Stop Anthropomorphizing Intermediate Tokens as Reasoning/Thinking Traces!

Subbarao Kambhampati,[*] Kaya Stechly, Karthik Valmeekam, Lucas Saldyt, Siddhant Bhambri, Vardhan Palod, Atharva Gundawar, Soumya Rani Samineni, Durgesh Kalwar, Upasana Biswas

School of Computing & AI
Arizona State University

Abstract

Intermediate token generation (ITG), where a model produces output before the solution, has been proposed as a method to improve the performance of language models on reasoning tasks. These intermediate tokens have been called "reasoning traces" or even "thoughts" – implicitly anthropomorphizing the model, implying these tokens are interpretable and resemble steps a human might take when solving a challenging problem. In this position paper, we present evidence that this anthropomorphization isn't a harmless metaphor, and instead is quite dangerous – it confuses the nature of these models and how to use them effectively, and leads to questionable research.

1 Introduction

Recent advances in general planning and problem solving have been spearheaded by so-called "Long Chain-of-Thought" models, most notably DeepSeek's R1 [13]. These transformer-based large language model are further post-trained using iterative fine-tuning and reinforcement learning methods. Following the now-standard teacher-forced pre-training, instruction fine-tuning, and preference alignment stages, they undergo additional training on reasoning tasks resulting in a Large Reasoning Model (LRM): at each step, the model is presented with a question; it generates a sequence of intermediate tokens (colloquially or perhaps fancifully called a "Chain of Thought" or "reasoning trace"); and it ends it with a specially delimited answer sequence. After verification of this answer sequence by a formal system, the model's parameters are updated so that it is more likely to output sequences that end in correct answers and less likely to output those that end in incorrect answers with no guarantees of trace correctness.

While (typically) no direct optimization pressure is applied to the intermediate tokens [3, 43], empirically it has been observed that language models perform better on many domains if they output such tokens first [27, 38, 42, 15, 12, 13, 28, 25, 21]. While the fact of the performance increase is well-known, the reasons for it are less clear. Much of the previous work has framed intermediate tokens in wishful anthropomorphic terms, claiming that these models are "thinking" before outputting their answers [27, 11, 13, 39, 43, 6]. The traces are thus seen both as giving insights to the end users about the solution quality, and capturing the model's "thinking effort."

Interpretability, as used in the context of the intermediate tokens produced by LRMs, often confounds two very different notions: (1) mechanistic interpretability of why the tokens seem to help LRMs, and (2) interpretability of these tokens to the end user. The first, i.e., mechanistic interpretability of

[*]Corresponding author: rao@asu.edu

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: CogInterp: Interpreting Cognition in Deep Learning Models.

why the tokens seem to help LRMs is reasonable, but these studies don't even have to be limited to linguistic intermediate tokens with the sole focus of improving model performance. On the other hand, the first begs the question *"Intermediate token generation helps language models, but must they help end users?"*

In this paper, we take the position that anthropomorphizing intermediate tokens as reasoning/thinking traces is (1) wishful (2) has little concrete supporting evidence (3) engenders false confidence and (4) may be pushing the community into fruitless research directions. This position is supported by work questioning the interpretation of intermediate tokens as reasoning/thinking traces [32, 4, 30] (Section 3) and by stronger alternate explanations for their effectiveness [36, 14] (Section 4).

Anthropomorphization has long been a contentious issue in AI research [23], and LLMs have certainly increased our anthropomorphization tendencies [16]. While some forms of anthropomorphization can be treated rather indulgently as harmless and metaphorical, our view is that viewing ITG as reasoning/thinking is more serious and may give a false sense of model capability and correctness.

LRMs have been built on insights from two broad but largely orthogonal classes of ideas:
(i) **test-time inference** scaling techniques, which involve getting LLMs to do more work than simply providing the most likely direct answer [37, 40, 1, 41, 18, 17]; and (ii) **post-training methods**, which complement simple auto-regressive training on web corpora, with additional training on intermediate tokens [44, 10, 9].

To be clear, our focus here is on the anthropomorphization of unfiltered intermediate tokens rather than the post-facto rationalizations generated by models such as OpenAI o1 or the more recent gpt-oss, which gives summaries after hiding the intermediate tokens. It is well known that for humans at least, such post-facto exercises are meant to teach/convince the listener, and may not shed much meaningful light on the thinking that went in [26]. Now, we will first highlight the downsides of anthropomorphizing intermediate tokens in LRMs in the following section.

