabinets in the background. On the table in front of her, there's this setup that seems to be a pendulum or some sort of apparatus that can move back and forth.

She starts by explaining what she's going to do. She mentions a pendulum and talks about its motion. Then, she begins to manipulate the apparatus, pushing it back and forth. I can see that the object is moving in a way that suggests it's a pendulum, swinging from one side to the other.

As she continues, she adds another component to the setup. It looks like she's introducing a new element into the experiment, maybe to observe how it affects the pendulum's motion. She keeps adjusting things and observing the results.

At one point, she steps away from the table briefly and then returns to continue her demonstration. Throughout the video, she's engaged in showing and explaining the principles behind the pendulum's behavior.

Now, at the end of the video, there's a question about how the energy of the pendulum changes after adding a new item to the experiment, assuming no friction is present. The options are:

A. Larger B. Smaller, then larger C. Smaller D. Keep same E. Larger, then smaller
I need to choose the correct answer from these options.

First, I should recall some basic physics about pendulums and energy conservation. A pendulum swings back and forth, converting potential energy to kinetic energy and back again. If there's no friction or other dissipative forces, the total mechanical energy of the pendulum should remain constant. That means that the sum of its potential energy and kinetic energy shouldn't change over time.

So, if the woman adds a new item to the experiment without considering friction, I need to think about how this addition affects the pendulum's energy.

Possible scenarios could be:

1. Adding mass to the pendulum: If she adds mass to the pendulum, the potential energy at the highest points would increase because potential energy depends on mass and height. Since total energy is conserved, the kinetic energy at the lowest point would also have to adjust accordingly. So, the overall energy might be larger if mass is added, but I'm not entirely sure.
2. Changing the length of the pendulum: Changing the length affects the period of the pendulum but doesn't directly change its energy if the amplitude remains the same. However, if the length is changed and the amplitude is adjusted, that could affect the energy.

Figure 6: A CoT Example from QvQ-72B-Preview on MMVU 2023.

Original CoT(continue):

3. Adding an obstacle or another object that the pendulum interacts with: This could transfer some of the pendulum's energy to the other object, possibly reducing the pendulum's energy.

But without knowing exactly what she added to the experiment, it's a bit tricky to be precise.

Let me think differently. Suppose she added something that doesn't directly interact with the pendulum's energy, like a support or another structural component that doesn't affect the motion. In that case, the energy of the pendulum should remain the same, assuming no energy is being lost or gained from external sources.

However, if she added mass to the pendulum, say by attaching something to it, that would increase its moment of inertia, potentially affecting its motion and energy.

Wait a minute, maybe I should consider the principle of conservation of energy. If no friction is present and no external work is done on the system, the total energy should remain constant. So, unless she's adding or removing energy in some way, the energy of the pendulum should stay the same.

But option A suggests that the energy becomes larger, which might be the case if she added mass without accounting for the increased potential energy.

Option B says smaller, then larger, which doesn't make much sense immediately.

Option C says smaller, which might be if she added something that absorbed energy, like a damping mechanism.

Option D says keep same, which aligns with the conservation of energy principle if no energy is added or removed.

Option E says larger, then smaller, which also seems unclear.

Given that the question specifies "without consideration of friction," and assuming that the added item doesn't dissipate energy, the most straightforward answer would be D. Keep same.

However, if adding mass to the pendulum changes its energy, then perhaps it could be A. Larger.

But I think the key here is that if no friction is considered, and no energy is being added or removed, the energy should remain constant. Therefore, the answer should be D. Keep same.

I'll go with that.

****Final Answer****

$\boxed{D}$

Figure 7: A CoT Example from QvQ-72B-Preview on MMVU 2023.

No Wait CoT:

So I'm watching this video where there's a woman conducting some kind of experiment or demonstration. She's in a room that looks like a laboratory or a classroom with a lot of scientific equipment and wooden cabinets in the background. On the table in front of her, there's this setup that seems to be a pendulum or some sort of apparatus that can move left and right.

