
POSITION: STOP ANTHROPOMORPHIZING INTERMEDIATE TOKENS AS REASONING/THINKING TRACES!

Subbarao Kambhampati*

Karthik Valmeekam†

Siddhant Bhambri‡

Vardhan Palod

Lucas Paul Saldyt

Kaya Stechly§

Soumya Rani Samineni

Durgesh Kalwar

Upasana Biswas

School of Computing and AI,
Arizona State University

ABSTRACT

Intermediate token generation (ITG), where a model produces output before the solution, has become a standard 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 traces, and implying that these traces resemble steps a human might take when solving a challenging problem, and as such can provide an interpretable window into the operation of the model’s thinking process to the end user. 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. We call on the community to avoid such anthropomorphization of intermediate tokens.

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 [22]. These transformer-based large language models 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: 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 [4, 71], empirically it has been observed that language models perform better on many domains if they output such tokens first [43, 62, 68, 24, 21, 22, 46, 41, 35]. 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 [16, 22, 63, 71, 8]. The traces are thus seen both as giving insights to the end users about the solution quality, and capturing the model’s “thinking effort.”

*Corresponding author: rao@asu.edu

†Work done while at ASU, currently at Amazon AGI

‡Work done while at ASU, currently at Samsung Research

§Work done while at ASU, currently at Yale University

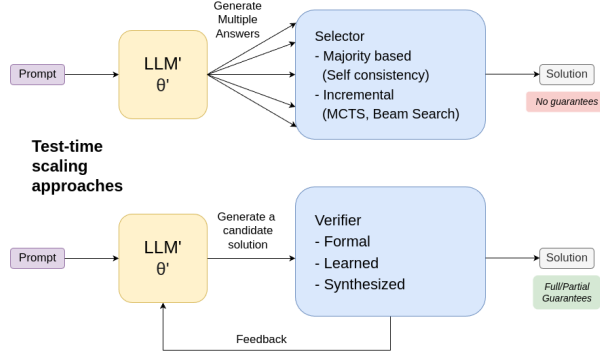


Figure 1: Test-time scaling approaches for teasing out reasoning

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. We support our position by collating significant body of emerging work questioning the interpretation of intermediate tokens as reasoning/thinking traces (Section 4). In Section 5, we will consider alternate views—that include expecting or hoping that intermediate tokens would give end users visibility into the operation of the model, and discuss how they affect our position. Finally, in Section 6, we will provide a call to action for the community that arises naturally from our position.

Anthropomorphization has long been a contentious issue in AI research [40], and LLMs have certainly increased our anthropomorphization tendencies [25]. 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.

The rest of the paper is organized as follows: We will start in Section 2 by giving some background on the main ideas behind reasoning models, with special attention to post-training on derivational traces.⁵ In Section 3, we will discuss the evidence for and ramifications of anthropomorphizing intermediate tokens as reasoning traces. In Section 4, we directly consider the question of whether intermediate tokens can be said to have any formal or human-interpretable semantics and also look at the pitfalls of viewing intermediate tokens as computation that is adaptive to problem complexity. In Section 5, we will consider alternate views, and discuss how they affect our position. Finally, in Section 6, we provide a call to action for the community that arises naturally from our position.

Before going forward, we should clarify some potential confusion regarding the “reasoning trace” terminology. By intermediate tokens, we refer to the unfiltered tokens emitted by the LLM before the solution. This should be distinguished from post-facto explanations or rationalizations of the process or the product of said “thinking.” For example, OpenAI o1 *hides* the intermediate tokens it produces (perhaps because they aren’t that interpretable to begin with?) but sometimes provides a sanitized summary/rationalization instead. In contrast, DeepSeek R1 [12] provides the full intermediate token sequences (which often *run for pages* even for simple problems; see Figure 3). To be clear, our focus here is on the anthropomorphization of unfiltered intermediate tokens rather than such post-facto rationalizations. 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 [42].

We should also clarify that our position and reservations are only about ascribing end user interpretability to intermediate tokens. This doesn’t extend to efforts that attempt to analyze why and how intermediate tokens help the model itself (e.g. [7]).

2 Background: Test Time Inference & Post-Training in Reasoning Models

Large Reasoning Models or 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; and (ii) **post-training methods**, which complement simple auto-regressive training on web corpora, with additional training on intermediate token data.

⁵We will use the term *derivational trace* as a neutral stand-in for intermediate tokens, whether generated by humans, formal solvers or other systems, rather than the more popular anthropomorphized phrases “Chains of thought” and “reasoning traces”.

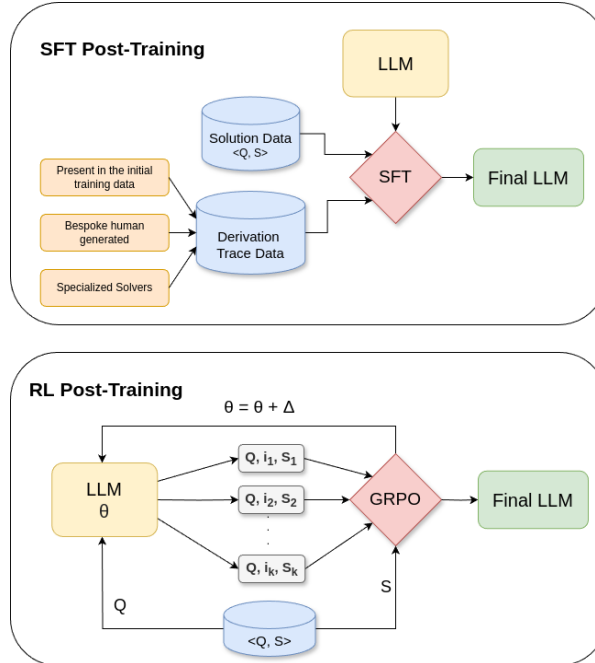


Figure 2: Post-training Approaches for teasing out reasoning

2.1 Test-time Inference

There is a rich history of approaches that use scalable online computation to improve upon faster initial guesses, including limited depth min-max, real-time A* search and dynamic programming, and Monte Carlo Tree Search [50, 19]. Test-time inference approaches (see Figure 1) mirror these ideas.