2 Consequences of Anthropomorphizing Intermediate Tokens

We list the various (unhealthy) ramifications of this anthropomorphization below:

1) Viewing intermediate tokens as reasoning/thinking traces has led to a drive to make them "interpretable" to humans-in-the-loop.[2] For example, DeepSeek [8] dabbled in training an RL-only model (R1-Zero) but released a final version (R1) that was trained with additional data and filtering steps specifically to reduce the model's default tendencies to produce intermediate token sequences that mix English and Chinese!

2) It has led to an implicit assumption that correctness/interpretability of the intermediate tokens has a strong correlation, or even causal connection, with the solution produced. This tendency is so pronounced that a major vendor's study showing that LRM's answers *are not always faithful* to their intermediate tokens was greeted with surprise [7].

3) Viewing intermediate tokens as traces of thinking/reasoning has naturally led to interpreting the *length* of the intermediate tokens as some sort of meaningful measure of problem [34, 35] difficulty/effort and techniques that increased the length of intermediate tokens were celebrated as "learning to reason" [8]. Simultaneously there were efforts to *shorten* intermediate traces produced and celebrate that as learning to reason efficiently [2].

4) There have been attempts to cast intermediate tokens as learning some "algorithm" that generated the training data. For example, the authors of SearchFormer [20] claim that their transformer learns to become "more optimal" than A* because it produces shorter intermediate token traces than A*'s derivational trace on the same problem.

These corollaries, in turn, have lead to research efforts, which, when viewed under the lens of our position, become questionable enterprises (as we shall discuss in the following sections).

[2]Never mind that interpretability mostly meant that the traces were in pseudo English.

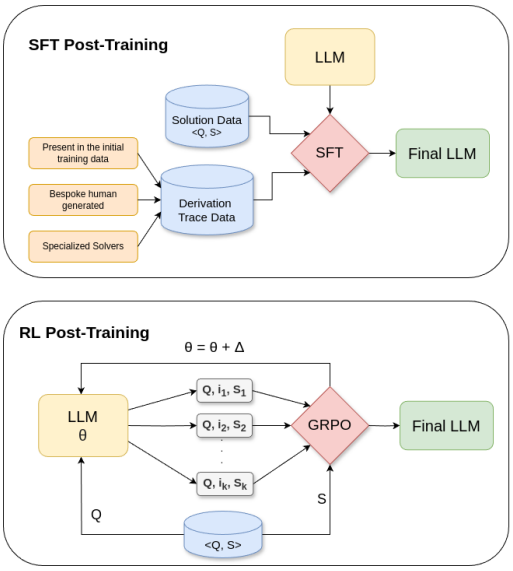

Figure 1: Post-training approaches for teasing out reasoning in LLMs.

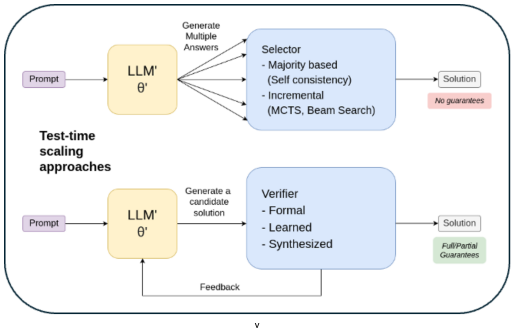

Figure 2: Test-time scaling approaches for teasing out reasoning

3 On the Amorphous Semantics of Intermediate Tokens

The fact that intermediate token sequences often reasonably look like better-formatted and spelled human scratch work – mumbling everything from *"hmm..."*, *"aha!"*, *"wait a minute"* to *"interesting"* along the way – doesn't tell us much about whether they are used for anywhere near the same purposes that humans use them for, let alone about whether they can be used as an interpretable window into what the LLM is "thinking," or as a reliable justification of the final answer. While a human may say "aha" to indicate exactly a sudden internal state change, this interpretation is unwarranted for models which do not have any such internal state, and which on the next forward pass will only differ from the pre-aha pass by the inclusion of that single token in their context. Interpreting the "aha" moment as meaningful exemplifies the long-neglected assumption about long CoT models – the false idea that derivational traces are semantically meaningful, either in resemblance to algorithm traces or to human reasoning. Further, there have also been works which attribute cognitive behaviors (like backtracking, self-verification etc.) to the models based on their reasoning traces and try to induce these kinds of behaviors through examples in the hope of improving the models' performance [11, 29].

