Reasoning Models Can Be Effective Without Thinking - Ma et. al.

Keno Bürger
Scientific Seminar Machine Learning
11 July 2025

Associate Professorship of Machine Learning
Technical University of Munich

Motivation

- Reasoning models have significantly improved the performance of LLM systems
- Include an explicit, lengthy Thinking process at the cost of increased token usage and latency
- Paper questions the necessity of this Thinking process
- Ongoing research to improve efficiency and effectiveness of reasoning models
- Investigates the effectiveness of reasoning models without the Thinking process

Background Knowledge: Reasoning Models

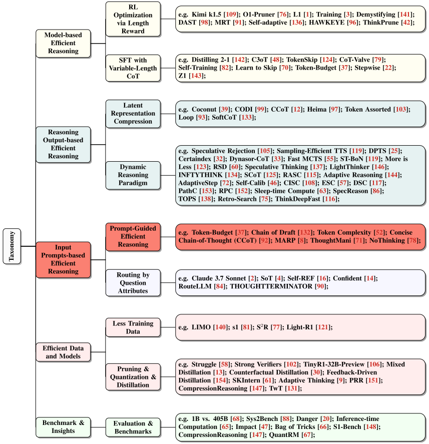
- Training Reasoning Models:
    - Chain-of-Thought (CoT) Fine-Tuning: Fine-tune models on datasets with explicit reasoning steps to improve interpretability and accuracy
    - Instruction Tuning: Train models on task-specific instructions to guide reasoning for complex tasks
    - Reinforcement Learning (RL): Use RL with human feedback (RLHF) to reward correct and interpretable reasoning steps
- Key Difference from non-Reasoning Models:
    - Reasoning models are explicitly trained to generate intermediate steps, while non-reasoning models rely on implicit patterns in the data

Background Knowledge: Efficiency in LLMs

- Efficiency Trade-offs:
  - Explicit reasoning improves accuracy but increases latency and token usage
  - Token-constrained settings require balancing reasoning depth and efficiency
- Implicit Reasoning:
  - Models generate answers without explicit reasoning steps
- Scaling Approaches:
  - Sequential Scaling: Processes reasoning steps one at a time, increasing latency but ensuring step-by-step accuracy
  - Parallel Scaling: Executes multiple reasoning paths simultaneously, reducing latency but requiring more computational resources

Related Work

- Contrasting prior work, this paper demonstrates competetive performance by disabling the Thinking process
- Techniques like CoT and instruction tuning to elicit reasoning abilities in LLMs [Huang+ 2023; Ott+ 2023; Xu+ 2025]
- Ongoing debate about whether LLMs truly reason or simply mimic reasoning patterns from training data [Hochlehnert+ 2025; Xu+ 2025]
  → necessity of explicit reasoning steps?



Taxonomy

Model-based Efficient Reasoning

RL Optimization via Length Reward

e.g. Kimi k1.5 [109]; O1-Pruner [76]; L1 [1]; Training [3]; Demystifying [141]; DAST [98]; MRT [91]; Self-adaptive [136]; HAWKEYE [96]; ThinkPrune [42];

SFT with Variable-Length CoT

e.g. Distilling 2-1 [142]; C3oT [48]; TokenSkip [124]; CoT-Valve [79]; Self-Training [82]; Learn to Skip [70]; Token-Budget [37]; Stepwise [22]; Z1 [143];

Reasoning Output-based Efficient Reasoning

Latent Representation Compression

e.g. Coconut [39]; CODI [99]; CCoT [12]; Heima [97]; Token Assorted [103]; Loop [93]; SoftCoT [133];

Dynamic Reasoning Paradigm

e.g. Speculative Rejection [105]; Sampling-Efficient TTS [119]; DPTS [25]; Certaindex [32]; Dynasor-CoT [33]; Fast MCTS [55]; ST-BoN [119]; More is Less [123]; RSD [60]; Speculative Thinking [137]; LightThinker [146]; INFTYTHINK [134]; SCoT [125]; RASC [115]; Adaptive Reasoning [144]; AdaptiveStep [72]; Self-Calib [46]; CISC [108]; ESC [57]; DSC [117]; PathC [153]; RPC [152]; Sleep-time Compute [63]; SpecReason [86]; TOPS [138]; Retro-Search [75]; ThinkDeepFast [116];

Input Prompts-based Efficient Reasoning

Prompt-Guided Efficient Reasoning

e.g. Token-Budget [37]; Chain of Draft [132]; Token Complexity [52]; Concise Chain-of-Thought (CCoT) [92]; MARP [8]; ThoughtMani [71]; NoThinking [78];

