Towards Reasoning Era: A Survey of Long Chain-of-Thought for Reasoning Large Language Models

Qiguang Chen[†] **Libo Qin**[‡] **Jinhao Liu**[†] **Dengyun Peng**[†] **Jiannan Guan**[†]
Peng Wang[‡] **Mengkang Hu**[◇] **Yuhang Zhou**[♡] **Te Gao**[‡] **Wanxiang Che**[†]
[†] Research Center for Social Computing and Interactive Robotics
[†] Harbin Institute of Technology
[‡] School of Computer Science and Engineering, Central South University
[◇] The University of Hong Kong   [♡] Fudan University
{qgchen,car}@ir.hit.edu.cn, lbqin@csu.edu.cn

https://long-cot.github.io/

Abstract

Recent advancements in reasoning with large language models (RLLMs), such as OpenAI-O1 and DeepSeek-R1, have demonstrated their impressive capabilities in complex domains like mathematics and coding. A central factor in their success lies in the application of long chain-of-thought (Long CoT) characteristics, which enhance reasoning abilities and enable the solution of intricate problems. However, despite these developments, a comprehensive survey on Long CoT is still lacking, limiting our understanding of its distinctions from traditional short chain-of-thought (Short CoT) and complicating ongoing debates on issues like "overthinking" and "test-time scaling." This survey seeks to fill this gap by offering a unified perspective on Long CoT. (1) We first distinguish Long CoT from Short CoT and introduce a novel taxonomy to categorize current reasoning paradigms. (2) Next, we explore the key characteristics of Long CoT: deep reasoning, extensive exploration, and feasible reflection, which enable models to handle more complex tasks and produce more efficient, coherent outcomes compared to the shallower Short CoT. (3) We then investigate key phenomena such as the emergence of Long CoT with these characteristics, including overthinking, and test-time scaling, offering insights into how these processes manifest in practice. (4) Finally, we identify significant research gaps and highlight promising future directions, including the integration of multi-modal reasoning, efficiency improvements, and enhanced knowledge frameworks. By providing a structured overview, this survey aims to inspire future research and further the development of logical reasoning in artificial intelligence[1].

1   Introduction

In recent years, the emergence of reasoning large language models (RLLMs) such as OpenAI O1 [208] and DeepSeek R1 [155] has sparked a growing body of research into Long Chain-of-Thought (Long CoT) reasoning, greatly improving their mathematical reasoning, programming tasks, and multidisciplinary knowledge reasoning capabilities [488, 686, 508, 50, 58, 673, 133, 776], as shown in Figure 1. This shift marks a significant departure from traditional approaches to task handling in large language models (LLMs) [798, 437, 439, 421]. Unlike the shorter chain-of-thought

---

[1]Our logo refers to a cute cartoon image - Snake Puppy.
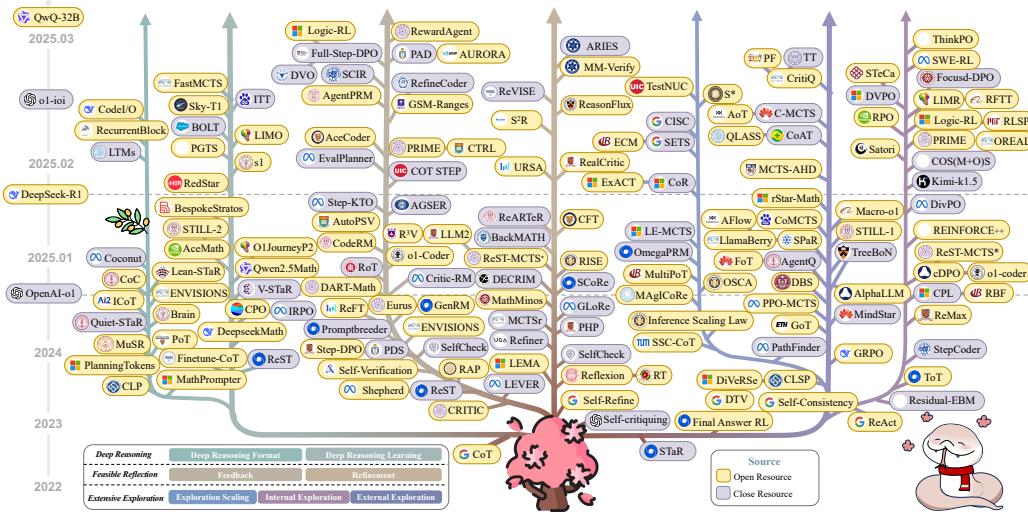
Figure 1: Evolution of selected Long CoT over the past three years, where colored branches represent different characteristics: deep reasoning, feasible reflection, and extensive exploration. Each characteristic is further divided into key areas: Deep reasoning includes its format and learning methods. Feasible reflection focuses on feedback and refinement techniques during reflection process as optimization strategies. Extensive exploration addresses scaling, internal, and external exploration as key improvements to Long CoT.

(Short CoT) used in traditional LLMs, Long CoT reasoning entails a more detailed, iterative process of exploration and reflection within a given problem space by test-time scaling [299, 520, 364]. This process has led to notable advancements in mathematical and logical reasoning, as well as in exploring how supervised fine-tuning (SFT) and reinforcement learning (RL) techniques can enhance the learning and exploration of extended reasoning chains [440, 385].

However, there is no comprehensive survey to systematically understand the main factors and recent efforts of Long CoT for RLLMs, which hinders the development of RLLMs. As a result, there are ongoing debates about the effectiveness of simple "test-time scaling" for Longer CoT [610, 343] versus the argument that "over-thinking" from excessively long scaling can harm LLMs and introduce unnecessary complexity [73, 96, 251]. Moreover, some researchers argue that, when solving specific problems, there is no clear relationship between length and accuracy [622].

To address this gap, we provide an extensive and comprehensive survey of Long CoT. Specifically, as illustrated in Figure 2, we first **define and examine the distinctions** between Long CoT and traditional Short CoT, focusing on the following key aspects: (1) *Deep Reasoning*, which requires a sufficient depth of logical processing to manage an extensive set of reasoning nodes; (2) *Extensive Exploration*, which involves generating parallel uncertain nodes and transitioning from known to unknown logic; and (3) *Feasible Reflection*, which involves feedback and refinement of logical connections. These characteristics enable Long CoT paradigms to integrate more intricate reasoning and accommodate a broader range of logical structures, ultimately leading to more efficient and coherent outcomes. Subsequently, we systematically **explore the underlying explanations for key phenomena associated with Long CoT**, such as its emergence, the overthinking phenomenon, inference time scaling during testing, and the "Aha Moment," among others. To our knowledge, This is the first comprehensive survey dedicated to these specific topics. Finally, considering the extensive body of literature, we **highlight promising areas for future research** and suggest valuable open-resource frameworks and datasets that can serve as a foundation for future investigations.

The main contributions of this work are as follows:

- **Systematic Distinction:** In this work, we first introduce the concept of Long CoT reasoning and distinguish it from the traditional Short CoT, thereby providing a clear framework for understanding both paradigms and their respective characteristics.
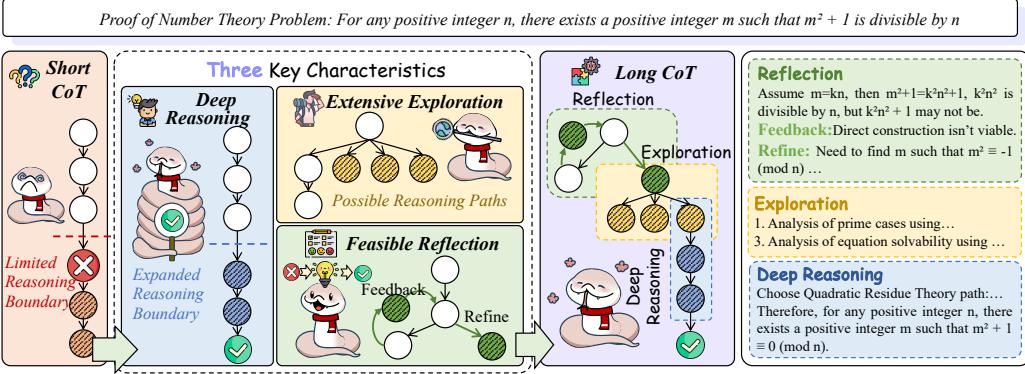
Figure 2: The differences between advanced Long CoT and traditional Short CoT are characterized by three key characteristics: deep reasoning, feasible reflection, and extensive exploration. Moreover, Long CoT integrates all these characteristics to achieve substantial logical efficacy.

- **Explanation of Hot Phenomena:** We systematically investigate the notable phenomena associated with Long CoT reasoning, such as overthinking, inference test-time scaling, and the "Aha Moment", offering valuable insights into the cognitive processes involved in complex reasoning.
- **Emerging Challenges and Frontiers:** We explore the emerging challenges within the field of Long CoT reasoning and identify key research frontiers. Given the vast body of literature, we highlight areas where further inquiry could significantly advance the development of Long CoT methodologies.

2 Discussion of Long CoT v.s. Short CoT

This section formalizes the key differences between Long Chain-of-Thought (Long CoT) and Short Chain-of-Thought (Short CoT), emphasizing reasoning depth, revisiting connections, and logical node exploration [607]. These distinctions are clearly separate from System 1 and System 2 thinking. The comparison between Long CoT and Short CoT is framed within System 2, with Long CoT involving more thorough reasoning, reflection, and exploration, while Short CoT generally prioritizes shallow and efficient logic over exhaustive reasoning.

2.1 Overview of Short CoT

As illustrated by Figure 2, Short CoT is typically characterized by a shallow, linear reasoning process, where conclusions are drawn sequentially, often relying on a limited number of logical nodes [386]. This reasoning is usually rapid and straightforward, with simple, surface-level transitions and minimal exploration of alternative paths, which restricts its generalizability [480]. Formally, given a reasoning model $\mathcal{R}$, we can define the rationale of Short CoT (CoT_S) as follows:

$$\text{CoT}_S = \mathcal{R}(\{n_i\}_{i=1}^k | (k \leq \mathcal{B}_s) \wedge (j=1 \Leftrightarrow \forall i \leq k, n_i \rightarrow n_{i+j}) \wedge (\forall i \neq j \leq k, n_i \neq n_j)), \quad (1)$$

where n_1 to n_k represent a sequence of logical nodes, which naturally satisfy that $\forall i, n_i \rightarrow n_{i+1}$. Here, $\mathcal{B}_s$ denotes the upper boundary on the number of reasoning nodes, as defined by Chen et al. [64]. In this paradigm, the reasoning progresses sequentially from one node to the next, with minimal revisitation of previous nodes and little exploration of alternative logical paths.

2.2 Overview of Long CoT

In contrast, Long CoT involves deeper reasoning, reflective analysis, and a broader exploration of logical structures. It facilitates reasoning across a wider range of logical steps, addressing both known and unknown elements of a problem [128]. Formally, Long CoT expands the constraints outlined in Equation 1 based on either an explicit or implicit tree structure. As shown in Figure 3, we will now discuss these key differences in detail.

Deep Reasoning for Long CoT (§4)	**Deep Reasoning Formation (§4.1)**	**Natural Language Deep Reasoning** — *e.g.,* CoT [594], Reflection of thought [781], MathPrompter [204], CLP [435], AutoCAP [749], PlanSearch [548], NaturalProgram [327], CodeI/O [285], Wang et al. [587], *etc.*
		Structured Language Deep Reasoning — *e.g.,*PoT [70], CoC [275], Brain [76], SIaM [684], ENVISIONS [631], SKIntern [314], QuaSAR [448], TinyGSM [330], MathDivide [483], Payoungkhamdee et al. [418], GPT-f [425], STP [104], *etc.*
		Latent Space Deep Reasoning — *e.g.,*Quiet-STaR [708], PlaningToken [575], Coconut [162], RecurrentBlock[136], MuSR [481], SERT [748], Heima [469], LTMs [250], ITT [77], Deng et al. [100], *etc.*
	Deep Reasoning Learning (§4.2)	**Deep Reasoning Imitation** — *e.g.,* GSM8K [95], AceMath [347], DART-Math [524], O1-Journey-P2[200], STILL-2 [385], LIMO [676], s1 [391], RedSTaR [635], Fine-tune-CoT [174], CoT-Collection [246], FastMCTS [293], *etc.*
		Deep Reasoning Self-Learning — *e.g.,* STaR [707], ReST [153], DynaThink [404], V-STaR [179], PGTS [302], CPO [746], TPO [316], OpenRFT [754], MCMC-EM [175], BOLT [409], Weak2Strong [663], Iterative RPO [410], *etc.*
Feasible Reflection for Long CoT (§5)	**Feedback (§5.1)**	**Overall Feedback** — *e.g.,*Self-Critique [458], SelfCheck [384], Self-Verification [600], Critic-CoT [772], Verifier [95], STaR [707],ReST [153],Critic [145], ReFT [531], AceCoder [709], DeepSeek-R1 [155], Critic-RM [692], o1-coder [753], RM [372], Logic-RL [622], Self-Contrast [744],AGSER [342],DeepSeek-Math [466], *etc.*
		Process Feedback — *e.g.,*ReAct [669],Reflexion [471],Math-Minos [130],Math-Shepherd [567], ER-PRM [729],Eurus [700], PAD [148], PAVs [462], CTRL [627], QwQ [517], Skywork-o1[400], AceMath [347], PRIME [97], AURORA [499], RewardAgent [419], PDS [640], COT STEP [89], Step-DPO [257], ORPS [696], *etc.*
		Hybrid Feedback — *e.g.,*Step-KTO [323], Zhang et al. [755], *etc.*
	Refinement (§5.2)	**Prompt-based Refinement Generation** — *e.g.,*Reflexion [471], SelfCheck [384], Self-Critique [458], Self-Refine [375], Refiner [417], MCTSr [726], ReST-MCTS* [725],LLM2 [652], GLoRe [164],Yang et al. [664],RR-MP [166], PAD [408], De-CRIM [122],StepCo [612], BackMATH [740], ReARTeR [493], START [196], PHP [768], *etc.*
		SFT-based Refinement Imitation — *e.g.,*RealCritic [501], Self-Debugging [74], ProgCo [479], S3c-Math [647], Math-Minos [130], LEMA [11], Reflection-tuning [289], CFT [586], Chen et al. [69],Xi et al. [616], rStar [434], RISE [442],R3V [83], MM-Verify [489], URSA [360], SIMA [577], CoT-based Synthesizer [722], Qin et al. [440], *etc.*
		RL-based Refinement Learning — *e.g.,*SCoRe [252], DeepSeek-r1 [155], Zeng et al. [712], S^2R [367], ReasonFlux [656], ReVISE [265], ARIES [713], *etc.*
Extensive Exploration for Long CoT (§6)	**Exploration Scaling (§6.1)**	**Vertical Scaling** — *e.g.,*OpenAI [208], Fu et al. [124], Jaech et al. [208], s1 [391], Geiping et al. [136], ITT [77], METAL [272], RaLU [277], Ji et al. [214], *etc.*
		Parallel Scaling — *e.g.,*Self-Consistency [580],Inference Scaling Law [610],Liu et al. [343], Zhao et al. [758],WoT [750], DIVERSE [304], DnA-Eval [291], SSC-CoT [766], ExACT [690], Kim et al. [248], S* [278], CISC [502], OmegaPRM [358], AFT [301],Seed-CTS [552],CLSP [435], MultiPoT [361], ECM [66], *etc.*
	Internal Exploration (§6.2)	**RL Strategies** — *e.g.,*PPO [460], GRPO [466], REINFORCE++ [181], OREO [555], DAPO [339], LIMR [300] DAPO [339], LIMR [300], TRPO [460], DVPO [188], RPO [490],PRIME [97], DivPO [261],COS(M+O)S [378],CPL [570], Focused-DPO [733],RFTT [735],OREO [555], DeepSeekMath [466],TPO [659], *etc.*
		Reward Strategies — *e.g.,*DeepSeek-R1 [155],Kimi-k1.5 [508], T1 [180],ReST-EM [475],SWE-RL [596], DeepScaleR [359], ReST-MCTS* [725], rStaR-Math [151], Logic-RL [622], OREAL [362], StepCoder [107], RLSP [674], Verifier [95], TS-LLM [540], STeCa [551], OREO[555], Chu et al. [91], Shen et al. [468], *etc.*
	External Exploration (§6.3)	**Human-driven Exploration** — *e.g.,*SPaR [82], Forest-of-thought [37],Scattered ForestSearch [318],Kang et al. [234],AlphaLLM [523], PATHFINDER [143],Least-to-Most [780], ToT [668], TreeBoN [441], CodeTree [284], Tree-of-Code [396] TouT [388], GoT [32], GraphReason [46], Besta et al. [33], AoT [520], Chen et al. [67], *etc.*
		Model-driven Exploration — *e.g.,*DBS [795], Lehnert et al. [269], MindStar [233], Residual-EBM [634], Mulberry [666], C-MCTS [322], PPO-MCTS [341], Llama-Berry [727], Marco-o1 [765], AtomThink [619], Puri et al. [432], LE-MCTS [414], rStar-Math [151], MC-NEST [443], CoAT [405], CoPlanner [547], CritiQ [156], *etc.*

Figure 3: Taxonomy of Long CoT, which includes deep reasoning, feasible reflection, and extensive exploration methodologies.

2.2.1 Deep Reasoning for Long CoT

As shown by Figure 2, deep reasoning refers to the capability to perform deep and thorough logical analysis across multiple interconnected logical nodes, where Short CoT generally can never achieve. This capability is essential when tackling complex problems that require a massive number of logical deductions to arrive at a valid conclusion. To better define and understand deep reasoning, we frame it as a capability that primarily relaxes the first constraint in Equation 1, as expressed by the following:

$$k \leq \mathcal{B}_s \rightarrowtail k \leq \mathcal{B}_l \wedge \mathcal{B}_s \ll \mathcal{B}_l, \tag{2}$$

where $\mathcal{B}_l$ represents the upper boundary for Long CoT reasoning, which can accommodate much more intricate reasoning nodes compared to the smaller boundary $\mathcal{B}_s$ for Short CoT. The larger boundary $\mathcal{B}_l$ alleviates issues related to insufficient depth in reasoning, thereby reducing the risk of generating unresolved answers or hallucinated responses in short-form reasoning.

Key Difference: Reasoning Depth

- Short CoT typically addresses a limited set of logical nodes, involving shallow reasoning, and struggles with problems requiring complex or intricate logical structures.
- Long CoT is designed to accommodate a significantly larger set of logical nodes, allowing for deeper logic and more thorough analysis during the reasoning process.

2.2.2 Extensive Exploration for Long CoT

As shown by Figure 2, Long CoT encourages branching out to extensively explore uncertain or unknown logical nodes, thereby expanding the potential set of reasoning paths. This exploration is particularly critical when solving problems characterized by ambiguity, incomplete information, or multiple possible solutions. More specifically, we describe how extensive exploration primarily addresses the relaxation of the second constraint in Equation 1, which can be formalized as follows:

$$j = 1 \Leftrightarrow \forall i \leq k, n_i \to n_{i+j} \rightarrowtail \exists m, \forall i, j \leq m, n_i \to n_{i+j}, \tag{3}$$

where the condition indicates that for a logical node n_i, there are m nodes that are explored in parallel. The acceptability of parallel exploration allows for a more systematic approach, enabling the exploration of previously unconsidered logical paths. This, in turn, helps maximize the understanding of all possible solutions, ultimately leading to the correct final answer.

> **Key Difference: Exploration of Logical Nodes**
>
> - Short CoT generally restricts exploration to a fixed set of logical nodes, often resulting in oversimplified reasoning and limited exploration.
> - Long CoT explores more various paths, including uncertain or uncharted areas, fostering more nuanced and comprehensive problem-solving.

2.2.3 Feasible Reflection for Long CoT

As shown by Figure 2, Long CoT involves revisiting previous logical nodes to verify their connections are valid and accurate, and then correcting them or selecting an alternative logical path. Formally, feasible reflection relaxes the third constraint in Equation 1, expressed as follows:

$$\forall i \neq j \leq k, n_i \neq n_j \rightarrowtail \exists i < j \leq k, n_i = n_j, \tag{4}$$

where this condition indicates that, for a logical node n_{j-1}, the subsequent node is not limited to the original next node $\hat{n}_j$. Instead, it may transition to n_i (i.e., the next reasoning node becomes n_j, where $n_j = n_i$). Practically, reflection implementation consists of two components:

Feedback Feedback refers to evaluating both overall and intermediate outputs for correctness and quality, also known as critique or verification. It can be derived from external sources, validation checks, or by reflecting on prior conclusions within the reasoning process. Formally, at each step n_i, a verification process $\mathcal{V}_i$ ensures the correctness, feasibility, and consistency of the reasoning. If an issue is identified, the process redirects n_i to the nearest correct node n_j, where $j < i$. This relationship is formalized as:

$$\mathcal{F}_i, n_j \leftarrow \text{Feedback}(\text{CoT}_L^i) \tag{5}$$

where $\text{CoT}_L^i = \{n_1, \ldots, n_i\}$ represents the current logic path up to the i-th logic node for Long CoT.

Refinement This involves adjusting intermediate steps or modifying the logical flow to correct inconsistencies or address gaps based on the given feedback. This process can be expressed mathematically as follows:

$$\widetilde{n}_{i+1} \leftarrow \text{Refine}(n_{i+1}|\text{CoT}_L^i, \mathcal{F}_i, n_j) \tag{6}$$

where $\widetilde{n}_{i+1}$ represents the refined version of the subsequent reasoning node n_{i+1}, according to the current logic CoT_L^i, feedback result $\mathcal{F}_i$, and previous reasoning node n_j.

Overall, incorporating reflection ensures that errors are identified and corrected promptly. This capability enables LLMs to quickly shift to alternative reasoning paths or correct their current trajectory. By doing so, error propagation is minimized, resulting in more accurate conclusions.

> **Key Difference: Feedback & Refinement**
>
> - Short CoT typically moves in a straightforward, non-repetitive manner from one node to the next, so that cannot correct their logic.
> - Long CoT allows for revisiting and revising earlier decisions by feedback and refinement, ensuring that optimizable and prior logical conclusions during the reasoning progress.

2.2.4 Unified Discussion

The Long CoT discussed here represents a unified reasoning system that integrates the three key characteristics outlined earlier. In contrast, during the Short CoT era, these abilities developed independently. As shown in Figure 2, early efforts focus on deep reasoning within traditional CoT paradigms, followed by the gradual introduction of reflective mechanisms based on human-designed pipeline. Exploration capabilities are then added, and these three components are merged to form the modern concept of Long CoT, which aims for a unified enhancement of reasoning.

The progression of Long CoT is gradual, rather than a sudden emergence through isolated models like O1 [208] and R1 [155]. Instead, it develops gradually. For example, earlier systems, such as ToT [668], enhance exploration but lack reflective mechanisms, disqualifying them as Long CoT [67]. While GoT [32] incorporates self-reflection based on ToT, its original model still lacked robust deep reasoning, preventing it from qualifying as Long CoT at that time. It is also notable that modern Long CoT systems, often neglect earlier technologies. This article addresses this gap by tracing the evolution of each capability, with the final section offering a comprehensive analysis of the integrated Long CoT system.

In summary, Long CoT and Short CoT represent distinct paradigms. Long CoT features a deeper, broader, and more reflective reasoning process, enhancing both accuracy and coherence. Short CoT, by contrast, is better suited to simpler, well-defined problems. This distinction highlights the scalability and adaptability of Long CoT, making it particularly effective for more complex reasoning.

> **Key Difference: Unified Capabilities**
>
> It is important to highlight that Long CoT integrates the following three distinct capabilities to perform complex reasoning. In contrast, traditional Short CoT optimization typically focuses on only one of these characteristics.

3 Long CoT Analysis & Evaluation

3.1 Analysis & Explanation for Long CoT

Research on Long CoT has significantly enhanced RLLMs by improving reasoning accuracy, reducing errors, and supporting dynamic decision-making. However, several phenomena and their corresponding mechanisms remain inadequately summarized. This section addresses key topics, including the mechanisms of Long CoT and their underlying principles [457, 45, 381, 455]. Methodologically, two main perspectives have emerged to explain Long CoT: (1) External Behavior Analysis (§ 3.1.1) and (2) Internal Mechanism Analysis (§ 3.1.2).

3.1.1 Long CoT External Behavior Analysis

The primary research stream focuses on explaining RLLM behaviors for Long CoT [20]. As illustrated in Figure 4, six key phenomena are identified and discussed for Long CoT in this part.

Long CoT Emergence Phenomenon   Research shows that contextual examples improve large models' generative abilities by guiding the formation of reasoning chains [707, 473, 297, 238, 369]. Wang et al. [543] demonstrate that these examples standardize reasoning chain generation relevant to the answers. In an experiment by Madaan et al. [374], removing problem-specific entities from contextual examples, while retaining only the logical structure, led to similar performance as using complete examples, highlighting the logical structure imitation of Long CoT during inference.

More recently, Stechly et al. [484] and Wang and Zhou [579] have shown that modifying the decoding process or designing specific prompts can activate the Long CoT within pre-trained models. They propose that CoT is embedded during pre-training and requires specific activation [658]. Further, Sadr et al. [455] focus the Long CoT source from the training data, and build on this with the notion of "model attribution", to specifically identify the training data most influential for specific outputs. Building on this, Guo et al. [155] and Xie et al. [622] investigate using rule-based reinforcement learning to directly activate Long CoT during pre-training, aiming to enhance performance [620]. Furthermore, Gandhi et al. [128] identify four key cognitive behaviors, including verification, backtracking, sub-target setting, and backlinking, which successfully facilitate Long CoT.
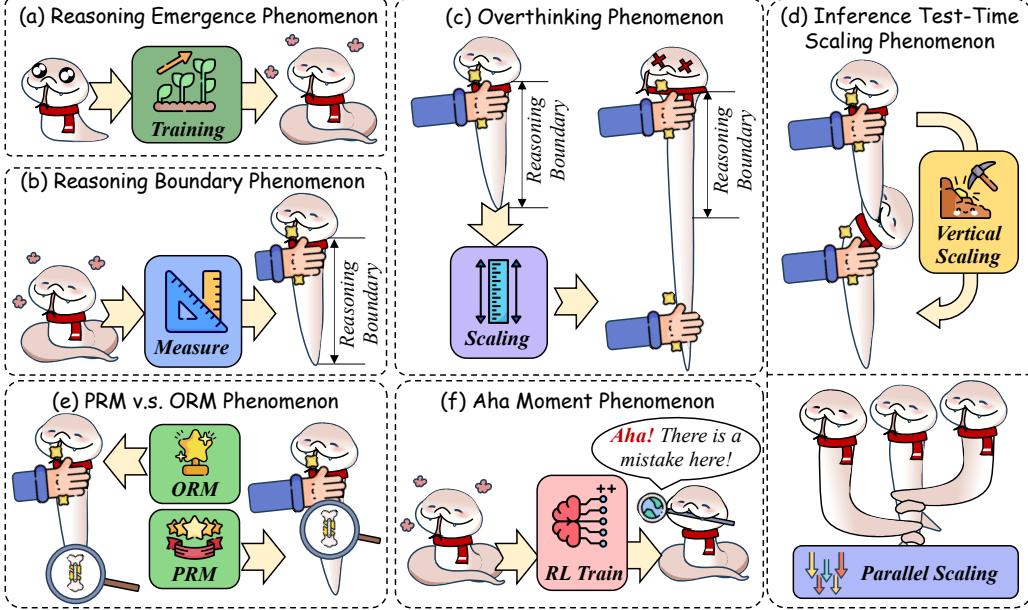
Figure 4: Analysis of the six classic phenomena of Long CoT external behavior: (a) emergence of Long CoT in current RLLMs; (b) reasoning boundaries and limitations of current Long CoT systems; (c) overthinking caused by scaling beyond RLLMs' reasoning boundaries, leading to performance decay; (d) test-time scaling, discussing mainstream scaling methods, corresponding scaling laws and their limitations; (e) use of process reward model (PRM) or outcome reward model (ORM); (f) exploration of the "aha" moment and its underlying causes.

Qwen [649] inherently demonstrates these behaviors, which can be easily triggered by rule-based reinforcement. In contrast, Llama [113] lacks these capabilities and thus requires example-based reinforcement learning to improve significantly.

Reasoning Boundary Phenomenon   Recent research has highlighted the upper bounds and limitations of RLLMs across various reasoning tasks [204, 191, 481, 178]. Specifically, Bi et al. [36] investigate these bounds in code generation, showing that RLLMs struggle with tasks that exceed certain complexity thresholds, especially when imitating Long CoT samples of varying complexity. In the context of upper-bound performance, Merrill and Sabharwal [383] and Li et al. [306] focus on single-step arithmetic tasks, concluding that model performance is constrained by input length. Moreover, Feng et al. [118] proposes a mathematical model indicating that fixed-size models cannot produce accurate numerical answers beyond specific limits. However, increasing the number of reasoning steps improves a model's ability to solve more complex problems.

Inspired by these explorations, Chen et al. [64] first define the "reasoning boundary" phenomenon and quantify these limits, showing that surpassing an RLLM's reasoning capacity leads to performance decline. Similarly, Zhou et al. [791] introduce GSM-Infinite, linking different upper limits to accuracy levels. Chen et al. [64] also examine the interaction between these boundaries across tasks of varying complexity, providing insights into the effectiveness of Long CoT strategies [759]. Moreover, Amiri et al. [9] propose a "tight lower bound" for Long CoT further guiding reasoning error reductions. Further, Baeumel et al. [22] suggest that due to its reliance on a single-digit lookahead heuristic, there are inherent boundaries in performing addition with multiple operands, which thus hinders the fundamental limitation of LLMs in scaling to more complex numerical reasoning.

Overthinking Phenomenon   Research has highlighted the overthinking phenomenon [73, 227, 404, 96, 251], where performance improves with longer reasoning chains up to a threshold, after which it declines. In contrast, Xie et al. [622] and Ma et al. [370] find no significant correlation between reasoning length and accuracy. To explain this, one line of research suggests that Long CoT strategies, like avoiding "snowball errors" [126]. Alternatively, Chen et al. [64], Wolf et al. [602] highlight a performance drop when the reasoning boundaries are exceeded, providing an explanation for the

overthinking phenomenon. This suggests that reasoning length and logical complexity should be kept below a certain boundary [756]. Building on this, Wu et al. [611] mathematically determine the feasible reasoning length for Long CoT. Finally, Chen et al. [66] introduces Faraday's law of Long CoT, which accurately predicts and controls performance.

Inference Test-Time Scaling Phenomenon Recent advances in inference-time scaling algorithms [364, 598] have garnered significant attention, particularly for their ability to extend reasoning length and improve performance [364]. Specifically, Brown et al. [40] identify a phenomenon called "Large Language Monkeys," in which a series of reasoning tasks show that with enough trials, a correct result can be achieved. Additionally, O1 [208] and R1 [155] demonstrated that directly scaling the length of model inference improves final performance.

To understand inference test-time scaling, we will discuss these two paradigms: (1) *Vertical Scaling:* Vertical scaling involves increasing the reasoning path length. While this can enhance performance, studies by Jin et al. [227] show that, beyond a certain point, longer reasoning paths can degrade performance due to error accumulation. They suggest an optimal path length that depends on the model's capabilities and task complexity [12, 463]. Furthermore, Chen et al. [64] and Wu et al. [611] explain that excessive exploration lengths beyond the RLLM's inherent reasoning boundary lead to performance decay, which guides RLLMs for deeper reasoning capabilities [25]. (2) *Parallel Scaling:* Parallel scaling involves performing multiple reasoning steps and verifying the results. While it shows promise, Parashar et al. [411] and Wang et al. [584] argue that simply increasing inference time does not guarantee improved performance. Wu et al. [610] show that the computational FLOPs N of inference are correlated with the lower bound of performance error, which scales with $\log N$. Additionally, Chen et al. [66] establish an upper bound for parallel scaling, showing that RLLMs cannot exceed Pass@k verification through various verifiers. They further argue that sampling optimization cannot exceed the model's internal reasoning limitations, demonstrating that for N samples, accuracy is proportional to $\frac{m}{(k/\log N+b)^2}$, where m, n, and b are model-dependent constants.

PRM v.s. ORM Phenomenon As RLLMs evolve, it is crucial to differentiate between process supervision and outcome supervision, two key reinforcement learning approaches for complex reasoning tasks, which is essential [632, 123]. While process supervision is intuitively advantageous for long-term reward assignments, the exact relationship between the two remains unclear. It is commonly believed that process supervision is more challenging due to the trajectory-level coverage problem, which demands significant effort to collect fine-grained supervision data [769, 478]. Additionally, PRM faces the issue of reward hacking [10, 101, 403, 24], where agents exploit flaws in the reward function to produce unintended behaviors [155]. Addressing this to surpass rule-based reward systems has become an important research area [155, 622, 419]. Furthermore, Lampinen et al. [260] and Tan [497] establish a causal link between intermediate steps and final answers in qualitative experiments. Building on this, Jia et al. [216] demonstrate that, under the standard data coverage assumption, reinforcement learning with outcome supervision is not statistically more challenging than process supervision, aside from polynomial factors.

Aha Moment Phenomenon Earlier, Guo et al. [155] demonstrated that direct RL using rule-based rewards can trigger the aha moment, fostering natural self-reflection without supervision. Following this, Team [507], Xie et al. [622] replicated this phenomenon. Further, Zhou et al. [782] and Meng et al. [382] further extend this phenomenon to multimodal scenarios. However, Liu et al. [346] argue that the aha moment may not emerge in R1-Zero-like training. Instead, they observe that self-reflection patterns, such as superficial self-reflection (SSR), appear at epoch 0, the stage of base models. In this case, self-reflections do not necessarily lead to correct answers. Upon closer examination of R1-Zero training via RL, they find that the increasing response length results not from self-reflection, but from RL optimizing well-designed rule-based rewards.

3.1.2 Long CoT Internal Mechanism Analysis

The second stream of research investigates the internal mechanisms of Long CoT-related RLLMs.

Reasoning Interal Mechanism Recent studies have explored the internal mechanisms underlying the coherent rationale outputs of Long CoT, with particular emphasis on attention mechanisms [476, 446]. These studies primarily examine neural substructures in RLLMs, framing CoT reasoning from a white-box perspective [583, 693, 159, 114]. Weston and Sukhbaatar [601] introduces the concept of

System 2 Attention (S2A), which demonstrates Long CoT generation by selectively focusing attention on relevant information. Additionally, Li et al. [290] explore gradient distributions between direct output and Long CoT layers, revealing that Long CoT layers help maintain stability by distinguishing relevant from irrelevant reasoning. Finally, Zhang et al. [747] conceptualize RLLMs as finite state automata, offering further insight into how internal dynamics influence external behavior. Despite Short CoT's struggles with self-correction, Bertolazzi et al. [31] show that these models rely on *consistency heads* (attention heads) to assess the alignment of numerical values in arithmetic solutions through internal shortcuts.

Knowledge Incorporating Mechanism   Current RLLMs primarily focus on mathematics and coding but have shown potential for generalization to other knowledge-rich domains, sparking growing interest in the mechanism for integrating domain-specific knowledge into Long CoT [608, 622]. Prystawski et al. [429] suggest that generative models store entity knowledge learned during pre-training independently, with the reasoning process in Long CoT linking this knowledge across entities. Radha and Goktas [444] recently introduced the Probabilistic Mixture Model (PMM), which categorizes model outputs into reasoning, memorization, and guessing. They also propose an Information-Theoretic Consistency (ITC) analysis to quantify the relationship between model confidence and strategy selection. Additionally, Jin et al. [228] define "Concept Depth" as the lowest layers at which complex concepts are understood, demonstrating varying levels of knowledge integration in RLLMs. Ou et al. [402] examine RLLM knowledge internalization through knowledge loop evolution, arguing that new knowledge acquisition is shaped by its connection to existing knowledge, with the loop evolving from formation to optimization and from shallow to deep.

3.2   Long CoT Evaluations

3.2.1   Metrics

In benchmarking, various metrics assess model performance across reasoning tasks, each focusing on different aspects of reasoning ability. These metrics evaluate both RLLMs' effectiveness in achieving desired outcomes and their learning efficiency. As a result, metrics for RLLMs have gained increasing attention in recent research. For mathematical or code-related tasks, three key metrics are commonly used: `Accuracy`, `Pass@k`, and `Cons@k` based on regex extraction:

- `Accuracy` measures the proportion of correct outputs.
- `Pass@k` evaluates the likelihood of generating at least one correct solution within k attempts.
- `Cons@k` assesses consistency by determining the model's ability to consistently produce correct or logically coherent solutions across multiple attempts.