One reason that this anthropomorphization continues unabated is because it is hard to either prove or disprove the correctness of these generated traces. DeepSeek's R1, even on very small and simple problems, will babble over 30 pages worth of text in response to each and every query, and it is far from clear how to check if these monologues constitute sound reasoning. Moreover, Deepseek R1 anthropomorphizes the increase in response length over the RL post training as test-time scaling/self reflection, but [30] shows this is an effect of length bias in the GRPO objective function that increase the response length for incorrect responses caused by the structural assumptions in the degenerate LLM-MDP framework for R1. While there have been some valiant efforts to make sense of these large-scale mumblings–e.g. [22]–the analyses here tend to be somewhat qualitative and suggestible reminiscent of "lines of code" analyses in software engineering. Unsurprisingly, few LRM evaluations assess pre-answer traces, focusing only on final answer correctness.

While evaluating intermediate tokens of general LRMs may be difficult, we *can* formally verify traces generated by format-constrained models trained to imitate domain-specific solvers. In [32], the authors challenge the assumption that intermediate tokens or "Chains of Thought" from models like DeepSeek's R1 are interpretable, semantically valid, and predictive of model behavior. Similarly, [5] investigates the correlation—and possible causation—between traces and final solution performance in QA tasks. Both studies find only a weak link between trace correctness and answer correctness [17]. Our ongoing work further shows that along with semantics, the cognitive interpretability of reasoning traces for end users can be an albatross from the perspective of LLM's task performance.

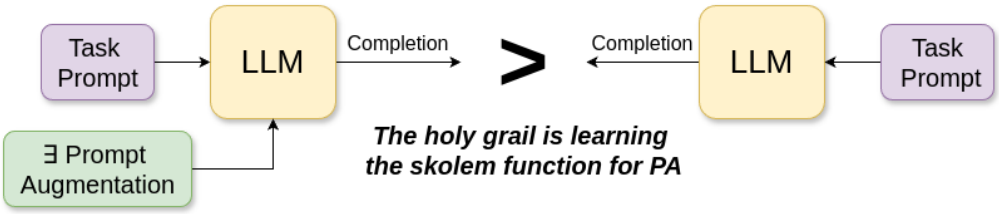

Figure 3: Augmenting a task prompt with additional tokens often seems to improve the accuracy of LLM completion even if the tokens don't have human-parseable meaning.

Given that these traces may not have any semantic import, deliberately making them *appear* more human-like is dangerous. In the end, LRMs are supposed to provide solutions that users don't already know (and which they may not even be capable of directly verifying). Engendering false confidence and trust by generating stylistically plausible ersatz reasoning traces seems ill-advised! After all, the last thing we want to do is to design powerful AI systems that potentially exploit the cognitive flaws of users to convince them of the validity of incorrect answers.

4 Understanding LRMs without Anthropomorphizing Intermediate Tokens

While the main focus of this position paper is to caution the community away from questionable explanations, rather than to provide complete explanations of the source of the power of LRMs we do present some plausible candidate explanations below:

1) Reasoning as Incremental Learning of Verifier Signal: Most documented advances of LRMs on reasoning problems have been on tasks for which there are formal verifiers from traditional AI and Computer Science. The *modus operandi* of current LRMs is leveraging these verifiers in a *generate-test* loop at test time, training time or distillation time in order to partially compile/internalize the verification signal into generation. In other words, post-training LRMs can be seen as iteratively compiling reasoning into retrieval via learning.

This general idea mirrors Marvin Minsky's insight that *intelligence is shifting the test part of generate-test into generation [24].* In particular, using verifiers at test time has already been advocated by the LLM-Modulo framework[18]. One way of seeing the training-, test-, and distillation-time verification is as a staged approach to compile the verification signal into an underlying LLM. This understanding is consistent with studies on the effectiveness of Chain of Thought[33], use of internal vs. external planning approaches for games[31], as well as self-improvement in transformers[19].