Perhaps the most popular and enduring class of test-time inference ideas involves generating many candidate solutions from an LLM and using some selection procedure to choose the final output. The simplest implementation is known as *self-consistency* [60]: choose the most common answer.

More sophisticated selection procedures attempt to verify that an LLM’s output is correct. When paired with an LLM in this manner, the combined system can be seen as a *generate-test* framework, and naturally raises questions about the verification process: *who does it*, and *with what guarantees*? A variety of approaches have been tried—including using LLMs themselves as verifiers[64] (although this is known to be problematic [54]), learning verifiers[2, 67], and using external sound verifiers that come with either full or partial guarantees. In cases where verifiers provide explanations or feedback when a guess is incorrect, these can be passed back to the LLM so it generates better subsequent guesses [49, 57, 11].

2.2 Post-Training on Derivational Traces

Unlike the test-time inference techniques, that augment the inference stage of standard LLMs, the post-training techniques are aimed at the LLM training stage. Standard LLMs are trained using a very simple objective: given a chunk of text, predict the most likely next token. This procedure, when employed with sufficiently high capacity models on web-scale corpora, has been surprisingly successful at capturing diverse text styles.

A variety of approaches have tried to generate derivational traces to post-train LLMs, ranging from paying annotators for step-by-step derivations to generating and selecting them with LLMs. We classify these in terms of (i) how candidate traces are generated and filtered, and (ii) how they are used to improve the underlying LLM through supervised fine tuning or reinforcement learning; see Figure 2.

Generating Candidate Derivational Traces: Several trace generation methods were considered:

Human-generated Traces: An obvious way to obtain additional derivational data is to have humans create it [36].

Solver-generated Traces: Searchformer [34], Stream of Search [15], as well as DeepMind’s work in [52, 39] use a much more scalable approach by using standard search algorithms to produce datasets containing not just answers but also the execution traces generated along the way.

LLM-generated Traces: Rather than creating high-quality traces from the start, an increasingly popular approach is to generate them from an LLM and filter afterwards [29].

Filtering Traces: Naively LLM-generated traces are often not useful unless they are filtered. Researchers have varied in how they approach this trace selection process, ranging from selecting only those that are correct at each step (according to human labelers), training process reward models that attempt to automate human verification[36], to selecting traces by formally verifying whether they lead to correct final solutions without considering the trace content [66, 12].

Improving LLMs Using Derivational Traces: Once derivational traces have been selected, they can be used to further train an LLM. Early approaches fine-tuned LLMs directly on such traces[66, 34, 15], but more recent advances have pivoted towards using reinforcement learning (RL) instead.

If we view the base model as a generator of plausible solutions to the reasoning problem, the test time inference/scaling techniques implement a “generate test” paradigm, improving the accuracy by checking the plausible solutions against a verifier. Post-training, in contrast, tries to shift the test part of this generate-test into the generator (model) itself⁶ Using DeepSeek R1 [12] as a case study, the model collects many synthetic problems, and for each generates plausible solution trajectories (comprising intermediate tokens followed by solution guesses). The solutions in these trajectories are evaluated by external problem-specific verifiers (DeepSeek calls them “rule-based reward models”). These trajectories with their rewards become the basis for a RL fine-tuning phase. The overall process has been termed RLVR—or RL with (externally) verified rewards [17, 32, 61]

3 Anthropomorphization of Intermediate Tokens

As we discussed, post-training can induce a model to first generate long strings of intermediate tokens before outputting its final answer. There has been a tendency in the field to view these intermediate tokens as the human-like “thoughts” of the model or to see them as *reasoning traces* which could reflect internal reasoning procedures. This is precisely the tendency our position paper argues against. We start by listing the various (unhealthy) ramifications of this anthropomorphization:

- Viewing intermediate tokens as reasoning/thinking traces has led to a drive to make them “interpretable” to humans in the loop (nevermind that interpretability mostly meant that the traces were in pseudo English). For example, DeepSeek [12] 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!
- 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 [9].
- 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 [55, 56] difficulty/effort and techniques that increased the length of intermediate tokens were celebrated as “learning to reason” [12]. Simultaneously there were efforts to *shorten* intermediate traces produced and celebrate that as learning to reason efficiently [3].
- There have been attempts to cast intermediate tokens as learning some “algorithm” that generated the training data. For example, the authors of Searchformer [33] 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).

⁶There is a famous dictum attributed to Marvin Minsky that *intelligence is shifting the test part of generate-test into generate part*.