In scientific or commonsense question-answering tasks, evaluation often uses `Exact Match (EM)` and `Accuracy` based on regex extraction, where `EM` determines whether the model's output exactly matches the expected solution.

For feedback techniques like ORM or PRM, `Rank` and `Best-of-N` metrics are often used:

- `Rank` measures whether the reward model correctly prioritizes the best reasoning processes from the top k candidates.
- `Best-of-N` selects the highest-scoring solution from N generated reasoning trajectories, indirectly measuring the reward model's effectiveness based on final outcomes.

3.2.2   Decoding Strategies

Decoding strategies are essential for controlling the inference process. Common approaches include `Greedy Decoding`, `Beam Search`, and `Major@k`. Both `Greedy Decoding` and `Beam Search` limit the sampling range to reduce randomness, guiding the model toward more consistent outputs. In contrast, `Major@k` identifies the most reliable solution by selecting the one with the highest consistency from a set of k candidate solutions.

3.2.3   Benchmarks

In the realm of Benchmarks, the focus lies on assessing the reasoning capabilities of RLLMs across diverse domains. There are two primary categories: (1) Outcome Benchmarks, which focus on the

holistic view of Long CoT reasoning, and (2) Process Benchmarks, which concentrate on the local view of the Long CoT process or individual capabilities.

Outcome Benchmarks   In the realm of Outcome Benchmarks, the first focus lies on evaluating the logical reasoning capabilities:

- **Complex Mathematics:** A central focus in complex mathematics is evaluating benchmarks like GSM8K [95] and MATH [172], which assess basic mathematical problem-solving abilities [788, 775]. Recent additions, such as AIME 2024 [6], AIME 2025 [401], MATH-500 [319], AMC 2023 [7], and OlympiadBench [165], expand the evaluation of LLM performance in mathematics. Moreover, Putnam-AXIOM [152] and FrontierMath [140] introduce more complex problems that challenge future reasoning systems. Additionally, ThinkBench [197] and MATH-Perturb [195] focus on robust evaluation for Long CoT.

- **Complex Coding:** Complex coding benchmarks are also vital, with competitions like Codeforces, SWEbench [224], CodeContests [305], and LiveCodeBench [210] evaluating LLM coding and problem-solving skills. Notable additions such as MHPP [99], ProBench [655], HumanEval Pro, MBPP Pro [694], and EquiBench [592] enhance the scope and complexity of coding challenges. Moreover, some studies have explored applying these benchmarks in real-world code development scenarios for automatic code generation and evaluation [167, 530].

- **Commonsense Puzzle:** Commonsense puzzle benchmarks, including BIG-Bench Hard [495] and ZebraLogic [320], assess models' ability to reason about commonsense situations. The ARC [87] is often viewed as a challenging commonsense-based AGI test. JustLogic [63] further contributes to the evaluation of deductive reasoning and commonsense problem-solving.

The second focus area concerns Knowledge Benchmarks, essential for evaluating a model's capability in complex reasoning across various domains:

- **Scientific Reasoning:** Scientific Reasoning benchmarks, such as GPQA Diamond [451], MMLU-Pro [585], and SuperGPQA [111], assess multi-domain reasoning in fields like chemistry, biology, and physics. These benchmarks test models' ability to not only accumulate knowledge but also integrate it for problem-solving. Humanity's Last Exam (HLE) [422] further challenges models by requiring deep interdisciplinary reasoning across scientific disciplines. Further, Chung et al. [94] propose TPBench to evaluate the effectiveness of RLLMs in solving theoretical physics problems.

- **Medical Reasoning:** In the realm of Medical Reasoning, the need for complex, domain-specific, and accurate reasoning is paramount [764, 719, 637]. Benchmarks, such as MedQA[225], JAMA Clinical Challenge [55], and Medbullets [55], simulate diagnostic and treatment decision-making processes, reflecting real-world medical practice. These benchmarks evaluate a model's handling of medical knowledge and reasoning, from diagnosis to treatment planning. Additionally, MedX-pertQA [801] introduces a comprehensive evaluation framework combining text and multimodal data, specifically assessing AI's reasoning capabilities in healthcare.

3.2.4   Process Evaluations

Deep Reasoning Benchmarks   Recent progress in RLLMs underscores the need for specialized benchmarks to evaluate their deep reasoning abilities in Long CoT [266]. Notably, Lin et al. [320] introduces ZebraLogic, a framework for assessing logical reasoning, especially in complex non-monotonic scenarios. Similarly, BigGSM [64] and GSM-Ranges [472] focus on perturbing numerical values to test logical and arithmetic reasoning in edge cases beyond the models' training distribution. ROSCOE [142], ReCEval [426], and DiVeRSe [304] are designed to assess each step in the deep reasoning process during Long CoT tasks.

Exploration Benchmarks   Several studies assess RLLMs' exploration capabilities in Long CoT tasks. Specifically, Sys2Bench [411] evaluates the exploration and scaling abilities of RLLMs, emphasizing generalization across diverse tasks. BanditBench [397] extends this by testing model performance in interactive environments, offering insights into practical applications. Additionally, Heyman and Zylberberg [173] introduce a graph coloring problem to assess reasoning and spatial exploration in complex problem-solving scenarios.

Reflection Benchmarks   Reflection benchmarks measure RLLMs' ability to identify, reflect upon, and correct errors in Long CoT reasoning. These benchmarks fall into two categories: feedback

and refinement. (1) ***Feedback Benchmark:*** These benchmarks assess the ability of LLMs to detect errors and respond to feedback for improvement. For example, Lambert et al. [259] introduces RewardBench to evaluate RLLMs' reward capabilities. This framework is extended by Yasunaga et al. [672] and Zhang et al. [720] to include multimodal and code contexts, respectively. Benchmarks such as ProcessBench [769], PRMBench [478], MR-Ben [716], and DeltaBench [171] focus on error detection and correction across various tasks at the step level. Additionally, ReaLMistake [232] and JudgeBench [498] address more real-world error evaluation. (2) ***Refinement Benchmark:*** These benchmarks focus on error correction in complex tasks. CriticBench å[324] assesses critique-correction capabilities, while ErrorRadar [645] specializes in multimodal error detection, particularly in mathematics. FinerReason [51] introduces a commonsense puzzle for broader feedback and refinement evaluations. Medec [1] adapts error correction to healthcare, addressing medical issues.

3.2.5 Advanced Evaluation

Agentic & Embodied Reasoning   Agentic and Embodied reasoning requires models to demonstrate an understanding of real-world interactions, tool use, and adaptive reasoning in response to change. To assess real-world understanding, Wang et al. [568] introduce a benchmark that evaluates agents' ability to reason about physical concepts. Zhang et al. [745] extend this by assessing agents' interactions with real-world physics. Additionally, realistic tasks often demand complex planning and tool usage, necessitating benchmarks to evaluate agent reasoning. These benchmarks assess agents' abilities to navigate and complete tasks in digital environments. Building on this, Huang et al. [191] propose a framework for evaluating decision-making in multi-agent, competitive settings. Nath et al. [393] introduce ToolComp, a benchmark designed to evaluate multi-step tool-use reasoning. To analyze adaptive reasoning in the face of real-world change, OSWorld [623], CogAgent [177], Mobile-Agent-E [589], WebShop [667], WebArena [789], and WebGames [522] assess AI systems across domains such as operating systems, mobile GUIs, browser tasks, and interactive entertainment [770, 559]. Hu et al. [186] present Text2World, which evaluates agents' ability to generate interactive environments from text to test agent adaptability [695].

Multimodal Reasoning   Multimodal reasoning refers to a system's ability to integrate and reason across diverse input types, including text, images, and occasionally code or graphs. This capability is crucial for solving complex problems that require information from diverse formats.

- **Complex Mathematics:** Mathematical reasoning often integrates both textual and visual components, such as equations, graphs, or diagrams. Specifically, challenges like MathVista [352], MathVision [561], MathVerse [739], M3CoT-Math [65], CMMaTH [309], EnigmaEval [546], CoMT-Geometry [85], and PGPS9K [737] aim to advance multimodal reasoning in mathematics, improving the evaluation of multimodal Long CoT logic.

- **Complex Code:** The second area of focus involves code-related reasoning, where systems interpret textual descriptions and code snippets. Benchmarks like HumanEval-V [728], Code-Vision [550], and Plot2Code [603] evaluate systems' capabilities to generate or interpret code from natural language and multimodal inputs for assessing systems that integrate natural language processing with programming tasks.

- **Complex Science:** This area involves integrating scientific texts with related diagrams or experimental data. Benchmarks like ScienceQA [351] and M3CoT-Science [65] evaluate how well models combine science information with Long CoT reasoning across various scientific domains.

- **Commonsense Puzzle:** This area focuses on commonsense reasoning, where systems combine reasoning cues and images to make deeper conclusions. Chen et al. [65] introduce M3CoT-Commensense, which incorporates commonsense Long CoT reasoning for complex multimodal interactions. Additionally, Wang et al. [544] propose two benchmarks: Clue-Visual Question Answering (CVQA), which tests visual comprehension through three task types, and Clue of Password-Visual Question Answering (CPVQA), which features two task types focusing on the interpretation and application of visual data.

AI for Research   Recent advancements in AI have significantly advanced scientific research [787, 581, 144], with platforms like SciWorld [568] improving the research process. Simultaneously, Pricope [428] and Chan et al. [48] introduce a machine-learning platform to evaluate the potential of RLLMs in automating experiments. Several studies also examine RLLMs' ability to generate innovative research ideas. For instance, Si et al. [474] conduct evaluations with over 100 NLP researchers to assess RLLMs' creativity, revealing notable limitations [287, 606, 512]. Additionally,
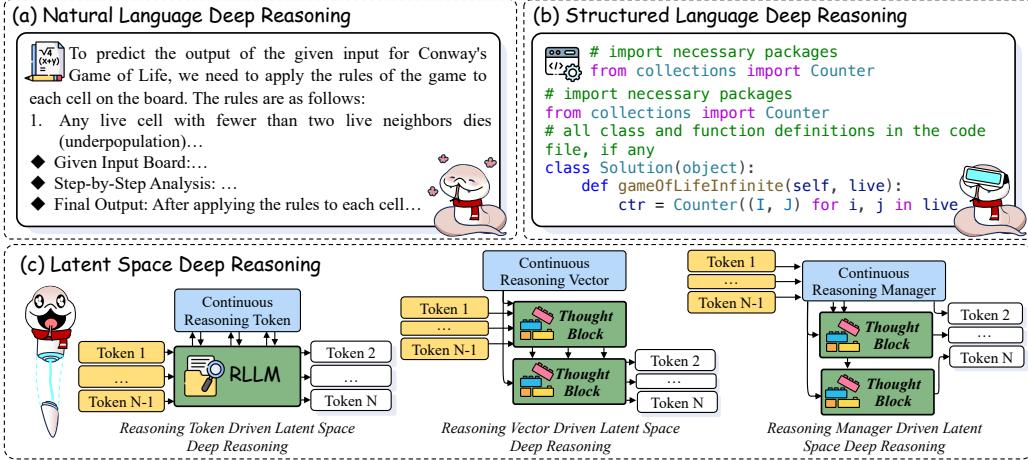
Figure 5: Three main categories of deep reasoning formats: natural language, structured language, and latent-space reasoning (subdivided into token-, vector-, and manager-driven latent reasoning), with examples drawn from Li et al. [285].

Li et al. [310] introduce SolutionBench, a benchmark for assessing systems' ability to generate feasible solutions for complex engineering problems.

4 Deep Reasoning for Long CoT

Deep reasoning capabilities primarily require profound depth and comprehensiveness in cognitive and reasoning processes. In the absence of such capabilities, RLLMs suffer significant performance declines [542, 587]. Current approaches to enhancing deep reasoning can be classified into two main categories: (1) ***Deep Reasoning Format*** (§ 4.1), which involves utilizing various reasoning execution formats, and (2) ***Deep Reasoning Learning*** (§ 4.2), which focuses on enabling the model to learn and strengthen its deep reasoning abilities.

4.1 Deep Reasoning Format

As illustrated in Figure 5, deep reasoning formats can be categorized into three main types: natural language (§ 4.1.1), structured language (§ 4.1.2), and latent-space reasoning (§ 4.1.3), the latter of which is further subdivided into token-, vector-, and manager-driven latent reasoning. The reasoning performance across these formats is presented in Table 1.

4.1.1 Natural Language Deep Reasoning

Traditionally, researchers have sought to adapt natural language for intuitive and free-flowing deep reasoning [594, 781, 204, 435, 749, 548]. Early work by Wei et al. [594] demonstrated that the use of natural language Long CoT significantly enhances the reasoning capabilities of RLLMs. Further, the Natural Program framework [327] allows RLLMs to engage in deeper natural language reasoning by ensuring a more structured and rigorous logical analysis. More recently, CodeI/O [285] has introduced a technique that reorganizes code-based reasoning patterns into natural language formats, further boosting the reasoning potential of RLLMs.

4.1.2 Structured Language Deep Reasoning

Structured language deep reasoning encompasses various approaches designed to program [70, 330, 483, 418, 132, 599] or symbolic language [425, 104, 321, 264, 654, 424] format for enhanced deep reasoning. In this context, most studies focus on utilizing code to better enhance the mathematical reasoning capabilities [275, 76, 684]. Xu et al. [631] propose a neural-symbol self-training framework guided by the environment, addressing both the scarcity of symbolic data and the limitations of symbolic processing in LLMs. Additionally, Liao et al. [314] present SKIntern, which refines symbolic RLLMs through curriculum learning and linear attenuation, enabling the internalization of

Model	Base Model	GSM8k	MATH	GPQA	OlympiadBench	LiveCodeBench
Latent Space Deep Reasoning						
No-CoT [100]	Mistral-7B [217]	38.0	-	-	-	-
SQ-VAE [575]	Llama-2-7B [529]	40.0	7.0	-	-	-
RecurrentBlock-3.5B [136]	-	42.1	-	-	-	-
ICoT-SI [100]	Mistral-7B [217]	51.0	-	-	-	-
Natural Language Deep Reasoning						
Self-Rewarding [80]	Llama-2-7B [529]	40.0	10.7	-	-	-
Llama-3.1-8B [113]	-	56.7	20.3	-	-	
MetaMath [688]	Llama-2-7B [529]	66.5	-	-	-	-
OVM [685]	Llama-2-7B [529]	73.7	-	-	-	-
NuminaMath-7B-CoT [283]	-	75.4	55.2	-	19.9	-
Qwen2-7B [648]	-	79.9	44.2	-	21.3	-
Qwen2-Math-7B [650]	-	80.4	50.4	-	38.2	-
Internlm2-math-plus-7B [682]	-	84.0	54.4	-	18.8	-
OMI2 [285]	Qwen2.5-Coder-7B [202]	84.1	72.3	36.2	-	27.2
Llama-3.1-70B [113]	-	85.5	41.4	-	-	-
CODEI/O++ [285]	Qwen2.5-Coder-7B [202]	85.7	72.1	40.6	-	29.1
CODEI/O [285]	Qwen2.5-Coder-7B [202]	86.4	71.9	43.3	-	28.5
WI [285]	Qwen2.5-Coder-7B [202]	87.0	71.4	39.1	-	26.0
WI (Full) [285]	Qwen2.5-Coder-7B [202]	87.0	71.1	42.9	-	27.6
OMI2 (Full) [285]	Qwen2.5-Coder-7B [202]	88.5	73.2	40.9	-	28.4
DeepSeekMath-7B-RL [466]	-	88.2	51.7	-	19.0	-
Llama-3.1-405B [113]	-	89.0	53.8	-	-	-
CoMAT [263]	GPT-4 [3]	93.7	-	40.4	-	-
CoT [448]	GPT-4 [3]	94.5	-	41.8	50.2	-
FCoT [363]	GPT-4 [3]	95.0	-	-	-	-
Qwen2.5-Math-7B-Instruct [650]	-	95.2	83.6	-	41.6	-
MathPrompter [204]	GPT-4 [3]	95.6	-	-	-	-
Qwen2.5-Math-72B-Instruct [650]	-	95.9	85.9	-	49.0	-
DeepSeek-R1-Distill-Qwen-7B [155]	-	-	92.8	-	49.1	37.6
DeepSeek-R1-Distill-Qwen-32B [155]	-	-	94.3	-	62.1	57.2
Structured Language Deep Reasoning						
STaR [707]	Llama-2-7B [529]	58.2	16.0	-	-	-
ENVISIONS [631]	Llama-2-7B [529]	59.0	19.0	-	-	-
MAmmoTH [704]	Code-Llama-7B [453]	59.4	-	-	-	-
MathCoder-CL [562]	Code-Llama-7B [453]	67.8	30.2	-	-	-
ToRA-Code [146]	Llama-2-7B [529]	72.6	-	-	-	-
Brain [76]	Code-Llama-7B [453]	74.0	-	-	-	-
DeepSeek-Coder-7B [154]	-	77.4	44.4	-	-	-
SIaM [684]	Qwen-2-Math-Base	81.5	50	-	-	-
OC-SFT-1 [285]	Qwen2.5-Coder-7B [202]	86.7	70.9	37.7	-	27.5
PyEdu [285]	Qwen2.5-Coder-7B [202]	85.8	71.4	40.9	-	25.8
Qwen2.5-Math-7B-Instruct [650]	-	94.6	85.2	-	55.6	-
Qwen2.5-Math-72B-Instruct [650]	-	95.8	88.1	-	60.6	-
QuaSAR [448]	GPT-4 [3]	96.5	-	55.4	44.6	-
MathDivide [483]	GPT-4 [3]	96.8	-	-	-	-

Table 1: Performance of various deep reasoning formats, sorted primarily by GSM8K scores. "-" indicates that the paper did not report this score.

symbolic knowledge with fewer examples, reducing computational costs, and accelerating inference. Furthermore, Ranaldi et al. [448] introduce QuaSAR, a CoT variant that directs LLMs to operate at higher abstraction levels through quasi-symbolic reasoning, thus improving natural language reasoning and providing more precise structural representations.

4.1.3 Latent Space Deep Reasoning

Latent space deep reasoning encompasses techniques designed to enhance the reasoning abilities of LLMs by leveraging operations within continuous latent spaces [481, 100]. These approaches can be categorized into three main paradigms: (1) ***Reasoning Token-Driven Latent Space Deep Reasoning:*** Early work [575, 708] introduce the concept of "planning tokens" or "thought tokens" to guide reasoning within latent spaces. Further, Coconut [162] expands on this through the maintenance of multiple alternative reasoning paths, increasing both complexity and efficiency [748, 496]. At the extreme, Heima [469] condenses the entire Long CoT process into a single token, yielding substantial computational savings. (2) ***Reasoning Vector Driven Latent Space Deep Reasoning:*** Building on the previous paradigm, LTM [250] conceptualizes the layers of LLMs as "thought blocks" and introduces the concept of "thought vectors" for each layer. This approach allows for the scaling of test-time computations by implicitly performing reasoning within the latent space through recurrent depth. (3) ***Reasoning Manager Driven Latent Space Deep Reasoning:*** Inspired by these, Geiping et al. [136] and Saunshi et al. [459] propose a mechanism similar to a continuous reasoning manager, which iteratively governs a trained "recurrent block" as a recurrent "thought block". This method
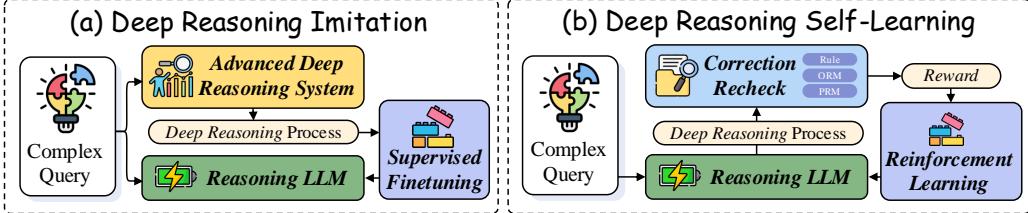
Figure 6: The different learning strategies of deep reasoning learning, including deep reasoning imitation of the data from advanced deep reasoning systems, like advanced RLLMs, MCTS, etc.; deep reasoning self-learning from preference-based RL by implicit reward.

integrates deeper model layers during reasoning, enhancing performance without needing specialized training data, and even outperforming larger RLLMs. Additionally, ITT [77] leverages the original transformer layer as a recurrent "thought block," selecting key tokens via adaptive token routing and controlling reasoning depth with residual thinking connections, enabling more efficient processing of critical tokens.

4.2 Deep Reasoning Learning

Insufficient deep reasoning in RLLMs can significantly degrade performance [542, 587]. As a result, research has focused on improving reasoning through training. Supervised fine-tuning (SFT) [741] stabilizes model outputs by serving as a memory process, while reinforcement learning (RL) enables generalization and self-learning [155, 91]. Recent studies for deep reasoning learning have explored using SFT to imitate advanced reasoning in RLLMs and applying RL to enhance self-improvement in reasoning. As illustrated in Figure 6, this section outlines two key approaches to improve deep reasoning: (1) *Deep Reasoning Imitation* (§ 4.2.1), which involves learning reasoning from human-annotated or distilled data through SFT, and (2) *Deep Reasoning Self-Learning* (§ 4.2.2), where models improve reasoning through preference-based RL with implicit rewards. The performance of these methods is shown in Table 2.

4.2.1 Deep Reasoning Imitation

Deep reasoning in RLLMs can be effectively achieved by mimicking advanced reasoning systems, such as human reasoning [390, 43, 81, 286], advanced RLLMs [155, 41, 670, 262, 72], and Scaling-augmented RLLMs [293, 702, 420, 792]. This approach enables the model to learn complex reasoning patterns and generalize across tasks [657]. Specifically, (1) *Imitation from Human:* Earlier, Cobbe et al. [95] first propose the deep reasoning imitation paradigm using human examples. ALT [390] improves RLLM reasoning by generating larger datasets of human-annotated logical templates, which fosters deeper reasoning. To enhance diversity, EIT [43] promotes simpler human-generated plans, while LLMs contribute more nuanced reasoning, facilitating collaboration between human input and AI. (2) *Imitation from Advanced RLLMs:* A body of work utilizes zero-shot prompting to guide large teacher RLLMs in generating reasoning rationale, which is then used to fine-tune smaller RLLMs, marking the beginning of deep reasoning imitation [174, 246, 679]. Additionally, AceMath [347] applies few-shot prompting to distill Long CoT samples from advanced LLMs, followed by multi-stage quality-guided SFT to enhance performance. Chen et al. [76] separate the data synthesis process into planning and reasoning stages, thereby improving reasoning quality. DART-Math [524] effectively distills complex queries requiring deeper reasoning during synthesis, advancing deep reasoning capabilities. (3) *Imitation from Scaling-augmented RLLMs:* Earlier, Bansal et al. [26] enhance data quality by scaling the sampling size and length, boosting imitation performance. Yang et al. [650] and Zhao et al. [761] further improve data quality by scaling sampling and selecting samples through a reward model. Additionally, Li et al. [293] identify optimal deep reasoning paths through MCTS, advancing imitation effectiveness.

Recent studies [200, 385] show that distilling knowledge from advanced RLLM APIs like O1 [208] and R1 [155] significantly enhances the performance of smaller LLMs. This method, employing supervised fine-tuning, boosts model performance on complex mathematical reasoning tasks, sometimes surpassing the teacher models' performance. Building on these findings, LIMO [676], S1 [391], and RedStar [635] argue that a large number of imitation samples is unnecessary. They demonstrate

14

Model	Data Size	Base Model	GSM8K	MATH	MATH-500	AIME2024	GPQA	OlympiadBench
Deep Reasoning Imitation								
SFT [679]	200K	Llama-3.1-8B [113]	-	-	-	54.1	3.5	-
Retro-Enh [81]	14M	Llama-3-8B [113]	45.1	21.7	-	-	-	-
Query-Exp [81]	24M	Llama-3-8B [113]	51.3	23.1	-	-	-	-
Res-Div [81]	14M	Llama-3-8B [113]	53.0	23.2	-	-	-	-
MetaMath [524]	0.40M	Mistral-7B [217]	76.5	29.8	-	-	-	5.9
ALT-FLDx2 [390]	100K	Llama-3.1-70B [113]	83.3	24.4	-	-	-	-
EIT [43]	15K	Llama-2-70B [529]	84.1	32.5	-	-	-	-
MathScale [524]	2.0M	Mistral-7B [217]	74.8	35.2	-	-	-	-
Tutor-Amp [81]	11M	Llama-3-8B [113]	64.4	35.9	-	-	-	-
MMIQC [524]	2.3M	Mistral-7B [217]	75.4	37.4	-	-	-	9.4
VRT [524]	0.59M	Mistral-7B [217]	82.3	38.7	-	-	-	8.7
KPMath-Plus [524]	1.6M	Mistral-7B [217]	82.1	46.8	-	-	-	-
Llama-2-70B-Xwin-Math-V1.1 [274]	1.4M	Llama-2-70B [529]	90.2	52.5	-	-	-	16.3
DART-Math-Mistral-7B [524]	591K	Mistral-7B [217]	81.1	45.5	-	-	-	14.7
DART-Math-Llama-3-70B [524]	591K	Llama-3-70B [113]	89.6	56.1	-	-	-	20.0
Rejection Sampling [293]	197K	Qwen2.5-7B [649]	87.1	70.0	-	10.0	-	27.1
Evol-Instruct-7B [356]	905K	Qwen2.5-Math-7B [650]	88.5	-	77.4	16.7	-	-
FastMCTS [293]	288K	Qwen2.5-7B [649]	88.9	74.0	-	20.0	-	27.5
KPDDS-7B [199]	800K	Qwen2.5-Math-7B [650]	89.9	-	76.0	10.0	-	-
DeepSeek-R1-Distill-Qwen-7B [155]	800K	Qwen2.5-7B-Instruct [649]	91.7	-	91.6	43.3	-	-
Openmathinstruct-7B [526]	14M	Qwen2.5-Math-7B [650]	92.0	-	79.6	10.0	-	-
NuminaMath [676]	100K	Qwen2.5-Math-7B [650]	92.9	-	81.8	20.0	-	-
PromptCoT-DS-7B [761]	115K	DeepSeek-R1-Distill-Qwen-7B [155]	92.6	-	93.0	60.0	-	-
PromptCoT-Qwen-7B [761]	905K	Qwen2.5-Math-7B [650]	93.3	-	84.0	26.7	-	-
AceMath-7B-Instruct [347]	1.2M	Qwen2-Math-7B-Instruct [650]	93.7	83.1	-	-	-	42.2
AceMath-72B-Instruct [347]	1.2M	Qwen2.5-Math-72B-Instruct [650]	96.4	86.1	-	-	-	48.4
NuminaMath [676]	100K	Qwen2.5-32B-Instruct [649]	-	-	59.2	6.5	25.8	36.7
OpenThoughts [676]	114K	Qwen2.5-32B-Instruct [649]	-	-	80.6	50.2	42.9	56.3
Sky-T1-32B-Preview [510]	17K	Qwen2.5-32B-Instruct [649]	-	-	82.4	43.3	56.8	-
Journey Learning [200]	5K	Qwen2.5-Math-72B [650]	-	-	87.2	43.3	-	-
STILL-2 [385]	3.9K	Qwen2.5-32B-Instruct [649]	-	-	90.2	46.7	55.1	-
Bespoke-32B [256]	17K	Qwen2.5-32B-Instruct [649]	-	-	93.0	63.3	58.1	-
s1 [391]	1K	Qwen2.5-32B-Instruct [649]	-	-	93.0	56.7	59.6	-
DeepSeek-R1-Distill-Qwen-32B [155]	800K	Qwen2.5-32B-Instruct [649]	-	-	94.3	72.6	62.1	-
LIMO [676]	817	Qwen2.5-32B-Instruct [649]	-	-	94.8	15.8	66.7	66.8
Deep Reasoning Self-Learning								
DPO [203]	40K	DeepSeek-Math-7B-Base [466]	74.8	34.9	-	-	-	-
ReFT [203]	40K	DeepSeek-Math-7B-Base [466]	71.4	36.0	-	-	-	-
Self-Explore [203]	40K	DeepSeek-Math-7B-Base [466]	78.6	37.7	-	-	-	-
SimPO [509]	10K	Qwen2.5-7B-Instruct [650]	88.8	40.0	56.6	-	-	-
DPO [316]	11K	DeepSeek-Math-7B-Instruct [466]	-	48.7	-	-	-	-
TPO [316]	11K	DeepSeek-Math-7B-Instruct [466]	-	51.3	-	-	-	-
DPO [316]	11K	Qwen2-7B-Instruct [648]	-	54.3	-	-	-	-
TPO [316]	11K	Qwen2-7B-Instruct [648]	-	55.5	-	-	-	-
MCTS [53]	15K	DeepSeek-Math-7B-Base [466]	83.2	64.0	-	-	-	-
SBS [53]	15K	DeepSeek-Math-7B-Base [466]	84.1	66.3	-	-	-	-
FastMCTS+Branch-DPO [293]	152K	FastMCTS-7B [293]	89.9	75.4	-	20.0	-	29.6

Table 2: Performance of various deep reasoning learning methods, sorted primarily by Math or Math-500 scores. "-" indicates that the paper did not report this score.

that even a minimal set of samples can activate deep reasoning capabilities in foundational LLMs. For practical applications, Turtel et al. [532] showcase how these techniques can predict future events beyond a model's knowledge cutoff.

4.2.2 Deep Reasoning Self-Learning

While simple imitation can yield strong performance, current models still rely heavily on human annotations or outputs from more advanced models for both imitation and distillation. To address this limitation, recent research has focused on enabling more advanced reasoning through techniques like self-play and self-learning [663, 754, 292]. Specifically, self-learning methods can be classified into two paradigms, differentiated by their sampling strategies:

(1) *Self-Learning from Direct Sampling:* The earliest method, STaR [707], utilizes In-Context Learning (ICL) to sample deep reasoning results and uses the correctness of the final answer as an implicit reward for self-learning [175, 409, 410, 742, 588, 328]. Further, ReST [153] extends this by introducing a Grow-Improve paradigm, where self-generated reasoning is first annotated with rewards and then enhanced via offline RL algorithms. However, these approaches can be fragile, especially when the reward process lacks robustness. Inspired by the Expectation-Maximization (EM) algorithm, Singh et al. [475] propose a method that generates rewards and iteratively optimizes LLMs to achieve the best performance on a validation set, significantly improving robustness. To further strengthen the reward process, Hosseini et al. [179] introduce a method to adapt incorrect solutions, training a verifier to refine the reward process and improve self-learning quality. (2) *Self-Learning from Tree Search:* Early deep learning methods, such as EXIT [15], combined MCTS with deep neural networks for reinforcement learning, iteratively self-training the network to guide the tree search and enhance reasoning. Building on this, CPO [746] and TPO [316] align each step of Long CoT reasoning with the corresponding tree search path, using Tree of Thoughts (ToT) [668] preference information to support deeper reasoning [665, 203]. Li [302] propose Policy-Guided Tree Search (PGTS),

integrating RL with structured tree exploration for more efficient navigation of reasoning paths. Further developments, such as AlphaMath [53], AlphaLLM-CPL [578], and TongGeometry [724], refine MCTS behavior through stepwise trajectory pair extraction and curriculum preference learning, boosting LLM reasoning abilities [431, 295].

Takeaways: Imitation & Self-Learning

- Imitating deep reasoning from advanced RLLMs, and scaling-augmented methods like MCTS can help models learn complex reasoning patterns with fewer samples.
- Self-learning techniques, including reinforcement learning and tree search, allow RLLMs to enhance their reasoning abilities over time.
- The combination of imitation from advanced RLLMs and self-learning techniques strengthens RLLM reasoning, leading to strong performance on complex tasks.

---

5 Feasible Reflection for Long CoT

5.1 Feedback

Feedback refers to the process of providing evaluations of both overall outputs and the processes that lead to them, with the goal of assessing their accuracy and quality [280, 282, 595, 149]. This process, also referred to as critique or verification, can be executed using either natural language or structured data formats, which serve as the foundation for tree-search methods [79]. Specifically, as shown in Figure 7, feedback can be categorized into three distinct types: (1) Overall Feedback (§ 5.1.1); (2) Process Feedback (§ 5.1.2); (3) Hybrid Feedback (§ 5.1.3).

5.1.1 Overall Feedback

The overall feedback focuses on providing a global view of the entire process and results, rather than assessing each step individually. This feedback significantly enhances reasoning skills and reward modeling in reinforcement learning for RLLMs. Specifically, as shown in Figure 7 (a), the overall feedback can be categorized into three main sources: Outcome Reward Model, Rule Extraction, and Critic Models Feedback. The performance across these categories is summarized in Table 3.
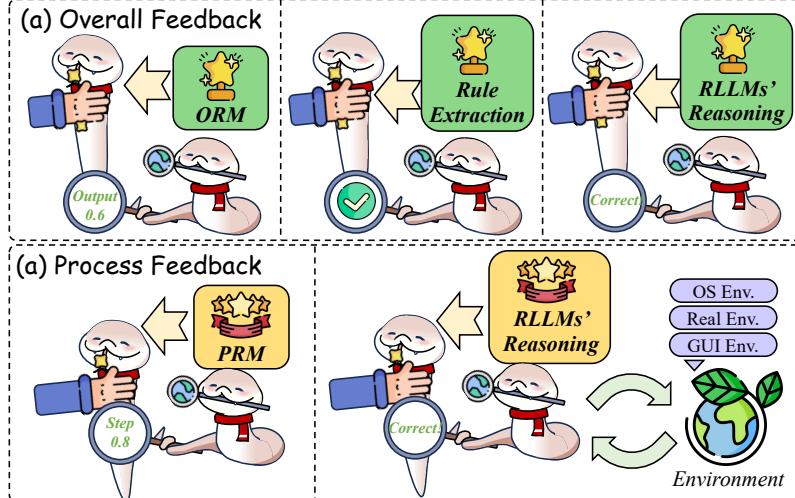
Figure 7: The feedback capabilities framework for feasible reflection consists of Overall Feedback and Process Feedback. Overall Feedback includes the Outcome Reward Model (ORM) in a value format, rule extraction for correctness judgment, and overall critic models based on RLLMs. Process Feedback includes the Process Reward Model (PRM) in a value format and step-level critic models, also based on RLLMs.

Model	Base Model	Chat	Chat_Hard	Safety	Reasoning	Overall
Critic Models						
GPT-4o-mini [3]	-	95.0	60.7	80.8	83.7	80.1
Llama3.1-70B-Instruct [113]	-	97.2	70.2	86.0	82.8	84.0
Llama3.1-405B-Instruct [113]	-	97.2	74.6	87.1	77.6	84.1
GPT-4 [3]	-	95.3	74.3	86.9	87.6	86.0
GPT-4o [3]	-	96.1	76.1	86.6	88.1	86.7
Gemini-1.5-pro [505]	-	92.3	80.6	87.9	92.0	88.2
Self-taught Evaluator [572]	Llama-3.1-70B-Instruct [113]	96.6	84.2	81.0	91.5	88.3
SFR-LLaMA-3.1-8B-Judge [566]	Llama-3.1-70B-Instruct [566]	95.5	77.7	86.2	95.1	88.7
SFR-NeMo-12B-Judge [566]	Mistral-NeMo-Instruct-12B [511]	97.2	82.2	86.5	95.1	90.3
SFR-LLaMA-3.1-70B-Judge [566]	Llama-3.1-70B-Instruct [113]	96.9	84.8	91.6	**97.6**	92.7
Skywork-Critic-Llama-3.1-70B [566]	Llama-3.1-70B-Instruct [113]	96.6	87.9	**93.1**	95.5	93.3
LMUnit [454]	Llama-3.1-70B-Instruct [113]	-	-	-	-	93.4
EvalPlanner [456]	Llama-3.1-70B-Instruct [113]	**97.5**	**89.4**	93.0	95.5	**93.9**
Outcome Reward Models						
tulu-v2.5-13b-uf-rm [207]	TULU-2-13B [206]	39.4	42.3	55.5	47.4	46.1
Prometheus-2-7B [247]	Mistral-7B-Instruct-v0.2 [217]	85.5	49.1	77.1	76.5	72.0
Prometheus-8x7b-v2 [247]	Mixtral-8x7B-Instruct [218]	93.0	47.1	80.5	77.4	74.5
Critic-RM-Rank [692]	Llama-3.1-70B-Instruct [113]	97.0	58.0	84.0	92.0	82.8
RM [485]	Llama-3.1-70B-Instruct [113]	98.3	74.5	83.8	88.0	86.4
SynRM [677]	Llama-3.1-70B-Instruct [113]	97.5	76.8	86.3	88.5	87.3
CLoud [14]	Llama-3-70B-Instruct [113]	98.0	75.6	87.6	89.0	87.6
FLAMe-RM-24B [538]	PaLM-2-24B [13]	92.2	75.7	89.6	93.8	87.8
SteerLM-RM 70B [590]	Llama-2-70B-chat [529]	91.3	80.3	90.6	92.8	88.8
Llama-3-OffsetBias-RM-8B [413]	Llama-3-8B-Instruct [113]	97.2	81.8	86.8	91.9	89.4
InternLM-20B-Reward [44]	InternLM2-8B-Instruct [44]	**98.9**	76.5	89.9	95.8	90.2
ArmoRM-Llama3-8B-v0.1 [553]	Llama-3-8B-Instruct [113]	96.9	76.8	92.2	**97.3**	90.8
Nemotron-4-340B-Reward [590]	Nemotron-4-340B [4]	95.8	87.1	**92.2**	93.6	92.2
Skywork-Reward-Llama-3.1-8B [331]	Llama-3.1-70B-Instruct [113]	95.8	87.3	90.6	96.2	92.5
Skywork-Reward-Gemma-2-27B [331]	Gemma-2-27B-it [506]	95.8	**91.4**	92.0	96.1	**93.8**

Table 3: Performance of various overall feedback methods, sorted primarily by Overall scores in RewardBench [259]. "-" indicates that the paper did not report this score.