She starts by explaining what she's going to do. She mentions a pendulum and talks about its motion. Then, she begins to manipulate the apparatus, pushing it and observing how it moves. She does this a few times, and it seems like she's demonstrating the natural motion of the pendulum without任何 external forces acting on it.

After that, she introduces a new element into the experiment. She adds something to the setup, possibly a magnet or另一个 pendulum, and continues to observe the motion. This time, the motion of the original pendulum is affected by this new addition, and she explains what's happening.

Throughout the video, she's very animated and engaged, using her hands to show what's going on and speaking clearly to the camera. The room is filled with various scientific instruments and models, which adds to the educational atmosphere of the setting.

At one point, she steps away from the table to retrieve something, and then returns to continue her demonstration. This suggests that she's prepared multiple parts to her experiment, each building upon the last to illustrate a particular concept or principle.

As the video progresses, she continues to interact with the apparatus, adjusting it and observing the effects. Her explanations are thorough, and she makes sure to point out key points in the motion and how they relate to the properties of pendulums.

Towards the end of the video, she summarizes what she's demonstrated and how it relates to the choices provided in the question. She emphasizes the importance of considering friction and otros factors that might affect the energy of the pendulum over time.

Overall, the video is a well-structured and informative demonstration of pendulum motion and the factors that influence its energy, particularly focusing on the changes in energy when additional elements are introduced into the system.

****Final Answer****

$\boxed{D}$

Figure 8: A CoT Example from QvQ-72B-Preview applied NOWAIT on MMVU 2023.

Original CoT:

Okay, so I need to solve this problem where positive real numbers x and y satisfy two equations: $y^3 = x^2$ and $(y - x)^2 = 4y^2$. And then find $x + y$. Let me think about how to approach this.

First, let me write down both equations to have them clear: 1. $y^3 = x^2$ 2. $(y - x)^2 = 4y^2$
Hmm. Since both equations involve x and y , maybe I can express one variable in terms of the other using the first equation and substitute into the second. Let me try that.
From equation 1: $y^3 = x^2$. Since we're dealing with positive real numbers, I can take square roots or cube roots without worrying about multiple solutions. Let me solve for x . Taking square roots on both sides:

$x = y^{(3/2)}$. Alternatively, $x = \sqrt{y^3}$. Either way. Maybe I can write it as $x = y\sqrt{y}$. That might be helpful later.

Alternatively, maybe solving for y in terms of x ? Let me see. If $y^3 = x^2$, then $y = x^{(2/3)}$. Hmm, that's another way. So maybe I can write y as x raised to $2/3$ power. But maybe substituting x in terms of y into the second equation is better. Let me try that. So equation 2: $(y - x)^2 = 4y^2$. Let me expand the left side:

$(y - x)^2 = y^2 - 2xy + x^2$. And the right side is $4y^2$. So: $y^2 - 2xy + x^2 = 4y^2$.

Let me rearrange terms:

$$y^2 - 2xy + x^2 - 4y^2 = 0 \Rightarrow -3y^2 - 2xy + x^2 = 0 \Rightarrow x^2 - 2xy - 3y^2 = 0.$$

Hmm, so $x^2 - 2xy - 3y^2 = 0$. Maybe this quadratic can be factored? Let me check.

Looking for factors of the form $(x + a y)(x + b y) = x^2 + (a + b)xy + ab y^2$. Comparing to $x^2 - 2xy - 3y^2$. So need $a + b = -2$ and $ab = -3$. Let me think. Factors of -3 that add up to -2 ...

Let's see:

Possible pairs for $ab = -3$:

1 and -3 : $1 + (-3) = -2$. Yes! Perfect. So $(x + (1)y)(x + (-3)y) = (x + y)(x - 3y) = x^2 - 2xy - 3y^2$. Correct.

So equation factors to $(x + y)(x - 3y) = 0$. Therefore, either $x + y = 0$ or $x - 3y = 0$.