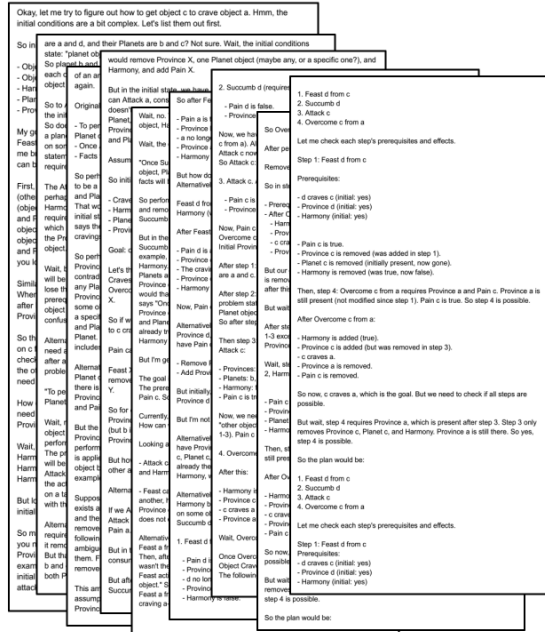


Figure 3: Intermediate tokens DeepSeek-R1 produces for a planning problem.

4 On the Questionable Semantic Status and Interpretability of Intermediate Tokens

While the fact that use of intermediate tokens during training and inference stages seem to improve LLM performance is beyond dispute, there are significant questions on whether these traces have any valid semantic import to the end user. 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”.

Famously, DeepSeek’s R1 paper claimed that one of the most impressive observed behaviors of their trained models was the so-called “aha” moment: as part of the chain of thought it was producing in order to answer some question, the model output the token “aha”, seeming to indicate that it had come upon a sudden realization. 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 [16, 48].

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 pages and pages of text in response to each and every query (a snippet is shown in Figure 3), and it is far from clear how to verify if these monologues constitute sound reasoning.⁷ Arguably, people examine the traces in places, see familiar phrases reminiscent of what a human solving such a problem might utter, and assume that the reasoning trace sounds plausible. While there have been some valiant efforts to make sense of these large-scale mumbblings—e.g. [38]—the analyses here tend to be somewhat qualitative and suggestible reminiscent of “lines of code” analyses in software engineering. It is no

⁷Before DeepSeek, the entire question was moot. OpenAI’s o1 model deliberately hides its intermediate tokens from end users, despite charging based on how many were produced!

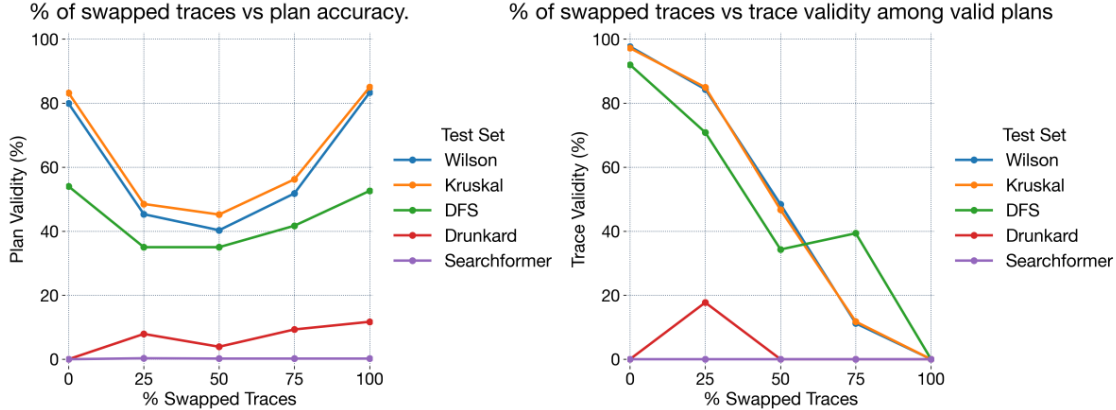


Figure 4: A study reproduced from [58] showing the curious phenomenon that as the models are trained with increasingly incorrect—in their case “swapped” traces—the inference time accuracy of the resulting model is high both with fully correct and fully swapped traces, dipping only in the middle.

wonder then that few if any LRM evaluations even try to check their pre-answer traces, and focus only on evaluating the correctness of their final answers.⁸

However, while evaluating the intermediate tokens produced by general LRMs may be out of direct reach, the traces generated by format-constrained models trained to imitate the derivational traces of domain-specific solvers can be formally verified. In [58] the authors systematically study the prevailing narrative that intermediate tokens or “Chains of Thought” generated by Large Reasoning Models like DeepSeek’s R1 are interpretable, semantically valid sequences with predictable effects on the model’s behavior. As they didn’t have access to any frontier LLM’s training data or even exact training procedure, and since the traces these models output are in multiply-interpretable natural language without a concrete ground truth, they design a series of controlled experiments where they trained smaller reasoning models on formally verifiable A* search traces, building on previous smaller model reasoning work.[14, 33]. Their findings show that there is only a loose correlation between the correctness of the trace and the correctness of the output plan, especially when the problem instances go out of the training distribution. They then report a causal intervention, training additional models on noisy or irrelevant traces and find that there are (nonsensical) trace formats that nevertheless maintain or even increase the model’s performance – all despite them being much less informative or connected to the problem at hand (see Figure 4). Their experiments suggest that what matters for a model to improve accuracy with intermediate tokens is not their semantic import, but rather a consistent pattern in the training data for the model to fit itself.

Other studies on training with noisy traces support these conclusions, with findings that show how LLMs remain robust to semantic noise in traces and similar performance gains can be achieved without semantic correctness [35, 55]. Li et al. [35] perform model distillation using noisy traces on math and coding problems and find that the smaller LLM that is being trained remains largely robust to the semantic noise in the trace. Even when trained on derivational trace containing largely incorrect mathematical operations, the LLM shows significant performance improvements as compared to the base model. Dualformer [55], an extension of Searchformer [33], which trains transformer models on truncated A* derivational traces (by arbitrarily removing steps from the original A* search process—and thus destroying any trace semantics) to improve solution accuracy, is another evidence for performance improvements with wrong traces!