Routing by Question Attributes

e.g. Claude 3.7 Sonnet [2]; SoT [4]; Self-REF [16]; Confident [14]; RouteLLM [84]; THOUGHTTERMINATOR [90];

Efficient Data and Models

Less Training Data

e.g. LIMO [140]; s1 [81]; S^{2}R [77]; Light-R1 [121];

Pruning & Quantization & Distillation

e.g. Struggle [58]; Strong Verifiers [102]; TinyR1-32B-Preview [106]; Mixed Distillation [13]; Counterfactual Distillation [30]; Feedback-Driven Distillation [154]; SKIntern [61]; Adaptive Thinking [9]; PRR [151]; CompressionReasoning [147]; TwT [131];

Benchmark & Insights

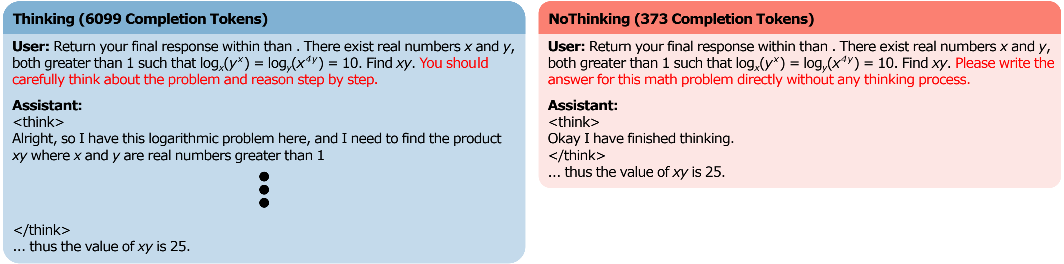
Evaluation & Benchmarks

e.g. 1B vs. 405B [68]; Sys2Bench [88]; Danger [20]; Inference-time Computation [65]; Impact [47]; Bag of Tricks [66]; S1-Bench [148]; CompressionReasoning [147]; QuantRM [67];

Main Contribution

- Identification of NoThinking's **low-budget superiority**, particularly in token-constrained settings
- Development of a **parallel scaling framework** combining NoThinking with best-of-N sampling, achieving **7× lower latency** and **4× less token usage** compared to sequential Thinking approaches
- comparable results can be achieved through **implicit reasoning**

Methods and Experimental Setup



Thinking (6099 Completion Tokens)

User: Return your final response within than . There exist real numbers x and y, both greater than 1 such that $\log_x(y^x) = \log_y(x^{4y}) = 10$. Find xy. You should carefully think about the problem and reason step by step.

Assistant:
<think>
Alright, so I have this logarithmic problem here, and I need to find the product xy where x and y are real numbers greater than 1

⋮

</think>
... thus the value of xy is 25.

NoThinking (373 Completion Tokens)

User: Return your final response within than . There exist real numbers x and y, both greater than 1 such that $\log_x(y^x) = \log_y(x^{4y}) = 10$. Find xy. Please write the answer for this math problem directly without any thinking process.

Assistant:
<think>
Okay I have finished thinking.
</think>
... thus the value of xy is 25.

- DeepSeek-R1-Distill-Qwen-32B (additional baseline: Qwen-32B-Instruct)
- Budget forcing technique [Muennighoff+ 2025]: average token usage from NoThinking applied to Thinking → at this limit model is prompeted to produce a final answer

DeepSeek-R1-Distill-Qwen-32B

- First-generation reasoning model from DeepSeek
- Distilled from DeepSeek-R1 based on Qwen2.5-32B architecture
- Created through sophisticated knowledge distillation process
- Two RL stages in teacher model to discover improved reasoning patterns
- Two SFT stages to improve reasoning and non-reasoning capabilities

Benchmarks

Use Case	Benchmark	Details
Mathematical Problem Solving	AIME 2024/AIME 2025	■ 15 questions, 3 hours ■ Answers are integers from 0 to 999
	AMC 2023	■ 25 multiple choice questions, 40 minutes ■ One correct answer out of 5 choices
	Olympiad Bench	■ Advanced reasoning benchmark containing math and physics problems ■ 8476 problems with an sophisticated automated scoring pipeline to verify solutions
Coding	LiveCodeBench	■ Contamination-free evaluation of coding abilities ■ Code execution framework evaluates generated programs against test case
Formal Theorem Proving	MiniF2F	■ Statements from olympiads (AMC, AIME, IMO) and high-school/undergraduate math classes ■ Formal system proof checkers
	ProofNet	■ Logic and theorem proving benchmark ■ 371 examples with formal theorem statements, natural language theorem statements, and proofs.