Overall Feedback from Outcome Reward Model   Since many tasks cannot be directly evaluated using accuracy or other standard metrics, research has increasingly focused on Outcome Reward Models (ORM), which provide value-based rewards for more general and quantifiable feedback. In 2021, OpenAI [95] has proposed a "Gen-Verifier" paradigm, which uses a specialized ORM to evaluate the accuracy of generated rationales, showing significant progress in feedback capabilities [466]. Ji et al. [215] introduce a trained knowledge scorer to analyze hallucinations in the reasoning process, providing feedback to RLLMs and improving the accuracy of their outputs over time. Moreover, Generative Reward Models [736] use next-token prediction for overall feedback, which seamlessly integrates with instruction adjustments, leveraging test-time calculations to improve ORM feedback.

However, specifically trained ORMs are often costly and not sufficiently robust. Building on this, Self-Rewarding Language Models (SRLMs) [790] incorporate a self-consistency framework, optimizing feedback to improve model alignment and consistency. Yu et al. [692] introduce Critic-RM, combining RLLM-generated natural language criticism with corresponding feedback. This method filters high-quality feedback while jointly fine-tuning reward prediction and criticism generation, optimizing ORM performance.

Overall Feedback from Rule Extraction   Although ORM has achieved significant improvements, its accuracy still falls short of 100%, preventing it from outperforming rule-based answer correction feedback [668, 160]. Previous studies, such as STaR [707], ReST [153], and ReFT Trung et al. [531], have demonstrated that feedback based on final answer rewards is more effective than both PRM and ORM in mathematical scenarios [131]. Furthermore, Guo et al. [155] and Xie et al. [622] introduce a multi-stage RL framework that incorporates rule-based rewards, significantly enhancing both output accuracy and length while mitigating reward hacking through simple yet robust rules [24], such as format validation and result verification. In coding scenarios where direct rule-based feedback is difficult, AceCoder [709], O1-Coder [753], and VerMCTS [39] address this challenge by implementing an automated test-case synthesis pipeline, deriving rewards based on program performance [395, 145, 778]. Additionally, Ma et al. [372] propose an automated approach to training a test case generator, which alleviates the scarcity of test cases and demonstrates that increasing the number of test cases correlates with improved reward quality. Moreover, Ma et al. [371] decompose problem-solving into structured subtasks: file localization, function localization, line localization, and code editing generation, and applies multi-viewed rule-based rewards.

Overall Feedback from Critic Models   Research on feedback from Critic Models centers on detecting errors and biases through natural language feedback, also known as self-reflection or self-critique [231, 23, 452, 384, 571, 701]. This method has led to significant improvements across various tasks, particularly in self-correction [600, 772, 137, 121, 752]. Huang et al. [194] contend that traditional LLMs struggle to generate effective feedback without external signals, requiring the development of RLLMs with enhanced feedback capabilities [458]. As a result, many studies leverage RLLMs' error-identification strengths, often stemming from their pretraining phase, to improve feedback generation and correction [675].

Earlier, McAleese et al. [380] found that training RLLMs to learn self-critique and deep reasoning can further boost performance. Zhang et al. [744] propose a self-contrast mechanism that compares multiple perspectives, identifies differences, and summarizes insights to resolve inconsistencies. However, these methods often offer task-independent feedback. To address this, Hao et al. [161] introduce AutoRace, which tailors evaluation criteria for specific tasks. The Reversal of Thought (RoT) framework [698] introduces a novel paradigm combining reverse reasoning with self-reflection, helping models identify the limits of their knowledge and enhance reasoning efficiency. Furthermore, ACR [779] implements a scoring system for coding tasks, using LLM-as-a-Judge for quality assessment and LLM-as-a-Critic for critiquing low-quality code, improving consistency across benchmarks. Zheng et al. [771] integrate code execution error data and feedback from RLLMs to improve code generation performance. Liu et al. [342] present AGSER, a method using attention-guided self-reflection to address hallucinations by splitting input queries into attentive and non-attentive components. Finally, Saha et al. [456] introduce EvalPlanner, which separates feedback into planning and reasoning components for more streamlined expression using existing RLLMs.

5.1.2   Process Feedback

Techniques combine process feedback with MCTS or RL rewards to provide automated, step-by-step guidance, reducing the need for labor-intensive annotations while enhancing reasoning capabilities [534, 239]. These techniques can be categorized into two main types based on the source of feedback: process reward models (PRMs) and prompted LLMs. The performance comparison are mainly shown in Table 4.

Process Feedback from Process Rewarded Model   Recent studies highlight the significance of feedback in developing effective PRMs for complex reasoning tasks, particularly in a step-level view [88, 303, 366]. (1) *Process Annotated PRM Training:* Earlier, Lightman et al. [319] demonstrate that training process feedback with human-annotated data (PRM800K) surpasses outcome supervision in creating reliable reward models. However, this approach requires significant human effort. To address this, Wang et al. [567] introduce Math-Shepherd, a dataset that generates step-by-step supervision using a Tree Search-inspired method [52, 700]. Following this, methods like QwQ [517], Skywork-o1 [400], AceMath [347], and PRIME [97] adopt similar techniques to enhance PRM performance. Additionally, Zhang et al. [729] propose entropy regularization to improve model convergence. Rather than focusing solely on the first error step, Full-Step-DPO [636] assigns rewards for the entire reasoning chain, including error steps. VersaPRM [710] extends PRMs across multiple domains, broadening their applicability. Similarly, Gu et al. [148] and Zhang et al. [751] suggest training models with student preferences aligned to teacher preferences, ensuring effective preference distillation. (2) *Outcome Annotated PRM Training:* Alternative approaches, such as OVM [685], Implicit PRM [699], AutoPSV [350], and DVO [730], leverage outcome supervision or implicit feedback to train PRMs, reducing the need for extensive human-annotated data [627, 456]. UAS [687] incorporates uncertainty-aware value models [187] into feedback predictions. Additionally, Aurora [499] utilizes ensemble prompting strategies and reference answers for reverse verification, training stronger PRMs that better align with the Long CoT data distribution. Furthermore, PAV [462] suggests that rewards should reflect reasoning progress, as measured by changes in the likelihood of producing a correct future response before and after each step. Yang et al. [653], Lee et al. [267], Yoon et al. [683] extend these paradigms to the token level.

Process Feedback from Critic Models   As PRM training remains heavily dependent on manually annotated data, recent research has explored methods for enabling models to generate their natural language feedback to optimize performance [640]. These approaches fall into two primary categories: (1) *Model-Driven Feedback Reasoning:* Earlier work such as React [669] and Reflexion [471] enhances RLLMs with natural language feedback at each action and reasoning step [130, 89],

		ProcessBench				PRMBench		
		GSM8K	MATH	OlympiadBench	OmniMATH	Simplicity	Soundness	Sensitivity
Process Reward Models								
Qwen2.5-Math-7B-PRM [769]	Qwen2.5-Math-7B [650]	39.4	52.2	39.4	33.1	-	-	-
Math-Shepherd-PRM-7B [567]	Mistral-7B [217]	47.9	29.5	24.8	23.8	47.1	45.7	60.7
RLHFlow-PRM-Mistral-8B [103]	Mistral-7B [217]	50.4	33.4	13.8	15.8	46.7	57.5	68.5
RLHFlow-PRM-DeepSeek-8B [103]	DeepSeek-7B [35]	38.8	33.8	16.9	16.9	47.6	57.5	68.1
Skywork-PRM-1.5B [331]	Qwen2.5-Math-1.5B-Instruct [649]	59.0	48.0	19.3	19.2	33.6	28.6	48.8
Skywork-PRM-7B [331]	Qwen2.5-Math-7B-Instruct [649]	70.8	53.6	22.9	21.0	38.4	32.7	54.3
Qwen2-1.5B-PRM800k [491]	Qwen2-Math-1.5B-Instruct [650]	34.0	55.3	34.2	41.0	-	-	-
Qwen2-1.5B-Math-Shepherd [491]	Qwen2-Math-1.5B-Instruct [650]	48.9	34.1	9.8	13.7	-	-	-
Qwen2-1.5B-Epic50k [491]	Qwen2-Math-1.5B-Instruct [650]	55.6	36.1	20.2	30.0	-	-	-
Qwen2.5-Math-7B-PRM800K	Qwen2.5-Math-7B-Instruct [650]	68.2	62.6	50.7	44.3	-	-	-
Qwen2.5-Math-PRM-7B [769]	Qwen2.5-Math-7B-Instruct [650]	82.4	77.6	67.5	66.3	-	-	-
Universal-PRM-7B [499]	Qwen2.5-Math-7B-Instruct [650]	85.8	77.7	67.6	66.4	-	-	-
Critic Model								
Llama-3.1-8B-Instruct [113]	-	27.5	26.7	18.5	19.2	-	-	-
GPT-4o [3]	-	61.9	53.9	48.3	44.6	59.7	70.9	75.8
QwQ-32B-Preview [517]	Qwen2.5-32B-Instruct [649]	62.3	52.7	46.2	43.9	-	-	-
DeepSeek-R1-Distill-Qwen-14B [155]	Qwen2.5-14B-Instruct [649]	67.3	38.8	29.9	32.1	-	-	-
Dyve-14B [774]	DeepSeek-R1-Distill-Qwen-14B [155]	68.5	58.3	49.0	47.2	-	-	-
Qwen2.5-72B-Instruct [649]	-	76.2	61.8	54.6	52.2	-	-	-
SCRIT [500]	Qwen2.5-72B-Instruct [649]	80.2	60.0	32.5	27.8	-	-	-
o1-mini [208]	-	93.2	88.9	87.2	82.4	64.6	72.1	75.5
Llemma-PRM800k-7B [478]	Llemma-7B [21]	-	-	-	-	51.4	50.9	66.0
Llemma-MetaMath-7B [478]	Llemma-7B [21]	-	-	-	-	50.3	49.0	66.0
Llemma-oprm-7B [478]	Llemma-7B [21]	-	-	-	-	49.0	49.8	64.1
MATHMinos-Mistral-7B [129]	Mistral-7B [217]	-	-	-	-	51.4	54.4	66.5
ReasonEval-7B [618]	Llemma-7B [21]	-	-	-	-	55.5	63.9	71.0
ReasonEval-34B [618]	Llemma-34B [21]	-	-	-	-	51.5	63.0	73.1
Gemini-2.0-flash-exp [478]	-	-	-	-	-	62.7	67.3	75.4
Gemini-2.0-thinking-exp-1219 [478]	-	-	-	-	-	66.2	71.8	75.3

Table 4: Performance of various process feedback methods on ProcessBench [769] and PRM-Bench [478]. "-" indicates that the paper did not report this score.

improving decision-making in diverse tasks. Similarly, Step-DPO [257] uses RLLM to self-verify step-level positive and negative pairs for training through the DPO paradigm, achieving strong performance. Additionally, Sun et al. [492] propose a dynamic error classification framework that adapts based on model outputs, improving performance in mathematical reasoning tasks by addressing specific error patterns in math word problems. Furthermore, Xie et al. [625] and He et al. [168] iteratively apply MCTS to collect preference data, utilizing its forward-looking capabilities to decompose instance-level rewards into more precise step-level signals, thereby enhancing feedback accuracy. However, step-wise feedback often suffers from reliability issues, which can be mitigated by uncertainty quantification [681, 678], improving the reliability of step-wise verification in reward models for mathematical reasoning tasks. Moreover, Fu et al. [123] define the CoT Average Causal Effect (CACE) to capture causal relationships between steps, resulting in a causalized Long CoT where all steps are both correct and comprehensible. (2) ***Environment-Driven Feedback Reasoning:*** Given the increasing complexity of large models, there is growing interest in combining prompt-based LLMs with external environments to generate more interpretable and controllable feedback. For example, ORPS [696] and Drori et al. [108] minimize dependence on human annotations by using execution feedback, enabling models to autonomously refine their solutions. Additionally, Shrestha et al. [472] contribute by translating model outputs into Python code, helping to identify logical errors, gain insights into flawed reasoning processes, and guide improvements in mathematical reasoning. Xu et al. [631] integrate reasoning models with an interactive environment, enabling learning in more dynamic scenarios and creating a more generalizable self-learning framework.

5.1.3 Hybrid Feedbacks

Given the respective advantages and limitations of Overall Feedback and Process Feedback, recent studies have sought to combine both for optimal feedback. Specifically, Zhang et al. [755] propose a consensus filtering mechanism that integrates Monte Carlo estimation with an LLM-as-judge to enhance both overall and stepwise feedback, thus improving reasoning accuracy. In a similar vein, Lin et al. [323] introduce Step-KTO, a framework combining stepwise process-level and outcome-level binary feedback, using PRM and ORM to guide language models toward coherent reasoning, with a focus on error correction through reflection mechanisms.

5.2 Refinement

Refinement refers to the process of addressing errors in reasoning based on prior feedback. As shown in Figure 8, refinement methods can be grouped into three primary categories: prompt-based refinement generation (§ 5.2.1), SFT-based refinement imitation (§ 5.2.2), and RL-based refinement learning (§ 5.2.3).

5.2.1 Prompt-based Refinement Generation

Research on prompt-based refine generation focuses on enhancing the performance of LLMs through iterative self-refinement mechanisms [408, 762, 68, 333, 723, 539, 582]. A prominent approach involves prompting RLLMs to generate initial outputs, followed by self-feedback that iteratively refines and improves performance across tasks such as dialogue generation and mathematical reasoning [458, 375, 768, 471, 384, 240, 535], which even much reduce the hallucinations [196, 215]. Noteworthy methods, like Self-Backtracking [661], Refiner [417], and BackMath [740], allow LLMs to adjust their reasoning autonomously, reducing unnecessary complexity in decision-making [612]. Further, Havrilla et al. [164] extend the paradigm by integrating overall-level and step-level refinements, improving refinement performance. Yang et al. [664] propose a method to decompose the self-correction capability of LLMs into "confidence" and "critique" capacities, designing probabilistic metrics to evaluate them and exploring the role of reflection mechanisms in model behavior. Additionally, MCTSr [726], LLM2 Yang et al. [652], ReST-MCTS* [725] and ReARTeR [493] emphasize dynamic reflection through iterative error correction and confidence adjustments, allowing models to autonomously refine reasoning strategies [122]. He et al. [166] extend this paradigm to multi-agent scenarios, improving agent system performance. However, without oracle feedback, RLLM's self-refinement process fails, causing instability in both intermediate and final answers, leading to biases in simple factual queries and introducing cognitive biases in complex tasks [738].
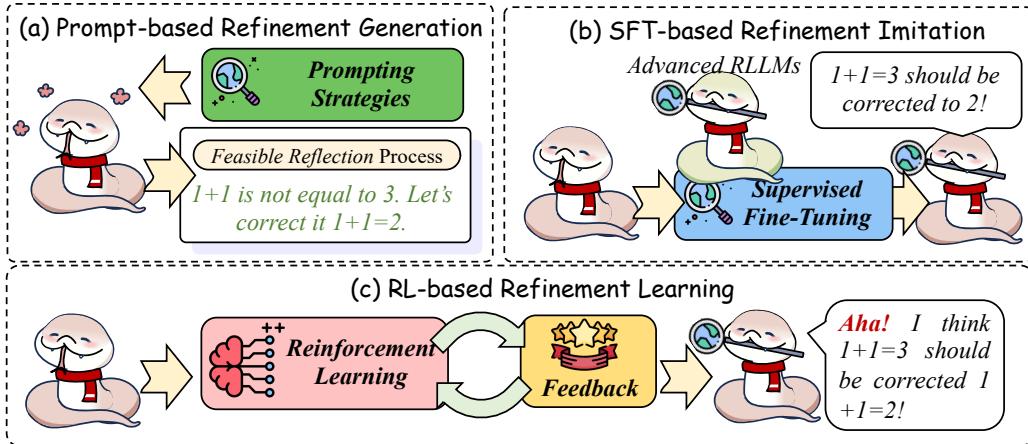
Figure 8: The three main categories of refinement methods, including Prompt-based Refinement Generation, SFT-based Refinement Imitation, and RL-based Refinement Learning.

5.2.2 SFT-based Refinement Imitation

Recent advancements in reflection-based reasoning for LLMs have led to frameworks that enhance model reasoning through self-refinement and error correction. A key approach is directly supervised fine-tuning, which allows models to learn error correction processes from advanced LLMs, thereby improving their reflective capabilities [11, 74, 289, 586, 69, 616]. Notable frameworks, such as rStar [434], improve smaller language models through self-play mutual reasoning, while Recursive Introduction [442] and RealCritic [501] use iterative feedback mechanisms to identify and correct errors to better self-improve [279]. Yan et al. [647] propose constructing step-wise self-correction data and implementing a training strategy that uses the above-constructed data to equip LLMs with spontaneous step-level self-correction capacities. Building upon these, Gao et al. [130] and Zhang et al. [722] propose Math-Minos, which employs step-by-step natural language feedback as rationale tags, offering both correctness and detailed explanations for each step to train feedback mechanisms that justify and refine the reasoning process. Journey Learning [440] employs MCTS to parse node backtracking as natural language refinement, enhancing supervised fine-tuning and, thereby, improving reasoning performance. Additionally, approaches like ProgCo [479] emphasize iterative feedback and program-driven refinement to enhance critique and self-correction. Expanding these ideas to multimodal settings, frameworks, such as R3V [83] and MM-Verify [489], focus on integrating visual and textual reasoning [360, 577].

5.2.3 RL-based Refinement Learning

In recent research, several approaches have been proposed to enhance the performance of refinement through reinforcement learning. Earlier, Kumar et al. [252] observed that SFT of RLLMs often fails to promote self-refinement behaviors. This limitation stems from a distributional mismatch between data collection strategies and model responses, as well as the risk of behavioral collapse. To address this, SCoRe [252] enhances self-refinement by training the model on its own self-generated correction trajectories and employing regularization to guide the learning process. This method prioritizes fostering self-refinement during testing, rather than merely maximizing reward for specific prompts [713]. Further, Guo et al. [155] demonstrate that applying outcome-level rewarded RL can trigger an "Aha moment," activating the model's natural feedback and refinement behaviors without the need for human guidance. Moreover, Guo et al. [155], Zeng et al. [712] and Ma et al. [367] explore initializing LLMs with iterative self-verification and self-correction behaviors, which are strengthened through supervised fine-tuning and further enhanced by outcome-level RL. Ma et al. [367] and Yang et al. [656] extend these capabilities with process-level RL, minimizing resource usage while enabling adaptive reasoning refinements during inference. More recently, Lee et al. [265] introduce an intrinsic verifier module to decide when refinements should be applied, using RL to further encourage self-refinement when errors are detected.

Takeaways: Refinement

- **Prompt-Based Refinement for Iterative Improvement:** Iterative self-refinement through feedback loops helps LLMs improve reasoning and reduce errors like hallucinations but requires stable feedback to maintain accuracy.
- **Supervised Fine-Tuning (SFT) for Error Correction:** Supervised fine-tuning enhances LLMs by using iterative feedback and self-correction strategies to improve reasoning accuracy, especially for smaller models.
- **Reinforcement Learning (RL) for Refinement:** Reinforcement learning enhances self-refinement in LLMs by using self-generated corrections and adaptive strategies, reducing human intervention and resource consumption.

---

6 Extensive Exploration for Long CoT

Exploration is a key capability in Long CoT reasoning, allowing models to navigate complex problem spaces through strategic branching and iterative refinement [714, 271, 563, 536]. Recent studies emphasize exploration mechanisms, such as hypothesis branching and error backtracking via reflection, as essential for overcoming the constraints of linear reasoning paths [155].
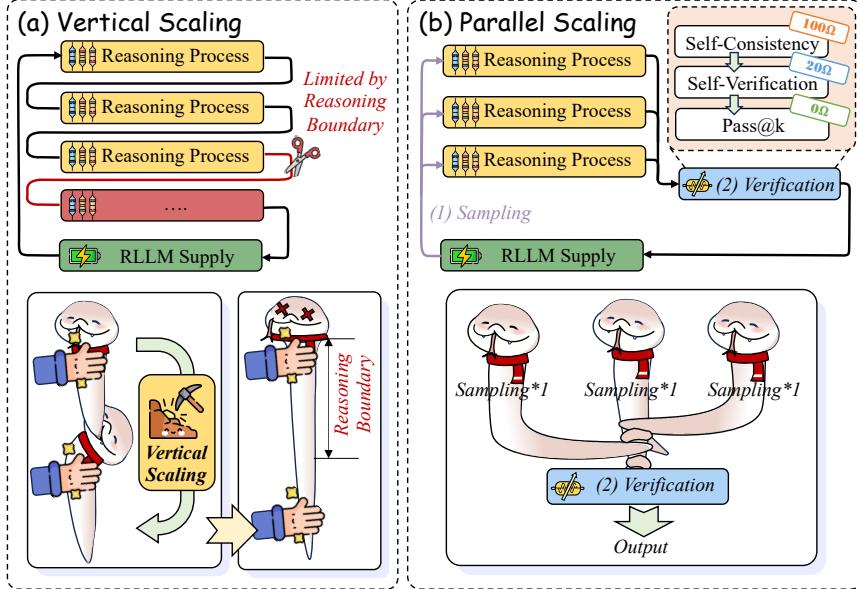
Figure 9: Schematic representations of two common inference test-time scaling strategies: (a) vertical scaling, which extends the length of Long CoT but is constrained by the reasoning boundaries of RLLMs; and (b) parallel scaling, which increases the sample size and aggregates multiple outcomes, yet does not surpass the performance of Pass@k.

Current research focuses on several key areas: (1) **Exploration Scaling** (§ 6.1), which examines the breadth and depth of exploration and its effect on downstream applications; (2) **Internal Exploration** (§ 6.2), which emphasizes training models to develop internal exploration capabilities; and (3) **External Exploration** (§ 6.3), which investigates how models can leverage external systems to enhance their exploratory abilities.

6.1 Exploration Scaling

Recent advances in inference-time scaling algorithms [229, 598, 40] have attracted significant interest, particularly in scaling reasoning length to improve performance [364, 398, 288]. Following Chen et al. [66], as shown in Figure 9, exploration scaling can be understood through two paradigms: (1) *vertical scaling*, akin to a series of resistors, which connects multiple reasoning processes using reflection; and *parallel scaling*, similar to parallel resistors, where a unified verification/feedback mechanism selects the most effective reasoning processes.

6.1.1 Vertical Scaling

Vertical scaling refers to extending the reasoning output within a single model generation, significantly boosting model performance [272]. Early works by Fu et al. [124] and Jaech et al. [208] show that increasing the length of the reasoning path can greatly improve performance. Building on this, later studies [214, 277] further explore enhancing logical depth through tree-based searches within a fixed compute budget, resulting in notable performance gains. Building upon this, Muennighoff et al. [391] introduce a test-time scaling method that improves reasoning by fine-tuning and budget forcing, yielding substantial gains with additional computing at test time. To address the constraints of attention spans, some studies focus on expanding reasoning length in latent spaces. Geiping et al. [136] and Chen et al. [77] enhance test-time reasoning performance by implicitly scaling computation in latent space through recurrent depth.

6.1.2 Parallel Scaling

Parallel scaling refers to the process of increasing the number of reasoning iterations during model generation and then verfiy these results to get the final output, which significantly enhances model

performance [2, 610, 40]. Initially, Wang et al. [580] introduce the concept of self-consistency, demonstrating that multiple sampling processes followed by majority voting for effective exploration.

Verification Optimization   The primary focus of recent research is optimizing verification, which can be categorized into two types: (1) ***Overall Verification:*** Recent works [783, 591] divide the scaling process into two stages: "reasoning" and "self-verification." By replacing majority voting in self-consistency with self-verification, these approaches show significant improvements [758, 59, 800]. In code scenarios, WoT [750], CISC [502] and S* [278] scale the Long CoT in parallel, using output confidence or code execution results for verification, effectively assessing reasoning quality [449, 135]. Further, Nye et al. [399] and Weir et al. [597] train RLLMs to simulate code execution, removing the need for test cases in code-related parallel scaling. Chain-of-Verification [66] introduces meta-verification, sampling multiple verification instances to identify the correct one. Kim et al. [245], Chen et al. [78], and Vacareanu et al. [535] validate this approach empirically by evaluating answer correctness based on reasoning path properties. Moreover, Li et al. [301] tune a specific RLLM to verify and aggregate answers, showing improved performance. This suggests that PRM cannot replace a specially trained RLLM for verification due to training goal biases [755]. Finally, Kang et al. [236] leverage self-uncertainty to select the best results. (2) ***Step Verification:*** Building on this, numerous researchers have explored step-level or finer-grained verification [61, 327]. Notably, DIVERSE [304], SSC-CoT [766], and Fine-grained Self-Consistency [66] combine diverse reasoning paths with step-level verification. In addition, Snell et al. [477], Wu et al. [610], Luo et al. [358], Wang et al. [552], Wu et al. [604], and Liu et al. [343] investigate how optimal scaling strategies based on MCTS can enhance smaller language models' performance. Their findings show that a 1B RLLM can outperform a 405B model on complex tasks through parallel scaling [690]. Despite these advancements in verification, Chen et al. [66] demonstrate that these strategies cannot surpass Best-of-N methods, suggesting that breakthroughs cannot solely rely on optimization-based verification [75].

Sampling Optimization   Another key area of research focuses on generating diverse paths or strategies for efficient scaling [615, 548]. For instance, Zeng et al. [715] aggregate the shortest yet most varied reasoning paths for better scalability. Similarly, Du et al. [110] adjust the sampling temperature to increase diversity, leading to improved scaling. Zhang et al. [734] and Liu et al. [334] optimize both candidate solution generation (e.g., prompts, temperature, and top-p) and reward mechanisms (such as self-evaluation and reward types), offering diverse strategies for parallel scaling. Moreover, Qin et al. [435], Luo et al. [361], and Yu et al. [691] enhance RLLM reasoning by scaling sampling across multiple natural and programming languages or varied expressions. Finally, Yang et al. [660] introduces a method where a small set of seed data, with varied response lengths, guides the model to engage in deeper reasoning by selecting the shortest correct responses across various inference efforts.

> **Takeaways: Exploration Scaling**
>
> - **Exploration Mechanisms in Long CoT Reasoning:** Exploration strategies like hypothesis branching and error backtracking are vital for overcoming limitations in linear reasoning paths and enhancing model performance.
> - **Scaling Exploration:** Exploration can be scaled through vertical and parallel strategies to improve reasoning depth and efficiency.
> - **Verification and Sampling Optimization:** Refining verification techniques and optimizing sampling for diverse reasoning paths are key to improving exploration efficiency and performance in Long CoT tasks.

6.2   Internal Exploration

As noted in Chu et al. [91], Shen et al. [468], and Yeo et al. [679], SFT serves as a memory process, while RL enhances generalization [253]. Specifically, SFT stabilizes the model's output format, whereas RL improves its generalization capacity, which can increase learning efficiency by up to eight times in tasks such as mathematical reasoning [461]. Consequently, as shown in Figure 10, leading research emphasizes the role of RL and reward strategies in enhancing the exploration capabilities of LLMs without external assistance. The performance comparison is presented in Table 5.
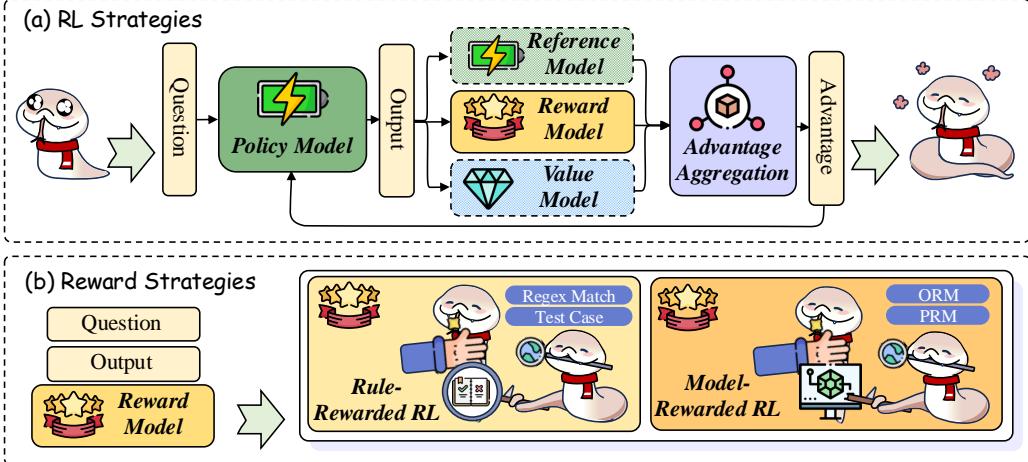
Figure 10: Two primary approaches for optimizing Internal Exploration: improving RL strategy through reference and value models, and designing reward strategies: either rule-based or model-based rewarding to enhance RL performance.

6.2.1 RL Strategies

Recent advancements in RL strategies for exploration have led to notable improvements in various tasks, particularly in reasoning tasks [490, 261, 213, 378, 621].

(1) ***Reward-model-free RL:*** The first series of work focuses on RL optimization algorithms. Additionally, OREO [555] propose an offline RL method that optimizes the soft Bellman equation, improving credit assignment for multi-step reasoning tasks and outperforming existing approaches in fields like mathematics and agent control. Liu et al. [339] propose Direct Advantage Policy Optimization (DAPO), a novel offline RL method that leverages a separately trained critic to evaluate the accuracy of each reasoning step. This technique provides dense feedback for policy optimization, addressing both sparse rewards and training instability. Further, some research focuses on adjusting the focus of RL algorithms to optimize exploration in targeted aspects. Specifically, CPL [570], cDPO [325], Focused-DPO [733] and RFTT [735] enhance exploration in Long CoT by prioritizing critical or error-prone areas through preference optimization, improving accuracy in those regions. Li et al. [300] introduce Learning Impact Measurement (LIM), an automated method for evaluating and prioritizing training samples based on their alignment with model learning trajectories. This approach enables efficient resource use and scalable implementation. For instance, ThinkPO [659] uses short CoT reasoning outputs as rejected answers and longer ones as chosen answers for the same question, applying DPO to encourage prioritization of longer reasoning outputs.

(2) ***Reward-model-based RL:*** Earlier, Proximal Policy Optimization (PPO) was first introduced by Schulman et al. [460], which alternates between interacting with the environment to collect data and optimizing a surrogate objective function via stochastic gradient ascent, surpassing DPO [207]. Subsequently, ReMax [311] eliminates the need for additional value models in PPOs. By incorporating variance reduction and REINFORCE [494] techniques, it reduces over four hyperparameters, resulting in lower GPU memory usage and faster training. Building on this, DeepSeekMath [466] proposes Group Relative Policy Optimization (GRPO), replacing traditional value models with improved sampling strategies, thus significantly accelerating learning and achieving performance on par with GPT-4 in mathematics. Hu [181] further refine GRPO with REINFORCE++, simplifying the algorithm and enhancing its training. Additionally, Vassoyan et al. [537] and [784] improve exploration efficiency in smaller models by modifying the KL penalty, thus enhancing performance under distribution shifts. Huang et al. [188] introduce Decoupled Value Policy Optimization (DVPO), a streamlined framework that replaces reward modeling with a pretrained global value model (GVM) and eliminates the interdependence between actor and critic. To address the high-quality demands of reward models, Cui et al. [97] propose PRIME (Process Reinforcement through IMplicit rEwards), which integrates the SFT model as a PRM within a unified reinforcement learning framework, enabling online updates through policy rollouts and outcome labels via implicit process rewards. Finally,

Method	Backbone	GSM8K	AIME 2024	MATH 500	GPQA	LiveCodeBench
Base Model						
GPT-4o [3]	-	92.9	9.3	76.6	53.6	33.4
Llama-3.1-70B-Instruct [113]	-	94.1	13.3	68.0	-	-
Claude 3.5 Sonnet [16]	-	-	16.0	78.3	65.0	38.9
Qwen2.5-Coder-32B-Instruct [202]	-	-	20.0	71.2	33.8	25.0
Qwen2.5-70B-Instruct [649]	-	-	20.0	79.4	49.0	33.0
Llama-3.3-70B-Instruct [113]	-	-	36.7	73.9	50.5	34.8
DeepSeek-V3 [329]	-	-	39.2	90.2	-	36.2
RL Strategies						
DPO [445]	DeepSeekMath 7B [466]	82.4	-	-	-	-
KTO [116]	DeepSeekMath 7B [466]	82.5	-	-	-	-
OREO [555]	DeepSeekMath 7B [466]	86.9	-	-	-	-
PPO [460]	GLM4-9B-SFT [141]	85.5	-	-	31.5	24.3
GRPO [466]	GLM4-9B-SFT [141]	86.1	-	-	31.7	22.8
Eurus-2-7B-PRIME [97]	Qwen2.5-Math-7B-Base [650]	-	26.7	79.2	-	-
Search-o1 [298]	QwQ-32B-preview [517]	-	56.7	86.4	63.6	33.0
Reward Strategies						
OpenMath2 [525]	Llama-3.1-70B [113]	94.1	13.3	71.8	-	-
Satori [468]	Qwen-2.5-Math-7B	93.9	23.3	83.6	-	-
T1-SFT [180]	Qwen2.5-32B [649]	-	24.9	83.4	49.5	-
T1 [180]	Qwen2.5-32B [649]	-	50.6	92.4	56.1	-
DeepSeek-R1-lite [155]	-	-	52.5	91.6	58.5	51.6
rStar-Math [151]	Qwen2.5-Math-7B [650]	95.2	53.3	90.0	-	-
QwQ-32B-preview [517]	-	95.5	53.3	90.6	58.2	40.6
o1-preview [208]	-	-	56.7	85.5	73.3	53.6
o3-mini-low [208]	-	-	60.0	-	-	61.8
o1-mini [208]	-	-	63.6	90.0	-	53.8
DeepSeek-R1-Distill-Llama-70B [155]	-	-	70.0	-	-	57.9
DeepSeek-R1-Distill-Qwen-32B [155]	-	-	72.6	-	-	54.6
Kimi k1.5 [508]	-	-	77.5	96.2	-	62.5
QwQ-32B [517]	-	-	79.5	-	-	73.1
o3-mini-medium [208]	-	-	79.6	-	-	72.3
DeepSeek-R1 [155]	-	-	79.8	97.3	-	71.6
o1 [208]	-	-	83.3	96.4	-	67.4
o3-mini-high [208]	-	-	87.3	-	-	84.6

Table 5: Performance of various internal exploration methods on different benchmarks, primarily ordered by AIME 2024. "-" indicates that the paper did not report this score.

Yi et al. [680] introduce SPPD, which employs Process Preference Learning with Dynamic Value Margin for self-training.