Presumably, natural language reasoning follows algorithmic structure, even if it does not correspond to a rigidly-defined algorithm. For example, see Polya’s “How to Solve It,” [47] which outlines the elements of mathematical problem solving in an algorithmic way, even if they are often implicit. Accordingly, we argue that studying algorithmic search traces, such as in [58], resembles a *model organism* for understanding systems like R1 (analogous to the roles of fruit flies or worms in biology). If a technique can learn to produce semantic reasoning traces for natural language problems, it ought to be able to do so for algorithmic traces as well, and vice-versa. Accordingly, evidence that models trained on algorithmic traces do not learn semantics applies to natural language problems and systems that apply to them, namely R1.

⁸Approaches like Process Reward Models [70] try to make the reasoning traces a bit more locally consistent—but they seem to have taken a back seat since the success of DeepSeek R1.

A similar investigation to test the correlation between intermediate traces and final solution performance was carried out by the authors in [6] in the Question-Answering (QA) domains. By decomposing the QA reasoning problems into verifiable sub-problems that can be evaluated at inference time, the authors first generated a Supervised Fine-Tuning (SFT) dataset with correct intermediate traces paired with correct final solutions. To carry out an intervention experiment, they generated another SFT dataset consisting of incorrect intermediate traces again paired with correct final solutions. For the first SFT experiment setting, the results show a large number of False Positives where the fine-tuned models output correct final solutions but incorrect intermediate traces. Interestingly, the intervention experiments with incorrect intermediate traces even outperforms the SFT with correct intermediate trace setting. The authors also show empirically that trace correctness does not guarantee final solution correctness. Similarly, final solution correctness also does not imply that they were preceded by semantically correct intermediate traces. In yet another study comparing the correlation between end-user interpretability and SFT performance, the authors in [5] showed via user studies that the cognitive interpretability of reasoning traces for end users can also be an albatross from the perspective of LLM’s task performance.

If the intermediate tokens produced by models that are explicitly trained on correct traces are still not guaranteed to be valid during inference time, then there seems to be little reason to believe that trace validity improves when these models are further post-trained with RL or incremental SFT. This is because such post-training techniques [12, 51] change the base model parameters to bias it more towards the trajectories that end up on solutions verified correct by the external verifiers during training. Authors in [58] examined the effects of post-training with RL, specifically GRPO, on semantic correctness of reasoning traces. They report that post-training reduces semantic correctness of the traces while simultaneously improving performance. They also find that models trained on irrelevant traces exhibit similar performance gain with post-training. This should not be surprising given that most works that do these types of post-training reward only the solution accuracy and ignore the content of intermediate tokens [12, 65].

Other works that demonstrate how reasoning traces are not reliable indicators of the model’s internal computations include [4, 31, 10, 20, 9, 1]

4.1 Intermediate Token Production and Problem Adaptive Computation

Although our main focus is on the anthropomorphization and semantics of derivational traces, a closely related aspect is the extent to which traces reflect learned procedures or problem adaptive computation. When an LRM is generating *more intermediate tokens* before providing the solution, it is clearly doing more computation, but the nature of this computation is questionable, as is interpreting it as a meaningful procedure. The question is whether this computation reflects an intended procedure, and then if the length of computation can be viewed meaningfully as adaptive to problem difficulty.

Interestingly, there has been a tendency to celebrate post-training techniques for *increasing* the intermediate token length. DeepSeek R1 [12], for example, claims that RL post-training is *learning to reason* as shown by the increased length of intermediate tokens over RL epochs. It is even more ironic that there have been subsequent efforts to *reign in* the intermediate token lengths, and claim that as a way to reduce compute while preserving task performance/accuracy (c.f. [3]).

While it is difficult to check whether the intermediate tokens generated by an LRM correspond to a meaning full procedure and whether they reflect problem adaptive computation, it is possible to examine the length of the traces where the model is trained on derivational traces generated by a classical algorithm. Authors in [45] examined the trace lengths of models trained on A* search traces on problems of varying difficulties. They found that trace lengths can look indicative of problem adaptive computation when tested on in-distribution problems, however, this correlation breaks down when the problem instances are out-of-distribution. In one of their experiments, they show that on trivial problems which would require minimal computation for A* search, the transformer models often produces extremely long derivation traces, in many cases even exhausting the context window. These findings indicate that the correlation is quite tenuous between the from-scratch computational complexity of the problem and the derivational trace produced by the LLM.

Part of this misconception comes from the simplistic MDP formulation adopted by DeepSeek R1 and subsequent work [22]. In [51, 13] the authors examine this formulation, showing that with the structural assumption of representing states as sequences of tokens, and uniformly distributing the terminal reward into intermediate tokens, RL is incentivized to generate longer intermediate token sequences—something that has been misattributed to “improved reasoning.” At some level, this shouldn’t be surprising given that the whole point of RL is to figure out credit assignment, and the division of final reward equally into intermediate tokens short circuits this process in an *ad hoc* way.

Given that the increased length of intermediate tokens is celebrated by DeepSeek R1 [12], the fact that these may be happening due to a rather simplistic way of equally dividing advantage over all tokens should temper the credibility of claims that longer intermediate tokens in systems like R1 [12] are automatically indicative of “thinking effort.”

5 Alternate Views

We have made it clear from the outset that there certainly are alternate views about the semantic status of the intermediate tokens—indeed their prevalence and popularity is the main reason motivating this position paper. To summarize, the phrase “chains of thought” originally arose as a way of prompting LLMs to elicit particular types of prompt completions (“behaviors”) [62, 28]. Originally such CoT’s were meant to be hand-crafted by the end users and include human interpretable advice that the LLMs were seen to be following. Later studies, such as [53, 59] pushed back on the alignment between the advice and the completions.