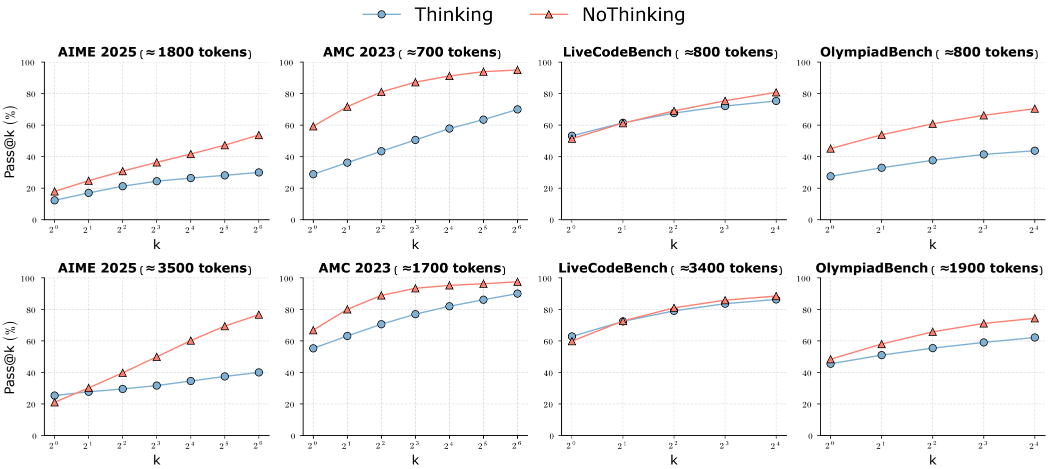
Metrics

- **pass@k** measuring the probability of obtaining at least one correct output
  - k randomly selected samples, n generated completions, c correct outputs
  - pass$@k = \mathbb{E}_{\text{problems}}\left[1 - \frac{\binom{n-c}{k}}{\binom{n}{k}}\right]$
- **Mean and standard deviation** of entropy
  - entropy: measure of uncertainty in a probability distribution (model prediction)
  - higher mean entropy (close to the theoretical maximum): high uncertainty, greater diversity across questions
  - low variance (small compared to the the mean): more consistent diversity

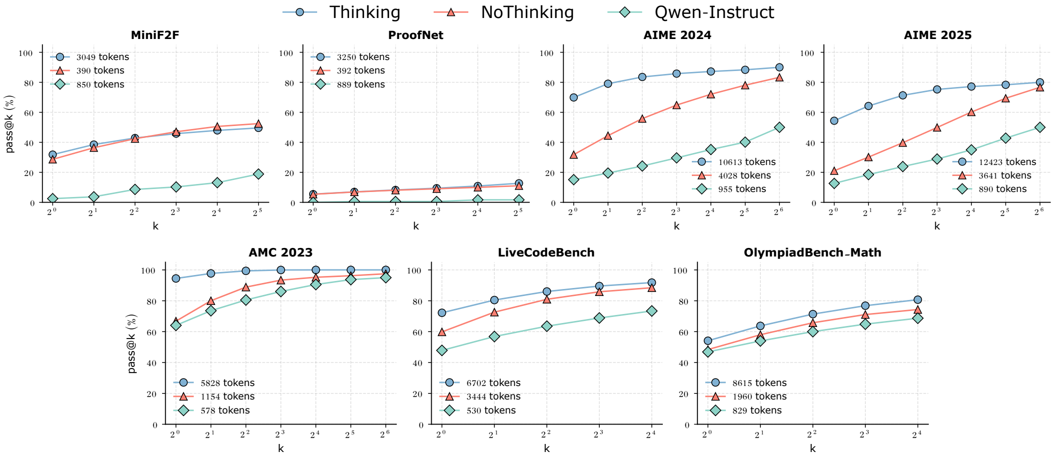
Scaling

- **Parallel Scaling**: Distributes computations across multiple devices to process tasks simultaneously
- **Sequential Scaling**: Executes computations one after another
- **Metrics**: Latency, maximum number of tokens.
- **Perfect verifiers not available**:
  - *Confidence-based*: Model's self-certainty (confidence score) in predictions. Methods include highest confidence selection and weighted voting across answers.
  - *Majority voting*: Selection of the most common response (consensus)