2) Embracing Reason-less Intermediate Tokens: One reasonable question about our position is *So what if the intermediate traces don't have semantics? We can just hide them from end user (like OpenAI o1/o3 do)* . We believe that a half-hearted lip service to human-legibility properties can not only engender false trust in the solutions (as already discussed), but also can become an albatross if our goal is increase task performance. This is already hinted by experiments in works such as [32, 4] that show that performance can improve when the model is trained on incorrect traces!

Reinforcement learning can potentially train LLMs to output any old intermediate token sequences – all that matters is that the bottom line improves. Indeed, we believe that de-anthropomorphization of intermediate tokens starts by acknowledging the common assumption across most "chain of thought" approaches: that an LLM will generate more accurate completions when provided with an appropriate *prompt augmentation* rather than just the base task prompt (see Figure 3). The big question then is how to get the right prompt augmentation. That is, given a task prompt T,

$$\exists PA s.t. Pr(Sol(LLM(T + PA), T)) > Pr(Sol(LLM(T), T)),$$

where PA is some appropriate prompt augmentation, $LLM(x)$ is the completion output by LLM given x as the prompt, and $Sol(y, T)$ checks, with the aid of a verifier, if y contains a solution for T.

The holy grail then is learning the Skolem function that supplies the right prompt augmentation that increases the probability of producing the correct answer in the succeeding tokens. The fact that we have an existential in the prompt augmentation inequality above means that in the most general case, the PA may be a function of both the task and the model. Note that *there is nothing here saying that PA must make any sense to the humans or be a correct trace of some algorithm.*

5 Summary

In this position paper, we argued against the prevalent tendency to anthropomorphize intermediate tokens as reasoning or "thinking". Anthropomorphization has been a part of AI research [23], and has significantly increased in the era of LLMs [16]. While some anthropomorphization has been harmless metaphors, we argued that viewing intermediate tokens as reasoning traces or "thinking" is actively harmful, because it engenders false trust and capability in these systems, and prevents researchers from understanding or improving how they actually work. We collated emerging evidence to support our position, and offered some more supported and balanced alternate ways of viewing LRM performance and the role of intermediate tokens. Our hope is that this position catalyzes the community towards more fruitful research directions to understand frontier models.

Acknowledgment

This research is supported in part by grants from ONR (N00014-25-1-2301 and N00014-23-1-2409), DARPA (HR00112520016), DoD RAI (via CMU subcontract 25-00306-SUB-000), an Amazon Research Award, and a generous gift from Qualcomm.

References

[1] Daman Arora and Subbarao Kambhampati. Learning and leveraging verifiers to improve planning capabilities of pre-trained language models. *ICML Workshop on Knowledge and Logical Reasoning in the Era of Data-driven Learning*, 2023.

[2] Daman Arora and Andrea Zanette. Training language models to reason efficiently, 2025. *URL https://arxiv. org/abs/2502.04463*, 2025.

[3] Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025.

[4] Siddhant Bhambri, Upasana Biswas, and Subbarao Kambhampati. Interpretable traces, unexpected outcomes: Investigating the disconnect in trace-based knowledge distillation, 2025.

[5] Siddhant Bhambri, Upasana Biswas, and Subbarao Kambhampati. Interpretable traces, unexpected outcomes: Investigating the disconnect in trace-based knowledge distillation. *arXiv preprint arXiv:2505.13792*, 2025.

[6] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[7] Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner Fabien Roger Vlad Mikulik, Sam Bowman, Jan Leike Jared Kaplan, et al. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025.

[8] DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, 2025.

[9] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature machine intelligence*, 5(3):220–235, 2023.

[10] Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*, 2020.

[11] Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*, 2025.

[12] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. *arXiv preprint arXiv:2306.08543*, 2023.

[13] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[14] Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space, 2024.

[15] Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*, 2023.

[16] Lujain Ibrahim and Myra Cheng. Thinking beyond the anthropomorphic paradigm benefits llm research, 2025.

[17] Subbarao Kambhampati. Can large language models reason and plan? *Annals of the New York Academy of Sciences*, 1534(1):15–18, 2024.

[18] Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Paul Saldyt, and Anil B Murthy. Position: LLMs can't plan, but can help planning in LLM-modulo frameworks. In *Forty-first International Conference on Machine Learning*, 2024.