With the advent of reasoning models such as DeepSeek R1, the CoT terminology has been repurposed to refer to the intermediate tokens that the models are trained to produce on their way to the solutions. These tokens have been analyzed for potentially human interpretable patterns. The DeepSeek R1 paper itself [12] helped this narrative along by analyzing the intermediate tokens for the presence of phrases that, when used by humans, typically suggest reflection and insight. In their paper, they talk about the *aha* moment in R1’s intermediate tokens. Latter work such as *thoughtology* [38] took this narrative further by looking for correlations between specific types of passages in the intermediate tokens (as extracted *post-facto* by another LLM) and the solution accuracy. More recently, another group [27] extended the same type of LLM-based analysis of the intermediate tokens generated by a reasoning model—this time in terms of shifting voices/perspectives—and claimed that *reasoning models generate societies of thought*, and implied that this is what explains their effectiveness. It should be noted that these analyses are often qualitative, and fail to establish direct connection between the narrative of the intermediate tokens and the final result.

Given that the current models are trained on large corpora of human data, the fact that they produce intermediate tokens (“chains of thought”) that sound plausibly like those that might be generated by humans may well be a form of imitating *cultural routines* (c.f. [18]) in the training data. Thus, our position is not that intermediate tokens will never have passages that might be interpretable by humans as corresponding to reasoning, but that such interpretability may be accidental and cannot be relied upon by the end users to assess their trust in the solutions provided by the models. Even such accidental interpretability might dissipate as models are increasingly post-trained with outcome reward-based RL [12]. Interestingly, some works such as [37] characterize this lack of connection between intermediate tokens and final solution as indication of models learning to cheat!

A related issue is that none of the major frontier model makers—OpenAI, Google, Anthropic—show their actual intermediate tokens for citing proprietary concerns. The model card for GPT-OSS [44], the open-weight reasoning models released by OpenAI, states that they use Harmony Response Format, which has three channels, *analysis*, *commentary* and *final*. The *analysis* part seems to correspond to the intermediate tokens (that are not shown in their production models), and the *final* part corresponds to the solution tokens. The *commentary* part typically has high level commentary interpretable for the end user, and is admittedly distinct from the *analysis* part that corresponds to intermediate tokens. It is not clear how and when the *commentary* part is generated. It is clear that their production models only show the summary part, and not the actual intermediate tokens that are the subject of post-training.

Ironically, the increasing realization that intermediate tokens may not have interpretable semantics has lead some researchers to issue public entreaties to the frontier model makers to preserve some semblance of interpretability in CoTs so the models can be monitored [30].

6 Summary and Call to Action

In this position paper, we argued against the popular tendency in the LLM research community to anthropomorphize intermediate tokens as reasoning or “thinking”. Anthropomorphization has been a part of AI research [40], and has significantly increased in the era of LLMs [25]. We collated emerging evidence to support our position that intermediate tokens are not guaranteed to have any end user semantics, and that their interpretability and solution accuracy are often at loggerheads (Section 4), and also discussed alternate views in the literature and why our position makes sense in spite of them (Section 5).

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.

To the extent the research community finds our position persuasive, our recommendation is to stop assuming (or looking for) end user semantics in the intermediate tokens produced by the reasoning models. Human interpretation of intermediate tokens should not be used as a proxy measure for the trustworthiness of the solutions.

Given that the intermediate tokens 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!⁹ 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.

Where trust in the final solution is needed, it should instead come from verification of the correctness of the solution itself by the end users or third party sources—including problem class specific verifiers (c.f. [26]).

Given that intermediate tokens are meant mostly to help LLMs, restricting them to some syntactic format with hopes that it will be more palatable to end users becomes quite an albatross. This is a lesson from DeepSeek R1 [12] that is often missed. When they re-trained their original R0 model—that happened to produce a combination of English and Chinese tokens, with a costly supervised fine tuning phase on carefully curated English intermediate tokens generated by humans, the performance (as measured in solution accuracy) worsened, without any concomitant measured improvements in the actual validity of the intermediate tokens generated!

Once we stop ascribing questionable interpretability to the intermediate tokens, and recognize that they are meant to help the LLM and not the end user, that would also free us to train models that optimize the intermediate tokens only for solution accuracy—even if the intermediate tokens themselves don't any longer look like plausible language utterances that humans might exhibit. This could, in theory, allow models to consider intermediate tokens made up of non-linguistic tokens—basically any vector from the embedding space, even if it doesn't correspond to a unique vocabulary item. Already there is some evidence that such methods can lead to further improvements in solution accuracy (c.f. [23, 69]).

As we have mentioned in Section 5, most frontier models, with the exception of DeepSeek R1, already seem to, in effect, abide by our position in that they are no longer showing the intermediate tokens anyways (citing proprietary considerations). Ironically it is the research community that still seems to entertain the possibility that tokens produced by intermediate tokens can provide an interpretable explanation of the model's operation to the end user. We hope this paper succeeds in dissuading them.

Acknowledgment

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

References

- [1] Arcuschin, I., Janiak, J., Krzyzanowski, R., Rajamanoharan, S., Nanda, N., and Conmy, A. Chain-of-thought reasoning in the wild is not always faithful. *arXiv preprint arXiv:2503.08679*, 2025.
- [2] Arora, D. and Kambhampati, S. 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.
- [3] Arora, D. and Zanette, A. Training language models to reason efficiently, 2025. URL <https://arxiv.org/abs/2502.04463>, 2025.
- [4] Baker, B., Huizinga, J., Gao, L., Dou, Z., Guan, M. Y., Madry, A., Zaremba, W., Pachocki, J., and Farhi, D. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025.