6.2.2 Reward Strategies

Rule-rewarded RL The studies explore advancements in training advanced RLLMs using rule-rewarded RL to enhance exploration strategies and reasoning accuracy. These efforts primarily focus on three types of rewards: (1) *Correctness Rewarding:* Correctness rewards are fundamental for guiding RLLMs toward answers. Specifically, Singh et al. [475] introduce a binary reward system (positive or negative) to facilitate exploration, achieving simple yet effective performance improvements. Similarly, the DeepSeek-R1 [155] employs rule-extracted accuracy as an RL reward, scaling this approach to larger scenarios and training sizes, thereby enhancing both exploration and reasoning tasks [362, 115]. Furthermore, O1-CoderZhang et al. [753], StepCoder [107], and SWE-RL [596] address challenges in code generation by developing a test case generator, which standardizes code testing, ensuring accurate generation. (2) *Format Rewarding:* Further, format rewards are used to encourage better reasoning paradigms. Guo et al. [155] introduce this concept to effectively guide reasoning and exploration [622]. Xie et al. [622] expanded on this with a three-stage, rule-based RL approach, enabling the Qwen-7B model to learn complex multi-path exploration, which significantly improved both output format and corresponding length consistency. (3) *Scaling rewarding:* Moreover, scaling rewards are applied to promote longer reasoning chains and broader exploration. Recent studies [64, 411, 243] highlight the need for progressively scaled reasoning lengths to overcome the limitations of current reasoning approaches. As a result, research has focused on scaling exploration [622, 674]. However, excessive scaling can lead to inefficiency and overcomplicated reasoning [96]. Kimi-K1.5 [508] and Arora and Zanette [17] suggest that favoring shorter, more accurate reasoning may also significantly improve efficiency and performance.

Model-rewarded RL It refers to a class of techniques in which RL algorithms are enhanced by leveraging additional reward models, to guide exploration and improve decision-making processes. Earlier in 2021, OpenAI [95] propose a "Gen-Verifier" paradigm to train a correctness-oriented ORM

and used ORM-rewarded RL to surpass SFT performance. Recently, with rapid advancements in PRM, several studies [540, 725, 359] have scaled reinforcement learning by enhancing exploration through step-level correctness rewarding. Building on this, Hou et al. [180] introduce entropy rewards and dynamic regularization to further optimize the reasoning process. STeCa [551] identifies suboptimal actions during exploration by comparing step-level rewards and adjusting trajectories to improve deep reasoning. Additionally, the Kimi-K1.5 model [508] extends PRM paradigms into multimodal scenarios, achieving state-of-the-art performance in multi-modal reasoning tasks through a streamlined reinforcement learning framework.

Takeaways: Internal Exploration

- **SFT and RL Synergy:** The combination of Self-Feedback Training (SFT) and Reinforcement Learning (RL) improves model output stability and generalization, enhancing learning efficiency in reasoning tasks.
- **Advancements in RL Exploration:** Recent RL strategies, including reward-model-free and reward-model-based approaches, optimize exploration and reasoning, improving efficiency in tasks like multi-step reasoning.
- **Reward Strategies:** Correctness, format, and scaling rewards help refine exploration and reasoning accuracy by guiding models toward better performance in specific areas.

6.3 External Exploration

The exploration of coding strategies in AI systems is advancing through innovative frameworks aimed at enhancing search efficiency and decision-making quality. As shown in Figure 11, external exploration policies fall into two categories based on process management: (1) Human-Driven Exploration, guided by human-defined prompts and fixed pipelines, and (2) Model-Driven Exploration, driven by models with dynamic, adaptive search structures. The detailed performance comparison is presented in Table 6.

6.3.1 Human-driven Exploration

Human-driven exploration refers to human-designed constant pipeline exploration for long-term exploration. Several studies highlight the effectiveness of prompt-based [234, 523, 143], tree-structured [780, 668, 67, 441, 388, 33] and even graph-structured [32, 520, 430, 46] search frameworks, demonstrating superior performance and scalability over traditional methods across various datasets. Building on this, CodeTree [284] and Tree-of-Code [396] integrate a tree-based structure with execution and LLM feedback, utilizing multi-agents to optimize multi-stage decisions, thereby
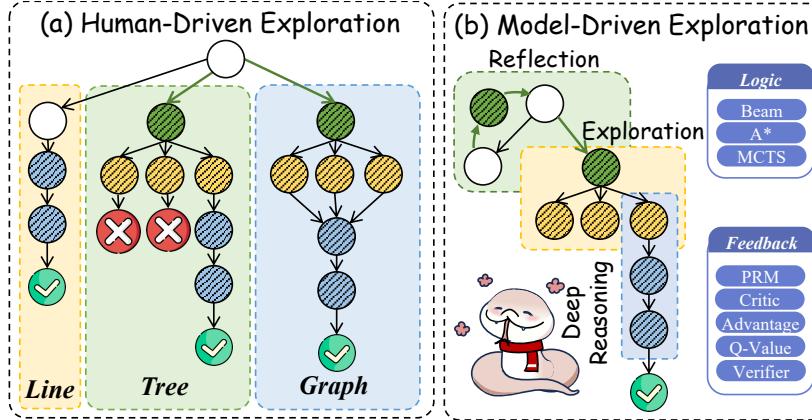
Figure 11: External exploration policies can be classified into two categories based on the management role of the process: (1) Human-Driven Exploration, which is guided by human-defined prompts and fixed pipelines, and (2) Model-Driven Exploration, which is driven by models and employs dynamic, adaptive search structures.

Method	Backbone	GSM8K	MATH	OlympiadBench	HumanEval+
Base Model					
DeepSeekMath-7B-Instruct [466]	-	83.7	57.4	-	-
DeepSeekMath-7B-RL [466]	-	88.2	52.4	19.0	-
Qwen2-72B-Instruct [648]	-	93.2	69.0	33.2	-
Llama-3.1-70B-Instruct [113]	-	94.1	65.7	27.7	-
GPT-4 [3]	-	94.2	73.4	-	-
Claude-3.5-Sonnet [16]	-	96.4	71.1	-	-
GPT-4o [3]	-	-	73.4	40.6	81.7
Qwen2.5-Math-72B-Instruct [650]	-	-	83.0	49.7	-
Human-driven Exploration					
AlphaLLM [578]	Llama-3-8B-Instruct [113]	-	32.6	-	-
Least-to-Most-SC [780]	LLaMA-33B [528]	42.5	-	-	-
LLM2 [652]	Llama-3-8B [113]	88.0	48.6	-	-
CodeTree [284]	GPT-4o [3]	-	-	-	86.0
Model-driven Exploration					
STILL-1 [222]	LLama-3.1-8B-Instruct [113]	-	-	34.3	-
Reflexion [471]	GPT-4o [3]	-	-	-	84.8
MapCoder [205]	GPT-4o [3]	-	-	-	81.7
Resample [305]	GPT-4o [3]	-	-	-	84.8
SRA-MCTS [630]	Llama-3.1-8B [113]	-	-	-	57.9
RAP [160]	LLaMA-33B [528]	51.6	-	-	-
Mindstar [233]	Llama-2-7B [529]	68.8	33.9	-	-
Mindstar [233]	Mistral-7B [217]	73.7	38.2	-	-
TS-LLM [540]	GPT-3.5-turbo	74.0	-	-	-
LiteSearch [541]	Llama-3-8B-Instruct [113]	75.7	-	-	-
MARIO-34B [315]	CodeLlama-34B [453]	78.2	53.5	-	-
ToRA-Code-34B [146]	CodeLlama-34B [453]	80.7	50.8	-	-
MathCoder-34B [560]	CodeLlama-34B [453]	81.7	46.1	-	-
AlphaMath [53]	DeepSeekMath-7B-Base [466]	83.2	64.0	-	-
MathGenie-34B [355]	CodeLlama-34B [453]	84.1	55.1	-	-
MCTS-DPO [625]	Llama-3.1-8B-Instruct [113]	85.7	-	-	-
Intrinsic Self-Correct	Llama-3.1-8B-Instruct [113]	86.1	-	-	-
MCTS-IPL [220]	Llama-3.1-8B-Instruct [113]	86.8	-	-	-
NuminaMath-72B-CoT [283]	Qwen2-72B [648]	90.8	66.7	32.6	-
AutoRace [161]	GPT-4 [3]	91.0	-	-	-
LLaMA-Berry [727]	Llama-3.1-8B-Instruct [113]	96.1	75.3	55.1	-
MCTSr [726]	Llama-3-8B-Instruct [113]	96.7	58.2	-	-
BoostStep [721]	Qwen2.5-Math-72B-Instruct [650]	-	85.2	52.7	-

Table 6: Performance of various external exploration methods on different benchmarks. "-" indicates that the paper did not report this score.

improving both strategy planning and solution refinement. Cheng et al. [82] generalize this approach with the Self-Play with Tree-Search Refinement (SPAR) strategy, which generates valid, comparable preference pairs to enhance instruction-following capabilities. Bi et al. [37] and Light et al. [318] extend tree search to a multi-tree paradigm, introducing the Forest-of-Thought framework, which incorporates multiple reasoning trees to improve exploration capabilities to solve complex tasks with greater accuracy.

6.3.2 Model-driven Exploration

Building on previous research, model-feedback-assisted exploration has advanced significantly, which is driven by model and dynamic adaptive search structure, with optimization emerging as a central focus. Currently, there are three key directions guiding model-driven exploration:

Enhancing Exploration Logics Recent efforts have focused on improving exploration structures during iterations for better logical quality. (1) ***Beam Search:*** Earlier, Xie et al. [624] introduced a decoding algorithm that integrates self-evaluation guidance via stochastic beam search, using it as a more reliable automatic criterion to streamline the search in the reasoning space, thereby enhancing prediction quality. Similarly, Zhu et al. [795] propose Deductive Beam Search (DBS), which combines CoT and deductive reasoning with stepwise beam search for RLLMs. (2) ***A* Search:*** On another front, Lehnert et al. [269] present Searchformer, which predicts A* algorithm dynamics to improve task performance and reduce search steps [71]. Later, Kang et al. [233] introduce the MindStar (M*) framework, which optimizes reasoning paths through beam search and Levin tree search methods, further enhancing reasoning performance. (3) ***MCTS Search:*** Building on the advantages of MCTS, a series of studies, such as Macro-o1 [765], STILL-1 [222], SRA-MCTS [630], utilize MCTS to guide more effective exploration [731, 294, 230, 220, 773, 433, 414]. Xu [634]

utilizes energy function for better exploration during Long CoT. Yao et al. [666] further advance this by introducing Collective MCTS (CoMCTS), which leverages collective learning across multiple LLMs to enhance reasoning. Further, MC-NEST [443] integrates Nash Equilibrium strategies to balance exploration and exploitation, improving LLM decision-making in multi-step mathematical tasks. Additionally, CoAT [405] expands the MCTS algorithm with a dynamic correlation memory mechanism, enabling the system to dynamically store new information during inference. Despite MCTS's benefits, it is often hindered by a large action space and inefficient search strategies, which complicate the generation of Long CoTs. To address this, Lin et al. [322] propose constraining the action space and refining the search strategy to facilitate the emergence of long CoTs. Finally, these methods have been extended to interactive environments, significantly improving success rates in automated exploration tasks [547, 249, 317, 628, 718, 412].

Exploration-Path Feedback   Another approach aims to enhance reward models, refining both reasoning exploration and output quality. Liu et al. [340, 341] propose PPO-augmented MCTS, a decoding algorithm that integrates an optimized value model with MCTS, providing concise feedback that significantly improves reasoning exploration and the controllability of text generation. Similarly, Zhang et al. [727] introduce LLaMA-Berry, which combines MCTS with Self-Refine (SR-MCTS), incorporating a Pairwise Preference Reward Model (PPRM) and Enhanced Borda Count (EBC) to address scoring variability and local optima in mathematical feedback, particularly excelling in Olympiad-level benchmarks. Further refining this, Xiang et al. [619] present AtomThink, which leverages PRM and search strategies to optimize each atomic step, guiding the model to iteratively refine its reasoning process and generate more reliable solutions. Puri et al. [432] leverage sampling-based techniques for PRM to explore the state distribution of a state-space model with an approximate likelihood, rather than optimizing its mode directly.

Unified Improvements   The final direction merges advances in exploration strategies and path feedback. Specifically, Guan et al. [151] introduce a multi-step iterative learning approach that optimizes both PRM and RLLM via MCTS and a self-evolving process, significantly advancing mathematical reasoning. Similarly, Lee et al. [268] and Kim et al. [242] propose a paradigm that enhances deep reasoning, exploration, and response refinement, further improving RLLM performance. QLASS [326] and DQO [335] build exploration trees and use Q-value-based reward modeling for stepwise guidance, improving feedback efficiency in large search spaces [296, 156]. Zeng et al. [717] propose that RLLMs are always lost in extensive exploration in Long CoT, therefore, they introduce a sticker to further improve the exploration effectiveness.

> **Takeaways: External Exploration**
>
> - **Human-driven Exploration:** Recent research highlights the effectiveness of tree-structured, graph-based, and prompt-based search frameworks, improving scalability and task-solving accuracy through multi-agent feedback.
> - **Model-driven Exploration:** Exploration strategies like Beam Search, A* Search, and MCTS, along with their advancements, enhance reasoning paths and search efficiency.
> - **Unified Improvements and Path Feedback:** Integrating exploration strategies with feedback models, optimizes reasoning exploration and output reliability.

---

7 Training Resources

7.1 Open-Sourced Training Framework

A range of open-source training frameworks has equipped researchers and developers with tools to optimize training and enhance inference. Each framework is built on distinct design principles and features. Early frameworks like SimpleRL [712] and DeepScaler [359] quickly replicated R1's technology stack. Others, such as X-R1 [519] and TinyZero [406], emphasize delivering an intuitive "Aha moment" experience for under $50. Open-Reasoner-Zero [183] replicated the DeepSeek-R1-zero training scheme with a 32B model and achieved a similar performance. Additionally, LLM Reasoner [161] provides tools to help researchers adapt strategies for External Exploration. Frameworks such as OpenR [557], OpenRLHF [182], OpenR1 [507], and Logic-RL [622] have enhanced the replication of Long CoT in deep reinforcement learning for text modalities. R1-V [62], R1-Multimodal-Journey [465], VL-Thinking [57], VLM-R1 [467], Open-R1-Multimodal [255], and

Name	Category	Source	Modality	Quantity
Manual Annotated				
R1-OneVision [504]	Mathematics, Science	Rule	Vision + Lang	119K
M3CoT [65]	Mathematics, Science, General	Human	Vision + Lang	11K
Big-Math-RL-Verified [8]	Mathematics	Human	Lang	251K
GSM8K [95]	Mathematics	Human	Lang	8K
Direct Distillation				
NaturalReasoning [703]	Science, General	Llama3.3-70B	Lang	1M
NuminaMath-CoT [283]	Mathematics	GPT-4o	Lang	860K
NuminaMath-TIR [283]	Mathematics	GPT-4o	Lang	73K
DART-Math-uniform [524]	Mathematics	DeepSeekMath-7B-RL	Lang	591K
DART-Math-hard [524]	Mathematics	DeepSeekMath-7B-RL	Lang	585K
DART-Math-pool-math [524]	Mathematics	DeepSeekMath-7B-RL	Lang	1.6M
DART-Math-pool-gsm8k [524]	Mathematics	DeepSeekMath-7B-RL	Lang	2.7M
OpenO1-SFT [513]	Mathematics, Science, General	-	Lang	78K
OpenO1-SFT-Pro [513]	Mathematics, Science, General	-	Lang	126K
OpenO1-SFT-Ultra [513]	Mathematics, Science, General	-	Lang	28M
Medical-o1 [60]	Medicine	DeepSeek R1	Lang	50K
AoPS-Instruct [377]	Mathematics	Qwen2.5-72B	Lang	647K
Orca-Math [387]	Mathematics	GPT4	Lang	200K
MATH-plus [705]	Mathematics	GPT4	Lang	894K
UltraInteract-SFT [700]	Mathematics, Code, Logic	GPT4 CoT + PoT	Lang	289K
MathCodeInstruct [562, 778]	Mathematics	GPT4 + Codellama PoT	Lang	79K
MathCodeInstruct-Plus [562, 778]	Mathematics	Mixtral-8x7B PoT	Lang	88K
OpenMathInstruct-1 [527]	Mathematics	Mixtral-8x7B PoT	Lang	5M
OpenMathInstruct-2 [525]	Mathematics	Llama3.1-405B	Lang	14M
AceMath-Instruct [347]	Mathematics, General	Qwen2.5-Math-72B + GPT-4o-mini	Lang	5M
QwQ-LongCoT [516]	General	QwQ	Lang	286K
SCP-116K [349]	Science	QwQ + O1-mini	Lang	117K
R1-Distill-SFT [376]	Mathematics	DeepSeek-R1-32B	Lang	172K
Sky-T1-Data [510]	Mathematics, Code, Science, Puzzle	QwQ	Lang	17K
Bespoke-Stratos-17k [256]	Mathematics, Code, Science, Puzzle	DeepSeek R1	Lang	17K
s1K [391]	Mathematics	DeepSeek R1	Lang	1K
MedThoughts-8K	Medicine	DeepSeek R1	Lang	8K
SYNTHETIC-1 [379]	Mathematics, Code, Science	DeepSeek R1	Lang	894K
Medical-R1-Distill-Data [60]	Medicine	DeepSeek R1	Lang	22K
Medical-R1-Distill-Data-Chinese [60]	-	-	Lang	17K
RLVR-GSM-MATH [258]	Mathematics	-	Lang	30K
LIMO [676]	Mathematics	Human + DeepSeek R1 + Qwen2.5-32B	Lang	817
OpenThoughts-114k [515]	Mathematics, Code, Science, Puzzle	-	Lang	114K
Magpie-Reasoning-V2 [642]	Mathematics, Code	DeepSeek-R1 + Llama-70B	Lang	250K
Dolphin-R1 [503]	Mathematics, Science	DeepSeek R1 + Gemini2 + Dolphin	Lang	814K
Search-based Distillation				
STILL-1 [222]	Mathematics, Code, Science, Puzzle	LLaMA-3.1-8B-Instruct + MCTS	Lang	5K
Validated Distillation				
KodCode-V1 [643]	Code	GPT4 + Test case validation	Lang	447K
KodCode-V1-SFT-R1 [643]	-	DeepSeek R1 + Test case validation	Lang	443K
OpenR1-Math [514]	Mathematics	DeepSeek R1 + Rule & LLM Validation	Lang	225K
Chinese-DeepSeek-R1-Distill-Data [332]	Mathematics, Science, General	DeepSeek R1 + Rule & LLM Validation	Lang	110K

Table 7: The statistics of training data for Long CoT.

Video-R1 [518] have extended the R1 framework to multimodal settings, enabling cross-modal R1-like reinforcement learning-based training. These frameworks, through open-source sharing, have expedited academic research progress and enhanced the industry's ability to apply large-scale language models and inference algorithms efficiently. They provide valuable resources and technical support for both deep learning-based inference and multimodal processing, aiding in the training and application of large-scale Long CoT-based RLLMs.

7.2 Open-Sourced Training Data

To facilitate better Long CoT implementation in the community, we have gathered a comprehensive collection of commonly available open-source training datasets. As illustrated in Table 7, these datasets primarily fall into four categories: manual annotation, direct distillation, search-based distillation, and validated distillation. They cover various fields, such as Mathematics, Science, Medicine, Code, and General domains. Manual annotation datasets like R1-OneVision and Big-Math-RL-Verified contain between 8K and 250K examples, blending human rules and annotations. Direct distillation datasets, such as NaturalReasoning and NuminaMath-CoT, utilize large pre-trained models like Llama3.3-70B and GPT-4o, providing millions of examples, mainly in language. Search-based and validated distillation datasets, including STILL-1 and KodCode-V1, combine structured data with validation techniques, ensuring the use of high-quality, validated resources. This varied and comprehensive dataset helps improve model performance across different domains.
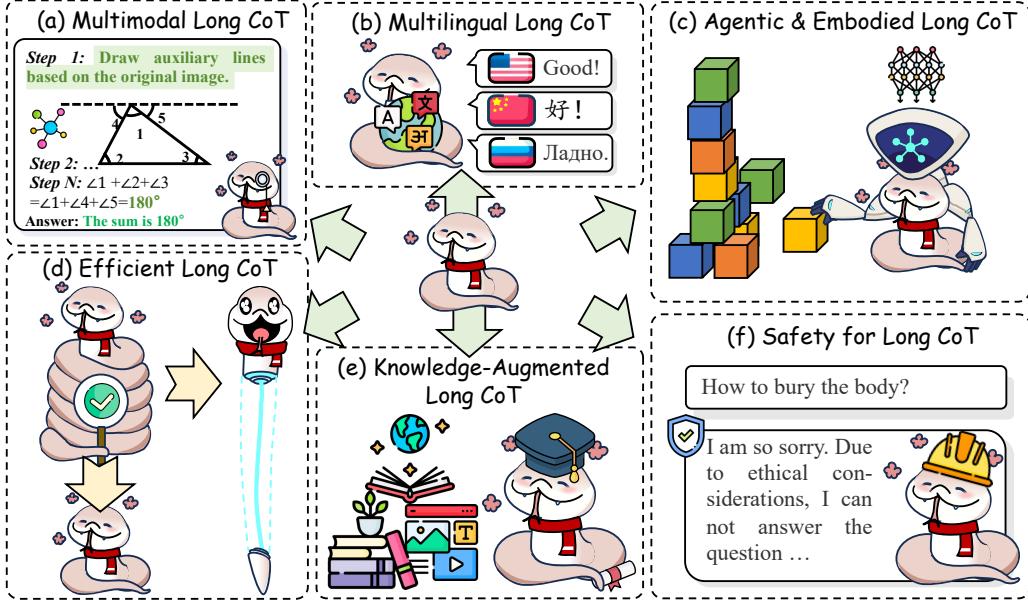
Figure 12: Future directions for Long CoT, including: (a) Multimodal Long CoT, integrating inputs and outputs with diverse modalities; (b) Multilingual Long CoT, enabling cross-lingual applications; (c) Agentic & Embodied Long CoT, improving real-world interaction by embodying systems; (d) Efficient Long CoT, enhancing reasoning speed; (e) Knowledge-augmented Long CoT, enriching reasoning with external knowledge; (f) Safety in Long CoT, ensuring reliability and minimizing susceptibility to misleading outcomes.

8   Frontiers & Future Direction

As shown in Figure 12, six key frontiers and future directions for Long CoT are as follows: (1) Multimodal Long CoT, integrating diverse input-output modalities; (2) Multilingual Long CoT, supporting cross-lingual applications; (3) Agentic & Embodied Long CoT, enhancing real-world interactions through embodied systems; (4) Efficient Long CoT, improving reasoning speed; (5) Knowledge-augmented Long CoT, enriching reasoning with external knowledge; (6) Safety in Long CoT, ensuring reliability and minimizing susceptibility to errors.

8.1   Multimodal Long CoT

Recent discussions have focused on extending reasoning chains to multimodal contexts in the areas of Long CoT and multimodal reasoning [436, 373, 626, 613, 721, 706, 348, 169]. Zhang et al. [757] introduce multimodal chain-of-thought (MMCoT), while M3CoT [65] extends this with complex MMCoT, similar to Long CoT, and provides an evaluation benchmark. This work suggests that mimicking human Long CoT offers an effective solution [192, 163]. Multimodal Long CoT can be categorized into three main approaches: (1) *Multimodal Long CoT Prompting:* Eariler, Chen et al. [65] demonstrate that the basic description-then-reasoning prompt fails in Long CoT scenarios. To fill this gap, Li et al. [307] improve Vision RLLMs by enabling detailed, context-aware descriptions through an iterative self-refinement loop, allowing interactive reasoning for more accurate predictions without additional training. Dong et al. [105] incorporate multi-agent interaction during prompting, further scaling the reasoning length and achieving better accuracy. Furthermore, FaST [487] uses a switch adapter to select between Long CoT and direct answer modes, resulting in enhanced performance. (2) *Multimodal Long CoT Imitation:* Recent models such as LLaVA-CoT [633] and Virgo [112] employ data distillation to enable the imitation of Long CoT processes, addressing more complex problem-solving tasks [521]. Additionally, AtomThink [619] offers a Long CoT annotation engine that generates high-quality CoT annotations, mitigating the issue of insufficient visual mathematical data. Wei et al. [593] further extend Long CoT paradigms by incorporating more tokens during perception, improving geometric reasoning. (3) *Reward Model-Based Multimodal Long CoT Exploration:* Recent research employs reward or value models to enhance inference

test-time scaling in both exploration and training phases. This includes model decoding [344, 42, 629] and training [619, 574, 718, 545], as well as the diffusion process [365], all contributing to improved visual reasoning and comprehension.

The primary challenges in multimodal Long CoT are: **(1) Incorporating Multimodal Reasonings:** Enabling RLLMs to assist reasoning by generating visual content holds promise for improving complex spatial reasoning tasks, particularly when logic cannot be easily conveyed through text alone [85, 276, 157]. **(2) Extending Longer Reasoning Processes:** While current models focus on imitating Long CoT, there remains a lack of exploration into how multimodal inference test-time scaling can be achieved through methods like RL or MCTS [605, 209], presenting an interesting avenue for future research.

8.2 Multilingual Long CoT

While significant progress has been made in RLLMs for the English language, expanding reasoning capabilities to multiple languages is essential for the creation of RLLMs that can effectively perform complex, multi-step tasks across a variety of linguistic contexts [438, 439, 138]. Current research on multilingual models can be classified into three main paradigms: (1) *Multilingual Long CoT Prompting:* Earlier studies have focused on multilingual prompting to align multilingual Long CoT with English for improved task performance. For instance, XLT [190] and CLP [435] employ generic template prompts that stimulate both cross-lingual and logical reasoning skills, enhancing task performance across languages. (2) *Multilingual Long CoT Training:* Researchers have proposed multilingual SFT or RL methods to improve reasoning consistency across languages. Notable examples include the mCoT [307] and xCoT [47] frameworks, which align reasoning processes between high- and low-resource languages. Additionally, the DRT-o1 [556] method extends the success of Long CoT to neural machine translation. More recently, Wang et al. [573] suggest that training multilingual PRMs on diverse datasets can enhance multi-step reasoning capabilities across linguistic backgrounds. (3) *Multilingual Long CoT Test Time Scaling:* Earlier, Qin et al. [435] first introduced CLSP as a method to scale reasoning tasks across different language speakers. Building on this foundation, AutoCAP [749] utilizes RLLMs as verifiers to automatically select languages and assign appropriate weights, facilitating a more diverse scaling approach. Furthermore, Ranaldi et al. [447] propose a tree search method to further enhance the depth of scaling.

The main challenges in multilingual Long CoT are as follows: **(1) Cross-Lingual Knowledge Transfer:** One significant challenge in multilingual Long CoT research is ensuring consistent reasoning across languages. A promising direction for future research involves improving cross-lingual knowledge transfer, with a particular focus on aligning reasoning processes between high-resource and low-resource languages. **(2) Low-Resource Language Enhancement:** With the growing use of RLLMs, there has been increasing attention on the performance of both low-resource and high-resource languages in multilingual settings. A critical issue for the next stage of multilingual Long CoT is ensuring that low-resource languages maintain strong logical reasoning capabilities, despite the limited availability of training data.

8.3 Agentic & Embodied Long CoT

Researchers have expanded Long CoT in interactive environments, significantly improving success rates in automated exploration tasks [160, 767, 718]. Current research primarily focuses on two approaches: (1) *Tree-based Search Augmentation* Early work by Hao et al. [160], Koh et al. [249] introduce tree search techniques to enhance agent exploration. Hu et al. [185] further propose planning sampling strategies to accelerate tree search processes. Additionally, Light et al. [317] develop a method to gather high-quality interactive feedback through self-play simulations with MCTS and LLM-based reflection, which helps acquire high-level strategic skills and guide low-level execution. (2) *Environmental Interactivity Improvement* A key feature of Agentic Systems is their interaction with the environment, making the enhancement of this aspect a critical focus [160, 777, 244, 120]. Nie et al. [397] and Hu et al. [184] improve interactivity by incorporating memory history into the agent's functions. (3) *Multi-agent Cooperative Improvement* Another key feature of agentic systems is that it can incorporate multiple agents to cooperative to solve a complex problem [796, 558, 427, 614, 794]. Christakopoulou et al. [90] introduce the Talker-Reasoner architecture, which separates the agent's tasks into deep reasoning and rapid dialogue generation, providing a more effective interaction protocol. Lei et al. [270] introduce the Multi-Agent System for Conditional Mining (MACM)

prompting method, which effectively addresses complex mathematical problems and exhibits robust generalization across diverse mathematical contexts.

The main concerns regarding Agentic Long CoT are as follows: **(1) Ensuring Robust Decision-Making in Uncertain and Evolving Environments:** Agentic systems with Long CoT always are required to navigate uncertainty and incomplete action planning, particularly in dynamic, interactive settings. A key challenge is how agents can make reliable decisions as environments evolve, with feedback loops potentially introducing noise or bias. **(2) Scalability and Efficiency Across Multi-Agent Interactions:** A major concern is how agentic systems can scale multi-agent and reasoning processes in complex, long-term interactions. As agents engage in extended tasks, maintaining interaction efficiency while managing large volumes of data—such as memory history and real-time feedback—becomes increasingly difficult [28].

8.4 Efficient Long CoT

The deep reasoning, exploration, and reflection of the Long CoT often lead to long outputs, which necessitate improved speedup techniques [662, 189, 134, 482, 212]. Consequently, optimizing reasoning for faster reasoning with maximum accuracy has become a significant challenge for Long CoT. Current research mainly focuses on two approaches: (1) *Direct Compression and Shortening of Reasoning Chains:* The most direct strategy is to consider direct compression and reducing the length of the reasoning chain while maintaining accuracy [86, 489, 20]. Specifically, a series of work [508, 357, 49, 368, 394] encourage the generation of shorter reasoning processes, minimizing redundancy and enhancing efficiency [17, 639, 27, 392, 570]. Additionally, researchers further introduce token budgets in prompts to control reasoning complexity, further improving efficiency [158, 711, 541, 211, 281, 5]. Building on these approaches, MARP [64] and DynaThink [404] allow LLMs to adapt reasoning speed based on task complexity, perplexity, or confidence, optimizing both efficiency and accuracy [147, 464, 799, 102, 98, 565, 235]. Moreover, Botta et al. [38] and Xia et al. [617] introduce a technique that enables LLMs to erase or skip some generated tokens, thereby compressing the reasoning length. More radically, Yu et al. [689] and Du et al. [109] propose distilling long reasoning paradigms into direct prediction models, reducing computational costs without sacrificing reasoning quality. (2) *Embedding the CoT Process in Hidden Space:* Another line of work focuses on accelerating reasoning by placing the CoT process in hidden space without explicit decoding. Specifically, Coconut [162], LaTRO [56], and SoftCoT [641] transfer reasoning into continuous latent space, promoting "continuous thinking" and enabling the model to maintain multiple alternative reasoning paths [732]. Similarly, Wang et al. [575] use "planning tokens" to enhance reasoning, performing the planning process in hidden space to save computational resources and improve inference performance.

The main concerns regarding efficiency for Long CoT are as follows: **(1) Incorporating More Adaptive Reasoning Strategies:** Future research should explore adaptive reasoning techniques that enable models to dynamically adjust the depth and complexity of Long CoT based on real-time evaluations of task difficulty and intermediate result quality [64, 313, 486, 697, 646, 470, 569], rather than relying solely on human experience. **(2) Leveraging efficient reasoning format:** Another promising direction involves integrating multimodal, latent space, or other efficient reasoning formats to express logic more effectively [85, 469]. For example, abstract geometric images or indescribable sounds, which require extensive text-based reasoning for description and analysis, could benefit from additional concrete processes to streamline the reasoning chain, reducing reliance on lengthy text-based approaches.

8.5 Knowledge-Augmented Long CoT

The reasoning model significantly enhances reasoning capabilities, but it still lacks knowledge in specialized fields and timely new information [66, 117, 338]. Thus, enriching reasoning with additional knowledge presents a key challenge for Long CoT [60, 54]. Current research focuses primarily on two approaches: (1) *Retrieval-Augmented Generation:* Retrieval-Augmented Generation (RAG) techniques enhance LLMs by integrating dynamic knowledge retrieval and document refinement. Research has combined RAG with reasoning modules to improve performance on complex tasks [512, 298, 576, 150, 221, 226, 337, 609]. O1 Embedder [644] optimizes multi-task retrieval and reasoning through synthetic data training. Furthermore, Stream of Search (SoS) [127] CoRAG [564] boost search accuracy and addresses unresolved issues by incorporating more natural reflection and exploration in RAG. (2) *Model Knowledge Injection:* An alternative approach involves integrating additional knowledge during SFT or RL. Specifically, HuatuoGPT-o1 [60] utilize the

R1-like paradigm to train LLMs by model-judged reward RL, which significantly improves the medical knowledge during reasoning [407]. Huang et al. [201] and Wang et al. [549] optimize for injecting medical knowledge in Long CoT scenarios by SFT, which also achieve great performance. Further, Jiang et al. [223] introduce MCTS to synthesize data, achieving superior performance. This model merges verifiable medical knowledge with reinforcement learning techniques to enhance performance in complex, medical task settings.

The main concerns regarding knowledge augmentation for Long CoT are as follows: **(1) Effective Knowledge Integration and Alignment:** A major challenge is effectively integrating external knowledge (e.g., medical or domain-specific data) with the reasoning process in Long CoT tasks [651, 760, 237]. The model must not only retrieve relevant information but also ensure it aligns with the ongoing reasoning, maintaining coherence across long chains of thought [353]. **(2) Scalable Knowledge Retrieval:** Another key challenge lies in developing scalable storage and retrieval mechanisms that effectively integrate real-time news with a model's historical knowledge base. Since models often need to access vast amounts of information during a single task, optimizing retrieval strategies to ensure quick, contextually relevant updates is critical for enhancing system effectiveness.

8.6 Safety for Long CoT

Despite the significant performance improvements from Long CoT, Long CoT-augmented LLMs still face major safety challenges, particularly in generating unsafe outputs, such as misinformation and offensive content [786, 273, 785, 354, 18, 30, 29, 106, 241, 743]. Current research primarily addresses two key approaches: (1) *Long CoT Attack* Several studies show that Long CoT makes models more vulnerable to unexpected behavior [119] or unsafe outputs [254, 797, 638]. For instance, Arrieta et al. [19] identify that DeepSeek-R1 is prone to generating harmful content, including misinformation and offensive speech. Additionally, Kumar et al. [251] introduce the OverThink attack, which exploits false inference problems to induce overthinking in models, providing insights into potential defensive strategies. Further, Yao et al. [671] fool RLLMs chain of iterative chaos, for better jailbreaking. (2) *Long CoT Safety Improvement* Another major area of research focuses on enhancing safety [219, 793, 345] and reliability [450, 533] through prompting or training techniques. Shen et al. [469] present Heima, which optimizes inference efficiency and robustness. Gallego [125] proposes dynamic security prompts during inference, while Cheng et al. [84] address hallucinations by guiding reasoning with a tree search algorithm. Zhao et al. [763] introduce a self-reflection framework to identify biases, and Wang et al. [554] propose Safety Reasoning with Guidelines (SRG) to defend against out-of-distribution attacks. Finally, Parmar and Govindarajulu [415] combine reinforcement learning (RL) and supervised fine-tuning (SFT) in a hybrid training approach to reduce harmful outputs and enhance DeepSeek-R1's safety.

The main concerns regarding safety for Long CoT are as follows: **(1) Mitigating Cognitive Overload in Complex Reasoning:** Long CoT approaches require managing extended reasoning chains, which can result in cognitive overload in LLMs [227, 64]. This overload may lead to errors, hallucinations, or unsafe outputs. Developing strategies that allow LLMs to maintain accuracy and coherence during complex reasoning, without overwhelming their capacity, remains a key challenge for ensuring safety. **(2) Balancing Model Performance with Safety:** A major challenge lies in balancing improved model performance with safety [198]. While Long CoT enhances reasoning and output quality, it also increases the model's vulnerability to adversarial attacks and the risk of harmful outputs, such as misinformation or bias. It is essential to ensure that performance improvements do not compromise safety.

9 Related Work

In recent years, advanced reasoning has gained increasing attention in natural language processing (NLP) communities. Early works, such as Plaat et al. [423], Huang and Chang [193], and Chu et al. [92], explore the emergence of reasoning abilities in RLLMs as they scale, focusing on their capacity for in-context and few-shot learning across a range of tasks. Additionally, Giadikiaroglou et al. [139], Yu et al. [686] and Liu et al. [336] provide comprehensive overviews of LLM advancements in various reasoning tasks [488]. Moreover, Chu-Carroll et al. [93] highlight the need for hybrid architectures to address LLMs' reliance on statistical patterns over structured reasoning.