But since x and y are positive real numbers, $x + y$ can't be zero. So we discard that solution. Therefore, $x - 3y = 0 \Rightarrow x = 3y$.

So from this, $x = 3y$. Now, substitute this into the first equation, which is $y^3 = x^2$. Let me do that.

$$y^3 = (3y)^2 \Rightarrow y^3 = 9y^2.$$

Since y is a positive real number, we can divide both sides by y^2 :

$$y^3 / y^2 = 9y^2 / y^2 \Rightarrow y = 9.$$

Therefore, $y = 9$. Then $x = 3y = 3 \cdot 9 = 27$.

Therefore, $x + y = 27 + 9 = 36$. So is the answer 36?

Wait, let me check if this satisfies both equations. Let me verify.

First equation: $y^3 = 9^3 = 729$. $x^2 = 27^2 = 729$. Yes, that works.

Second equation: $(y - x)^2 = (9 - 27)^2 = (-18)^2 = 324$. $4y^2 = 4 \cdot (9)^2 = 4 \cdot 81 = 324$. That also works. So seems correct.

Figure 9: A CoT Example from Qwen3-32B on AMC 2023.

Original CoT(continue):

Wait, but let me check if there are other solutions. Because when we squared or did substitutions, sometimes extraneous solutions can come up, but here we had only one valid solution from the factoring. Since the other solution $x + y = 0$ is invalid because x and y are positive, so only $x = 3y$. So substituting back gives unique solution. But let me just think again. Is there another way this could be approached?

Alternatively, since we had $(y - x)^2 = 4y^2$, taking square roots on both sides. Since both sides are positive, we can write $|y - x| = 2y$. Since y and x are positive, but we don't know which is bigger. So $|y - x| = 2y$ implies that either $y - x = 2y$ or $y - x = -2y$.

Case 1: $y - x = 2y \Rightarrow -x = y \Rightarrow x = -y$. But since x and y are positive, this is impossible. So discard.

Case 2: $y - x = -2y \Rightarrow -x = -3y \Rightarrow x = 3y$. Which is same as before. So same result. So that gives $x = 3y$. Then same steps as before. So that seems consistent.

Therefore, the only solution is $x = 27$, $y = 9$, so $x + y = 36$. Therefore, answer is 36.

But let me think again if there's another possible case? For example, maybe when I took square roots, I considered absolute value, but since $y - x$ is negative, leading to $x = 3y$. But is there a case where $y - x$ is positive? Let me check.

If $|y - x| = 2y$, then if $y - x$ is positive, then $y - x = 2y \Rightarrow -x = y \Rightarrow x = -y$. Which is impossible. So only the case where $y - x$ is negative gives us a valid solution. So seems like the only solution.

Alternatively, maybe I can use substitution earlier. For example, since from equation 1, $x = y^{3/2}$. Let me try substituting into equation 2.

Equation 2: $(y - x)^2 = 4y^2$. So substituting x : $(y - y^{3/2})^2 = 4y^2$. Let me write that: $(y - y^{3/2})^2 = 4y^2$. Let me take square roots? But maybe expanding.

First, expand left side: $= [y - y^{3/2}]^2 = y^2 - 2 * y * (y^{3/2}) + y^3$.

Compute each term: First term: y^2 . Second term: $-2 * y * y^{3/2} = -2y^{1 + 3/2} = -2y^{5/2}$. Third term: $[y^{3/2}]^2 = y^3$.

Therefore, left side is $y^2 - 2y^{5/2} + y^3$. And this equals $4y^2$. Therefore: $y^3 - 2y^{5/2} + y^2 - 4y^2 = 0 \Rightarrow y^3 - 2y^{5/2} - 3y^2 = 0$.

Hmm. Let me factor out y^2 : $y^2(y - 2y^{1/2} - 3) = 0$.

Since y is positive, $y^2 \neq 0$, so: $y - 2\sqrt{y} - 3 = 0$.

Let me set $t = \sqrt{y}$. Since y is positive, t is positive. Then $y = t^2$. Substitute into equation: $t^2 - 2t - 3 = 0$.

