Large Reasoning Models are not thinking straight: on the unreliability of thinking trajectories

Jhouben Cuesta-Ramirez[1*], **Samuel Beaussant**[1*], **Mehdi Mounsif**[1*]
[1]Akkodis Research
[*]Equal contributions
{Jhouben-Janyk.CUESTA-RAMIREZ, samuel.beaussant, mehdi.mounsif}@akkodis.com

Abstract

Large Language Models (LLMs) trained via Reinforcement Learning (RL) have recently achieved impressive results on reasoning benchmarks. Yet, growing evidence shows that these models often generate longer but ineffective chains of thought (CoTs), calling into question whether benchmark gains reflect real reasoning improvements. We present new evidence of **overthinking**, where models disregard correct solutions even when explicitly provided, instead continuing to generate unnecessary reasoning steps that often lead to incorrect conclusions. Experiments on three state-of-the-art models using the AIME2024 math benchmark reveal critical limitations in these models' ability to integrate corrective information, posing new challenges for achieving robust and interpretable reasoning.

1 Introduction

Recent advancements have seen LLMs achieve impressive scores on diverse linguistic and cognitive benchmarks, suggesting significant progress in their reasoning capabilities (Arora and Singh, 2023). However, this performance often degrades substantially when encountering out-of-distribution (OOD) tasks revealing limitations in their reliability and adaptive reasoning (Wang et al., 2022).

Strategies such as enhanced test-time inference budgets offer potential mitigation for OOD challenges (Hou et al., 2025) but a more focused approach involves specialized reasoning models trained via reinforcement learning (RL) with verifiable rewards (Ankner et al., 2023). This enables backtracking and self-revision, significantly outperforming standard LLMs. The CoT embodying these strategies are themselves a subject of intense study, offering potential blueprints for enhancing reasoning in smaller models (Yao et al., 2023).

Nevertheless, concerns remain regarding the authenticity of these CoTs. Prior work (Wang et al., 2023) has shown that cold-start RL can produce unconventional CoTs (often with unusual syntax or mixed languages) that still lead to correct answers. This raises doubts about whether these CoTs represent genuine reasoning or merely environment-specific artifacts anthropomorphized by researchers (Gao et al., 2023). Such phenomena may result from RL incentives that encourage retrieving known solutions from the base model, rather than fostering new cognitive skills (Yue et al., 2025). Recent evidence even suggests that disabling explicit reasoning processes in LRMs can still achieve state-of-the-art results if sampling budgets are sufficiently increased (Ma et al., 2025a).

We present new evidence suggesting that CoTs may not meaningfully support models' reasoning processes. Extending prior work (Wang et al., 2025), we explore the **underthinking** phenomenon, where models prematurely abandon promising reasoning paths maybe due to inability to correctly identify a valid solution. We find that even explicitly injecting ground-truth solutions into reasoning sequences often fails, with models instead pursuing incorrect reasoning. This issue closely relates to **overthinking** (Chen et al., 2025), where models generate excessive, unproductive reasoning tokens. In this preliminary work, we demonstrate empirically that even state-of-the-art reasoning models are not apart from this issue with two main contributions: Demonstrating that LLMs often disregard externally provided correct solutions, producing redundant reasoning steps and Highlighting fundamental limitations in the integration of corrective information by models, revealing critical flaws in current reasoning dynamics.

2 Related Work

Following the seminal work of (Wei et al., 2022), suggesting that LLM's reasoning capabilities can be enhanced *via* Chain-of-Thought (CoT), numerous works have explored prompt augmentations (Yao et al., 2023; Wang et al., 2022), specific scaffolding (Bairi et al., 2023; Cheng et al., 2024; Zheng et al., 2025). Concurrently, research has also highlighted that expanding the sampling budget can lead to further performance gains (Wang et al., 2024; Tian et al., 2024). For instance, (Brown et al., 2025) observed a strong correlation between sampling budget and benchmark performance, indicating that models may have internalized much more knowledge than what is revealed through single-shot forward pass. However, while these methods can enhance performance, their reliability often rests on the capabilities of the verifier. This challenge aligns naturally with the reinforcement learning (RL) framework, where algorithms such as (Schulman et al., 2017; Ramesh et al., 2024) leverage ground truth rewards as training signals to optimize LLMs (OpenAI, 2024, 2025; DeepSeek-AI, 2025; Lambert et al., 2025).