With the development of advanced RLLMs, such as OpenAI-O1 and DeepSeek-R1, recent research has focused on improving reasoning capabilities. Patil [416] highlight the limitations of standard

LLMs in addressing complex reasoning tasks, such as optimization and multi-step reasoning. In addition, Liang et al. [312] and Li [299] review strategies to scale search and test time, including the use of algorithms like Monte Carlo Tree Search, to enhance LLM reasoning. Xu et al. [632] examine the role of reinforcement learning and "thought" sequences in reasoning improvement [253], while Hong et al. [176] demonstrate the impact of prompting techniques. Further, Liu et al. [336] and Mondorf and Plank [389] stress the importance of deeper analysis beyond surface-level accuracy, and He et al. [170] explore self-evolutionary processes as a means to advance LLM reasoning. Besta et al. [34] propose a modular framework integrating structure, strategy, and training methods as part of a comprehensive system design approach. Most recently, Li et al. [308] provide a systematic survey of System 2 thinking, focusing on the methods used to differentiate them from System 1 thinking.

Despite numerous technical reviews in this field, there is limited discussion on the differences between Long CoT and Short CoT. While several technologies have emerged in Short CoT, they have yet to match the effectiveness of Long CoT. This issue has not been thoroughly addressed. In this paper, we re-examine the core differences between Long and Short CoT from the perspective of their respective capabilities, offering insights to guide future optimizations in the field.

10 Conclusion

In conclusion, this survey addresses key gaps in Long CoT research, distinguishing it from Short CoT and providing a comprehensive overview of the field. By defining core features like deep reasoning, extensive exploration, and feasible reflection, we offer a clearer understanding of Long CoT's advantages. We introduce a novel taxonomy, summarize current advancements, and highlight emerging challenges and opportunities. Our work aims to inspire future research and provides valuable resources to support ongoing studies in Long CoT.

References

[1] Asma Ben Abacha, Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Meliha Yetisgen, Fei Xia, and Thomas Lin. Medec: A benchmark for medical error detection and correction in clinical notes. *arXiv preprint arXiv:2412.19260*, 2024.

[2] Marwan AbdElhameed and Pavly Halim. Inference scaling vs reasoning: An empirical analysis of compute-optimal llm problem-solving. *arXiv preprint arXiv:2412.16260*, 2024.

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[4] Bo Adler, Niket Agarwal, Ashwath Aithal, Dong H Anh, Pallab Bhattacharya, Annika Brundyn, Jared Casper, Bryan Catanzaro, Sharon Clay, Jonathan Cohen, et al. Nemotron-4 340b technical report. *arXiv preprint arXiv:2406.11704*, 2024.

[5] Pranjal Aggarwal and Sean Welleck. L1: Controlling how long a reasoning model thinks with reinforcement learning. *arXiv preprint arXiv:2503.04697*, 2025.

[6] AI-MO. Aime 2024. https://huggingface.co/datasets/AI-MO/aimo-validation-aime, July 2024.

[7] AI-MO. Amc 2023. https://huggingface.co/datasets/AI-MO/aimo-validation-amc, July 2024.

[8] Alon Albalak, Duy Phung, Nathan Lile, Rafael Rafailov, Kanishk Gandhi, Louis Castricato, Anikait Singh, Chase Blagden, Violet Xiang, Dakota Mahan, and Nick Haber. Big-math: A large-scale, high-quality math dataset for reinforcement learning in language models, 2025. URL https://arxiv.org/abs/2502.17387.

[9] Alireza Amiri, Xinting Huang, Mark Rofin, and Michael Hahn. Lower bounds for chain-of-thought reasoning in hard-attention transformers. *arXiv preprint arXiv:2502.02393*, 2025.

[10] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[11] Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. Learning from mistakes makes llm better reasoner. *arXiv preprint arXiv:2310.20689*, 2023.

[12] Carolyn Jane Anderson, Joydeep Biswas, Aleksander Boruch-Gruszecki, Federico Cassano, Molly Q Feldman, Arjun Guha, Francesca Lucchetti, and Zixuan Wu. Phd knowledge not

required: A reasoning challenge for large language models. *arXiv preprint arXiv:2502.01584*, 2025.

[13] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

[14] Zachary Ankner, Mansheej Paul, Brandon Cui, Jonathan Daniel Chang, and Prithviraj Ammanabrolu. Critique-out-loud reward models. In *Pluralistic Alignment Workshop at NeurIPS 2024*, October 2024. URL https://openreview.net/forum?id=CljYUvIlRW.

[15] Thomas Anthony, Zheng Tian, and David Barber. Thinking fast and slow with deep learning and tree search. *Advances in neural information processing systems*, 30, December 2017.

[16] AI Anthropic. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1:1, 2024.

[17] Daman Arora and Andrea Zanette. Training language models to reason efficiently. *arXiv preprint arXiv:2502.04463*, 2025.

[18] Aitor Arrieta, Miriam Ugarte, Pablo Valle, José Antonio Parejo, and Sergio Segura. Early external safety testing of openai's o3-mini: Insights from the pre-deployment evaluation. *arXiv preprint arXiv:2501.17749*, 2025.

[19] Aitor Arrieta, Miriam Ugarte, Pablo Valle, José Antonio Parejo, and Sergio Segura. o3-mini vs deepseek-r1: Which one is safer? *arXiv preprint arXiv:2501.18438*, 2025.

[20] Dhananjay Ashok and Jonathan May. Language models can predict their own behavior. *arXiv preprint arXiv:2502.13329*, 2025.

[21] Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen Marcus McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. Llemma: An open language model for mathematics. In *The Twelfth International Conference on Learning Representations*, January 2024. URL https://openreview.net/forum?id=4WnqRR915j.

[22] Tanja Baeumel, Josef van Genabith, and Simon Ostermann. The lookahead limitation: Why multi-operand addition is hard for llms. *arXiv preprint arXiv:2502.19981*, 2025.

[23] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[24] Bowen Baker, Joost Huizinga, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. March 2025. URL https://openai.com/index/chain-of-thought-monitoring/.

[25] Marthe Ballon, Andres Algaba, and Vincent Ginis. The relationship between reasoning and performance in large language models–o3 (mini) thinks harder, not longer. *arXiv preprint arXiv:2502.15631*, 2025.

[26] Hritik Bansal, Arian Hosseini, Rishabh Agarwal, Vinh Q. Tran, and Mehran Kazemi. Smaller, weaker, yet better: Training LLM reasoners via compute-optimal sampling. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*, January 2025. URL https://openreview.net/forum?id=HuYSURUxs2.

[27] Hieu Tran Bao, Nguyen Cong Dat, Nguyen Duc Anh, and Hoang Thanh Tung. Learning to stop overthinking at test time. *arXiv preprint arXiv:2502.10954*, 2025.

[28] Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. *arXiv preprint arXiv:2501.00663*, 2024.

[29] Yoshua Bengio, Michael Cohen, Damiano Fornasiere, Joumana Ghosn, Pietro Greiner, Matt MacDermott, Sören Mindermann, Adam Oberman, Jesse Richardson, Oliver Richardson, et al. Superintelligent agents pose catastrophic risks: Can scientist ai offer a safer path? *arXiv preprint arXiv:2502.15657*, 2025.

[30] Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Philip Fox, Ben Garfinkel, Danielle Goldfarb, et al. International ai safety report. *arXiv preprint arXiv:2501.17805*, 2025.

[31] Leonardo Bertolazzi, Philipp Mondorf, Barbara Plank, and Raffaella Bernardi. The validation gap: A mechanistic analysis of how language models compute arithmetic but fail to validate it. *arXiv preprint arXiv:2502.11771*, 2025.

[32] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690, Mar. 2024. doi: 10.1609/aaai.v38i16.29720. URL https://ojs.aaai.org/index.php/AAAI/article/view/29720.

[33] Maciej Besta, Florim Memedi, Zhenyu Zhang, Robert Gerstenberger, Guangyuan Piao, Nils Blach, Piotr Nyczyk, Marcin Copik, Grzegorz Kwaśniewski, Jürgen Müller, et al. Demystifying chains, trees, and graphs of thoughts. *arXiv preprint arXiv:2401.14295*, 2024.

[34] Maciej Besta, Julia Barth, Eric Schreiber, Ales Kubicek, Afonso Catarino, Robert Gerstenberger, Piotr Nyczyk, Patrick Iff, Yueling Li, Sam Houliston, et al. Reasoning language models: A blueprint. *arXiv preprint arXiv:2501.11223*, 2025.

[35] Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

[36] Zhen Bi, Ningyu Zhang, Yinuo Jiang, Shumin Deng, Guozhou Zheng, and Huajun Chen. When do program-of-thought works for reasoning? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17691–17699, 2024.

[37] Zhenni Bi, Kai Han, Chuanjian Liu, Yehui Tang, and Yunhe Wang. Forest-of-thought: Scaling test-time compute for enhancing llm reasoning. *arXiv preprint arXiv:2412.09078*, 2024.

[38] Edoardo Botta, Yuchen Li, Aashay Mehta, Jordan T Ash, Cyril Zhang, and Andrej Risteski. On the query complexity of verifier-assisted language generation. *arXiv preprint arXiv:2502.12123*, 2025.

[39] David Brandfonbrener, Simon Henniger, Sibi Raja, Tarun Prasad, Chloe Loughridge, Federico Cassano, Sabrina Ruixin Hu, Jianang Yang, William E Byrd, Robert Zinkov, et al. Vermcts: Synthesizing multi-step programs using a verifier, a large language model, and tree search. *arXiv preprint arXiv:2402.08147*, 2024.

[40] Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

[41] Dan Busbridge, Amitis Shidani, Floris Weers, Jason Ramapuram, Etai Littwin, and Russ Webb. Distillation scaling laws. *arXiv preprint arXiv:2502.08606*, 2025.

[42] Ju-Seung Byun, Jiyun Chun, Jihyung Kil, and Andrew Perrault. ARES: Alternating reinforcement learning and supervised fine-tuning for enhanced multi-modal chain-of-thought reasoning through diverse AI feedback. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4410–4430, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.252. URL https://aclanthology.org/2024.emnlp-main.252/.

[43] Huanqia Cai, Yijun Yang, and Zhifeng Li. System-2 mathematical reasoning via enriched instruction tuning. *arXiv preprint arXiv:2412.16964*, 2024.

[44] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024.

[45] Erik Cambria, Lorenzo Malandri, Fabio Mercorio, Navid Nobani, and Andrea Seveso. Xai meets llms: A survey of the relation between explainable ai and large language models. *arXiv preprint arXiv:2407.15248*, 2024.

[46] Lang Cao. GraphReason: Enhancing reasoning capabilities of large language models through a graph-based verification approach. In Bhavana Dalvi Mishra, Greg Durrett, Peter Jansen, Ben Lipkin, Danilo Neves Ribeiro, Lionel Wong, Xi Ye, and Wenting Zhao, editors, *Proceedings of the 2nd Workshop on Natural Language Reasoning and Structured Explanations (@ACL 2024)*,

pages 1–12, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.nlrse-1.1/.

[47] Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xiannian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, et al. xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. *arXiv preprint arXiv:2401.07037*, 2024.

[48] Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, et al. Mle-bench: Evaluating machine learning agents on machine learning engineering. *arXiv preprint arXiv:2410.07095*, 2024.

[49] Hyeong Soo Chang. On the convergence rate of mcts for the optimal value estimation in markov decision processes. *IEEE Transactions on Automatic Control*, pages 1–6, February 2025. doi: 10.1109/TAC.2025.3538807.

[50] Andong Chen, Yuchen Song, Wenxin Zhu, Kehai Chen, Muyun Yang, Tiejun Zhao, et al. Evaluating o1-like llms: Unlocking reasoning for translation through comprehensive analysis. *arXiv preprint arXiv:2502.11544*, 2025.

[51] Guizhen Chen, Weiwen Xu, Hao Zhang, Hou Pong Chan, Chaoqun Liu, Lidong Bing, Deli Zhao, Anh Tuan Luu, and Yu Rong. Finereason: Evaluating and improving llms' deliberate reasoning through reflective puzzle solving. *arXiv preprint arXiv:2502.20238*, 2025.

[52] Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. Step-level value preference optimization for mathematical reasoning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7889–7903, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.463. URL https://aclanthology.org/2024.findings-emnlp.463/.

[53] Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. Alphamath almost zero: Process supervision without process. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, September 2024. URL https://openreview.net/forum?id=VaXnxQ3UKo.

[54] Haibin Chen, Kangtao Lv, Chengwei Hu, Yanshi Li, Yujin Yuan, Yancheng He, Xingyao Zhang, Langming Liu, Shilei Liu, Wenbo Su, et al. Chineseecomqa: A scalable e-commerce concept evaluation benchmark for large language models. *arXiv preprint arXiv:2502.20196*, 2025.

[55] Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. Benchmarking large language models on answering and explaining challenging medical questions. *arXiv preprint arXiv:2402.18060*, 2024.

[56] Haolin Chen, Yihao Feng, Zuxin Liu, Weiran Yao, Akshara Prabhakar, Shelby Heinecke, Ricky Ho, Phil Mui, Silvio Savarese, Caiming Xiong, et al. Language models are hidden reasoners: Unlocking latent reasoning capabilities via self-rewarding. *arXiv preprint arXiv:2411.04282*, 2024.

[57] Hardy Chen, Haoqin Tu, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Vl-thinking: An r1-derived visual instruction tuning dataset for thinkable lvlms. https://github.com/UCSC-VLAA/VL-Thinking, 2025.

[58] Jian Chen, Guohao Tang, Guofu Zhou, and Wu Zhu. Chatgpt and deepseek: Can they predict the stock market and macroeconomy? *arXiv preprint arXiv:2502.10008*, 2025.

[59] Jiefeng Chen, Jie Ren, Xinyun Chen, Chengrun Yang, Ruoxi Sun, and Sercan Ö Arık. Sets: Leveraging self-verification and self-correction for improved test-time scaling. *arXiv preprint arXiv:2501.19306*, 2025.

[60] Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*, 2024.

[61] Justin Chih-Yao Chen, Archiki Prasad, Swarnadeep Saha, Elias Stengel-Eskin, and Mohit Bansal. Magicore: Multi-agent, iterative, coarse-to-fine refinement for reasoning. *arXiv preprint arXiv:2409.12147*, 2024.

37

[62] Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. R1-v: Reinforcing super gener-alization ability in vision-language models with less than \$3. https://github.com/Deep-Agent/R1-V, 2025. Accessed: 2025-02-02.

[63] Michael K Chen, Xikun Zhang, and Dacheng Tao. Justlogic: A comprehensive benchmark for evaluating deductive reasoning in large language models. *arXiv preprint arXiv:2501.14851*, 2025.

[64] Qiguang Chen, Libo Qin, Jiaqi WANG, Jingxuan Zhou, and Wanxiang Che. Unlocking the capabilities of thought: A reasoning boundary framework to quantify and optimize chain-of-thought. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, September 2024. URL https://openreview.net/forum?id=pC44UMwy2v.

[65] Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M^{3}CoT: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8199–8221, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.446. URL https://aclanthology.org/2024.acl-long.446/.

[66] Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiaqi Wang, Mengkang Hu, Zhi Chen, Wanxiang Che, and Ting Liu. Ecm: A unified electronic circuit model for explaining the emergence of in-context learning and chain-of-thought in large language model. *arXiv preprint arXiv:2502.03325*, 2025.

[67] Qiqi Chen, Xinpeng Wang, Philipp Mondorf, Michael A Hedderich, and Barbara Plank. Understanding when tree of thoughts succeeds: Larger models excel in generation, not discrimination. *arXiv preprint arXiv:2410.17820*, 2024.

[68] Sijia Chen and Baochun Li. Toward adaptive reasoning in large language models with thought rollback. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 7033–7056. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/chen24y.html.

[69] Weizhe Chen, Sven Koenig, and Bistra Dilkina. Iterative deepening sampling for large language models. *arXiv preprint arXiv:2502.05449*, 2025.

[70] Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*, November 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=YfZ4ZPt8zd.

[71] Wenxiang Chen, Wei He, Zhiheng Xi, Honglin Guo, Boyang Hong, Jiazheng Zhang, Rui Zheng, Nijun Li, Tao Gui, Yun Li, et al. Better process supervision with bi-directional rewarding signals. *arXiv preprint arXiv:2503.04618*, 2025.

[72] Xinghao Chen, Zhijing Sun, Wenjin Guo, Miaoran Zhang, Yanjun Chen, Yirong Sun, Hui Su, Yijie Pan, Dietrich Klakow, Wenjie Li, et al. Unveiling the key factors for distilling chain-of-thought reasoning. *arXiv preprint arXiv:2502.18001*, 2025.

[73] Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, et al. Do not think that much for 2+ 3=? on the overthinking of o1-like llms. *arXiv preprint arXiv:2412.21187*, 2024.

[74] Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. Teaching large language models to self-debug. In *The Twelfth International Conference on Learning Representations*, January 2024. URL https://openreview.net/forum?id=KuPixIqPiq.

[75] Yanxi Chen, Xuchen Pan, Yaliang Li, Bolin Ding, and Jingren Zhou. A simple and provable scaling law for the test-time compute of large language models. *arXiv preprint arXiv:2411.19477*, 2024.

[76] Yezeng Chen, Zui Chen, and Yi Zhou. Brain-inspired two-stage approach: Enhancing mathematical reasoning by imitating human thought processes. *arXiv preprint arXiv:2403.00800*, 2024.

[77] Yilong Chen, Junyuan Shang, Zhenyu Zhang, Yanxi Xie, Jiawei Sheng, Tingwen Liu, Shuo-huan Wang, Yu Sun, Hua Wu, and Haifeng Wang. Inner thinking transformer: Leveraging dynamic depth scaling to foster adaptive internal thinking. *arXiv preprint arXiv:2502.13842*, 2025.

[78] Zhi Chen, Qiguang Chen, Libo Qin, Qipeng Guo, Haijun Lv, Yicheng Zou, Wanxiang Che, Hang Yan, Kai Chen, and Dahua Lin. What are the essential factors in crafting effective long context multi-hop instruction datasets? insights and best practices. *arXiv preprint arXiv:2409.01893*, 2024.

[79] Ziru Chen, Michael White, Ray Mooney, Ali Payani, Yu Su, and Huan Sun. When is tree search useful for LLM planning? it depends on the discriminator. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13659–13678, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.738. URL https://aclanthology.org/2024.acl-long.738/.

[80] Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 6621–6642. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/chen24j.html.

[81] Zui Chen, Tianqiao Liu, Mi Tian, Qing Tong, Weiqi Luo, and Zitao Liu. Advancing math reasoning in language models: The impact of problem-solving data, data synthesis methods, and training stages. *arXiv preprint arXiv:2501.14002*, 2025.

[82] Jiale Cheng, Xiao Liu, Cunxiang Wang, Xiaotao Gu, Yida Lu, Dan Zhang, Yuxiao Dong, Jie Tang, Hongning Wang, and Minlie Huang. Spar: Self-play with tree-search refinement to improve instruction-following in large language models. *arXiv preprint arXiv:2412.11605*, 2024.

[83] Kanzhi Cheng, Yantao Li, Fangzhi Xu, Jianbing Zhang, Hao Zhou, and Yang Liu. Vision-language models can self-improve reasoning via reflection. *arXiv preprint arXiv:2411.00855*, 2024.

[84] Xiaoxue Cheng, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. Think more, hallucinate less: Mitigating hallucinations via dual process of fast and slow thinking. *arXiv preprint arXiv:2501.01306*, 2025.

[85] Zihui Cheng, Qiguang Chen, Jin Zhang, Hao Fei, Xiaocheng Feng, Wanxiang Che, Min Li, and Libo Qin. Comt: A novel benchmark for chain of multi-modal thought on large vision-language models. *arXiv preprint arXiv:2412.12932*, 2024.

[86] Daiki Chijiwa, Taku Hasegawa, Kyosuke Nishida, Kuniko Saito, and Susumu Takeuchi. Portable reward tuning: Towards reusable fine-tuning across different pretrained models. *arXiv preprint arXiv:2502.12776*, 2025.

[87] François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

[88] Sanjiban Choudhury. Process reward models for llm agents: Practical framework and directions. *arXiv preprint arXiv:2502.10325*, 2025.

[89] Jishnu Ray Chowdhury and Cornelia Caragea. Zero-shot verification-guided chain of thoughts. *arXiv preprint arXiv:2501.13122*, 2025.

[90] Konstantina Christakopoulou, Shibl Mourad, and Maja Mataric. Agents thinking fast and slow: A talker-reasoner architecture. In *NeurIPS 2024 Workshop on Open-World Agents*, October 2024. URL https://openreview.net/forum?id=xPhcP6rbI4.

[91] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025.

[92] Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 1173–1203, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.65. URL https://aclanthology.org/2024.acl-long.65/.

[93] Jennifer Chu-Carroll, Andrew Beck, Greg Burnham, David OS Melville, David Nachman, A Erdem Özcan, and David Ferrucci. Beyond llms: Advancing the landscape of complex reasoning. *arXiv preprint arXiv:2402.08064*, 2024.

[94] Daniel JH Chung, Zhiqi Gao, Yurii Kvasiuk, Tianyi Li, Moritz Münchmeyer, Maja Rudolph, Frederic Sala, and Sai Chaitanya Tadepalli. Theoretical physics benchmark (tpbench)–a dataset and study of ai reasoning capabilities in theoretical physics. *arXiv preprint arXiv:2502.15815*, 2025.

[95] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[96] Alejandro Cuadron, Dacheng Li, Wenjie Ma, Xingyao Wang, Yichuan Wang, Siyuan Zhuang, Shu Liu, Luis Gaspar Schroeder, Tian Xia, Huanzhi Mao, et al. The danger of overthinking: Examining the reasoning-action dilemma in agentic tasks. *arXiv preprint arXiv:2502.08235*, 2025.

[97] Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025.

[98] Yingqian Cui, Pengfei He, Jingying Zeng, Hui Liu, Xianfeng Tang, Zhenwei Dai, Yan Han, Chen Luo, Jing Huang, Zhen Li, et al. Stepwise perplexity-guided refinement for efficient chain-of-thought reasoning in large language models. *arXiv preprint arXiv:2502.13260*, 2025.

[99] Jianbo Dai, Jianqiao Lu, Yunlong Feng, Dong Huang, Guangtao Zeng, Rongju Ruan, Ming Cheng, Haochen Tan, and Zhijiang Guo. Mhpp: Exploring the capabilities and limitations of language models beyond basic code generation. *arXiv preprint arXiv:2405.11430*, 2024.

[100] Yuntian Deng, Yejin Choi, and Stuart Shieber. From explicit cot to implicit cot: Learning to internalize cot step by step. *arXiv preprint arXiv:2405.14838*, 2024.

[101] Lauro Langosco Di Langosco, Jack Koch, Lee D Sharkey, Jacob Pfau, and David Krueger. Goal misgeneralization in deep reinforcement learning. In *International Conference on Machine Learning*, pages 12004–12019. PMLR, October 2022.

[102] Yifu Ding, Wentao Jiang, Shunyu Liu, Yongcheng Jing, Jinyang Guo, Yingjie Wang, Jing Zhang, Zengmao Wang, Ziwei Liu, Bo Du, et al. Dynamic parallel tree search for efficient llm reasoning. *arXiv preprint arXiv:2502.16235*, 2025.

[103] Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*, 2024.

[104] Kefan Dong and Tengyu Ma. Stp: Self-play llm theorem provers with iterative conjecturing and proving. *arXiv e-prints*, pages arXiv–2502, January 2025.

[105] Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. *arXiv preprint arXiv:2411.14432*, 2024.

[106] Zhichen Dong, Zhanhui Zhou, Zhixuan Liu, Chao Yang, and Chaochao Lu. Emergent response planning in llm. *arXiv preprint arXiv:2502.06258*, 2025.

[107] Shihan Dou, Yan Liu, Haoxiang Jia, Limao Xiong, Enyu Zhou, Wei Shen, Junjie Shan, Caishuang Huang, Xiao Wang, Xiaoran Fan, et al. Stepcoder: Improve code generation with reinforcement learning from compiler feedback. *arXiv preprint arXiv:2402.01391*, 2024.

[108] Iddo Drori, Gaston Longhitano, Mao Mao, Seunghwan Hyun, Yuke Zhang, Sungjun Park, Zachary Meeks, Xin-Yu Zhang, Ben Segev, Howard Yong, et al. Diverse inference and verification for advanced reasoning. *arXiv preprint arXiv:2502.09955*, 2025.

[109] Kounianhua Du, Hanjing Wang, Jianxing Liu, Jizheng Chen, Xinyi Dai, Yasheng Wang, Ruiming Tang, Yong Yu, Jun Wang, and Weinan Zhang. Boost, disentangle, and customize: A robust system2-to-system1 pipeline for code generation. *arXiv preprint arXiv:2502.12492*, 2025.

[110] Weihua Du, Yiming Yang, and Sean Welleck. Optimizing temperature for language models with multi-sample inference. *arXiv preprint arXiv:2502.05234*, 2025.

[111] Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, Kang Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, et al. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*, 2025.

[112] Yifan Du, Zikang Liu, Yifan Li, Wayne Xin Zhao, Yuqi Huo, Bingning Wang, Weipeng Chen, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. Virgo: A preliminary exploration on reproducing o1-like mllm. *arXiv preprint arXiv:2501.01904*, 2025.

[113] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[114] Subhabrata Dutta, Joykirat Singh, Soumen Chakrabarti, and Tanmoy Chakraborty. How to think step-by-step: A mechanistic understanding of chain-of-thought reasoning. *Transactions on Machine Learning Research*, July 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=uHLDkQVtyC.

[115] Ahmed El-Kishky, Alexander Wei, Andre Saraiva, Borys Minaev, Daniel Selsam, David Dohan, Francis Song, Hunter Lightman, Ignasi Clavera, Jakub Pachocki, et al. Competitive programming with large reasoning models. *arXiv preprint arXiv:2502.06807*, 2025.

[116] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

[117] Yi Fang, Wenjie Wang, Yang Zhang, Fengbin Zhu, Qifan Wang, Fuli Feng, and Xiangnan He. Large language models for recommendation with deliberative user preference alignment. *arXiv preprint arXiv:2502.02061*, 2025.

[118] Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. Towards revealing the mystery behind chain of thought: A theoretical perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*, September 2023. URL https://openreview.net/forum?id=qHrADgAdYu.

[119] Xiachong Feng, Longxu Dou, and Lingpeng Kong. Reasoning does not necessarily improve role-playing ability. *arXiv preprint arXiv:2502.16940*, 2025.

[120] Xueyang Feng, Bo Lan, Quanyu Dai, Lei Wang, Jiakai Tang, Xu Chen, Zhenhua Dong, and Ji-Rong Wen. Improving retrospective language agents via joint policy gradient optimization. *arXiv preprint arXiv:2503.01490*, 2025.

[121] Chrisantha Fernando, Dylan Sunil Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 13481–13544. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/fernando24a.html.

[122] Thomas Palmeira Ferraz, Kartik Mehta, Yu-Hsiang Lin, Haw-Shiuan Chang, Shereen Oraby, Sijia Liu, Vivek Subramanian, Tagyoung Chung, Mohit Bansal, and Nanyun Peng. LLM self-correction with deCRIM: Decompose, critique, and refine for enhanced following of instructions with multiple constraints. In *The First Workshop on System-2 Reasoning at Scale, NeurIPS'24*, October 2024. URL https://openreview.net/forum?id=RQ6Ff8lso0.

[123] Jiarun Fu, Lizhong Ding, Hao Li, Pengqi Li, Qiuning Wei, and Xu Chen. Unveiling and causalizing cot: A causal pespective. *arXiv preprint arXiv:2502.18239*, 2025.

[124] Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations*, February 2023. URL https://openreview.net/forum?id=yf1icZHC-l9.

[125] Víctor Gallego. Metasc: Test-time safety specification optimization for language models. *arXiv preprint arXiv:2502.07985*, 2025.

[126] Zeyu Gan, Yun Liao, and Yong Liu. Rethinking external slow-thinking: From snowball errors to probability of correct reasoning. *arXiv preprint arXiv:2501.15602*, 2025.

[127] Kanishk Gandhi, Denise HJ Lee, Gabriel Grand, Muxin Liu, Winson Cheng, Archit Sharma, and Noah Goodman. Stream of search (sos): Learning to search in language. In *First Conference on Language Modeling*, July 2024.

[128] 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.

[129] Bofei Gao, Zefan Cai, Runxin Xu, Peiyi Wang, Ce Zheng, Runji Lin, Keming Lu, Junyang Lin, Chang Zhou, Tianyu Liu, and Baobao Chang. The reason behind good or bad: Towards a better mathematical verifier with natural language feedback, 2024. URL https://arxiv.org/abs/2406.14024.

[130] Bofei Gao, Zefan Cai, Runxin Xu, Peiyi Wang, Ce Zheng, Runji Lin, Keming Lu, Dayiheng Liu, Chang Zhou, Wen Xiao, et al. Llm critics help catch bugs in mathematics: Towards a better mathematical verifier with natural language feedback. *arXiv preprint arXiv:2406.14024*, 2024.

[131] Jiaxuan Gao, Shusheng Xu, Wenjie Ye, Weilin Liu, Chuyi He, Wei Fu, Zhiyu Mei, Guangju Wang, and Yi Wu. On designing effective rl reward at training time for llm reasoning. *arXiv preprint arXiv:2410.15115*, 2024.

[132] Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. PAL: Program-aided language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10764–10799. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/gao23f.html.

[133] Tianchen Gao, Jiashun Jin, Zheng Tracy Ke, and Gabriel Moryoussef. A comparison of deepseek and other llms. *arXiv preprint arXiv:2502.03688*, 2025.

[134] Zitian Gao, Boye Niu, Xuzheng He, Haotian Xu, Hongzhang Liu, Aiwei Liu, Xuming Hu, and Lijie Wen. Interpretable contrastive monte carlo tree search reasoning. *arXiv preprint arXiv:2410.01707*, 2024.

[135] Jonas Gehring, Kunhao Zheng, Jade Copet, Vegard Mella, Taco Cohen, and Gabriel Synnaeve. Rlef: Grounding code llms in execution feedback with reinforcement learning. *arXiv preprint arXiv:2410.02089*, 2024.

[136] Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. *arXiv preprint arXiv:2502.05171*, 2025.

[137] Zelalem Gero, Chandan Singh, Hao Cheng, Tristan Naumann, Michel Galley, Jianfeng Gao, and Hoifung Poon. Self-verification improves few-shot clinical information extraction. In *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*, June 2023. URL https://openreview.net/forum?id=SBbJICrglS.

[138] Akash Ghosh, Debayan Datta, Sriparna Saha, and Chirag Agarwal. The multilingual mind: A survey of multilingual reasoning in language models. *arXiv preprint arXiv:2502.09457*, 2025.

[139] Panagiotis Giadikiaroglou, Maria Lymperaiou, Giorgos Filandrianos, and Giorgos Stamou. Puzzle solving using reasoning of large language models: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11574–11591, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.646. URL https://aclanthology.org/2024.emnlp-main.646/.

[140] Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, et al. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. *arXiv preprint arXiv:2411.04872*, 2024.

[141] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

[142] Olga Golovneva, Moya Peng Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. ROSCOE: A suite of metrics for scoring step-by-step reasoning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=xYlJRpzZtsY.

[143] Olga Golovneva, Sean O'Brien, Ramakanth Pasunuru, Tianlu Wang, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. PATHFINDER: Guided search over multi-step reasoning paths. In *R0-FoMo:Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*, December 2023. URL https://openreview.net/forum?id=5TsfEEwRsu.

[144] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

[145] Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*, 2023.

[146] Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Minlie Huang, Nan Duan, and Weizhu Chen. Tora: A tool-integrated reasoning agent for mathematical problem solving. *arXiv preprint arXiv:2309.17452*, 2023.

[147] Julia Grosse, Ruotian Wu, Ahmad Rashid, Philipp Hennig, Pascal Poupart, and Agustinus Kristiadi. Uncertainty-guided optimization on large language model search trees. *arXiv preprint arXiv:2407.03951*, 2024.

[148] Yanggan Gu, Junzhuo Li, Sirui Huang, Xin Zou, Zhenghua Li, and Xuming Hu. Capturing nuanced preferences: Preference-aligned distillation for small language models. *arXiv preprint arXiv:2502.14272*, 2025.

[149] Xinyan Guan, Yanjiang Liu, Xinyu Lu, Boxi Cao, Ben He, Xianpei Han, Le Sun, Jie Lou, Bowen Yu, Yaojie Lu, et al. Search, verify and feedback: Towards next generation post-training paradigm of foundation models via verifier engineering. *arXiv preprint arXiv:2411.11504*, 2024.

[150] Xinyan Guan, Jiali Zeng, Fandong Meng, Chunlei Xin, Yaojie Lu, Hongyu Lin, Xianpei Han, Le Sun, and Jie Zhou. Deeprag: Thinking to retrieval step by step for large language models. *arXiv preprint arXiv:2502.01142*, 2025.

[151] Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. rstar-math: Small llms can master math reasoning with self-evolved deep thinking. *arXiv preprint arXiv:2501.04519*, 2025.

[152] Aryan Gulati, Brando Miranda, Eric Chen, Emily Xia, Kai Fronsdal, Bruno de Moraes Dumont, and Sanmi Koyejo. Putnam-AXIOM: A functional and static benchmark for measuring higher level mathematical reasoning. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*, 2024. URL https://openreview.net/forum?id=YXnwlZe0yf.

[153] Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, et al. Reinforced self-training (rest) for language modeling. *arXiv preprint arXiv:2308.08998*, 2023.

[154] Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yu Wu, YK Li, et al. Deepseek-coder: When the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*, 2024.

[155] 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.

[156] Honglin Guo, Kai Lv, Qipeng Guo, Tianyi Liang, Zhiheng Xi, Demin Song, Qiuyinzhe Zhang, Yu Sun, Kai Chen, Xipeng Qiu, et al. Critiq: Mining data quality criteria from human preferences. *arXiv preprint arXiv:2502.19279*, 2025.

[157] Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. Can we generate images with cot? let's verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*, 2025.

[158] Tingxu Han, Chunrong Fang, Shiyu Zhao, Shiqing Ma, Zhenyu Chen, and Zhenting Wang. Token-budget-aware llm reasoning. *arXiv preprint arXiv:2412.18547*, 2024.

[159] Michael Hanna, Ollie Liu, and Alexandre Variengien. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. September 2023. URL https://openreview.net/forum?id=p4PckNQR8k.

[160] Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8154–8173, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.507. URL https://aclanthology.org/2023.emnlp-main.507/.

[161] Shibo Hao, Yi Gu, Haotian Luo, Tianyang Liu, Xiyan Shao, Xinyuan Wang, Shuhua Xie, Haodi Ma, Adithya Samavedhi, Qiyue Gao, Zhen Wang, and Zhiting Hu. LLM reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models. In *First Conference on Language Modeling*, July 2024. URL https://openreview.net/forum?id=b0y6fbSUG0.

[162] 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. *arXiv preprint arXiv:2412.06769*, 2024.

[163] Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. *arXiv preprint arXiv:2501.05444*, 2025.

[164] Alexander Havrilla, Sharath Chandra Raparthy, Christoforos Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravinskyi, Eric Hambro, and Roberta Raileanu. GLore: When, where, and how to improve LLM reasoning via global and local refinements. In *Forty-first International Conference on Machine Learning*, May 2024. URL https://openreview.net/forum?id=LH6R06NxdB.

[165] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.211. URL https://aclanthology.org/2024.acl-long.211/.

[166] Chengbo He, Bochao Zou, Xin Li, Jiansheng Chen, Junliang Xing, and Huimin Ma. Enhancing llm reasoning with multi-path collaborative reactive and reflection agents. *arXiv preprint arXiv:2501.00430*, 2024.

[167] Junda He, Jieke Shi, Terry Yue Zhuo, Christoph Treude, Jiamou Sun, Zhenchang Xing, Xiaoning Du, and David Lo. From code to courtroom: Llms as the new software judges. *arXiv preprint arXiv:2503.02246*, 2025.

[168] Mingqian He, Yongliang Shen, Wenqi Zhang, Zeqi Tan, and Weiming Lu. Advancing process verification for large language models via tree-based preference learning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2086–2099, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.125. URL https://aclanthology.org/2024.emnlp-main.125/.

[169] Qiangqiang He, Shuwei Qian, Jie Zhang, and Chongjun Wang. Inference retrieval-augmented multi-modal chain-of-thoughts reasoning for language models. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10888701.