⁹To illustrate how false confidence can be engendered, consider the following thought experiment. Suppose you have a question Q for which you truly don't know the answer. You take two separate sheets of paper, write the question at the top of each of the sheets, and put a box for the answer at the bottom of the sheet. You give these sheets to two of your friends—Tom and Mary—and ask them to answer. Mary writes an answer a_m on the sheet and gives it back to you. Tom writes a different answer a_t ($a_t \neq a_m$), but also writes a bunch of plausible sounding platitudes in the scratch area. Consider now the possibility that despite you not knowing which answer is correct, you are likely to be drawn to Tom's answer just because it has these plausible intermediate tokens.

- [5] Bhambri, S., Biswas, U., and Kambhampati, S. Do cognitively interpretable reasoning traces improve llm performance? *arXiv preprint arXiv:2508.16695*, 2025.
- [6] Bhambri, S., Biswas, U., and Kambhampati, S. Interpretable traces, unexpected outcomes: Investigating the disconnect in trace-based knowledge distillation. *arXiv preprint arXiv:2505.13792*, 2025.
- [7] Bogdan, P. C., Macar, U., Nanda, N., and Conmy, A. Thought anchors: Which llm reasoning steps matter?, 2025. URL <https://arxiv.org/abs/2506.19143>.
- [8] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.
- [9] Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., Somani, A., Hase, P., Wagner, M., Roger, F., et al. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025.
- [10] Chua, J., Betley, J., Taylor, M., and Evans, O. Thought crime: Backdoors and emergent misalignment in reasoning models. *arXiv preprint arXiv:2506.13206*, 2025.
- [11] DeepMind, G. AlphaEvolve: a coding agent for scientific and algorithmic discovery. URL <https://deepmind.google/discover/blog/alphaevolve-a-gemini-powered-coding-agent-for-designing-advanced-algorithms/>, 2025.
- [12] DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, 2025. URL <https://arxiv.org/abs/2501.12948>.
- [13] Fatemi, M., Rafee, B., Tang, M., and Talamadupula, K. Concise reasoning via reinforcement learning. *arXiv preprint arXiv:2504.05185*, 2025.
- [14] Gandhi, K., Lee, D., Grand, G., Liu, M., Cheng, W., Sharma, A., and Goodman, N. D. Stream of search (sos): Learning to search in language. *arXiv preprint arXiv:2404.03683*, 2024.
- [15] Gandhi, K., Lee, D., Grand, G., Liu, M., Cheng, W., Sharma, A., and Goodman, N. D. Stream of Search (SoS): Learning to Search in Language. In *Conference on Language Modeling (COLM)*, 2024.
- [16] Gandhi, K., Chakravarthy, A., Singh, A., Lile, N., and Goodman, N. D. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*, 2025.
- [17] Gao, J., Xu, S., Ye, W., Liu, W., He, C., Fu, W., Mei, Z., Wang, G., and Wu, Y. On designing effective rl reward at training time for llm reasoning. *arXiv preprint arXiv:2410.15115*, 2024.
- [18] Gopnik, A. *The gardener and the carpenter: What the new science of child development tells us about the relationship between parents and children*. Macmillan, 2016.
- [19] Graves, A. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*, 2016.
- [20] Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S. R., and Hubinger, E. Alignment faking in large language models, 2024. URL <https://arxiv.org/abs/2412.14093>.
- [21] Gu, Y., Dong, L., Wei, F., and Huang, M. Minillm: Knowledge distillation of large language models. *arXiv preprint arXiv:2306.08543*, 2023.
- [22] Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [23] Hao, S., Sukhbaatar, S., Su, D., Li, X., Hu, Z., Weston, J., and Tian, Y. Training large language models to reason in a continuous latent space, 2024. URL <https://arxiv.org/abs/2412.06769>.
- [24] Hsieh, C.-Y., Li, C.-L., Yeh, C.-K., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C.-Y., and Pfister, T. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*, 2023.