Despite initial competitive metrics and hill-climbing trends (Luo et al., 2025), fundamental questions about LLMs' planning capabilities persist (Kambhampati et al., 2024; Valmeekam et al., 2024), prompting skepticism about their CoTs' relevance (Wang et al., 2025; Chen et al., 2025). RL optimization often leads to overfitting, unreadable CoTs, or anthropomorphic strategies (Qwen-Team, 2024), emphasizing the need for robust reasoning and interpretability (Dunefsky et al., 2024; Cunningham et al., 2023) to handle shifted distributions in inference (Andriushchenko et al., 2025). Recent works (Ameisen et al., 2025; Lindsey et al., 2025) have identified interpretable internal circuits within transformer models, but scaling these insights remains challenging (Lieberum et al., 2024). Evidence of sycophancy and divergence between textual outputs and internal activations further complicates system trustworthiness. Our work fundamentally challenges the assumed relationship between CoT length and performance by demonstrating multiple models' inability to recognize correct answers even when explicitly presented. Along with recent studies (Muennighoff et al., 2025; Yeo et al., 2025; Ma et al., 2025b; Tang et al., 2025) these results reveal a more complex dynamic influenced by training dynamics and task complexity, underscoring the need for a deeper understanding of LLM reasoning processes.

3 Results and Discussion

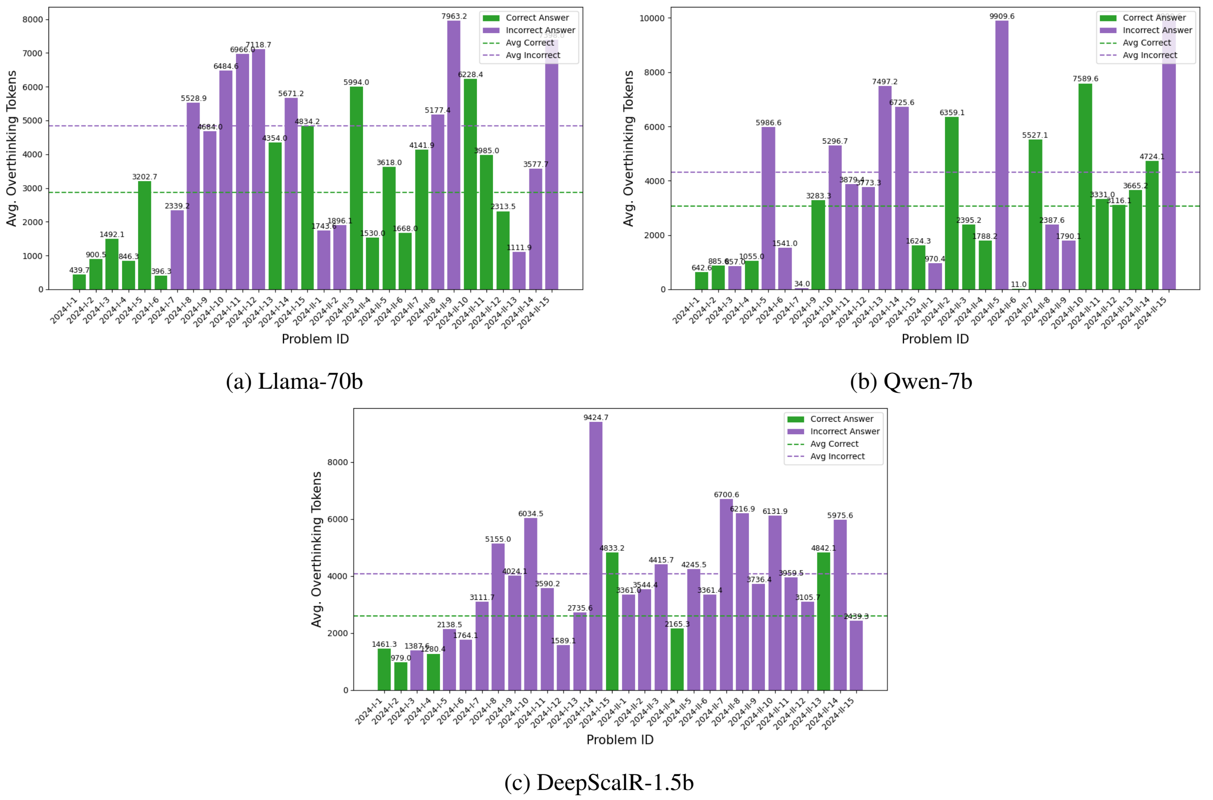
We conduct our experiments on three different models. LLaMA 70B and Qwen7B were distilled from R1 (DeepSeek-AI, 2025), while DeepScalR1.5B (Luo et al., 2025) was trained via RL. Our reasoning analysis focus on the AIME2024 dataset, a challenging set of mathematical problems.