[170] Tao He, Hao Li, Jingchang Chen, Runxuan Liu, Yixin Cao, Lizi Liao, Zihao Zheng, Zheng Chu, Jiafeng Liang, Ming Liu, et al. A survey on complex reasoning of large language models through the lens of self-evolution. February 2025.

[171] Yancheng He, Shilong Li, Jiaheng Liu, Weixun Wang, Xingyuan Bu, Ge Zhang, Zhongyuan Peng, Zhaoxiang Zhang, Wenbo Su, and Bo Zheng. Can large language models detect errors in long chain-of-thought reasoning? *arXiv preprint arXiv:2502.19361*, 2025.

[172] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, October 2021. URL https://openreview.net/forum?id=7Bywt2mQsCe.

[173] Alex Heyman and Joel Zylberberg. Evaluating the systematic reasoning abilities of large language models through graph coloring. *arXiv preprint arXiv:2502.07087*, 2025.

[174] Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14852–14882, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.830. URL https://aclanthology.org/2023.acl-long.830/.

[175] Matthew Douglas Hoffman, Du Phan, david dohan, Sholto Douglas, Tuan Anh Le, Aaron T Parisi, Pavel Sountsov, Charles Sutton, Sharad Vikram, and Rif A. Saurous. Training chain-of-thought via latent-variable inference. In *Thirty-seventh Conference on Neural Information Processing Systems*, September 2023. URL https://openreview.net/forum?id=a147pIS2Co.

[176] Ruixin Hong, Xinyu Pang, and Changshui Zhang. Advances in reasoning by prompting large language models: A survey. *Cybernetics and Intelligence*, pages 1–15, 2024. doi: 10.26599/CAI.2024.9390004.

[177] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, and Jie Tang. Cogagent: A visual language model for gui agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14281–14290, June 2024.

[178] Arian Hosseini, Alessandro Sordoni, Daniel Kenji Toyama, Aaron Courville, and Rishabh Agarwal. Not all LLM reasoners are created equal. In *The First Workshop on System-2 Reasoning at Scale, NeurIPS'24*, October 2024. URL https://openreview.net/forum?id=aPAWbip1xV.

[179] Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordoni, and Rishabh Agarwal. V-STar: Training verifiers for self-taught reasoners. In *First Conference on Language Modeling*, July 2024. URL https://openreview.net/forum?id=stmqBSW2dV.

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

[181] Jian Hu. Reinforce++: A simple and efficient approach for aligning large language models. *arXiv preprint arXiv:2501.03262*, 2025.

[182] Jian Hu, Xibin Wu, Zilin Zhu, Xianyu, Weixun Wang, Dehao Zhang, and Yu Cao. Openrlhf: An easy-to-use, scalable and high-performance rlhf framework. *arXiv preprint arXiv:2405.11143*, 2024.

[183] Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, and Heung-Yeung Shum Xiangyu Zhang. Open-reasoner-zero: An open source approach to scaling reinforcement learning on the base model. https://github.com/Open-Reasoner-Zero/Open-Reasoner-Zero, February 2025.

[184] Mengkang Hu, Tianxing Chen, Qiguang Chen, Yao Mu, Wenqi Shao, and Ping Luo. Hiagent: Hierarchical working memory management for solving long-horizon agent tasks with large language model. *arXiv preprint arXiv:2408.09559*, 2024.

[185] Mengkang Hu, Yao Mu, Xinmiao Chelsey Yu, Mingyu Ding, Shiguang Wu, Wenqi Shao, Qiguang Chen, Bin Wang, Yu Qiao, and Ping Luo. Tree-planner: Efficient close-loop task

planning with large language models. In *The Twelfth International Conference on Learning Representations*, January 2024. URL https://openreview.net/forum?id=Glcsog6zOe.

[186] Mengkang Hu, Tianxing Chen, Yude Zou, Yuheng Lei, Qiguang Chen, Ming Li, Hongyuan Zhang, Wenqi Shao, and Ping Luo. Text2world: Benchmarking large language models for symbolic world model generation. *arXiv preprint arXiv:2502.13092*, 2025.

[187] Zhiyuan Hu, Chumin Liu, Xidong Feng, Yilun Zhao, See-Kiong Ng, Anh Tuan Luu, Junxian He, Pang Wei Koh, and Bryan Hooi. Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in large language models. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, March 2024. URL https://openreview.net/forum?id=ZWyLjimciT.

[188] Chenghua Huang, Lu Wang, Fangkai Yang, Pu Zhao, Zhixu Li, Qingwei Lin, Dongmei Zhang, Saravan Rajmohan, and Qi Zhang. Lean and mean: Decoupled value policy optimization with global value guidance. *arXiv preprint arXiv:2502.16944*, 2025.

[189] Haiduo Huang, Fuwei Yang, Zhenhua Liu, Yixing Xu, Jinze Li, Yang Liu, Xuanwu Yin, Dong Li, Pengju Ren, and Emad Barsoum. Jakiro: Boosting speculative decoding with decoupled multi-head via moe. *arXiv preprint arXiv:2502.06282*, 2025.

[190] Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.826. URL https://aclanthology.org/2023.findings-emnlp.826/.

[191] Jen-tse Huang, Eric John Li, Man Ho Lam, Tian Liang, Wenxuan Wang, Youliang Yuan, Wenxiang Jiao, Xing Wang, Zhaopeng Tu, and Michael R Lyu. How far are we on the decision-making of llms? evaluating llms' gaming ability in multi-agent environments. *arXiv preprint arXiv:2403.11807*, 2024.

[192] Jiaxing Huang and Jingyi Zhang. A survey on evaluation of multimodal large language models. *arXiv preprint arXiv:2408.15769*, 2024.

[193] Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.67. URL https://aclanthology.org/2023.findings-acl.67/.

[194] Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations*, January 2024. URL https://openreview.net/forum?id=IkmD3fKBPQ.

[195] Kaixuan Huang, Jiacheng Guo, Zihao Li, Xiang Ji, Jiawei Ge, Wenzhe Li, Yingqing Guo, Tianle Cai, Hui Yuan, Runzhe Wang, et al. Math-perturb: Benchmarking llms' math reasoning abilities against hard perturbations. *arXiv preprint arXiv:2502.06453*, 2025.

[196] Lei Huang, Xiaocheng Feng, Weitao Ma, Liang Zhao, Yuchun Fan, Weihong Zhong, Dongliang Xu, Qing Yang, Hongtao Liu, and Bing Qin. Advancing large language model attribution through self-improving. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3822–3836, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.223. URL https://aclanthology.org/2024.emnlp-main.223/.

[197] Shulin Huang, Linyi Yang, Yan Song, Shuang Chen, Leyang Cui, Ziyu Wan, Qingcheng Zeng, Ying Wen, Kun Shao, Weinan Zhang, et al. Thinkbench: Dynamic out-of-distribution evaluation for robust llm reasoning. *arXiv preprint arXiv:2502.16268*, 2025.

[198] Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, Zachary Yahn, Yichang Xu, and Ling Liu. Safety tax: Safety alignment makes your large reasoning models less reasonable. *arXiv preprint arXiv:2503.00555*, 2025.

[199] Yiming Huang, Xiao Liu, Yeyun Gong, Zhibin Gou, Yelong Shen, Nan Duan, and Weizhu Chen. Key-point-driven data synthesis with its enhancement on mathematical reasoning. *arXiv preprint arXiv:2403.02333*, 2024.

[200] Zhen Huang, Haoyang Zou, Xuefeng Li, Yixiu Liu, Yuxiang Zheng, Ethan Chern, Shijie Xia, Yiwei Qin, Weizhe Yuan, and Pengfei Liu. O1 replication journey–part 2: Surpassing o1-preview through simple distillation, big progress or bitter lesson? *arXiv preprint arXiv:2411.16489*, 2024.

[201] Zhongzhen Huang, Gui Geng, Shengyi Hua, Zhen Huang, Haoyang Zou, Shaoting Zhang, Pengfei Liu, and Xiaofan Zhang. O1 replication journey–part 3: Inference-time scaling for medical reasoning. *arXiv preprint arXiv:2501.06458*, 2025.

[202] Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2.5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

[203] Hyeonbin Hwang, Doyoung Kim, Seungone Kim, Seonghyeon Ye, and Minjoon Seo. Self-explore: Enhancing mathematical reasoning in language models with fine-grained rewards. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1444–1466, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.78. URL https://aclanthology.org/2024.findings-emnlp.78/.

[204] Shima Imani, Liang Du, and Harsh Shrivastava. Mathprompter: Mathematical reasoning using large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 37–42, July 2023.

[205] Md Ashraful Islam, Mohammed Eunus Ali, and Md Rizwan Parvez. Mapcoder: Multi-agent code generation for competitive problem solving. *arXiv preprint arXiv:2405.11403*, 2024.

[206] Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. Camels in a changing climate: Enhancing lm adaptation with tulu 2, 2023.

[207] Hamish Ivison, Yizhong Wang, Jiacheng Liu, Zeqiu Wu, Valentina Pyatkin, Nathan Lambert, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. Unpacking DPO and PPO: Disentangling best practices for learning from preference feedback. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, September 2024. URL https://openreview.net/forum?id=JMBWTlazjW.

[208] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[209] Eeshaan Jain, Johann Wenckstern, Benedikt von Querfurth, and Charlotte Bunne. Test-time view selection for multi-modal decision making. In *ICLR 2025 Workshop on Machine Learning for Genomics Explorations*, March 2025. URL https://openreview.net/forum?id=aNmZ9s6BZV.

[210] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations*, January 2025. URL https://openreview.net/forum?id=chfJJYC3iL.

[211] Ke Ji, Jiahao Xu, Tian Liang, Qiuzhi Liu, Zhiwei He, Xingyu Chen, Xiaoyuan Liu, Zhijie Wang, Junying Chen, Benyou Wang, et al. The first few tokens are all you need: An efficient and effective unsupervised prefix fine-tuning method for reasoning models. *arXiv preprint arXiv:2503.02875*, 2025.

[212] Tao Ji, Bin Guo, Yuanbin Wu, Qipeng Guo, Lixing Shen, Zhan Chen, Xipeng Qiu, Qi Zhang, and Tao Gui. Towards economical inference: Enabling deepseek's multi-head latent attention in any transformer-based llms. *arXiv preprint arXiv:2502.14837*, 2025.

[213] Yichao Ji. A small step towards reproducing openai o1: Progress report on the steiner open source models, October 2024. URL https://medium.com/@peakji/b9a756a00855.

[214] Yixin Ji, Juntao Li, Hai Ye, Kaixin Wu, Jia Xu, Linjian Mo, and Min Zhang. Test-time computing: from system-1 thinking to system-2 thinking. *arXiv preprint arXiv:2501.02497*, 2025.

[215] Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. Towards mitigating LLM hallucination via self reflection. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.123. URL https://aclanthology.org/2023.findings-emnlp.123/.

[216] Zeyu Jia, Alexander Rakhlin, and Tengyang Xie. Do we need to verify step by step? rethinking process supervision from a theoretical perspective. *arXiv preprint arXiv:2502.10581*, 2025.

[217] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, October 2023. URL https://arxiv.org/abs/2310.06825.

[218] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[219] Fengqing Jiang, Zhangchen Xu, Yuetai Li, Luyao Niu, Zhen Xiang, Bo Li, Bill Yuchen Lin, and Radha Poovendran. Safechain: Safety of language models with long chain-of-thought reasoning capabilities. *arXiv preprint arXiv:2502.12025*, 2025.

[220] Huchen Jiang, Yangyang Ma, Chaofan Ding, Kexin Luan, and Xinhan Di. Towards intrinsic self-correction enhancement in monte carlo tree search boosted reasoning via iterative preference learning. *arXiv preprint arXiv:2412.17397*, 2024.

[221] Jinhao Jiang, Jiayi Chen, Junyi Li, Ruiyang Ren, Shijie Wang, Wayne Xin Zhao, Yang Song, and Tao Zhang. Rag-star: Enhancing deliberative reasoning with retrieval augmented verification and refinement. *arXiv preprint arXiv:2412.12881*, 2024.

[222] Jinhao Jiang, Zhipeng Chen, Yingqian Min, Jie Chen, Xiaoxue Cheng, Jiapeng Wang, Yiru Tang, Haoxiang Sun, Jia Deng, Wayne Xin Zhao, et al. Technical report: Enhancing llm reasoning with reward-guided tree search. *arXiv preprint arXiv:2411.11694*, 2024.

[223] Shuyang Jiang, Yusheng Liao, Zhe Chen, Ya Zhang, Yanfeng Wang, and Yu Wang. Meds^3: Towards medical small language models with self-evolved slow thinking. *arXiv preprint arXiv:2501.12051*, 2025.

[224] Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*, January 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

[225] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14), July 2021. ISSN 2076-3417. doi: 10.3390/app11146421. URL https://www.mdpi.com/2076-3417/11/14/6421.

[226] Mingyu Jin, Weidi Luo, Sitao Cheng, Xinyi Wang, Wenyue Hua, Ruixiang Tang, William Yang Wang, and Yongfeng Zhang. Disentangling memory and reasoning ability in large language models. *arXiv preprint arXiv:2411.13504*, 2024.

[227] Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. The impact of reasoning step length on large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1830–1842, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.108. URL https://aclanthology.org/2024.findings-acl.108/.

[228] Mingyu Jin, Qinkai Yu, Jingyuan Huang, Qingcheng Zeng, Zhenting Wang, Wenyue Hua, Haiyan Zhao, Kai Mei, Yanda Meng, Kaize Ding, Fan Yang, Mengnan Du, and Yongfeng Zhang. Exploring concept depth: How large language models acquire knowledge and concept

at different layers? In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 558–573, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.37/.

[229] Andy L Jones. Scaling scaling laws with board games. *arXiv preprint arXiv:2104.03113*, 2021.

[230] Prashank Kadam. Gpt-guided monte carlo tree search for symbolic regression in financial fraud detection. *arXiv preprint arXiv:2411.04459*, 2024.

[231] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

[232] Ryo Kamoi, Sarkar Snigdha Sarathi Das, Renze Lou, Jihyun Janice Ahn, Yilun Zhao, Xiaoxin Lu, Nan Zhang, Yusen Zhang, Haoran Ranran Zhang, Sujeeth Reddy Vummanthala, Salika Dave, Shaobo Qin, Arman Cohan, Wenpeng Yin, and Rui Zhang. Evaluating LLMs at detecting errors in LLM responses. In *First Conference on Language Modeling*, July 2024. URL https://openreview.net/forum?id=dnwRScljXr.

[233] Jikun Kang, Xin Zhe Li, Xi Chen, Amirreza Kazemi, Qianyi Sun, Boxing Chen, Dong Li, Xu He, Quan He, Feng Wen, et al. Mindstar: Enhancing math reasoning in pre-trained llms at inference time. *arXiv preprint arXiv:2405.16265*, 2024.

[234] Liwei Kang, Zirui Zhao, David Hsu, and Wee Sun Lee. On the empirical complexity of reasoning and planning in LLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2897–2936, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.164. URL https://aclanthology.org/2024.findings-emnlp.164/.

[235] Yu Kang, Xianghui Sun, Liangyu Chen, and Wei Zou. C3ot: Generating shorter chain-of-thought without compromising effectiveness. *arXiv preprint arXiv:2412.11664*, 2024.

[236] Zhewei Kang, Xuandong Zhao, and Dawn Song. Scalable best-of-n selection for large language models via self-certainty. *arXiv preprint arXiv:2502.18581*, 2025.

[237] Manuj Kant, Sareh Nabi, Manav Kant, Roland Scharrer, Megan Ma, and Marzieh Nabi. Towards robust legal reasoning: Harnessing logical llms in law. *arXiv preprint arXiv:2502.17638*, 2025.

[238] Mehran Kazemi, Najoung Kim, Deepti Bhatia, Xin Xu, and Deepak Ramachandran. LAMBADA: Backward chaining for automated reasoning in natural language. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6547–6568, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.361. URL https://aclanthology.org/2023.acl-long.361/.

[239] Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. Vineppo: Unlocking rl potential for llm reasoning through refined credit assignment. *arXiv preprint arXiv:2410.01679*, 2024.

[240] Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. Grace: Discriminator-guided chain-of-thought reasoning. *arXiv preprint arXiv:2305.14934*, 2023.

[241] Artyom Kharinaev, Viktor Moskvoretskii, Egor Shvetsov, Kseniia Studenikina, Bykov Mikhail, and Evgeny Burnaev. Investigating the impact of quantization methods on the safety and reliability of large language models. *arXiv preprint arXiv:2502.15799*, 2025.

[242] Hyunwoo Kim, Melanie Sclar, Tan Zhi-Xuan, Lance Ying, Sydney Levine, Yang Liu, Joshua B Tenenbaum, and Yejin Choi. Hypothesis-driven theory-of-mind reasoning for large language models. *arXiv preprint arXiv:2502.11881*, 2025.

[243] Juno Kim, Denny Wu, Jason Lee, and Taiji Suzuki. Metastable dynamics of chain-of-thought reasoning: Provable benefits of search, rl and distillation. *arXiv preprint arXiv:2502.01694*, 2025.

[244] Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025.

[245] Naryeong Kim, Sungmin Kang, Gabin An, and Shin Yoo. Lachesis: Predicting llm inference accuracy using structural properties of reasoning paths. *arXiv preprint arXiv:2412.08281*, 2024.

[246] Seungone Kim, Se Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. The CoT collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12685–12708, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.782. URL https://aclanthology.org/2023.emnlp-main.782/.

[247] Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.248. URL https://aclanthology.org/2024.emnlp-main.248/.

[248] Sunnie SY Kim, Jennifer Wortman Vaughan, Q Vera Liao, Tania Lombrozo, and Olga Russakovsky. Fostering appropriate reliance on large language models: The role of explanations, sources, and inconsistencies. *arXiv preprint arXiv:2502.08554*, 2025.

[249] Jing Yu Koh, Stephen McAleer, Daniel Fried, and Ruslan Salakhutdinov. Tree search for language model agents. *arXiv preprint arXiv:2407.01476*, 2024.

[250] Deqian Kong, Minglu Zhao, Dehong Xu, Bo Pang, Shu Wang, Edouardo Honig, Zhangzhang Si, Chuan Li, Jianwen Xie, Sirui Xie, et al. Scalable language models with posterior inference of latent thought vectors. *arXiv preprint arXiv:2502.01567*, 2025.

[251] Abhinav Kumar, Jaechul Roh, Ali Naseh, Marzena Karpinska, Mohit Iyyer, Amir Houmansadr, and Eugene Bagdasarian. Overthink: Slowdown attacks on reasoning llms. *arXiv e-prints*, pages arXiv–2502, February 2025.

[252] Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. Training language models to self-correct via reinforcement learning. *arXiv preprint arXiv:2409.12917*, 2024.

[253] Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip H. S. Torr, Salman Khan, and Fahad Shahbaz Khan. Llm post-training: A deep dive into reasoning large language models, 2025. URL https://arxiv.org/abs/2502.21321.

[254] Martin Kuo, Jianyi Zhang, Aolin Ding, Qinsi Wang, Louis DiValentin, Yujia Bao, Wei Wei, Da-Cheng Juan, Hai Li, and Yiran Chen. H-cot: Hijacking the chain-of-thought safety reasoning mechanism to jailbreak large reasoning models, including openai o1/o3, deepseek-r1, and gemini 2.0 flash thinking. *arXiv preprint arXiv:2502.12893*, 2025.

[255] EvolvingLMMs Lab. Open-r1-multimodal. https://github.com/EvolvingLMMs-Lab/open-r1-multimodal, February 2025.

[256] Bespoke Labs. Bespoke-stratos: The unreasonable effectiveness of reasoning distillation. https://www.bespokelabs.ai/blog/bespoke-stratos-the-unreasonable-effectiveness-of-reasoning-distillation, January 2025. Accessed: 2025-01-22.

[257] Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. Stepdpo: Step-wise preference optimization for long-chain reasoning of llms. *arXiv preprint arXiv:2406.18629*, 2024.

[258] 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, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training, 2024. URL https://arxiv.org/abs/2411.15124.

[259] Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024.

[260] Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. Can language models learn from explanations in context? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 537–563, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.38. URL https://aclanthology.org/2022.findings-emnlp.38.

[261] Jack Lanchantin, Angelica Chen, Shehzaad Dhuliawala, Ping Yu, Jason Weston, Sainbayar Sukhbaatar, and Ilia Kulikov. Diverse preference optimization. *arXiv preprint arXiv:2501.18101*, 2025.

[262] Anh Duc Le, Tu Vu, Nam Le Hai, Nguyen Thi Ngoc Diep, Linh Ngo Van, Trung Le, and Thien Huu Nguyen. Cot2align: Cross-chain of thought distillation via optimal transport alignment for language models with different tokenizers. *arXiv preprint arXiv:2502.16806*, 2025.

[263] Joshua Ong Jun Leang, Aryo Pradipta Gema, and Shay B Cohen. Comat: Chain of mathematically annotated thought improves mathematical reasoning. *arXiv preprint arXiv:2410.10336*, 2024.

[264] Joshua Ong Jun Leang, Giwon Hong, Wenda Li, and Shay B Cohen. Theorem prover as a judge for synthetic data generation. *arXiv preprint arXiv:2502.13137*, 2025.

[265] Hyunseok Lee, Seunghyuk Oh, Jaehyung Kim, Jinwoo Shin, and Jihoon Tack. Revise: Learning to refine at test-time via intrinsic self-verification. *arXiv preprint arXiv:2502.14565*, 2025.

[266] Jinu Lee and Julia Hockenmaier. Evaluating step-by-step reasoning traces: A survey. *arXiv preprint arXiv:2502.12289*, 2025.

[267] Jung Hyun Lee, June Yong Yang, Byeongho Heo, Dongyoon Han, and Kang Min Yoo. Token-supervised value models for enhancing mathematical reasoning capabilities of large language models. *arXiv preprint arXiv:2407.12863*, 2024.

[268] Kuang-Huei Lee, Ian Fischer, Yueh-Hua Wu, Dave Marwood, Shumeet Baluja, Dale Schuurmans, and Xinyun Chen. Evolving deeper llm thinking. *arXiv preprint arXiv:2501.09891*, 2025.

[269] 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 *First Conference on Language Modeling*, July 2024. URL https://openreview.net/forum?id=SGoVIC0u0f.

[270] Bin Lei, Yi Zhang, Shan Zuo, Ali Payani, and Caiwen Ding. MACM: Utilizing a multi-agent system for condition mining in solving complex mathematical problems. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, September 2024. URL https://openreview.net/forum?id=VR2RdSxtzs.

[271] Adam Lerer, Hengyuan Hu, Jakob Foerster, and Noam Brown. Improving policies via search in cooperative partially observable games. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7187–7194, Apr. 2020. doi: 10.1609/aaai.v34i05.6208. URL https://ojs.aaai.org/index.php/AAAI/article/view/6208.

[272] Bingxuan Li, Yiwei Wang, Jiuxiang Gu, Kai-Wei Chang, and Nanyun Peng. Metal: A multi-agent framework for chart generation with test-time scaling. *arXiv preprint arXiv:2502.17651*, 2025.

[273] Bohan Li, Jiannan Guan, Longxu Dou, Yunlong Feng, Dingzirui Wang, Yang Xu, Enbo Wang, Qiguang Chen, Bichen Wang, Xiao Xu, et al. Can large language models understand you better? an mbti personality detection dataset aligned with population traits. *arXiv preprint arXiv:2412.12510*, 2024.

[274] Chen Li, Weiqi Wang, Jingcheng Hu, Yixuan Wei, Nanning Zheng, Han Hu, Zheng Zhang, and Houwen Peng. Common 7b language models already possess strong math capabilities. *arXiv preprint arXiv:2403.04706*, 2024.

[275] Chengshu Li, Jacky Liang, Andy Zeng, Xinyun Chen, Karol Hausman, Dorsa Sadigh, Sergey Levine, Li Fei-Fei, Fei Xia, and Brian Ichter. Chain of code: Reasoning with a language model-augmented code emulator. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 28259–28277. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/li24ar.html.

[276] Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*, 2025.

[277] Cheryl Li, Tianyuan Xu, and Yiwen Guo. Reasoning-as-logic-units: Scaling test-time reasoning in large language models through logic unit alignment. *arXiv preprint arXiv:2502.07803*, 2025.

[278] Dacheng Li, Shiyi Cao, Chengkun Cao, Xiuyu Li, Shangyin Tan, Kurt Keutzer, Jiarong Xing, Joseph E Gonzalez, and Ion Stoica. S*: Test time scaling for code generation. *arXiv preprint arXiv:2502.14382*, 2025.

[279] Dacheng Li, Shiyi Cao, Tyler Griggs, Shu Liu, Xiangxi Mo, Shishir G Patil, Matei Zaharia, Joseph E Gonzalez, and Ion Stoica. Llms can easily learn to reason from demonstrations structure, not content, is what matters! *arXiv preprint arXiv:2502.07374*, 2025.

[280] Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*, 2024.

[281] Gengxu Li, Tingyu Xia, Yi Chang, and Yuan Wu. Length-controlled margin-based preference optimization without reference model. *arXiv preprint arXiv:2502.14643*, 2025.

[282] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024.

[283] Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. Numina-math. [https://huggingface.co/AI-MO/NuminaMath-CoT](https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf), 2024.

[284] Jierui Li, Hung Le, Yinbo Zhou, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Codetree: Agent-guided tree search for code generation with large language models. *arXiv preprint arXiv:2411.04329*, 2024.

[285] Junlong Li, Daya Guo, Dejian Yang, Runxin Xu, Yu Wu, and Junxian He. Codei/o: Condensing reasoning patterns via code input-output prediction. *arXiv preprint arXiv:2502.07316*, 2025.

[286] Kechen Li, Wenqi Zhu, Coralia Cartis, Tianbo Ji, and Shiwei Liu. Sos1: O1 and r1-like reasoning llms are sum-of-square solvers. *arXiv preprint arXiv:2502.20545*, 2025.

[287] Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xingxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, et al. Chain of ideas: Revolutionizing research via novel idea development with llm agents. *arXiv preprint arXiv:2410.13185*, 2024.

[288] Margaret Li, Sneha Kudugunta, and Luke Zettlemoyer. (mis) fitting: A survey of scaling laws. *arXiv preprint arXiv:2502.18969*, 2025.

[289] Ming Li, Lichang Chen, Jiuhai Chen, Shwai He, Heng Huang, Jiuxiang Gu, and Tianyi Zhou. Reflection-tuning: Data recycling improves llm instruction-tuning. *arXiv preprint arXiv:2310.11716*, 2023.

[290] Ming Li, Yanhong Li, and Tianyi Zhou. What happened in llms layers when trained for fast vs. slow thinking: A gradient perspective. *arXiv preprint arXiv:2410.23743*, 2024.

[291] Minzhi Li, Zhengyuan Liu, Shumin Deng, Shafiq Joty, Nancy Chen, and Min-Yen Kan. Dna-eval: Enhancing large language model evaluation through decomposition and aggregation.

In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2277–2290, January 2025.

[292] Moxin Li, Yuantao Zhang, Wenjie Wang, Wentao Shi, Zhuo Liu, Fuli Feng, and Tat-Seng Chua. Self-improvement towards pareto optimality: Mitigating preference conflicts in multi-objective alignment. *arXiv preprint arXiv:2502.14354*, 2025.

[293] Peiji Li, Kai Lv, Yunfan Shao, Yichuan Ma, Linyang Li, Xiaoqing Zheng, Xipeng Qiu, and Qipeng Guo. Fastmcts: A simple sampling strategy for data synthesis. *arXiv preprint arXiv:2502.11476*, 2025.

[294] Qingyao Li, Wei Xia, Kounianhua Du, Xinyi Dai, Ruiming Tang, Yasheng Wang, Yong Yu, and Weinan Zhang. Rethinkmcts: Refining erroneous thoughts in monte carlo tree search for code generation. *arXiv preprint arXiv:2409.09584*, 2024.

[295] Shuangtao Li, Shuaihao Dong, Kexin Luan, Xinhan Di, and Chaofan Ding. Enhancing reasoning through process supervision with monte carlo tree search. In *The First Workshop on Neural Reasoning and Mathematical Discovery at AAAI'2025*, January 2025. URL https://openreview.net/forum?id=OupEEi1341.

[296] Wendi Li and Yixuan Li. Process reward model with q-value rankings. *arXiv preprint arXiv:2410.11287*, 2024.

[297] Xiaonan Li and Xipeng Qiu. MoT: Memory-of-thought enables ChatGPT to self-improve. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6354–6374, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.392. URL https://aclanthology.org/2023.emnlp-main.392/.

[298] Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models. *arXiv preprint arXiv:2501.05366*, 2025.

[299] Xinzhe Li. A survey on llm test-time compute via search: Tasks, llm profiling, search algorithms, and relevant frameworks. *arXiv preprint arXiv:2501.10069*, 2025.

[300] Xuefeng Li, Haoyang Zou, and Pengfei Liu. Limr: Less is more for rl scaling. *arXiv preprint arXiv:2502.11886*, 2025.

[301] Yafu Li, Zhilin Wang, Tingchen Fu, Ganqu Cui, Sen Yang, and Yu Cheng. From drafts to answers: Unlocking llm potential via aggregation fine-tuning. *arXiv preprint arXiv:2501.11877*, 2025.

[302] Yang Li. Policy guided tree search for enhanced llm reasoning. *arXiv preprint arXiv:2502.06813*, 2025.

[303] Yang Li, Dong Du, Linfeng Song, Chen Li, Weikang Wang, Tao Yang, and Haitao Mi. Hunyuanprover: A scalable data synthesis framework and guided tree search for automated theorem proving. *arXiv preprint arXiv:2412.20735*, 2024.

[304] Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. Making language models better reasoners with step-aware verifier. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5315–5333, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.291. URL https://aclanthology.org/2023.acl-long.291/.

[305] Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. Competition-level code generation with alphacode. *arXiv preprint arXiv:2203.07814*, 2022.

[306] Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. Chain of thought empowers transformers to solve inherently serial problems. In *The Twelfth International Conference on Learning Representations*, January 2023.

[307] Zhiyuan Li, Dongnan Liu, Chaoyi Zhang, Heng Wang, Tengfei Xue, and Weidong Cai. Enhancing advanced visual reasoning ability of large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1915–1929, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.114. URL https://aclanthology.org/2024.emnlp-main.114/.

[308] Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, et al. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*, 2025.

[309] Zhongzhi Li, Ming-Liang Zhang, Pei-Jie Wang, Jian Xu, Rui-Song Zhang, Yin Fei, Zhi-Long Ji, Jin-Feng Bai, Zhen-Ru Pan, Jiaxin Zhang, and Cheng-Lin Liu. CMMaTH: A Chinese multi-modal math skill evaluation benchmark for foundation models. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2690–2726, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.184/.

[310] Zhuoqun Li, Haiyang Yu, Xuanang Chen, Hongyu Lin, Yaojie Lu, Fei Huang, Xianpei Han, Yongbin Li, and Le Sun. Deepsolution: Boosting complex engineering solution design via tree-based exploration and bi-point thinking. *arXiv preprint arXiv:2502.20730*, 2025.

[311] Ziniu Li, Tian Xu, Yushun Zhang, Zhihang Lin, Yang Yu, Ruoyu Sun, and Zhi-Quan Luo. Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models. In *Forty-first International Conference on Machine Learning*, May 2024. URL https://openreview.net/forum?id=Stn8hXkpe6.

[312] Xun Liang, Shichao Song, Zifan Zheng, Hanyu Wang, Qingchen Yu, Xunkai Li, Rong-Hua Li, Yi Wang, Zhonghao Wang, Feiyu Xiong, et al. Internal consistency and self-feedback in large language models: A survey. *arXiv preprint arXiv:2407.14507*, 2024.

[313] Baohao Liao, Yuhui Xu, Hanze Dong, Junnan Li, Christof Monz, Silvio Savarese, Doyen Sahoo, and Caiming Xiong. Reward-guided speculative decoding for efficient llm reasoning. *arXiv preprint arXiv:2501.19324*, 2025.

[314] Huanxuan Liao, Shizhu He, Yupu Hao, Xiang Li, Yuanzhe Zhang, Jun Zhao, and Kang Liu. Skintern: Internalizing symbolic knowledge for distilling better cot capabilities into small language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3203–3221, January 2025.

[315] Minpeng Liao, Wei Luo, Chengxi Li, Jing Wu, and Kai Fan. Mario: Math reasoning with code interpreter output–a reproducible pipeline. *arXiv preprint arXiv:2401.08190*, 2024.

[316] Weibin Liao, Xu Chu, and Yasha Wang. Tpo: Aligning large language models with multi-branch & multi-step preference trees. *arXiv preprint arXiv:2410.12854*, 2024.

[317] Jonathan Light, Min Cai, Weiqin Chen, Guanzhi Wang, Xiusi Chen, Wei Cheng, Yisong Yue, and Ziniu Hu. Strategist: Learning strategic skills by LLMs via bi-level tree search. In *Automated Reinforcement Learning: Exploring Meta-Learning, AutoML, and LLMs*, June 2024. URL https://openreview.net/forum?id=UHWBmZuJPF.

[318] Jonathan Light, Yue Wu, Yiyou Sun, Wenchao Yu, Xujiang Zhao, Ziniu Hu, Haifeng Chen, Wei Cheng, et al. Scattered forest search: Smarter code space exploration with llms. *arXiv preprint arXiv:2411.05010*, 2024.

[319] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, January 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

[320] Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. Zebralogic: On the scaling limits of llms for logical reasoning. *arXiv preprint arXiv:2502.01100*, 2025.

[321] Haohan Lin, Zhiqing Sun, Yiming Yang, and Sean Welleck. Lean-star: Learning to interleave thinking and proving. *arXiv preprint arXiv:2407.10040*, 2024.

[322] Qingwen Lin, Boyan Xu, Zijian Li, Zhifeng Hao, Keli Zhang, and Ruichu Cai. Leveraging constrained monte carlo tree search to generate reliable long chain-of-thought for mathematical reasoning. *arXiv preprint arXiv:2502.11169*, 2025.

[323] Yen-Ting Lin, Di Jin, Tengyu Xu, Tianhao Wu, Sainbayar Sukhbaatar, Chen Zhu, Yun He, Yun-Nung Chen, Jason Weston, Yuandong Tian, et al. Step-kto: Optimizing mathematical reasoning through stepwise binary feedback. *arXiv preprint arXiv:2501.10799*, 2025.

[324] Zicheng Lin, Zhibin Gou, Tian Liang, Ruilin Luo, Haowei Liu, and Yujiu Yang. CriticBench: Benchmarking LLMs for critique-correct reasoning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1552–1587, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.91. URL https://aclanthology.org/2024.findings-acl.91/.

[325] Zicheng Lin, Tian Liang, Jiahao Xu, Xing Wang, Ruilin Luo, Chufan Shi, Siheng Li, Yujiu Yang, and Zhaopeng Tu. Critical tokens matter: Token-level contrastive estimation enhence llm's reasoning capability. *arXiv preprint arXiv:2411.19943*, 2024.

[326] Zongyu Lin, Yao Tang, Xingcheng Yao, Da Yin, Ziniu Hu, Yizhou Sun, and Kai-Wei Chang. Qlass: Boosting language agent inference via q-guided stepwise search. *arXiv preprint arXiv:2502.02584*, 2025.

[327] Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. Deductive verification of chain-of-thought reasoning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 36407–36433. Curran Associates, Inc., September 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/72393bd47a35f5b3bee4c609e7bba733-Paper-Conference.pdf.

[328] Aiwei Liu, Haoping Bai, Zhiyun Lu, Xiang Kong, Xiaoming Wang, Jiulong Shan, Meng Cao, and Lijie Wen. Direct large language model alignment through self-rewarding contrastive prompt distillation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9688–9712, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.523. URL https://aclanthology.org/2024.acl-long.523/.

[329] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[330] Bingbin Liu, Sebastien Bubeck, Ronen Eldan, Janardhan Kulkarni, Yuanzhi Li, Anh Nguyen, Rachel Ward, and Yi Zhang. Tinygsm: achieving> 80% on gsm8k with small language models. *arXiv preprint arXiv:2312.09241*, 2023.

[331] Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. Skywork-reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*, 2024.

[332] Cong Liu, Zhong Wang, ShengYu Shen, Jialiang Peng, Xiaoli Zhang, ZhenDong Du, and YaFang Wang. The chinese dataset distilled from deepseek-r1-671b. https://huggingface.co/datasets/Congliu/Chinese-DeepSeek-R1-Distill-data-110k, 2025.