[19] Nayoung Lee, Ziyang Cai, Avi Schwarzschild, Kangwook Lee, and Dimitris Papailiopoulos. Self-improving transformers overcome easy-to-hard and length generalization challenges, 2025.

[20] Lucas Lehnert, Sainbayar Sukhbaatar, DiJia Su, Qinqing Zheng, Paul Mcvay, Michael Rabbat, and Yuandong Tian. Beyond A*: Better Planning with Transformers via Search Dynamics Bootstrapping. In *Conference on Language Models (COLM)*, 2024.

[21] Dacheng Li, Shiyi Cao, Tyler Griggs, Shu Liu, Xiangxi Mo, Eric Tang, Sumanth Hegde, Kourosh Hakhamaneshi, Shishir G Patil, Matei Zaharia, et al. Llms can easily learn to reason from demonstrations structure, not content, is what matters! *arXiv preprint arXiv:2502.07374*, 2025.

[22] Sara Vera Marjanović, Arkil Patel, Vaibhav Adlakha, Milad Aghajohari, Parishad BehnamGhader, Mehar Bhatia, Aditi Khandelwal, Austin Kraft, Benno Krojer, Xing Han Lù, Nicholas Meade, Dongchan Shin, Amirhossein Kazemnejad, Gaurav Kamath, Marius Mosbach, Karolina Stańczak, and Siva Reddy. Deepseek-r1 thoughtology: Let's think about llm reasoning, 2025.

[23] Drew McDermott. Artificial intelligence meets natural stupidity. *SIGART Newsl.*, 57:4–9, 1976.

[24] Marvin Minsky. *Society of mind*. Simon and Schuster, 1986.

[25] 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. *arXiv preprint arXiv:2501.19393*, 2025.

[26] Richard E Nisbett and Timothy D Wilson. Telling more than we can know: Verbal reports on mental processes. *Psychological review*, 84(3):231, 1977.

[27] Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.

[28] Jacob Pfau, William Merrill, and Samuel R Bowman. Let's think dot by dot: Hidden computation in transformer language models. *arXiv preprint arXiv:2404.15758*, 2024.

[29] Tian Qin, David Alvarez-Melis, Samy Jelassi, and Eran Malach. To backtrack or not to backtrack: When sequential search limits model reasoning. *arXiv preprint arXiv:2504.07052*, 2025.

[30] Soumya Rani Samineni, Durgesh Kalwar, Karthik Valmeekam, Kaya Stechly, and Subbarao Kambhampati. Rl in name only? analyzing the structural assumptions in rl post-training for llms, 2025.

[31] John Schultz, Jakub Adamek, Matej Jusup, Marc Lanctot, Michael Kaisers, Sarah Perrin, Daniel Hennes, Jeremy Shar, Cannada Lewis, Anian Ruoss, Tom Zahavy, Petar Veličković, Laurel Prince, Satinder Singh, Eric Malmi, and Nenad Tomašev. Mastering board games by external and internal planning with language models, 2024.

[32] Kaya Stechly, Karthik Valmeekam, Atharva Gundawar, Vardhan Palod, and Subbarao Kambhampati. Beyond semantics: The unreasonable effectiveness of reasonless intermediate tokens, 2025.

[33] Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. Chain of Thoughtlessness: An Analysis of CoT in Planning. In *Proc. NeurIPS*, 2024.

[34] DiJia Su, Sainbayar Sukhbaatar, Michael Rabbat, Yuandong Tian, and Qinqing Zheng. Dualformer: Controllable fast and slow thinking by learning with randomized reasoning traces. In *The Thirteenth International Conference on Learning Representations*, 2024.

[35] Jinyan Su, Jennifer Healey, Preslav Nakov, and Claire Cardie. Between underthinking and overthinking: An empirical study of reasoning length and correctness in llms, 2025.

[36] Karthik Valmeekam, Kaya Stechly, Atharva Gundawar, and Subbarao Kambhampati. A systematic evaluation of the planning and scheduling abilities of the reasoning model o1. *Transactions on Machine Learning Research*, 2025.

[37] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

[38] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[39] Shu Yang, Junchao Wu, Xin Chen, Yunze Xiao, Xinyi Yang, Derek F Wong, and Di Wang. Understanding aha moments: from external observations to internal mechanisms. *arXiv preprint arXiv:2504.02956*, 2025.

[40] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[41] Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction, 2024.

[42] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022.

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

[44] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.