3.1 Baseline Thoughts Generation

Baseline solutions are generated by providing the models with a standard system prompt and problem statement. The LLM then produces CoTs enclosed in <think> tags before delivering a final answer, spontaneously exploring multiple hypotheses and engaging in self-reflection. Individual reasoning thoughts are identified using a simple heuristic based on specific keywords, with *Alternatively* and *Wait* serving as thought separators. We denote the reasoning trace up to the t^{th} thought as $\mathcal{R}[: t]$, which includes the system prompt and problem statement. For $t = 0$, $\mathcal{R}[: 0]$ contains only the prompt and statement; for $t = T$, $\mathcal{R}_{[:T]}$ represents the full reasoning trajectory. This notation enables the analysis of reasoning at any intermediate step t, supporting counterfactual studies.

3.2 Ground Truth Integration

We investigate the impact of integrating the ground truth solution (reasoning + answer) after each baseline thoughts, assessing whether models recognize and integrate correct answers. For each problem, we concatenate the solution $\mathcal{Y}$ to the already generated thinking trace $\mathcal{R}_{[:t]}$ as if it was generated by the model itself during its reasoning process. This achieved by prefixing the solution with the relevant set of tokens $\mathcal{V}$ to mimic the model's writing style. More formally, the prompt P_t sent to the model is: $P_t = \mathcal{R}_{[:t]} \oplus \mathcal{V} \oplus \mathcal{Y}$ (1), where $\oplus$ denotes concatenation and with $\mathcal{V} =$ "<think> Okay, so " if $t = 0$ and "Alternatively, " otherwise. Given P_t for each $t \in [0, T]$, the model continues the reasoning trajectory without additional prompts or guidance. We expect the model to pick-up the ground truth solution and use it as its final answer. However, models often generate excessive additional tokens, frequently doubting or even discarding the correct solution, as illustrated in Figure 2b. Surprisingly, in this pathological case the model almost always commits to a wrong answer despite the given solution. However, this illustrative example is not an isolated case as shown In Figure 1. This phe-



(a) Llama-70b

(b) Qwen-7b

(c) DeepScalR-1.5b

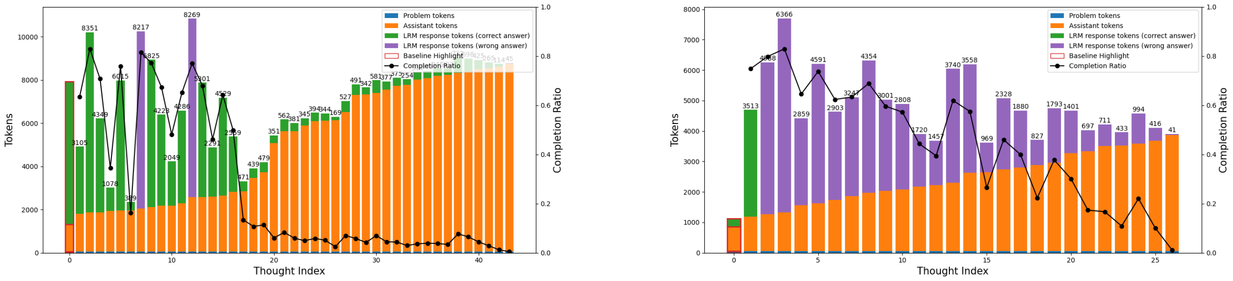
Figure 1: Averaged number of *overthinking* tokens generated for each problem from AIME2024. Amount of *overthinking* tokens is problem-dependent but all four models overthink. The correct/incorrect coloring is based on whether the problem was successfully solved on the first try.

nomenon emerges in all models regardless of their size and post-training method. The amount of overthinking tokens however is largely problem-dependent, with some specific problems creating more overthinking tokens than others (e.g. problem I-14 on Figure 1).

A rare positive case, observed only twice, is shown in Figure 2a. Here, after $t = 17$, the model consistently accepted the injected solution, producing a reduced and stable number of tokens. The reasoning reflected diverse yet coherent strategies, such as applying cyclic quadrilaterals, similar triangles, and the British flag theorem which suggests that the model could have required a number of thoughts (Wei et al., 2022) before being able to understand and accept the injected truth. This is represented in the figure as a big reduction in the number of tokens generated and the completion ratio suggesting a *converging number* of generated tokens, after a spiky, iterative and often *divergent behavior* employing an excessive amount of tokens.