- [25] Ibrahim, L. and Cheng, M. Thinking beyond the anthropomorphic paradigm benefits llm research, 2025. URL <https://arxiv.org/abs/2502.09192>.
- [26] Kambhampati, S., Valmeekam, K., Guan, L., Verma, M., Stechly, K., Bhambri, S., Saldyt, L. P., and Murthy, A. B. Position: LLMs can’t plan, but can help planning in LLM-modulo frameworks. In *Forty-first International Conference on Machine Learning*, 2024.
- [27] Kim, J., Lai, S., Scherrer, N., Evans, J., et al. Reasoning models generate societies of thought. *arXiv preprint arXiv:2601.10825*, 2026.
- [28] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.
- [29] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.
- [30] Korbak, T., Balesni, M., Barnes, E., Bengio, Y., Benton, J., Bloom, J., Chen, M., Cooney, A., Dafoe, A., Dragan, A., Emmons, S., Evans, O., Farhi, D., Greenblatt, R., Hendrycks, D., Hobbhahn, M., Hubinger, E., Irving, G., Jenner, E., Kokotajlo, D., Krakovna, V., Legg, S., Lindner, D., Luan, D., Madry, A., Michael, J., Nanda, N., Orr, D., Pachocki, J., Perez, E., Phuong, M., Roger, F., Saxe, J., Shlegeris, B., Soto, M., Steinberger, E., Wang, J., Zaremba, W., Baker, B., Shah, R., and Mikulik, V. Chain of thought monitorability: A new and fragile opportunity for ai safety, 2025. URL <https://arxiv.org/abs/2507.11473>.
- [31] Korbak, T., Balesni, M., Barnes, E., Bengio, Y., Benton, J., Bloom, J., Chen, M., Cooney, A., Dafoe, A., Dragan, A., et al. Chain of thought monitorability: A new and fragile opportunity for ai safety. *arXiv preprint arXiv:2507.11473*, 2025.
- [32] Lambert, N., Morrison, J., Pyatkin, V., Huang, S., Ivison, H., Brahman, F., Miranda, L. J. V., Liu, A., Dziri, N., Lyu, S., et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.
- [33] Lehnert, L., Sukhbaatar, S., Su, D., Zheng, Q., Mcvay, P., Rabbat, M., and Tian, Y. Beyond a*: Better planning with transformers via search dynamics bootstrapping. *arXiv preprint arXiv:2402.14083*, 2024.
- [34] Lehnert, L., Sukhbaatar, S., Su, D., Zheng, Q., Mcvay, P., Rabbat, M., and Tian, Y. Beyond A*: Better Planning with Transformers via Search Dynamics Bootstrapping. In *Conference on Language Models (COLM)*, 2024.
- [35] Li, D., Cao, S., Griggs, T., Liu, S., Mo, X., Tang, E., Hegde, S., Hakhamaneshi, K., Patil, S. G., Zaharia, M., et al. Llms can easily learn to reason from demonstrations structure, not content, is what matters! *arXiv preprint arXiv:2502.07374*, 2025.
- [36] Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let’s verify step by step, 2023. URL <https://arxiv.org/abs/2305.20050>.
- [37] MacDiarmid, M., Wright, B., Uesato, J., Benton, J., Kutasov, J., Price, S., Bouscal, N., Bowman, S., Bricken, T., Cloud, A., Denison, C., Gasteiger, J., Greenblatt, R., Leike, J., Lindsey, J., Mikulik, V., Perez, E., Rodrigues, A., Thomas, D., Webson, A., Ziegler, D., and Hubinger, E. Natural emergent misalignment from reward hacking in production rl, 2025. URL <https://arxiv.org/abs/2511.18397>.
- [38] Marjanović, S. V., Patel, A., Adlakha, V., Aghajohari, M., BehnamGhader, P., Bhatia, M., Khandelwal, A., Kraft, A., Krojer, B., Lù, X. H., Meade, N., Shin, D., Kazemnejad, A., Kamath, G., Mosbach, M., Stańczyk, K., and Reddy, S. Deepseek-r1 thoughtology: Let’s think about llm reasoning, 2025. URL <https://arxiv.org/abs/2504.07128>.
- [39] Markeeva, L., Mcleish, S., Ibarz, B., Bounsi, W., Kozlova, O., Vitvitskyi, A., Blundell, C., Goldstein, T., Schwarzschild, A., and Veličkovi´veličkovi´c, P. The CLRS-Text Algorithmic Reasoning Language Benchmark. Technical report, 2024. URL <https://github.com/google-deepmind/>.
- [40] McDermott, D. Artificial intelligence meets natural stupidity. *SIGART Newsl.*, 57:4–9, 1976. URL <https://api.semanticscholar.org/CorpusID:28619965>.
- [41] Muennighoff, N., Yang, Z., Shi, W., Li, X. L., Fei-Fei, L., Hajishirzi, H., Zettlemoyer, L., Liang, P., Candès, E., and Hashimoto, T. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