Token Efficiency



Accuracy



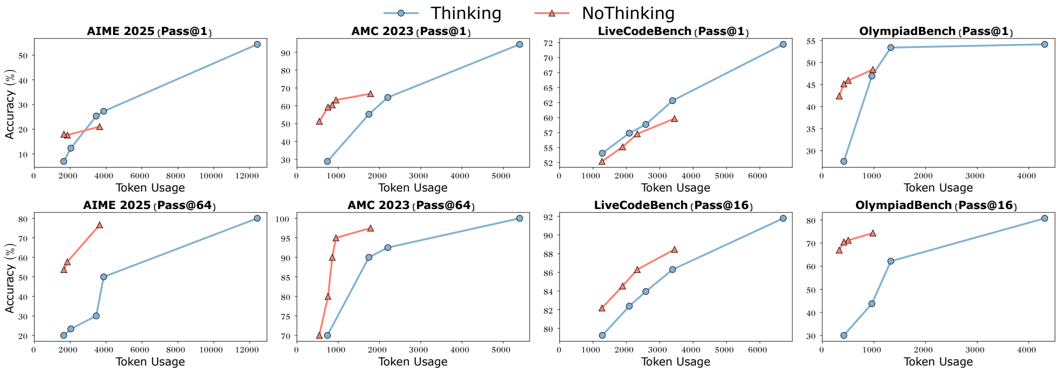
Thinking
NoThinking
Qwen-Instruct
MiniF2F
3049 tokens
390 tokens
850 tokens
ProofNet
3250 tokens
392 tokens
889 tokens
AIME 2024
10613 tokens
4028 tokens
955 tokens
AIME 2025
12423 tokens
3641 tokens
890 tokens
AMC 2023
5828 tokens
1154 tokens
578 tokens
LiveCodeBench
6702 tokens
3444 tokens
530 tokens
OlympiadBench_Math
8615 tokens
1960 tokens
829 tokens
pass@k (%)
k
0
20
40
60
80
100
2^0
2^1
2^2
2^3
2^4
2^5
2^6

Parallel Scaling

Task	**Thinking**	**BF (tokens)**	**Pass@K**	**Selection Methods (Pass@1)**		
				Majority Voting	**Confidence + Highest**	**Confidence + Voting**
AIME 2024	*Thinking*	3500	73.33	43.33	40.00	**46.67**
	NoThinking	3500	77.30	46.67	20.00	**50.00**
AIME 2025	*Thinking*	3500	40.00	**30.00**	**30.00**	**30.00**
	NoThinking	3500	53.73	**33.33**	20.00	**33.33**
AMC 2023	*Thinking*	2400	92.50	**77.50**	65.00	**77.50**
	NoThinking	2400	95.00	77.50	57.50	**85.00**

- Improved pass@1 results at similar or lower latency
- Reduced token usage for tasks with perfect verifiers

Analysis



- In low-budgt settings, NoThinking outperforms Thinking
- Parallel scaling → improved accuracy-latency tradeoff

Discussion

Strengths	Limitations
■ Broad evaluation ■ Analysis of different scaled models ■ Efficiency gains ■ Practical Applications with parallel scaling	■ Poor performance at pass@1 and Coding ■ Results limited to one model ■ No interpretive/subjective/humaties-style tasks ■ No ablation studies on different components ■ Raw performance still favors large models

⇒ Peer review and test on general conversation/common sense reasoning tasks

⇒ Applications with strict latency or cost constraints

References I

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu and Maosong Sun.
OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems. 2024. eprint: 2402.14008.

Andreas Hochlehnert, Hardik Bhatnagar, Vishaal Udandarao, Samuel Albanie, Ameya Prabhu and Matthias Bethge.
A Sober Look at Progress in Language Model Reasoning: Pitfalls and Paths to Reproducibility. Apr. 2025. DOI: 10.48550/arXiv.2504.07086.

Jie Huang and Kevin Chen-Chuan Chang. **Towards Reasoning in Large Language Models: A Survey**. May 2023. DOI: 10.48550/arXiv.2212.10403.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen and Ion Stoica.
LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code. June 2024. DOI: 10.48550/arXiv.2403.07974.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès and Tatsunori Hashimoto. **s1: Simple test-time scaling**. Mar. 2025. DOI: 10.48550/arXiv.2501.19393.

Simon Ott, Konstantin Hebenstreit, Valentin Liévin, Christoffer Egeberg Hother, Milad Moradi, Maximilian Mayrhauser, Robert Praas, Ole Winther and Matthias Samwald. **ThoughtSource: A central hub for large language model reasoning data**. In: *Scientific Data* 10.1 (Aug. 2023). ISSN: 2052-4463. DOI: 10.1038/s41597-023-02433-3.

References II

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Chenyang Shao, Yuwei Yan, Qinglong Yang, Yiwen Song, Sijian Ren, Xinyuan Hu, Yu Li, Jie Feng, Chen Gao and Yong Li. **Towards Large Reasoning Models: A Survey of Reinforced Reasoning with Large Language Models**. Jan. 2025. DOI: 10.48550/arXiv.2501.09686.

Kunhao Zheng, Jesse Michael Han and Stanislas Polu. **MiniF2F: a cross-system benchmark for formal Olympiad-level mathematics**. Feb. 2022. DOI: 10.48550/arXiv.2109.00110.