This quadratic equation can be solved:

$$t = [2 \pm \sqrt{4 + 12}] / 2 = [2 \pm \sqrt{16}] / 2 = [2 \pm 4] / 2.$$

So $t = (2 + 4) / 2 = 6 / 2 = 3$, or $t = (2 - 4) / 2 = (-2) / 2 = -1$. Since $t = \sqrt{y}$ is positive, discard -1 .

So $t = 3$. Therefore, $\sqrt{y} = 3 \Rightarrow y = 9$. Which is same result as before. Then $x = y^{3/2} = 9^{3/2} = (\sqrt{9})^3 = 3^3 = 27$. So same result. Therefore, $x + y = 36$. So that's consistent.

Figure 10: A CoT Example from Qwen3-32B on AMC 2023.

Original CoT(continue):

Original equations: First: $y^3 = 9^3 = 729$. $x^2 = 27^2 = 729$. Correct.

Second equation: $(y - x)^2 = (9 - 27)^2 = (-18)^2 = 324$. $4y^2 = 4 \cdot (81) = 324$. Correct. So both equations are satisfied. So answer is 36. Seems solid.

But let me think if there's another approach. Maybe parametric?

Alternatively, since $y^3 = x^2$, so we can parametrize x and y in terms of another variable. For example, let me set $x = t^2$ and $y = t^3$ for some positive real t . Wait, because if $x = t^2$, then $y^3 = x^2 = t^4 \Rightarrow y = t^{4/3}$. Alternatively, maybe better to think of exponents. Let me think.

If $y^3 = x^2$, then let me write x and y as powers of some variable. Let me suppose that $x = k^m$ and $y = k^n$ for some $k > 0$. Then equation becomes $(k^n)^3 = (k^m)^2 \Rightarrow k^{3n} = k^{2m} \Rightarrow 3n = 2m$. So $m = (3/2)n$. So for example, if I set $n = 2$, then $m = 3$. Therefore, $x = k^3$, $y = k^2$. So then, maybe parametrize x and y as $x = k^3$, $y = k^2$. Let me try this.

Let me set $y = k^2$, so that $y^3 = (k^2)^3 = k^6$. Then $x^2 = k^6 \Rightarrow x = k^3$ (since x positive). Therefore, $x = k^3$, $y = k^2$. Therefore, now substitute into second equation: $(y - x)^2 = 4y^2$. So $(k^2 - k^3)^2 = 4(k^2)^2 \Rightarrow (k^2(1 - k))^2 = 4k^4 \Rightarrow k^4(1 - k)^2 = 4k^4$.

Since $k > 0$, divide both sides by k^4 : $(1 - k)^2 = 4 \Rightarrow 1 - k = \pm 2 \Rightarrow 1 - k = 2$ or $1 - k = -2$.

Case 1: $1 - k = 2 \Rightarrow -k = 1 \Rightarrow k = -1$. Disregard, since $k > 0$.

Case 2: $1 - k = -2 \Rightarrow -k = -3 \Rightarrow k = 3$. So $k = 3$. Therefore, $y = k^2 = 9$, $x = k^3 = 27$. So same result again. Therefore, $x + y = 36$. So same answer. So this method also gives 36. So seems solid.

Therefore, regardless of substitution method or parametrization, answer is 36.

Therefore, I think confident that 36 is correct.

But just to ensure, let me think if there's a possibility that when I divided by y^2 or k^4 , I might have missed something? But since y and k are positive, they can't be zero, so division is allowed. So no problem there.

Alternatively, maybe think of ratios. Let me see.

Since we have $x = 3y$ from earlier. Then, since $y^3 = x^2$, substituting $x = 3y$ gives $y^3 = (3y)^2 \Rightarrow y^3 = 9y^2 \Rightarrow y = 9$. So that's straightforward.