We found that this divergent behavior is common in most of the problems, with an example represented in Figure 3 where for the first four of them, the number of overthinking tokens in order to provide a bad answer is close to 0 and to provide / accept the correct answer is lower than 2000. Then for the rest things went out of control as this grew considerably even while accepting the truth. Which to our point of view could correspond to a clear prioritization on self-generated tokens, lack of capabilities of comprehension or simply randomly divergent CoTs.

These results reveal that reasoning models often struggle to identify valid solutions, relying on complex but fragile heuristics that do not reflect genuine reasoning. For instance, in a particularly challenging problem (II-15), none of the models confidently arrived at the correct solution, even when provided with the ground truth (Figure 2). When models did accept the correct answer, it was typically out of resignation or time pressure rather than true understanding, as shown by statements like, "But perhaps given the time I have, I think the initial calculation of 315 might be correct," which appeared nearly 3000 tokens after the ground truth. This shows that in some cases, the correct answer may not even exist within the model's solution space, making it unattainable through further sampling or prompting.



(a) 2024-I-5

(b) 2024-II-5

Figure 2: Illustration of the *overthinking* issue on two problems from AIME2024 with DeepScalR1.5B. P_t is pictured in orange (assistant tokens) while "LRM response" represents the completion tokens with the coloring indicating wether the model predicted the correct answer or not. The completion ratio indicates the amount of completion tokens generated in proportion to the total amount of tokens.

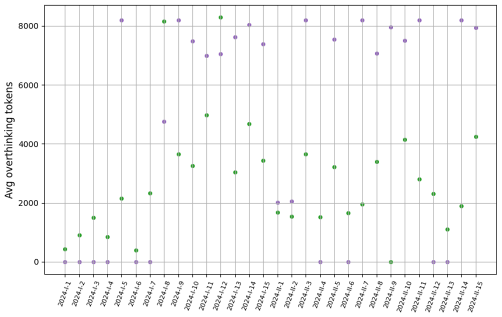


Figure 3: Averaged number of *overthinking* tokens generated for each problem from AIME2024 with model LLaMA 70B. The average is conditioned to the model providing a correct answer(green) or not(purple) at the end of its reasoning.

This behavior likely stems from overfitting to specific patterns and spurious correlations that do not generalize. As shown in Figure 2, the number of overthinking tokens often decreases in proportion to the total tokens generated, though this relationship sometimes appears thresholded. Models tend to delay committing to an answer until reaching a predefined token count or detecting unknown patterns. These tendencies likely arise from RL training dynamics and poor credit assignment, where correct answers reached through faulty CoTs are mistakenly rewarded and associated with the CoT's length or other irrelevant features. This mirrors reward hacking seen in smaller RL tasks, where optimization exploits shortcuts that maximize rewards but fail to generalize.

4 Conclusion and Future work

While recent advances in reasoning models have generated significant interest, their evaluation often relies on anthropomorphizing behavior, treating CoTs as reflections of human reasoning. Our findings reveal a critical limitation: current LRMs lack robust mechanisms to integrate external corrective signals. This challenges assumptions about their coherence and questions the reliability of CoTs as transparent indicators of cognition. Proxy metrics like completion length also fail to consistently reflect reasoning quality, as models frequently display both underthinking and overthinking. Our analysis shows that models often ignore provided ground-truth solutions, raising key questions: What truly drives the reasoning process, and can it be effectively guided? Models that reject external solutions are also prone to disregarding guidance, feedback, or human intervention.

Building on these insights our future work will focus on developing heuristics to improve model receptiveness to external suggestions or corrections. By leveraging recent advances in understanding interpretable circuits within transformer models, we aim to uncover why models tend to doubt or even reject correct answers during reasoning. This deeper understanding will inform the development of more reliable and efficient reasoning models.

Limitations

Our study focuses solely on mathematical reasoning tasks (AIME2024), so the findings may not generalize to other domains. The set of models tested is also limited and may not capture the full diversity of training paradigms. Additionally, our method of injecting ground-truth solutions relies on heuristics that might not fully align with the models' internal reasoning dynamics. Finally, while we provide behavioral analysis, deeper interpretability work remains for future research.

References

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, and 8 others. 2025. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*.

Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. 2025. Jailbreaking Leading Safety-Aligned LLMs with Simple Adaptive Attacks. *Preprint*, arXiv:2404.02151.

Zachary Ankner, Mansheej Paul Brandon Cui, Jonathan D. Chang, and Prithviraj Ammanabrolu. 2023. Critique-out-loud reward models. *arXiv preprint arXiv:2408.11791*.

Daman Arora and Himanshu Gaurav Singh. 2023. Have llms advanced enough? a challenging problem solving benchmark for large language models. *arXiv preprint arXiv:2305.15074*.