[333] Dancheng Liu, Amir Nassereldine, Ziming Yang, Chenhui Xu, Yuting Hu, Jiajie Li, Utkarsh Kumar, Changjae Lee, Ruiyang Qin, Yiyu Shi, et al. Large language models have intrinsic self-correction ability. *arXiv preprint arXiv:2406.15673*, 2024.

[334] Fan Liu, Wenshuo Chao, Naiqiang Tan, and Hao Liu. Bag of tricks for inference-time computation of llm reasoning. *arXiv preprint arXiv:2502.07191*, 2025.

[335] Guanlin Liu, Kaixuan Ji, Renjie Zheng, Zheng Wu, Chen Dun, Quanquan Gu, and Lin Yan. Enhancing multi-step reasoning abilities of language models through direct q-function optimization. *arXiv preprint arXiv:2410.09302*, 2024.

[336] Hanmeng Liu, Zhizhang Fu, Mengru Ding, Ruoxi Ning, Chaoli Zhang, Xiaozhang Liu, and Yue Zhang. Logical reasoning in large language models: A survey. *arXiv preprint arXiv:2502.09100*, 2025.

[337] Hao Liu, Zhengren Wang, Xi Chen, Zhiyu Li, Feiyu Xiong, Qinhan Yu, and Wentao Zhang. Hoprag: Multi-hop reasoning for logic-aware retrieval-augmented generation. *arXiv preprint arXiv:2502.12442*, 2025.

[338] Hongxuan Liu, Zhiyao Luo, and Tingting Zhu. Best of both worlds: Harmonizing LLM capabilities in decision-making and question-answering for treatment regimes. In *Advancements In Medical Foundation Models: Explainability, Robustness, Security, and Beyond*, 2024. URL https://openreview.net/forum?id=afu9qhp7md.

[339] Jiacai Liu, Chaojie Wang, Chris Yuhao Liu, Liang Zeng, Rui Yan, Yiwen Sun, Yang Liu, and Yahui Zhou. Improving multi-step reasoning abilities of large language models with direct advantage policy optimization. *arXiv preprint arXiv:2412.18279*, 2024.

[340] Jiacheng Liu, Andrew Cohen, Ramakanth Pasunuru, Yejin Choi, Hannaneh Hajishirzi, and Asli Celikyilmaz. Don't throw away your value model! generating more preferable text with value-guided monte-carlo tree search decoding. In *First Conference on Language Modeling*, July 2024. URL https://openreview.net/forum?id=kh9Zt2Ldmn.

[341] Jiacheng Liu, Andrew Cohen, Ramakanth Pasunuru, Yejin Choi, Hannaneh Hajishirzi, and Asli Celikyilmaz. Making PPO even better: Value-guided monte-carlo tree search decoding, September 2024. URL https://openreview.net/forum?id=QaODpeRaOK.

[342] Qiang Liu, Xinlong Chen, Yue Ding, Shizhen Xu, Shu Wu, and Liang Wang. Attention-guided self-reflection for zero-shot hallucination detection in large language models. *arXiv preprint arXiv:2501.09997*, 2025.

[343] Runze Liu, Junqi Gao, Jian Zhao, Kaiyan Zhang, Xiu Li, Biqing Qi, Wanli Ouyang, and Bowen Zhou. Can 1b llm surpass 405b llm? rethinking compute-optimal test-time scaling. *arXiv preprint arXiv:2502.06703*, 2025.

[344] Wei Liu, Junlong Li, Xiwen Zhang, Fan Zhou, Yu Cheng, and Junxian He. Diving into self-evolving training for multimodal reasoning. *arXiv preprint arXiv:2412.17451*, 2024.

[345] Yue Liu, Hongcheng Gao, Shengfang Zhai, Jun Xia, Tianyi Wu, Zhiwei Xue, Yulin Chen, Kenji Kawaguchi, Jiaheng Zhang, and Bryan Hooi. Guardreasoner: Towards reasoning-based llm safeguards. *arXiv preprint arXiv:2501.18492*, 2025.

[346] Zichen Liu, Changyu Chen, Wenjun Li, Tianyu Pang, Chao Du, and Min Lin. There may not be aha moment in r1-zero-like training — a pilot study. https://oatllm.notion.site/oat-zero, 2025. Notion Blog.

[347] Zihan Liu, Yang Chen, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Acemath: Advancing frontier math reasoning with post-training and reward modeling. *arXiv preprint arXiv:2412.15084*, 2024.

[348] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025.

[349] Dakuan Lu, Xiaoyu Tan, Rui Xu, Tianchu Yao, Chao Qu, Wei Chu, Yinghui Xu, and Yuan Qi. Scp-116k: A high-quality problem-solution dataset and a generalized pipeline for automated extraction in the higher education science domain, 2025. URL https://arxiv.org/abs/2501.15587.

[350] Jianqiao Lu, Zhiyang Dou, Hongru WANG, Zeyu Cao, Jianbo Dai, Yunlong Feng, and Zhijiang Guo. Autopsv: Automated process-supervised verifier. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 79935–79962. Curran Associates, Inc., December 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/9246aa822579d9b29a140ecdac36ad60-Paper-Conference.pdf.

[351] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, November 2022. URL https://openreview.net/forum?id=HjwK-Tc_Bc.

[352] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical

reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*, January 2024. URL https://openreview.net/forum?id=KUNzEQMWU7.

[353] Pan Lu, Bowen Chen, Sheng Liu, Rahul Thapa, Joseph Boen, and James Zou. Octo-tools: An agentic framework with extensible tools for complex reasoning. *arXiv preprint arXiv:2502.11271*, 2025.

[354] Rubing Lu, João Sedoc, and Arun Sundararajan. Reasoning and the trusting behavior of deepseek and gpt: An experiment revealing hidden fault lines in large language models. *arXiv preprint arXiv:2502.12825*, 2025.

[355] Zimu Lu, Aojun Zhou, Houxing Ren, Ke Wang, Weikang Shi, Junting Pan, Mingjie Zhan, and Hongsheng Li. Mathgenie: Generating synthetic data with question back-translation for enhancing mathematical reasoning of llms. *arXiv preprint arXiv:2402.16352*, 2024.

[356] Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*, 2023.

[357] Haotian Luo, Li Shen, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Naiqiang Tan, Xiaochun Cao, and Dacheng Tao. O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning. *arXiv preprint arXiv:2501.12570*, 2025.

[358] Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, et al. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*, 2024.

[359] Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Tianjun Zhang, Li Erran Li, Raluca Ada Popa, and Ion Stoica. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl, February 2025. Notion Blog.

[360] Ruilin Luo, Zhuofan Zheng, Yifan Wang, Yiyao Yu, Xinzhe Ni, Zicheng Lin, Jin Zeng, and Yujiu Yang. Ursa: Understanding and verifying chain-of-thought reasoning in multimodal mathematics. *arXiv preprint arXiv:2501.04686*, 2025.

[361] Xianzhen Luo, Qingfu Zhu, Zhiming Zhang, Libo Qin, Xuanyu Zhang, Qing Yang, Dongliang Xu, and Wanxiang Che. Python is not always the best choice: Embracing multilingual program of thoughts. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7185–7212, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.408. URL https://aclanthology.org/2024.emnlp-main.408/.

[362] Chengqi Lyu, Songyang Gao, Yuzhe Gu, Wenwei Zhang, Jianfei Gao, Kuikun Liu, Ziyi Wang, Shuaibin Li, Qian Zhao, Haian Huang, et al. Exploring the limit of outcome reward for learning mathematical reasoning. *arXiv preprint arXiv:2502.06781*, 2025.

[363] Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi, editors, *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 305–329, Nusa Dua, Bali, November 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.ijcnlp-main.20. URL https://aclanthology.org/2023.ijcnlp-main.20/.

[364] Alexander Lyzhov, Yuliya Molchanova, Arsenii Ashukha, Dmitry Molchanov, and Dmitry Vetrov. Greedy policy search: A simple baseline for learnable test-time augmentation. In Jonas Peters and David Sontag, editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 1308–1317. PMLR, 03–06 Aug 2020. URL https://proceedings.mlr.press/v124/lyzhov20a.html.

[365] Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*, 2025.

[366] Qianli Ma, Haotian Zhou, Tingkai Liu, Jianbo Yuan, Pengfei Liu, Yang You, and Hongxia Yang. Let's reward step by step: Step-level reward model as the navigators for reasoning. *arXiv preprint arXiv:2310.10080*, 2023.

[367] Ruotian Ma, Peisong Wang, Cheng Liu, Xingyan Liu, Jiaqi Chen, Bang Zhang, Xin Zhou, Nan Du, and Jia Li. S^{2}r: Teaching llms to self-verify and self-correct via reinforcement learning. *arXiv preprint arXiv:2502.12853*, 2025.

[368] Xinyin Ma, Guangnian Wan, Runpeng Yu, Gongfan Fang, and Xinchao Wang. Cot-valve: Length-compressible chain-of-thought tuning. *arXiv preprint arXiv:2502.09601*, 2025.

[369] Xuetao Ma, Wenbin Jiang, and Hua Huang. Problem-solving logic guided curriculum in-context learning for llms complex reasoning. *arXiv preprint arXiv:2502.15401*, 2025.

[370] Yiran Ma, Zui Chen, Tianqiao Liu, Mi Tian, Zhuo Liu, Zitao Liu, and Weiqi Luo. What are step-level reward models rewarding? counterintuitive findings from mcts-boosted mathematical reasoning. *arXiv preprint arXiv:2412.15904*, 2024.

[371] Zexiong Ma, Chao Peng, Pengfei Gao, Xiangxin Meng, Yanzhen Zou, and Bing Xie. Sorft: Issue resolving with subtask-oriented reinforced fine-tuning. *arXiv preprint arXiv:2502.20127*, 2025.

[372] Zeyao Ma, Xiaokang Zhang, Jing Zhang, Jifan Yu, Sijia Luo, and Jie Tang. Dynamic scaling of unit tests for code reward modeling. *arXiv preprint arXiv:2501.01054*, 2025.

[373] Ziyang Ma, Zhuo Chen, Yuping Wang, Eng Siong Chng, and Xie Chen. Audio-cot: Exploring chain-of-thought reasoning in large audio language model. *arXiv preprint arXiv:2501.07246*, 2025.

[374] Aman Madaan, Katherine Hermann, and Amir Yazdanbakhsh. What makes chain-of-thought prompting effective? a counterfactual study. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1448–1535, Singapore, December 2023.

[375] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc., March 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91edff07232fb1b55a505a9e9f6c0ff3-Paper-Conference.pdf.

[376] Sathwik Tejaswi Madhusudhan, Shruthan Radhakrishna, Jash Mehta, and Toby Liang. Millions scale dataset distilled from r1-32b. https://huggingface.co/datasets/ServiceNow-AI/R1-Distill-SFT, February 2025.

[377] Sadegh Mahdavi, Muchen Li, Kaiwen Liu, Christos Thrampoulidis, Leonid Sigal, and Renjie Liao. Leveraging online olympiad-level math problems for llms training and contamination-resistant evaluation. *arXiv preprint arXiv:2501.14275*, 2025.

[378] Tobias Materzok. Cos (m+ o) s: Curiosity and rl-enhanced mcts for exploring story space via language models. *arXiv preprint arXiv:2501.17104*, 2025.

[379] Justus Mattern, Sami Jaghouar, Manveer Basra, Jannik Straube, Matthew Di Ferrante, Felix Gabriel, Jack Min Ong, Vincent Weisser, and Johannes Hagemann. Synthetic-1: Two million collaboratively generated reasoning traces from deepseek-r1, 2025. URL https://www.primeintellect.ai/blog/synthetic-1-release.

[380] Nat McAleese, Rai Michael Pokorny, Juan Felipe Ceron Uribe, Evgenia Nitishinskaya, Maja Trebacz, and Jan Leike. Llm critics help catch llm bugs. *arXiv preprint arXiv:2407.00215*, 2024.

[381] R Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D Hardy, and Thomas L Griffiths. When a language model is optimized for reasoning, does it still show embers of autoregression? an analysis of openai o1. *arXiv preprint arXiv:2410.01792*, 2024.

[382] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang,

and Wenqi Shao. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

[383] William Merrill and Ashish Sabharwal. The expressive power of transformers with chain of thought. In *The Twelfth International Conference on Learning Representations*, January 2023.

[384] Ning Miao, Yee Whye Teh, and Tom Rainforth. Selfcheck: Using LLMs to zero-shot check their own step-by-step reasoning. In *The Twelfth International Conference on Learning Representations*, January 2024. URL https://openreview.net/forum?id=pTHfApDakA.

[385] Yingqian Min, Zhipeng Chen, Jinhao Jiang, Jie Chen, Jia Deng, Yiwen Hu, Yiru Tang, Jiapeng Wang, Xiaoxue Cheng, Huatong Song, et al. Imitate, explore, and self-improve: A reproduction report on slow-thinking reasoning systems. *arXiv preprint arXiv:2412.09413*, 2024.

[386] Seyed Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *The Thirteenth International Conference on Learning Representations*, January 2025. URL https://openreview.net/forum?id=AjXkRZIvjB.

[387] Arindam Mitra, Hamed Khanpour, Corby Rosset, and Ahmed Awadallah. Orca-math: Unlocking the potential of slms in grade school math. *arXiv preprint arXiv:2402.14830*, 2024.

[388] Shentong Mo and Miao Xin. Tree of uncertain thoughts reasoning for large language models. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12742–12746, April 2024. doi: 10.1109/ICASSP48485.2024. 10448355.

[389] Philipp Mondorf and Barbara Plank. Beyond accuracy: Evaluating the reasoning behavior of large language models–a survey. *arXiv preprint arXiv:2404.01869*, 2024.

[390] Terufumi Morishita, Gaku Morio, Atsuki Yamaguchi, and Yasuhiro Sogawa. Enhancing reasoning capabilities of llms via principled synthetic logic corpus. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 73572–73604. Curran Associates, Inc., September 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/8678da90126aa58326b2fc0254b33a8c-Paper-Conference.pdf.

[391] 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.

[392] Tergel Munkhbat, Namgyu Ho, Seohyun Kim, Yongjin Yang, Yujin Kim, and Se-Young Yun. Self-training elicits concise reasoning in large language models. *arXiv preprint arXiv:2502.20122*, 2025.

[393] Vaskar Nath, Pranav Raja, Claire Yoon, and Sean Hendryx. Toolcomp: A multi-tool reasoning & process supervision benchmark. *arXiv preprint arXiv:2501.01290*, 2025.

[394] Sania Nayab, Giulio Rossolini, Marco Simoni, Andrea Saracino, Giorgio Buttazzo, Nicolamaria Manes, and Fabrizio Giacomelli. Concise thoughts: Impact of output length on llm reasoning and cost. *arXiv preprint arXiv:2407.19825*, 2024.

[395] Ansong Ni, Srini Iyer, Dragomir Radev, Veselin Stoyanov, Wen-Tau Yih, Sida Wang, and Xi Victoria Lin. LEVER: Learning to verify language-to-code generation with execution. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 26106–26128. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/ni23b.html.

[396] Ziyi Ni, Yifan Li, Ning Yang, Dou Shen, Pin Lv, and Daxiang Dong. Tree-of-code: A tree-structured exploring framework for end-to-end code generation and execution in complex task handling. *arXiv preprint arXiv:2412.15305*, 2024.

[397] Allen Nie, Yi Su, Bo Chang, Jonathan N Lee, Ed H Chi, Quoc V Le, and Minmin Chen. Evolve: Evaluating and optimizing llms for exploration. *arXiv preprint arXiv:2410.06238*, 2024.

[398] Harsha Nori, Naoto Usuyama, Nicholas King, Scott Mayer McKinney, Xavier Fernandes, Sheng Zhang, and Eric Horvitz. From medprompt to o1: Exploration of run-time strategies for medical challenge problems and beyond. *arXiv preprint arXiv:2411.03590*, 2024.

[399] Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. Show your work: Scratchpads for intermediate computation with language models. In *Deep Learning for Code Workshop*, March 2022. URL https://openreview.net/forum?id=HBlx2idbkbq.

[400] Skywork o1 Team. Skywork-o1 open series. https://huggingface.co/Skywork, November 2024.

[401] OpenCompass. Aime 2025. https://huggingface.co/datasets/opencompass/AIME2025, February 2025.

[402] Yixin Ou, Yunzhi Yao, Ningyu Zhang, Hui Jin, Jiacheng Sun, Shumin Deng, Zhenguo Li, and Huajun Chen. How do llms acquire new knowledge? a knowledge circuits perspective on continual pre-training. *arXiv preprint arXiv:2502.11196*, 2025.

[403] Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. *arXiv preprint arXiv:2201.03544*, 2022.

[404] Jiabao Pan, Yan Zhang, Chen Zhang, Zuozhu Liu, Hongwei Wang, and Haizhou Li. DynaThink: Fast or slow? a dynamic decision-making framework for large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14686–14695, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.814. URL https://aclanthology.org/2024.emnlp-main.814/.

[405] Jianfeng Pan, Senyou Deng, and Shaomang Huang. Coat: Chain-of-associated-thoughts framework for enhancing large language models reasoning. *arXiv preprint arXiv:2502.02390*, 2025.

[406] Jiayi Pan, Junjie Zhang, Xingyao Wang, Lifan Yuan, Hao Peng, and Alane Suhr. Tinyzero. https://github.com/Jiayi-Pan/TinyZero, 2025. Accessed: 2025-01-24.

[407] Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *arXiv preprint arXiv:2502.19634*, 2025.

[408] Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *arXiv preprint arXiv:2308.03188*, 2023.

[409] Bo Pang, Hanze Dong, Jiacheng Xu, Silvio Savarese, Yingbo Zhou, and Caiming Xiong. Bolt: Bootstrap long chain-of-thought in language models without distillation. *arXiv preprint arXiv:2502.03860*, 2025.

[410] Richard Yuanzhe Pang, Weizhe Yuan, He He, Kyunghyun Cho, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 116617–116637. Curran Associates, Inc., September 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/d37c9ad425fe5b65304d500c6edcba00-Paper-Conference.pdf.

[411] Shubham Parashar, Blake Olson, Sambhav Khurana, Eric Li, Hongyi Ling, James Caverlee, and Shuiwang Ji. Inference-time computations for llm reasoning and planning: A benchmark and insights. *arXiv preprint arXiv:2502.12521*, 2025.

[412] Chanwoo Park, Seungju Han, Xingzhi Guo, Asuman Ozdaglar, Kaiqing Zhang, and Joo-Kyung Kim. Maporl: Multi-agent post-co-training for collaborative large language models with reinforcement learning. *arXiv preprint arXiv:2502.18439*, 2025.

[413] Junsoo Park, Seungyeon Jwa, Meiying Ren, Daeyoung Kim, and Sanghyuk Choi. Offsetbias: Leveraging debiased data for tuning evaluators, 2024.

[414] Sungjin Park, Xiao Liu, Yeyun Gong, and Edward Choi. Ensembling large language models with process reward-guided tree search for better complex reasoning. *arXiv preprint arXiv:2412.15797*, 2024.

[415] Manojkumar Parmar and Yuvaraj Govindarajulu. Challenges in ensuring ai safety in deepseek-r1 models: The shortcomings of reinforcement learning strategies. *arXiv preprint arXiv:2501.17030*, 2025.

[416] Avinash Patil. Advancing reasoning in large language models: Promising methods and approaches. *arXiv preprint arXiv:2502.03671*, 2025.

[417] Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. REFINER: Reasoning feedback on intermediate representations. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1100–1126, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.67/.

[418] Patomporn Payoungkhamdee, Pume Tuchinda, Jinheon Baek, Samuel Cahyawijaya, Can Udomcharoenchaikit, Potsawee Manakul, Peerat Limkonchotiwat, Ekapol Chuangsuwanich, and Sarana Nutanong. Towards better understanding of program-of-thought reasoning in cross-lingual and multilingual environments. *arXiv preprint arXiv:2502.17956*, 2025.

[419] Hao Peng, Yunjia Qi, Xiaozhi Wang, Zijun Yao, Bin Xu, Lei Hou, and Juanzi Li. Agentic reward modeling: Integrating human preferences with verifiable correctness signals for reliable reward systems. *arXiv preprint arXiv:2502.19328*, 2025.

[420] Miao Peng, Nuo Chen, Zongrui Suo, and Jia Li. Rewarding graph reasoning process makes llms more generalized reasoners. *arXiv preprint arXiv:2503.00845*, 2025.

[421] Rolf Pfister and Hansueli Jud. Understanding and benchmarking artificial intelligence: Openai's o3 is not agi. *arXiv preprint arXiv:2501.07458*, 2025.

[422] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

[423] Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. Reasoning with large language models, a survey. *arXiv preprint arXiv:2407.11511*, 2024.

[424] Gabriel Poesia, Kanishk Gandhi, Eric Zelikman, and Noah Goodman. Certified deductive reasoning with language models. *Transactions on Machine Learning Research*, May 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=yXnwrs2Tl6.

[425] Stanislas Polu and Ilya Sutskever. Generative language modeling for automated theorem proving. *arXiv preprint arXiv:2009.03393*, 2020.

[426] Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. ReCEval: Evaluating reasoning chains via correctness and informativeness. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10066–10086, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.622. URL https://aclanthology.org/2023.emnlp-main.622/.

[427] Archiki Prasad, Alexander Koller, Mareike Hartmann, Peter Clark, Ashish Sabharwal, Mohit Bansal, and Tushar Khot. ADaPT: As-needed decomposition and planning with language models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4226–4252, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.264. URL https://aclanthology.org/2024.findings-naacl.264/.

[428] Tidor-Vlad Pricope. Hardml: A benchmark for evaluating data science and machine learning knowledge and reasoning in ai. *arXiv preprint arXiv:2501.15627*, 2025.

[429] Ben Prystawski, Michael Li, and Noah Goodman. Why think step by step? reasoning emerges from the locality of experience. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 70926–70947. Curran Associates, Inc., September 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/e0af79ad53a336b4c4b4f7e2a68eb609-Paper-Conference.pdf.

[430] Israel Puerta-Merino, Carlos Núñez-Molina, Pablo Mesejo, and Juan Fernández-Olivares. A roadmap to guide the integration of llms in hierarchical planning. *arXiv preprint arXiv:2501.08068*, 2025.

[431] Haritz Puerto, Tilek Chubakov, Xiaodan Zhu, Harish Tayyar Madabushi, and Iryna Gurevych. Fine-tuning with divergent chains of thought boosts reasoning through self-correction in language models. *arXiv preprint arXiv:2407.03181*, 2024.

[432] Isha Puri, Shivchander Sudalairaj, Guangxuan Xu, Kai Xu, and Akash Srivastava. A probabilistic inference approach to inference-time scaling of llms using particle-based monte carlo methods. *arXiv preprint arXiv:2502.01618*, 2025.

[433] Pranav Putta, Edmund Mills, Naman Garg, Sumeet Motwani, Chelsea Finn, Divyansh Garg, and Rafael Rafailov. Agent q: Advanced reasoning and learning for autonomous ai agents. *arXiv preprint arXiv:2408.07199*, 2024.

[434] Zhenting Qi, Mingyuan Ma, Jiahang Xu, Li Lyna Zhang, Fan Yang, and Mao Yang. Mutual reasoning makes smaller llms stronger problem-solvers. *arXiv preprint arXiv:2408.06195*, 2024.

[435] Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709, July 2023.

[436] Libo Qin, Qiguang Chen, Hao Fei, Zhi Chen, Min Li, and Wanxiang Che. What factors affect multi-modal in-context learning? an in-depth exploration. *arXiv preprint arXiv:2410.20482*, 2024.

[437] Libo Qin, Qiguang Chen, Xiachong Feng, Yang Wu, Yongheng Zhang, Yinghui Li, Min Li, Wanxiang Che, and Philip S Yu. Large language models meet nlp: A survey. *arXiv preprint arXiv:2405.12819*, 2024.

[438] Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. Multilingual large language model: A survey of resources, taxonomy and frontiers. *arXiv preprint arXiv:2404.04925*, 2024.

[439] Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and S Yu Philip. A survey of multilingual large language models. *Patterns*, 6(1), January 2025.

[440] Yiwei Qin, Xuefeng Li, Haoyang Zou, Yixiu Liu, Shijie Xia, Zhen Huang, Yixin Ye, Weizhe Yuan, Hector Liu, Yuanzhi Li, et al. O1 replication journey: A strategic progress report–part 1. *arXiv preprint arXiv:2410.18982*, 2024.

[441] Jiahao Qiu, Yifu Lu, Yifan Zeng, Jiacheng Guo, Jiayi Geng, Huazheng Wang, Kaixuan Huang, Yue Wu, and Mengdi Wang. Treebon: Enhancing inference-time alignment with speculative tree-search and best-of-n sampling. *arXiv preprint arXiv:2410.16033*, 2024.

[442] Yuxiao Qu, Tianjun Zhang, Naman Garg, and Aviral Kumar. Recursive introspection: Teaching language model agents how to self-improve. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, September 2024. URL https://openreview.net/forum?id=DRC9pZwBwR.

[443] Gollam Rabby, Farhana Keya, Parvez Zamil, and Sören Auer. Mc-nest–enhancing mathematical reasoning in large language models with a monte carlo nash equilibrium self-refine tree. *arXiv preprint arXiv:2411.15645*, 2024.

[444] Santosh Kumar Radha and Oktay Goktas. On the reasoning capacity of ai models and how to quantify it. *arXiv preprint arXiv:2501.13833*, 2025.

[445] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

[446] Daking Rai and Ziyu Yao. An investigation of neuron activation as a unified lens to explain chain-of-thought eliciting arithmetic reasoning of LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7174–7193, Bangkok, Thailand,

August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.387. URL https://aclanthology.org/2024.acl-long.387/.

[447] Leonardo Ranaldi, Giulia Pucci, Federico Ranaldi, Elena Sofia Ruzzetti, and Fabio Massimo Zanzotto. A tree-of-thoughts to broaden multi-step reasoning across languages. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1229–1241, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.78. URL https://aclanthology.org/2024.findings-naacl.78/.

[448] Leonardo Ranaldi, Marco Valentino, Alexander Polonsky, and Andrè Freitas. Improving chain-of-thought reasoning via quasi-symbolic abstractions. *arXiv preprint arXiv:2502.12616*, 2025.

[449] Mohammad Raza and Natasa Milic-Frayling. Instantiation-based formalization of logical reasoning tasks using language models and logical solvers. *arXiv preprint arXiv:2501.16961*, 2025.

[450] Ali Razghandi, Seyed Mohammad Hadi Hosseini, and Mahdieh Soleymani Baghshah. Cer: Confidence enhanced reasoning in llms. *arXiv preprint arXiv:2502.14634*, 2025.

[451] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, July 2024. URL https://openreview.net/forum?id=Ti67584b98.

[452] Matthew Renze and Erhan Guven. Self-reflection in llm agents: Effects on problem-solving performance. *arXiv preprint arXiv:2405.06682*, 2024.

[453] Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.

[454] Jon Saad-Falcon, Rajan Vivek, William Berrios, Nandita Shankar Naik, Matija Franklin, Bertie Vidgen, Amanpreet Singh, Douwe Kiela, and Shikib Mehri. Lmunit: Fine-grained evaluation with natural language unit tests. *arXiv preprint arXiv:2412.13091*, 2024.

[455] Nikta Gohari Sadr, Sangmitra Madhusudan, and Ali Emami. Think or step-by-step? unzipping the black box in zero-shot prompts. *arXiv preprint arXiv:2502.03418*, 2025.

[456] Swarnadeep Saha, Xian Li, Marjan Ghazvininejad, Jason Weston, and Tianlu Wang. Learning to plan & reason for evaluation with thinking-llm-as-a-judge. *arXiv preprint arXiv:2501.18099*, 2025.

[457] S Sauhandikaa, R Bhagavath Narenthranath, and R Sathya Bama Krishna. Explainable ai in large language models: A review. In *2024 International Conference on Emerging Research in Computational Science (ICERCS)*, pages 1–6. IEEE, 2024.

[458] William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*, 2022.

[459] Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, and Sashank J Reddi. Reasoning with latent thoughts: On the power of looped transformers. *arXiv preprint arXiv:2502.17416*, 2025.

[460] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[461] Amrith Setlur, Saurabh Garg, Xinyang Geng, Naman Garg, Virginia Smith, and Aviral Kumar. Rl on incorrect synthetic data scales the efficiency of llm math reasoning by eight-fold. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 43000–43031. Curran Associates, Inc., September 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/4b77d5b896c321a29277524a98a50215-Paper-Conference.pdf.

[462] Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for LLM reasoning. In *The Thirteenth International Conference on*

Learning Representations, January 2025. URL https://openreview.net/forum?id=A6Y7AqlzLW.

[463] Amrith Setlur, Nived Rajaraman, Sergey Levine, and Aviral Kumar. Scaling test-time compute without verification or rl is suboptimal. *arXiv preprint arXiv:2502.12118*, 2025.

[464] Yu Shang, Yu Li, Fengli Xu, and Yong Li. Synergy-of-thoughts: Eliciting efficient reasoning in hybrid language models. *arXiv preprint arXiv:2402.02563*, 2024.

[465] Wenqi Shao, Qiaosheng Zhang, Lingxiao Du, Xiangyan Liu, and Fanqing Meng. R1-multimodal-journey. https://github.com/FanqingM/R1-Multimodal-Journey, February 2025.

[466] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[467] Haozhan Shen, Zilun Zhang, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. Vlm-r1: A stable and generalizable r1-style large vision-language model. https://github.com/om-ai-lab/VLM-R1, 2025. Accessed: 2025-02-15.

[468] Maohao Shen, Guangtao Zeng, Zhenting Qi, Zhang-Wei Hong, Zhenfang Chen, Wei Lu, Gregory Wornell, Subhro Das, David Cox, and Chuang Gan. Satori: Reinforcement learning with chain-of-action-thought enhances llm reasoning via autoregressive search. *arXiv preprint arXiv:2502.02508*, 2025.

[469] Xuan Shen, Yizhou Wang, Xiangxi Shi, Yanzhi Wang, Pu Zhao, and Jiuxiang Gu. Efficient reasoning with hidden thinking. *arXiv preprint arXiv:2501.19201*, 2025.

[470] Yi Shen, Jian Zhang, Jieyun Huang, Shuming Shi, Wenjing Zhang, Jiangze Yan, Ning Wang, Kai Wang, and Shiguo Lian. Dast: Difficulty-adaptive slow-thinking for large reasoning models. *arXiv preprint arXiv:2503.04472*, 2025.

[471] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 8634–8652. Curran Associates, Inc., December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1b44b878bb782e6954cd888628510e90-Paper-Conference.pdf.

[472] Safal Shrestha, Minwu Kim, and Keith Ross. Mathematical reasoning in large language models: Assessing logical and arithmetic errors across wide numerical ranges. *arXiv preprint arXiv:2502.08680*, 2025.

[473] Kashun Shum, Shizhe Diao, and Tong Zhang. Automatic prompt augmentation and selection with chain-of-thought from labeled data. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12113–12139, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.811. URL https://aclanthology.org/2023.findings-emnlp.811/.

[474] Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*, 2024.

[475] Avi Singh, John D Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J Liu, James Harrison, Jaehoon Lee, Kelvin Xu, et al. Beyond human data: Scaling self-training for problem-solving with language models. *Transactions on Machine Learning Research*, April 2024.

[476] Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. *arXiv preprint arXiv:2502.02013*, 2025.

[477] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

[478] Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. Prmbench: A fine-grained and challenging benchmark for process-level reward models. *arXiv preprint arXiv:2501.03124*, 2025.

[479] Xiaoshuai Song, Yanan Wu, Weixun Wang, Jiaheng Liu, Wenbo Su, and Bo Zheng. Progco: Program helps self-correction of large language models. *arXiv preprint arXiv:2501.01264*, 2025.

[480] Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *arXiv preprint arXiv:2409.12183*, 2024.

[481] Zayne Rea Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. MuSR: Testing the limits of chain-of-thought with multistep soft reasoning. In *The Twelfth International Conference on Learning Representations*, January 2024. URL https://openreview.net/forum?id=jenyYQzue1.

[482] Gaurav Srivastava, Shuxiang Cao, and Xuan Wang. Towards reasoning ability of small language models. *arXiv preprint arXiv:2502.11569*, 2025.

[483] Saksham Sahai Srivastava and Ashutosh Gandhi. Mathdivide: Improved mathematical reasoning by large language models. *arXiv preprint arXiv:2405.13004*, 2024.

[484] Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. Chain of thoughtlessness? an analysis of cot in planning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, September 2024. URL https://openreview.net/forum?id=kPBEAZU5Nm.

[485] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc., December 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf.

[486] DiJia Su, Sainbayar Sukhbaatar, Michael Rabbat, Yuandong Tian, and Qinqing Zheng. Dualformer: Controllable fast and slow thinking by learning with randomized reasoning traces. *arXiv preprint arXiv:2410.09918*, 2024.

[487] Guangyan Sun, Mingyu Jin, Zhenting Wang, Cheng-Long Wang, Siqi Ma, Qifan Wang, Tong Geng, Ying Nian Wu, Yongfeng Zhang, and Dongfang Liu. Visual agents as fast and slow thinkers. In *The Thirteenth International Conference on Learning Representations*, January 2025. URL https://openreview.net/forum?id=ncCuiD3KJQ.

[488] Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, et al. A survey of reasoning with foundation models. *arXiv preprint arXiv:2312.11562*, 2023.

[489] Linzhuang Sun, Hao Liang, Jingxuan Wei, Bihui Yu, Tianpeng Li, Fan Yang, Zenan Zhou, and Wentao Zhang. Mm-verify: Enhancing multimodal reasoning with chain-of-thought verification. *arXiv preprint arXiv:2502.13383*, 2025.

[490] Shengyang Sun, Yian Zhang, Alexander Bukharin, David Mosallanezhad, Jiaqi Zeng, Soumye Singhal, Gerald Shen, Adi Renduchintala, Tugrul Konuk, Yi Dong, et al. Reward-aware preference optimization: A unified mathematical framework for model alignment. *arXiv preprint arXiv:2502.00203*, 2025.

[491] Wei Sun, Qianlong Du, Fuwei Cui, and Jiajun Zhang. An efficient and precise training data construction framework for process-supervised reward model in mathematical reasoning. *arXiv preprint arXiv:2503.02382*, 2025.

[492] Yuhong Sun, Zhangyue Yin, Xuanjing Huang, Xipeng Qiu, and Hui Zhao. Error classification of large language models on math word problems: A dynamically adaptive framework. *arXiv preprint arXiv:2501.15581*, 2025.

[493] Zhongxiang Sun, Qipeng Wang, Weijie Yu, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Song Yang, and Han Li. Rearter: Retrieval-augmented reasoning with trustworthy process rewarding. *arXiv preprint arXiv:2501.07861*, 2025.

[494] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT

Press, November 1999. URL https://proceedings.neurips.cc/paper_files/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf.

[495] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.824. URL https://aclanthology.org/2023.findings-acl.824/.

[496] Jihoon Tack, Jack Lanchantin, Jane Yu, Andrew Cohen, Ilia Kulikov, Janice Lan, Shibo Hao, Yuandong Tian, Jason Weston, and Xian Li. Llm pretraining with continuous concepts. *arXiv preprint arXiv:2502.08524*, 2025.

[497] Juanhe (TJ) Tan. Causal abstraction for chain-of-thought reasoning in arithmetic word problems. In Yonatan Belinkov, Sophie Hao, Jaap Jumelet, Najoung Kim, Arya McCarthy, and Hosein Mohebbi, editors, *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 155–168, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.blackboxnlp-1.12. URL https://aclanthology.org/2023.blackboxnlp-1.12.

[498] Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. Judgebench: A benchmark for evaluating llm-based judges. *arXiv preprint arXiv:2410.12784*, 2024.

[499] Xiaoyu Tan, Tianchu Yao, Chao Qu, Bin Li, Minghao Yang, Dakuan Lu, Haozhe Wang, Xihe Qiu, Wei Chu, Yinghui Xu, et al. Aurora: Automated training framework of universal process reward models via ensemble prompting and reverse verification. *arXiv preprint arXiv:2502.11520*, 2025.

[500] Zhengyang Tang, Ziniu Li, Zhenyang Xiao, Tian Ding, Ruoyu Sun, Benyou Wang, Dayiheng Liu, Fei Huang, Tianyu Liu, Bowen Yu, et al. Enabling scalable oversight via self-evolving critic. *arXiv preprint arXiv:2501.05727*, 2025.

[501] Zhengyang Tang, Ziniu Li, Zhenyang Xiao, Tian Ding, Ruoyu Sun, Benyou Wang, Dayiheng Liu, Fei Huang, Tianyu Liu, Bowen Yu, et al. Realcritic: Towards effectiveness-driven evaluation of language model critiques. *arXiv preprint arXiv:2501.14492*, 2025.