- [42] Nisbett, R. E. and Wilson, T. D. Telling more than we can know: Verbal reports on mental processes. *Psychological review*, 84(3):231, 1977.
- [43] Nye, M., Andreassen, A. J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D., et al. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.
- [44] OpenAI, :, Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., Arbus, E., Arora, R. K., Bai, Y., Baker, B., Bao, H., Barak, B., Bennett, A., Bertao, T., Brett, N., Brevdo, E., Brockman, G., Bubeck, S., Chang, C., Chen, K., Chen, M., Cheung, E., Clark, A., Cook, D., Dukhan, M., Dvorak, C., Fives, K., Fomenko, V., Garipov, T., Georgiev, K., Glaese, M., Gogineni, T., Goucher, A., Gross, L., Guzman, K. G., Hallman, J., Hehir, J., Heidecke, J., Helyar, A., Hu, H., Huet, R., Huh, J., Jain, S., Johnson, Z., Koch, C., Kofman, I., Kundel, D., Kwon, J., Kyrylov, V., Le, E. Y., Leclerc, G., Lennon, J. P., Lessans, S., Lezcano-Casado, M., Li, Y., Li, Z., Lin, J., Liss, J., Lily, Liu, Liu, J., Lu, K., Lu, C., Martinovic, Z., McCallum, L., McGrath, J., McKinney, S., McLaughlin, A., Mei, S., Mostovoy, S., Mu, T., Myles, G., Neitz, A., Nichol, A., Pachocki, J., Paino, A., Palmie, D., Pantuliano, A., Parascandolo, G., Park, J., Pathak, L., Paz, C., Peran, L., Pimenov, D., Pokrass, M., Proehl, E., Qiu, H., Raila, G., Raso, F., Ren, H., Richardson, K., Robinson, D., Rotsted, B., Salman, H., Sanjeev, S., Schwarz, M., Sculley, D., Sikchi, H., Simon, K., Singhal, K., Song, Y., Stuckey, D., Sun, Z., Tillet, P., Toizer, S., Tsimpourlas, F., Vyas, N., Wallace, E., Wang, X., Wang, M., Watkins, O., Weil, K., Wendling, A., Whinnery, K., Whitney, C., Wong, H., Yang, L., Yang, Y., Yasunaga, M., Ying, K., Zaremba, W., Zhan, W., Zhang, C., Zhang, B., Zhang, E., and Zhao, S. gpt-oss-120b & gpt-oss-20b model card, 2025. URL <https://arxiv.org/abs/2508.10925>.
- [45] Palod, V., Valmeekam, K., Stechly, K., and Kambhampati, S. Performative thinking? the brittle correlation between cot length and problem complexity. *arXiv preprint arXiv:2509.07339*, 2025.
- [46] Pfau, J., Merrill, W., and Bowman, S. R. Let’s think dot by dot: Hidden computation in transformer language models. *arXiv preprint arXiv:2404.15758*, 2024.
- [47] Polya, G. How to solve it: A new aspect of mathematical method. In *How to solve it*. Princeton university press, 2014.
- [48] Qin, T., Alvarez-Melis, D., Jelassi, S., and Malach, E. To backtrack or not to backtrack: When sequential search limits model reasoning. *arXiv preprint arXiv:2504.07052*, 2025.
- [49] Romera-Paredes, B., Barekatin, M., Novikov, A., Balog, M., Kumar, M. P., Dupont, E., Ruiz, F. J., Ellenberg, J. S., Wang, P., Fawzi, O., et al. Mathematical discoveries from program search with large language models. *Nature*, pp. 1–3, 2023.
- [50] Russell, S. J. and Norvig, P. *Artificial Intelligence: A Modern Approach*. London, 2010.
- [51] Samineni, S. R., Kalwar, D., Valmeekam, K., Stechly, K., and Kambhampati, S. Rl in name only? analyzing the structural assumptions in rl post-training for llms, 2025. URL <https://arxiv.org/abs/2505.13697>.
- [52] Schultz, J., Adamek, J., Jusup, M., Lanctot, M., Kaisers, M., Perrin, S., Hennes, D., Shar, J., Lewis, C., Ruoss, A., Zahavy, T., Veličković, P., Prince, L., Singh, S., Malmi, E., and Tomašev, N. Mastering board games by external and internal planning with language models, 2024. URL <https://arxiv.org/abs/2412.12119>.
- [53] Stechly, K., Valmeekam, K., and Kambhampati, S. Chain of Thoughtlessness: An Analysis of CoT in Planning. In *Proc. NeurIPS*, 2024.
- [54] Stechly, K., Valmeekam, K., and Kambhampati, S. On the Self-Verification Limitations of Large Language Models on Reasoning and Planning Tasks. In *Proc. ICLR*, 2025.
- [55] Su, D., Sukhbaatar, S., Rabbat, M., Tian, Y., and Zheng, Q. Dualformer: Controllable fast and slow thinking by learning with randomized reasoning traces. In *The Thirteenth International Conference on Learning Representations*, 2024.
- [56] Su, J., Healey, J., Nakov, P., and Cardie, C. Between underthinking and overthinking: An empirical study of reasoning length and correctness in llms, 2025. URL <https://arxiv.org/abs/2505.00127>.
- [57] Trinh, T. H., Wu, Y., Le, Q. V., He, H., and Luong, T. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024.

- [58] Valmeekam, K., Stechly, K., Palod, V., Gundawar, A., and Kambhampati, S. Beyond semantics: The unreasonable effectiveness of reasonless intermediate tokens, 2025.
- [59] Wang, B., Min, S., Deng, X., Shen, J., Wu, Y., Zettlemoyer, L., and Sun, H. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, July 2023.
- [60] Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=1PL1NIMMrw>.
- [61] Wang, Y., Yang, Q., Zeng, Z., Ren, L., Liu, L., Peng, B., Cheng, H., He, X., Wang, K., Gao, J., et al. Reinforcement learning for reasoning in large language models with one training example. *arXiv preprint arXiv:2504.20571*, 2025.
- [62] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837, 2022.
- [63] Yang, S., Wu, J., Chen, X., Xiao, Y., Yang, X., Wong, D. F., and Wang, D. Understanding aha moments: from external observations to internal mechanisms. *arXiv preprint arXiv:2504.02956*, 2025.
- [64] Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., and Narasimhan, K. R. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=5Xc1ecx01h>.
- [65] Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Fan, T., Liu, G., Liu, L., Liu, X., et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.
- [66] Zelikman, E., Wu, Y., Mu, J., and Goodman, N. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.
- [67] Zhang, L., Hosseini, A., Bansal, H., Kazemi, M., Kumar, A., and Agarwal, R. Generative verifiers: Reward modeling as next-token prediction, 2024. URL <https://arxiv.org/abs/2408.15240>.
- [68] Zhang, Z., Zhang, A., Li, M., and Smola, A. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022.
- [69] Zhang, Z., He, X., Yan, W., Shen, A., Zhao, C., Wang, S., Shen, Y., and Wang, X. E. Soft thinking: Unlocking the reasoning potential of llms in continuous concept space, 2025. URL <https://arxiv.org/abs/2505.15778>.
- [70] Zhang, Z., Zheng, C., Wu, Y., Zhang, B., Lin, R., Yu, B., Liu, D., Zhou, J., and Lin, J. The lessons of developing process reward models in mathematical reasoning, 2025. URL <https://arxiv.org/abs/2501.07301>.
- [71] Zhou, H., Li, X., Wang, R., Cheng, M., Zhou, T., and Hsieh, C.-J. R1-zero’s” aha moment” in visual reasoning on a 2b non-sft model. *arXiv preprint arXiv:2503.05132*, 2025.