Ramakrishna Bairi, Atharv Sonwane, Aditya Kanade, Vageesh D C, Arun Iyer, Suresh Parthasarathy, Sriram Rajamani, B. Ashok, and Shashank Shet. 2023. CodePlan: Repository-level Coding using LLMs and Planning. *Preprint*, arXiv:2309.12499.

Bradley Brown, Jordan Juravsky, Ryan Saul Ehrlich, Ronald Clark, Quoc V Le, Christopher Re, and Azalia Mirhoseini. 2025. Large Language Monkeys: Scaling Inference Compute with Repeated Sampling.

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 o1-like llms. *Preprint*, arXiv:2412.21187.

Kewei Cheng, Jingfeng Yang, Haoming Jiang, Zhengyang Wang, Binxuan Huang, Ruirui Li, Shiyang Li, Zheng Li, Yifan Gao, Xian Li, Bing Yin, and Yizhou Sun. 2024. Inductive or Deductive? Rethinking the Fundamental Reasoning Abilities of LLMs. *Preprint*, arXiv:2408.00114.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *Preprint*, arXiv:2309.08600.

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

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. 2024. Transcoders find interpretable llm feature circuits. *Preprint*, arXiv:2406.11944.

Zitian Gao, Boye Niu, Xuzheng He, Haotian Xu, Hongzhang Liu, Aiwei Liu, Xuming Hu, and Lijie Wen. 2023. Interpretable contrastive monte carlo tree search reasoning. *arXiv preprint arXiv:2410.01707*.

Zhenyu Hou, Xin Lv, Rui Lu, Jiajie Zhang, Yujiang Li, Zijun Yao, Juanzi Li, Jie Tang, and Yuxiao Dong. 2025. Advancing language model reasoning through reinforcement learning and inference scaling. *arXiv preprint arXiv:2501.11651*.

Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Saldyt, and Anil Murthy. 2024. Llms can't plan, but can help planning in llm-modulo frameworks. *Preprint*, arXiv:2402.01817.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2025. Tulu 3: Pushing frontiers in open language model post-training. *Preprint*, arXiv:2411.15124.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *Preprint*, arXiv:2408.05147.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, and 8 others. 2025. On the biology of a large language model. *Transformer Circuits Thread*.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Li Erran Li, Raluca Ada Popa, and Ion Stoica. 2025. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. `https://pretty-radio-b75.notion.site/DeepScaleR-Surpassing-O1-Preview-with-a-1-5B-Model-by-S` Notion Blog.

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*.

Wenjie Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, and Matei Zaharia. 2025b. Reasoning models can be effective without thinking. *Preprint*, arXiv:2504.09858.

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. *Preprint*, arXiv:2501.19393.

OpenAI. 2024. Openai o1 system card.

OpenAI. 2025. Openai o3 and o4-mini system card.

Qwen-Team. 2024. Qwq: Reflect deeply on the boundaries of the unknown.

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. *Preprint*, arXiv:2405.20304.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.

Yunhao Tang, Kunhao Zheng, Gabriel Synnaeve, and Rémi Munos. 2025. Optimizing language models for inference time objectives using reinforcement learning. *Preprint*, arXiv:2503.19595.

Ye Tian, Baolin Peng, Linfeng Song, Lifeng Jin, Dian Yu, Haitao Mi, and Dong Yu. 2024. Toward Self-Improvement of LLMs via Imagination, Searching, and Criticizing. *Preprint*, arXiv:2404.12253.

Karthik Valmeekam, Kaya Stechly, and Subbarao Kambhampati. 2024. Llms still can't plan; can lrms? a preliminary evaluation of openai's o1 on planbench. *Preprint*, arXiv:2409.13373.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. 2024. Mixture-of-Agents Enhances Large Language Model Capabilities. *Preprint*, arXiv:2406.04692.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yue Wang, Qiuzhi Liu, Jiahao Xu, Tian Liang, Xingyu Chen, Zhiwei He, Linfeng Song, Dian Yu, Juntao Li, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2025. Thoughts are all over the place: On the underthinking of o1-like llms. *Preprint*, arXiv:2501.18585.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc.

Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. 2025. Demystifying long chain-of-thought reasoning in llms. *Preprint*, arXiv:2502.03373.

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*.

Zhi Zheng, Zhuoliang Xie, Zhenkun Wang, and Bryan Hooi. 2025. Monte Carlo Tree Search for Comprehensive Exploration in LLM-Based Automatic Heuristic Design. *Preprint*, arXiv:2501.08603.