Alternatively, let me think of ratios between x and y . Suppose that $x/y = r$, some positive real number. Then $x = r y$. Then, from equation 1: $y^3 = x^2 = (r y)^2 = r^2 y^2 \Rightarrow y^3 = r^2 y^2 \Rightarrow y = r^2$. Therefore, $y = r^2$. Then since $x = r y$, $x = r \cdot r^2 = r^3$. Therefore, $x = r^3$ and $y = r^2$. Then, substitute into equation 2: $(y - x)^2 = 4y^2$. Therefore:

$(y - x)^2 = (r^2 - r^3)^2 = r^4(1 - r)^2$. On the other hand, $4y^2 = 4r^4$. Therefore: $r^4(1 - r)^2 = 4r^4 \Rightarrow$ divide both sides by r^4 (allowed since $r > 0$): $(1 - r)^2 = 4 \Rightarrow$ same equation as before. Therefore, $1 - r = \pm 2 \Rightarrow r = 1 \pm 2$. So $r = 3$ or $r = -1$. Since $r = x/y > 0$, $r = 3$. Therefore, $r = 3$. Therefore, $y = r^2 = 9$, $x = r^3 = 27$. So again same result. Therefore, $x + y = 36$. So all methods lead to same answer. So I think that's definitely correct.

Therefore, the answer is 36. I think that's solid.

****Final Answer****

$\boxed{36}$

Figure 11: A CoT Example from Qwen3-32B on AMC 2023.

No Wait CoT:

<think> Okay, so I need to solve this problem where positive real numbers x and y satisfy two equations: $y^3 = x^2$ and $(y - x)^2 = 4y^2$. Then find the value of $x + y$. Let me try to figure it out step by step.

First, let's write down what we know:

1. $y^3 = x^2$. 2. $(y - x)^2 = 4y^2$.

We are looking for $x + y$ with both x and y being positive real numbers. Starting with equation 2: $(y - x)^2 = 4y^2$. Let me expand that left side first.

Expanding $(y - x)^2$ gives us $y^2 - 2xy + x^2$. So, $y^2 - 2xy + x^2 = 4y^2$

Let me rearrange terms here:

Bring all terms to one side: $y^2 - 2xy + x^2 - 4y^2 = 0$ Simplify like terms: $-3y^2 - 2xy + x^2 = 0$.

So that simplifies to: $x^2 - 2xy - 3y^2 = 0$

Looking at this quadratic in x or perhaps factorable?

Yes! This looks similar to a quadratic equation if we consider variables as coefficients related to either x or y . Let me see...

Treating it as a quadratic in x : $ax^2 + bx + c = 0$ form, which would be: $(1)x^2 - (2y)x - 3y^2 = 0$

Trying to factor this quadratic equation. To factor, look for two numbers multiplying to $(-3y^2)$ and adding up to $(-2y)$. Let me think... The factors could involve constants multiplied by y .

Suppose they multiply to $-3y^2$ and add to $-2y$. If we take $-3y$ and $+y$: $(-3y)(+y) = -3y^2$, and their sum is $-3y + y = -2y$. Yes!

Therefore, factoring: $(x - 3y)(x + y) = 0$

Set each factor equal to zero: Either: $x - 3y = 0 \rightarrow x = 3y$, or $x + y = 0 \rightarrow$ since x and y are positive real numbers, $x + y$ can't be zero. Therefore, discard second solution.

Thus, only valid solution from equation (2): $x = 3y$.

Alright, so we have x expressed in terms of y : $x = 3y$.

Given also equation (1): $y^3 = x^2$. Substitute x into there: $y^3 = (3y)^2 \Rightarrow y^3 = 9y^2$.

Since y is a positive real number, we can divide both sides by y^2 without issues: Divide both sides by y^2 :

$(y^3)/(y^2) = 9*(y^2)/y^2 \Rightarrow y = 9*1 \Rightarrow y = 9$.

Then substitute $y = 9$ into $x = 3y$: $x = 3 * 9 = 27$. Therefore, $x = 27$ and $y = 9$. Thus $x + y = 27 + 9 = 36$.

Hold on, before finalizing answer, let me confirm whether these values actually satisfy both original equations given.

First, checking equation (1): y^3 should equal x^2 . Compute y^3 : $9^3 = 729$; compute x^2 : $27^2 = 729$. Okay, good, matches.

Equation (2): $(y - x)^2$ vs. $4y^2$. Calculate $y - x$: $9 - 27 = -18$; square it: $(-18)^2 = 324$.

Compute $4y^2$: $4*(9^2)=4*81=324$. Also matches perfectly. Perfect. Seems correct then.

Answer must be 36.

BUT hold on—let me make sure that when solving the quadratic, there isn't some missing case or something else going on?

Original expansion was okay. Factored correctly leading to $x = 3y$ or $x = -y$. Since $x, y > 0$, $x = -y$ impossible. Hence unique soln. From substitution got $y=9$ etc., confirmed everything works. Still seems solid.

Figure 12: A CoT Example from Qwen3-32B applied NoWait on AMC 2023.

No Wait CoT(continue):

ALTERNATIVE APPROACH: Perhaps express variable substitutions using exponents based on relation between them.

From $y^3 = x^2$. We might want to define x in terms of y or vice versa. For example, taking $\sqrt{y^3} = |x|$, which because $x > 0$ becomes $\sqrt{y^3}$, hence $x = y^{3/2}$. Similarly, or expressing y in terms of x .

Suppose instead of doing that, though, perhaps my previous approach suffices already—but anyway, trying an alternate route may help cross-validate.

If $x = y^{3/2}$, plug into equation (2): $(y - x)^2 = 4y^2 \Rightarrow$ Take square roots?

Well, more straightforwardly plugging in expression for x : $(y - y^{3/2})^2 = 4y^2$.

Take square root on both sides? Well, note that squaring often introduces extraneous solutions, so care needed. Although earlier method led directly via algebraic manipulation. Still, proceeding differently—

Expand same way: Left-hand-side expands to $y^2 - 2*y*y^{3/2} + y^3$ equals $4y^2$?

Which would be $y^2 - 2 y^{5/2} + y^3 = 4y^2$. Rearranged: $y^3 - 2 y^{5/2} + y^2 - 4y^2 = 0 \Rightarrow y^3 - 2 y^{5/2} - 3y^2 = 0$

Factor out y^2 : $y^2 [y - 2 y^{1/2} - 3] = 0$ Because $y^2 \neq 0$ since $y > 0$, set inside bracket to zero: $y - 2\sqrt{y} - 3 = 0$.

This is equivalent equation after substituting. Solving this equation for $\sqrt{y}$. Let's set $z = \sqrt{y}$. Note that $z > 0$.

Then since $y = z^2$, rewrite above equation: $z^2 - 2z - 3 = 0$. Quadratic equation: $z^2 - 2z - 3 = 0$. Solve using quadratic formula:

Discriminant $D = b^2 - 4ac = 4 + 12 = 16$.

Roots $z = [2 \pm \sqrt{16}]/2 = [2 \pm 4]/2$. Which is $(2 + 4)/2 = 6/2 = 3$, and $(2 - 4)/2 = (-2)/2 = -1$.

Since $z = \sqrt{y} \geq 0$, reject negative root $z = -1$. So $z = 3$. Then $\sqrt{y} = 3$ implies $y = (3)^2 = 9$. As previously found. Therefore $x = y^{3/2} = 9^{3/2}$. Square root of 9 is 3, cubed is 27.

Same result. Therefore confirms our prior conclusion.

Alternate path leads same place.

So yes, consistent results across methods. Seems very confident about $x=27$, $y=9$, thus $x + y = 36$.

I don't see anything wrong with reasoning steps. Both approaches lead to the same solution and verified against initial conditions. Probably safe to conclude the answer is 36.

****Final Answer****

$\boxed{36}$

</think>

Figure 13: A CoT Example from Qwen3-32B applied NoWait on AMC 2023.