[502] Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. Confidence improves self-consistency in llms. *arXiv preprint arXiv:2502.06233*, 2025.

[503] DolphinR1 Team. Dolphin R1. https://huggingface.co/datasets/cognitivecomputations/dolphin-r1, February 2025.

[504] Fancy-MLLM Team. R1 Onevision. https://huggingface.co/datasets/Fancy-MLLM/R1-Onevision, February 2025.

[505] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[506] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[507] Huggingface Team. Open r1. https://github.com/huggingface/open-r1, January 2025.

[508] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

[509] NovaSky Team. Think less, achieve more: Cut reasoning costs by 50 https://novasky-ai.github.io/posts/reduce-overthinking, January 2025. Accessed: 2025-01-23.

[510] NovaSky Team. Sky-t1: Train your own o1 preview model within $ 450. https://novasky-ai.github.io/posts/sky-t1, January 2025. Accessed: 2025-01-09.

[511] NVIDIA Team. Mistral-nemo-12b-instruct. https://huggingface.co/nvidia/Mistral-NeMo-12B-Instruct, July 2024.

[512] OpenDeepResearch Team. Open deep research. https://github.com/nickscamara/open-deep-research, February 2025.

[513] OpenO1 Team. Open o1. https://github.com/Open-Source-O1/Open-O1, February 2025.

[514] OpenR1 Team. Open r1 math 200k. https://huggingface.co/datasets/open-r1/OpenR1-Math-220k, February 2025.

[515] OpenThoughts Team. Open Thoughts. https://open-thoughts.ai, January 2025.

[516] PowerInfer Team. QwQ LongCoT 500k. https://huggingface.co/datasets/PowerInfer/QWQ-LONGCOT-500K, January 2025.

[517] QwQ Team. Qwq: Reflect deeply on the boundaries of the unknown. https://qwenlm.github.io/blog/qwq-32b-preview/, November 2025.

[518] Video-R1 Team. Video-r1. https://github.com/tulerfeng/Video-R1, February 2025.

[519] X-R1 Team. X-r1. https://github.com/dhcode-cpp/X-R1, February 2025.

[520] Fengwei Teng, Zhaoyang Yu, Quan Shi, Jiayi Zhang, Chenglin Wu, and Yuyu Luo. Atom of thoughts for markov llm test-time scaling. *arXiv preprint arXiv:2502.12018*, 2025.

[521] Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *arXiv preprint arXiv:2501.06186*, 2025.

[522] George Thomas, Alex J Chan, Jikun Kang, Wenqi Wu, Filippos Christianos, Fraser Greenlee, Andy Toulis, and Marvin Purtorab. Webgames: Challenging general-purpose web-browsing ai agents. *arXiv preprint arXiv:2502.18356*, 2025.

[523] Ye Tian, Baolin Peng, Linfeng Song, Lifeng Jin, Dian Yu, Lei Han, Haitao Mi, and Dong Yu. Toward self-improvement of llms via imagination, searching, and criticizing. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 52723–52748. Curran Associates, Inc., September 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/5e5853f35164e434015716a8c2a66543-Paper-Conference.pdf.

[524] Yuxuan Tong, Xiwen Zhang, Rui Wang, Ruidong Wu, and Junxian He. Dart-math: Difficulty-aware rejection tuning for mathematical problem-solving. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 7821–7846. Curran Associates, Inc., September 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/0ef1afa0daa888d695dcd5e9513bafa3-Paper-Conference.pdf.

[525] Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. Openmathinstruct-2: Accelerating ai for math with massive open-source instruction data. *arXiv preprint arXiv:2410.01560*, 2024.

[526] Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. Openmathinstruct-2: Accelerating ai for math with massive open-source instruction data. *arXiv preprint arXiv:2410.01560*, 2024.

[527] Shubham Toshniwal, Ivan Moshkov, Sean Narenthiran, Daria Gitman, Fei Jia, and Igor Gitman. Openmathinstruct-1: A 1.8 million math instruction tuning dataset. *arXiv preprint arXiv:Arxiv-2402.10176*, 2024.

[528] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[529] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[530] Christoph Treude and Raula Gaikovina Kula. Interacting with ai reasoning models: Harnessing" thoughts" for ai-driven software engineering. *arXiv preprint arXiv:2503.00483*, 2025.

[531] Luong Trung, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. ReFT: Reasoning with reinforced fine-tuning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7601–7614, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.410. URL https://aclanthology.org/2024.acl-long.410/.

[532] Benjamin Turtel, Danny Franklin, and Philipp Schoenegger. Llms can teach themselves to better predict the future. *arXiv preprint arXiv:2502.05253*, 2025.

[533] Martin Tutek, Fateme Hashemi Chaleshtori, Ana Marasović, and Yonatan Belinkov. Measuring faithfulness of chains of thought by unlearning reasoning steps. *arXiv preprint arXiv:2502.14829*, 2025.

[534] Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

[535] Robert Vacareanu, Anurag Pratik, Evangelia Spiliopoulou, Zheng Qi, Giovanni Paolini, Neha Anna John, Jie Ma, Yassine Benajiba, and Miguel Ballesteros. General purpose verification for chain of thought prompting. *arXiv preprint arXiv:2405.00204*, 2024.

[536] Karthik Valmeekam, Kaya Stechly, and Subbarao Kambhampati. LLMs still can't plan; can LRMs? a preliminary evaluation of openAI's o1 on planbench. In *NeurIPS 2024 Workshop on Open-World Agents*, October 2024. URL https://openreview.net/forum?id=Gcr1Lx4Koz.

[537] Jean Vassoyan, Nathanaël Beau, and Roman Plaud. Ignore the kl penalty! boosting exploration on critical tokens to enhance rl fine-tuning. *arXiv preprint arXiv:2502.06533*, 2025.

[538] Tu Vu, Kalpesh Krishna, Salaheddin Alzubi, Chris Tar, Manaal Faruqui, and Yun-Hsuan Sung. Foundational autoraters: Taming large language models for better automatic evaluation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17086–17105, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.949. URL https://aclanthology.org/2024.emnlp-main.949/.

[539] Guangya Wan, Yuqi Wu, Jie Chen, and Sheng Li. Cot rerailer: Enhancing the reliability of large language models in complex reasoning tasks through error detection and correction. *arXiv preprint arXiv:2408.13940*, 2024.

[540] Ziyu Wan, Xidong Feng, Muning Wen, Stephen Marcus McAleer, Ying Wen, Weinan Zhang, and Jun Wang. Alphazero-like tree-search can guide large language model decoding and training. In *Forty-first International Conference on Machine Learning*, May 2024. URL https://openreview.net/forum?id=C4OpREezgj.

[541] Ante Wang, Linfeng Song, Ye Tian, Baolin Peng, Dian Yu, Haitao Mi, Jinsong Su, and Dong Yu. Litesearch: Efficacious tree search for llm. *arXiv preprint arXiv:2407.00320*, 2024.

[542] Ante Wang, Linfeng Song, Ye Tian, Dian Yu, Haitao Mi, Xiangyu Duan, Zhaopeng Tu, Jinsong Su, and Dong Yu. Don't get lost in the trees: Streamlining llm reasoning by overcoming tree search exploration pitfalls. *arXiv preprint arXiv:2502.11183*, 2025.

[543] Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. Towards understanding chain-of-thought prompting: An empirical study of what matters. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2739, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.153. URL https://aclanthology.org/2023.acl-long.153/.

[544] Chao Wang, Luning Zhang, Zheng Wang, and Yang Zhou. Can large language models unveil the mysteries? an exploration of their ability to unlock information in complex scenarios. *arXiv preprint arXiv:2502.19973*, 2025.

[545] Chaojie Wang, Yanchen Deng, Zhiyi Lyu, Liang Zeng, Jujie He, Shuicheng Yan, and Bo An. Q*: Improving multi-step reasoning for llms with deliberative planning. *arXiv preprint arXiv:2406.14283*, 2024.

[546] Clinton J Wang, Dean Lee, Cristina Menghini, Johannes Mols, Jack Doughty, Adam Khoja, Jayson Lynch, Sean Hendryx, Summer Yue, and Dan Hendrycks. Enigmaeval: A benchmark of long multimodal reasoning challenges. *arXiv preprint arXiv:2502.08859*, 2025.

[547] Danqing Wang, Zhuorui Ye, Fei Fang, and Lei Li. Cooperative strategic planning enhances reasoning capabilities in large language models. *arXiv preprint arXiv:2410.20007*, 2024.

[548] Evan Z Wang, Federico Cassano, Catherine Wu, Yunfeng Bai, William Song, Vaskar Nath, Ziwen Han, Sean M. Hendryx, Summer Yue, and Hugh Zhang. Planning in natural language improves LLM search for code generation. In *The First Workshop on System-2 Reasoning at Scale, NeurIPS'24*, October 2024. URL https://openreview.net/forum?id=B2iSfPNj49.

[549] Guoxin Wang, Minyu Gao, Shuai Yang, Ya Zhang, Lizhi He, Liang Huang, Hanlin Xiao, Yexuan Zhang, Wanyue Li, Lu Chen, et al. Citrus: Leveraging expert cognitive pathways in a medical language model for advanced medical decision support. *arXiv preprint arXiv:2502.18274*, 2025.

[550] Hanbin Wang, Xiaoxuan Zhou, Zhipeng Xu, Keyuan Cheng, Yuxin Zuo, Kai Tian, Jingwei Song, Junting Lu, Wenhui Hu, and Xueyang Liu. Code-vision: Evaluating multimodal llms logic understanding and code generation capabilities. *arXiv preprint arXiv:2502.11829*, 2025.

[551] Hanlin Wang, Jian Wang, Chak Tou Leong, and Wenjie Li. Steca: Step-level trajectory calibration for llm agent learning. *arXiv preprint arXiv:2502.14276*, 2025.

[552] Hao Wang, Boyi Liu, Yufeng Zhang, and Jie Chen. Seed-cts: Unleashing the power of tree search for superior performance in competitive coding tasks. *arXiv preprint arXiv:2412.12544*, 2024.

[553] Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10582–10592, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.620. URL https://aclanthology.org/2024.findings-emnlp.620/.

[554] Haoyu Wang, Zeyu Qin, Li Shen, Xueqian Wang, Minhao Cheng, and Dacheng Tao. Leveraging reasoning with guidelines to elicit and utilize knowledge for enhancing safety alignment. *arXiv preprint arXiv:2502.04040*, 2025.

[555] Huaijie Wang, Shibo Hao, Hanze Dong, Shenao Zhang, Yilin Bao, Ziran Yang, and Yi Wu. Offline reinforcement learning for llm multi-step reasoning. *arXiv preprint arXiv:2412.16145*, 2024.

[556] Jiaan Wang, Fandong Meng, Yunlong Liang, and Jie Zhou. Drt-o1: Optimized deep reasoning translation via long chain-of-thought. *arXiv preprint arXiv:2412.17498*, 2024.

[557] Jun Wang, Meng Fang, Ziyu Wan, Muning Wen, Jiachen Zhu, Anjie Liu, Ziqin Gong, Yan Song, Lei Chen, Lionel M Ni, et al. Openr: An open source framework for advanced reasoning with large language models. *arXiv preprint arXiv:2410.09671*, 2024.

[558] Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*, 2024.

[559] Junyang Wang, Haiyang Xu, Xi Zhang, Ming Yan, Ji Zhang, Fei Huang, and Jitao Sang. Mobile-agent-v: Learning mobile device operation through video-guided multi-agent collaboration. *arXiv preprint arXiv:2502.17110*, 2025.

[560] Ke Wang, Houxing Ren, Aojun Zhou, Zimu Lu, Sichun Luo, Weikang Shi, Renrui Zhang, Linqi Song, Mingjie Zhan, and Hongsheng Li. Mathcoder: Seamless code integration in llms for enhanced mathematical reasoning. *arXiv preprint arXiv:2310.03731*, 2023.

[561] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MATH-vision dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and*

Benchmarks Track, September 2024. URL https://openreview.net/forum?id=QWTCcxMpPA.

[562] Ke Wang, Houxing Ren, Aojun Zhou, Zimu Lu, Sichun Luo, Weikang Shi, Renrui Zhang, Linqi Song, Mingjie Zhan, and Hongsheng Li. Mathcoder: Seamless code integration in LLMs for enhanced mathematical reasoning. In *The Twelfth International Conference on Learning Representations*, January 2024. URL https://openreview.net/forum?id=z8TW0ttBPp.

[563] Kevin Wang, Junbo Li, Neel P Bhatt, Yihan Xi, Qiang Liu, Ufuk Topcu, and Zhangyang Wang. On the planning abilities of openai's o1 models: Feasibility, optimality, and generalizability. *arXiv preprint arXiv:2409.19924*, 2024.

[564] Liang Wang, Haonan Chen, Nan Yang, Xiaolong Huang, Zhicheng Dou, and Furu Wei. Chain-of-retrieval augmented generation. *arXiv preprint arXiv:2501.14342*, 2025.

[565] Libo Wang. Dynamic chain-of-thought: Towards adaptive deep reasoning. *arXiv preprint arXiv:2502.10428*, 2025.

[566] Peifeng Wang, Austin Xu, Yilun Zhou, Caiming Xiong, and Shafiq Joty. Direct judgement preference optimization. *arXiv preprint arXiv:2409.14664*, 2024.

[567] Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.510. URL https://aclanthology.org/2024.acl-long.510/.

[568] Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. Science-World: Is your agent smarter than a 5th grader? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11279–11298, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.775. URL https://aclanthology.org/2022.emnlp-main.775/.

[569] Siyuan Wang, Enda Zhao, Zhongyu Wei, and Xiang Ren. Stepwise informativeness search for improving llm reasoning. *arXiv preprint arXiv:2502.15335*, 2025.

[570] Tianlong Wang, Junzhe Chen, Xueting Han, and Jing Bai. Cpl: Critical plan step learning boosts llm generalization in reasoning tasks. *arXiv preprint arXiv:2409.08642*, 2024.

[571] Tianlu Wang, Ping Yu, Xiaoqing Ellen Tan, Sean O'Brien, Ramakanth Pasunuru, Jane Dwivedi-Yu, Olga Golovneva, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. Shepherd: A critic for language model generation. *arXiv preprint arXiv:2308.04592*, 2023.

[572] Tianlu Wang, Ilia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu, Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. Self-taught evaluators. *arXiv preprint arXiv:2408.02666*, 2024.

[573] Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. Demystifying multilingual chain-of-thought in process reward modeling. *arXiv preprint arXiv:2502.12663*, 2025.

[574] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, et al. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024.

[575] Xinyi Wang, Lucas Caccia, Oleksiy Ostapenko, Xingdi Yuan, William Yang Wang, and Alessandro Sordoni. Guiding language model reasoning with planning tokens. *arXiv preprint arXiv:2310.05707*, 2023.

[576] Xinyi Wang, Alfonso Amayuelas, Kexun Zhang, Liangming Pan, Wenhu Chen, and William Yang Wang. Understanding reasoning ability of language models from the perspective of reasoning paths aggregation. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 50026–50042. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/wang24a.html.

[577] Xiyao Wang, Jiuhai Chen, Zhaoyang Wang, Yuhang Zhou, Yiyang Zhou, Huaxiu Yao, Tianyi Zhou, Tom Goldstein, Parminder Bhatia, Furong Huang, et al. Enhancing visual-language modality alignment in large vision language models via self-improvement. *arXiv preprint arXiv:2405.15973*, 2024.

[578] Xiyao Wang, Linfeng Song, Ye Tian, Dian Yu, Baolin Peng, Haitao Mi, Furong Huang, and Dong Yu. Towards self-improvement of llms via mcts: Leveraging stepwise knowledge with curriculum preference learning. *arXiv preprint arXiv:2410.06508*, 2024.

[579] Xuezhi Wang and Denny Zhou. Chain-of-thought reasoning without prompting. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, September 2024. URL https://openreview.net/forum?id=4Zt7S0B0Jp.

[580] 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*, February 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

[581] Yao Wang, Mingxuan Cui, and Arthur Jiang. Enabling ai scientists to recognize innovation: A domain-agnostic algorithm for assessing novelty. *arXiv preprint arXiv:2503.01508*, 2025.

[582] Yifei Wang, Yuyang Wu, Zeming Wei, Stefanie Jegelka, and Yisen Wang. A theoretical understanding of self-correction through in-context alignment. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, September 2024. URL https://openreview.net/forum?id=OtvNLTWYww.

[583] Yiqun Wang, Sile Hu, Yonggang Zhang, Xiang Tian, Xuesong Liu, Yaowu Chen, Xu Shen, and Jieping Ye. How large language models implement chain-of-thought? September 2023.

[584] Yu Wang, Nan Yang, Liang Wang, and Furu Wei. Examining false positives under inference scaling for mathematical reasoning. *arXiv preprint arXiv:2502.06217*, 2025.

[585] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, September 2024. URL https://openreview.net/forum?id=y10DM6R2r3.

[586] Yubo Wang, Xiang Yue, and Wenhu Chen. Critique fine-tuning: Learning to critique is more effective than learning to imitate. *arXiv preprint arXiv:2501.17703*, 2025.

[587] Yue Wang, Qiuzhi Liu, Jiahao Xu, Tian Liang, Xingyu Chen, Zhiwei He, Linfeng Song, Dian Yu, Juntao Li, Zhuosheng Zhang, et al. Thoughts are all over the place: On the underthinking of o1-like llms. *arXiv preprint arXiv:2501.18585*, 2025.

[588] Zhaoyang Wang, Weilei He, Zhiyuan Liang, Xuchao Zhang, Chetan Bansal, Ying Wei, Weitong Zhang, and Huaxiu Yao. Cream: Consistency regularized self-rewarding language models. In *Neurips Safe Generative AI Workshop 2024*, October 2024. URL https://openreview.net/forum?id=oaWajnM93y.

[589] Zhenhailong Wang, Haiyang Xu, Junyang Wang, Xi Zhang, Ming Yan, Ji Zhang, Fei Huang, and Heng Ji. Mobile-agent-e: Self-evolving mobile assistant for complex tasks. *arXiv preprint arXiv:2501.11733*, 2025.

[590] Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. Helpsteer 2: Open-source dataset for training top-performing reward models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, September 2024. URL https://openreview.net/forum?id=PvVKUFhaNy.

[591] Zihan Wang, Yunxuan Li, Yuexin Wu, Liangchen Luo, Le Hou, Hongkun Yu, and Jingbo Shang. Multi-step problem solving through a verifier: An empirical analysis on model-induced process supervision. *arXiv preprint arXiv:2402.02658*, 2024.

[592] Anjiang Wei, Jiannan Cao, Ran Li, Hongyu Chen, Yuhui Zhang, Ziheng Wang, Yaofeng Sun, Yuan Liu, Thiago SFX Teixeira, Diyi Yang, et al. Equibench: Benchmarking code reasoning capabilities of large language models via equivalence checking. *arXiv preprint arXiv:2502.12466*, 2025.

[593] Haoran Wei, Youyang Yin, Yumeng Li, Jia Wang, Liang Zhao, Jianjian Sun, Zheng Ge, and Xiangyu Zhang. Slow perception: Let's perceive geometric figures step-by-step. *arXiv preprint arXiv:2412.20631*, 2024.

[594] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc., November 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

[595] Ting-Ruen Wei, Haowei Liu, Xuyang Wu, and Yi Fang. A survey on feedback-based multi-step reasoning for large language models on mathematics. *arXiv preprint arXiv:2502.14333*, 2025.

[596] Yuxiang Wei, Olivier Duchenne, Jade Copet, Quentin Carbonneaux, Lingming Zhang, Daniel Fried, Gabriel Synnaeve, Rishabh Singh, and Sida I. Wang. Swe-rl: Advancing llm reasoning via reinforcement learning on open software evolution. *arXiv preprint arXiv:2502.18449*, 2025.

[597] Nathaniel Weir, Muhammad Khalifa, Linlu Qiu, Orion Weller, and Peter Clark. Learning to reason via program generation, emulation, and search. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, September 2024. URL https://openreview.net/forum?id=te6VagJf6G.

[598] Sean Welleck, Amanda Bertsch, Matthew Finlayson, Hailey Schoelkopf, Alex Xie, Graham Neubig, Ilia Kulikov, and Zaid Harchaoui. From decoding to meta-generation: Inference-time algorithms for large language models. *Transactions on Machine Learning Research*, November 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=eskQMcIbMS. Survey Certification.

[599] Jiaxin Wen, Jian Guan, Hongning Wang, Wei Wu, and Minlie Huang. Codeplan: Unlocking reasoning potential in large language models by scaling code-form planning. In *The Thirteenth International Conference on Learning Representations*, January 2025. URL https://openreview.net/forum?id=dCPF1wlqj8.

[600] Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. Large language models are better reasoners with self-verification. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2550–2575, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.167. URL https://aclanthology.org/2023.findings-emnlp.167/.

[601] Jason Weston and Sainbayar Sukhbaatar. System 2 attention (is something you might need too). *arXiv preprint arXiv:2311.11829*, 2023.

[602] Yotam Wolf, Binyamin Rothberg, Dorin Shteyman, and Amnon Shashua. Compositional hardness of code in large language models–a probabilistic perspective. *arXiv preprint arXiv:2409.18028*, 2024.

[603] Chengyue Wu, Yixiao Ge, Qiushan Guo, Jiahao Wang, Zhixuan Liang, Zeyu Lu, Ying Shan, and Ping Luo. Plot2code: A comprehensive benchmark for evaluating multi-modal large language models in code generation from scientific plots. *arXiv preprint arXiv:2405.07990*, 2024.

[604] Jinyang Wu, Mingkuan Feng, Shuai Zhang, Feihu Che, Zengqi Wen, and Jianhua Tao. Beyond examples: High-level automated reasoning paradigm in in-context learning via mcts. *arXiv preprint arXiv:2411.18478*, 2024.

[605] Jinyang Wu, Mingkuan Feng, Shuai Zhang, Ruihan Jin, Feihu Che, Zengqi Wen, and Jianhua Tao. Boosting multimodal reasoning with mcts-automated structured thinking. *arXiv preprint arXiv:2502.02339*, 2025.

[606] Junde Wu, Jiayuan Zhu, and Yuyuan Liu. Agentic reasoning: Reasoning llms with tools for the deep research. *arXiv preprint arXiv:2502.04644*, 2025.

[607] Siwei Wu, Zhongyuan Peng, Xinrun Du, Tuney Zheng, Minghao Liu, Jialong Wu, Jiachen Ma, Yizhi Li, Jian Yang, Wangchunshu Zhou, et al. A comparative study on reasoning patterns of openai's o1 model. *arXiv preprint arXiv:2410.13639*, 2024.

[608] Tianhao Wu, Janice Lan, Weizhe Yuan, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar. Thinking llms: General instruction following with thought generation. *arXiv preprint arXiv:2410.10630*, 2024.

[609] Wenjie Wu, Yongcheng Jing, Yingjie Wang, Wenbin Hu, and Dacheng Tao. Graph-augmented reasoning: Evolving step-by-step knowledge graph retrieval for llm reasoning. *arXiv preprint arXiv:2503.01642*, 2025.

[610] Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models. *arXiv preprint arXiv:2408.00724*, January 2024.

[611] Yuyang Wu, Yifei Wang, Tianqi Du, Stefanie Jegelka, and Yisen Wang. When more is less: Understanding chain-of-thought length in llms. *arXiv preprint arXiv:2502.07266*, 2025.

[612] Zhenyu Wu, Qingkai Zeng, Zhihan Zhang, Zhaoxuan Tan, Chao Shen, and Meng Jiang. Enhancing mathematical reasoning in llms by stepwise correction. *arXiv preprint arXiv:2410.12934*, 2024.

[613] Zhenyu Wu, Qingkai Zeng, Zhihan Zhang, Zhaoxuan Tan, Chao Shen, and Meng Jiang. Large language models can self-correct with minimal effort. In *AI for Math Workshop @ ICML 2024*, May 2024. URL https://openreview.net/forum?id=mmZLMs4l3d.

[614] Zirui Wu, Xiao Liu, Jiayi Li, Lingpeng Kong, and Yansong Feng. Haste makes waste: Evaluating planning abilities of llms for efficient and feasible multitasking with time constraints between actions. *arXiv preprint arXiv:2503.02238*, 2025.

[615] Zongqian Wu, Tianyu Li, Jiaying Yang, Mengmeng Zhan, Xiaofeng Zhu, and Lei Feng. Is depth all you need? an exploration of iterative reasoning in llms. *arXiv preprint arXiv:2502.10858*, 2025.

[616] Zhiheng Xi, Dingwen Yang, Jixuan Huang, Jiafu Tang, Guanyu Li, Yiwen Ding, Wei He, Boyang Hong, Shihan Do, Wenyu Zhan, et al. Enhancing llm reasoning via critique models with test-time and training-time supervision. *arXiv preprint arXiv:2411.16579*, 2024.

[617] Heming Xia, Yongqi Li, Chak Tou Leong, Wenjie Wang, and Wenjie Li. Tokenskip: Controllable chain-of-thought compression in llms. *arXiv preprint arXiv:2502.12067*, 2025.

[618] Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. Evaluating mathematical reasoning beyond accuracy. *arXiv preprint arXiv:2404.05692*, 2024.

[619] Kun Xiang, Zhili Liu, Zihao Jiang, Yunshuang Nie, Runhui Huang, Haoxiang Fan, Hanhui Li, Weiran Huang, Yihan Zeng, Jianhua Han, et al. Atomthink: A slow thinking framework for multimodal mathematical reasoning. *arXiv preprint arXiv:2411.11930*, 2024.

[620] Violet Xiang, Charlie Snell, Kanishk Gandhi, Alon Albalak, Anikait Singh, Chase Blagden, Duy Phung, Rafael Rafailov, Nathan Lile, Dakota Mahan, et al. Towards system 2 reasoning in llms: Learning how to think with meta chain-of-though. *arXiv preprint arXiv:2501.04682*, 2025.

[621] Wenyi Xiao, Zechuan Wang, Leilei Gan, Shuai Zhao, Wanggui He, Luu Anh Tuan, Long Chen, Hao Jiang, Zhou Zhao, and Fei Wu. A comprehensive survey of direct preference optimization: Datasets, theories, variants, and applications. *arXiv preprint arXiv:2410.15595*, 2024.

[622] Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2502.14768*, February 2025.

[623] Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, September 2024. URL https://openreview.net/forum?id=tN61DTr4Ed.

[624] Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, Xu Zhao, Min-Yen Kan, Junxian He, and Qizhe Xie. Self-evaluation guided beam search for reasoning. In *Thirty-seventh Conference on Neural Information Processing Systems*, September 2023. URL https://openreview.net/forum?id=Bw82hwg5Q3.

[625] Yuxi Xie, Anirudh Goyal, Wenyue Zheng, Min-Yen Kan, Timothy P Lillicrap, Kenji Kawaguchi, and Michael Shieh. Monte carlo tree search boosts reasoning via iterative preference learning. *arXiv preprint arXiv:2405.00451*, 2024.

[626] Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. Audio-reasoner: Improving reasoning capability in large audio language models. *arXiv preprint arXiv:2503.02318*, 2025.

[627] Zhihui Xie, Liyu Chen, Weichao Mao, Jingjing Xu, Lingpeng Kong, et al. Teaching language models to critique via reinforcement learning. *arXiv preprint arXiv:2502.03492*, 2025.

[628] Siheng Xiong, Ali Payani, Yuan Yang, and Faramarz Fekri. Deliberate reasoning for llms as structure-aware planning with accurate world model. *arXiv preprint arXiv:2410.03136*, 2024.

[629] Wang Xiyao, Yang Zhengyuan, Li Linjie, Lu Hongjin, Xu Yuancheng, Lin Chung-Ching Lin, Lin Kevin, Huang Furong, and Wang Lijuan. Scaling inference-time search with vision value model for improved visual comprehension. *arXiv preprint arXiv:2412.03704*, 2024.

[630] Bin Xu, Yiguan Lin, Yinghao Li, et al. Sra-mcts: Self-driven reasoning aurmentation with monte carlo tree search for enhanced code generation. *arXiv preprint arXiv:2411.11053*, 2024.

[631] Fangzhi Xu, Qiushi Sun, Kanzhi Cheng, Jun Liu, Yu Qiao, and Zhiyong Wu. Interactive evolution: A neural-symbolic self-training framework for large language models. *arXiv preprint arXiv:2406.11736*, 2024.

[632] Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, et al. Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*, 2025.

[633] Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024.

[634] Haotian Xu. No train still gain. unleash mathematical reasoning of large language models with monte carlo tree search guided by energy function. *arXiv preprint arXiv:2309.03224*, 2023.

[635] Haotian Xu, Xing Wu, Weinong Wang, Zhongzhi Li, Da Zheng, Boyuan Chen, Yi Hu, Shijia Kang, Jiaming Ji, Yingying Zhang, et al. Redstar: Does scaling long-cot data unlock better slow-reasoning systems? *arXiv preprint arXiv:2501.11284*, 2025.

[636] Huimin Xu, Xin Mao, Feng-Lin Li, Xiaobao Wu, Wang Chen, Wei Zhang, and Anh Tuan Luu. Full-step-dpo: Self-supervised preference optimization with step-wise rewards for mathematical reasoning. *arXiv preprint arXiv:2502.14356*, 2025.

[637] Pusheng Xu, Yue Wu, Kai Jin, Xiaolan Chen, Mingguang He, and Danli Shi. Deepseek-r1 outperforms gemini 2.0 pro, openai o1, and o3-mini in bilingual complex ophthalmology reasoning. *arXiv preprint arXiv:2502.17947*, 2025.

[638] Rongwu Xu, Xiaojian Li, Shuo Chen, and Wei Xu. " nuclear deployed!": Analyzing catastrophic risks in decision-making of autonomous llm agents. *arXiv preprint arXiv:2502.11355*, 2025.

[639] Silei Xu, Wenhao Xie, Lingxiao Zhao, and Pengcheng He. Chain of draft: Thinking faster by writing less. *arXiv preprint arXiv:2502.18600*, 2025.

[640] Xin Xu, Shizhe Diao, Can Yang, and Yang Wang. Can we verify step by step for incorrect answer detection? *arXiv preprint arXiv:2402.10528*, 2024.

[641] Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. Softcot: Soft chain-of-thought for efficient reasoning with llms. *arXiv preprint arXiv:2502.12134*, 2025.

[642] Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. 2024.

[643] Zhangchen Xu, Yang Liu, Yueqin Yin, Mingyuan Zhou, and Radha Poovendran. Kodcode: A diverse, challenging, and verifiable synthetic dataset for coding. February 2025.

[644] Ruin Yan, Zheng Liu, and Defu Lian. O1 embedder: Let retrievers think before action. *arXiv preprint arXiv:2502.07555*, 2025.

[645] Yibo Yan, Shen Wang, Jiahao Huo, Hang Li, Boyan Li, Jiamin Su, Xiong Gao, Yi-Fan Zhang, Tianlong Xu, Zhendong Chu, et al. Errorradar: Benchmarking complex mathematical reasoning of multimodal large language models via error detection. *arXiv preprint arXiv:2410.04509*, 2024.

[646] Yibo Yan, Shen Wang, Jiahao Huo, Jingheng Ye, Zhendong Chu, Xuming Hu, Philip S Yu, Carla Gomes, Bart Selman, and Qingsong Wen. Position: Multimodal large language models can significantly advance scientific reasoning. *arXiv preprint arXiv:2502.02871*, 2025.

[647] Yuchen Yan, Jin Jiang, Yang Liu, Yixin Cao, Xin Xu, Xunliang Cai, Jian Shao, et al. S ³ c-math: Spontaneous step-level self-correction makes large language models better mathematical reasoners. *arXiv preprint arXiv:2409.01524*, 2024.

[648] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[649] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[650] An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

[651] Chen Yang, Chenyang Zhao, Quanquan Gu, and Dongruo Zhou. Cops: Empowering llm agents with provable cross-task experience sharing. *arXiv preprint arXiv:2410.16670*, 2024.

[652] Cheng Yang, Chufan Shi, Siheng Li, Bo Shui, Yujiu Yang, and Wai Lam. Llm2: Let large language models harness system 2 reasoning. *arXiv preprint arXiv:2412.20372*, 2024.

[653] Kailai Yang, Zhiwei Liu, Qianqian Xie, Jimin Huang, Erxue Min, and Sophia Ananiadou. Selective preference optimization via token-level reward function estimation. *arXiv preprint arXiv:2408.13518*, 2024.

[654] Kaiyu Yang, Gabriel Poesia, Jingxuan He, Wenda Li, Kristin Lauter, Swarat Chaudhuri, and Dawn Song. Formal mathematical reasoning: A new frontier in ai. *arXiv preprint arXiv:2412.16075*, 2024.

[655] Lei Yang, Renren Jin, Ling Shi, Jianxiang Peng, Yue Chen, and Deyi Xiong. Probench: Benchmarking large language models in competitive programming. *arXiv preprint arXiv:2502.20868*, 2025.

[656] Ling Yang, Zhaochen Yu, Bin Cui, and Mengdi Wang. Reasonflux: Hierarchical llm reasoning via scaling thought templates. *arXiv preprint arXiv:2502.06772*, 2025.

[657] Sherry Yang, Dale Schuurmans, Pieter Abbeel, and Ofir Nachum. Chain of thought imitation with procedure cloning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, November 2022. URL https://openreview.net/forum?id=ZJqqSa8FsH9.

[658] Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. Do large language models latently perform multi-hop reasoning? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10210–10229, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.550. URL https://aclanthology.org/2024.acl-long.550/.

[659] Wang Yang, Hongye Jin, Jingfeng Yang, Vipin Chaudhary, and Xiaotian Han. Thinking preference optimization. *arXiv preprint arXiv:2502.13173*, 2025.

[660] Wenkai Yang, Shuming Ma, Yankai Lin, and Furu Wei. Towards thinking-optimal scaling of test-time compute for llm reasoning. *arXiv preprint arXiv:2502.18080*, 2025.

[661] Xiao-Wen Yang, Xuan-Yi Zhu, Wen-Da Wei, Ding-Chu Zhang, Jie-Jing Shao, Zhi Zhou, Lan-Zhe Guo, and Yu-Feng Li. Step back to leap forward: Self-backtracking for boosting reasoning of language models. *arXiv preprint arXiv:2502.04404*, 2025.

[662] Yifei Yang, Zouying Cao, Qiguang Chen, Libo Qin, Dongjie Yang, Hai Zhao, and Zhi Chen. Kvsharer: Efficient inference via layer-wise dissimilar kv cache sharing. *arXiv preprint arXiv:2410.18517*, 2024.

[663] Yuqing Yang, Yan Ma, and Pengfei Liu. Weak-to-strong reasoning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8350–8367, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.490. URL https://aclanthology.org/2024.findings-emnlp.490/.

[664] Zhe Yang, Yichang Zhang, Yudong Wang, Ziyao Xu, Junyang Lin, and Zhifang Sui. Confidence vs critique: A decomposition of self-correction capability for llms. *arXiv preprint arXiv:2412.19513*, 2024.

[665] Zonghan Yang, Peng Li, Ming Yan, Ji Zhang, Fei Huang, and Yang Liu. React meets actre: Autonomous annotation of agent trajectories for contrastive self-training. In *First Conference on Language Modeling*, July 2024. URL https://openreview.net/forum?id=0VLBwQGWpA.

[666] Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*, 2024.

[667] Shunyu Yao, Howard Chen, John Yang, and Karthik R Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=R9KnuFlvnU.

[668] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc., September 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/271db9922b8d1f4dd7aaef84ed5ac703-Paper-Conference.pdf.

[669] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, February 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

[670] Xinhao Yao, Ruifeng Ren, Yun Liao, and Yong Liu. Unveiling the mechanisms of explicit cot training: How chain-of-thought enhances reasoning generalization. *arXiv preprint arXiv:2502.04667*, 2025.

[671] Yang Yao, Xuan Tong, Ruofan Wang, Yixu Wang, Lujundong Li, Liang Liu, Yan Teng, and Yingchun Wang. A mousetrap: Fooling large reasoning models for jailbreak with chain of iterative chaos. *arXiv preprint arXiv:2502.15806*, 2025.

[672] Michihiro Yasunaga, Luke Zettlemoyer, and Marjan Ghazvininejad. Multimodal rewardbench: Holistic evaluation of reward models for vision language models. *arXiv preprint arXiv:2502.14191*, 2025.

[673] Nicolas Yax, Hernán Anlló, and Stefano Palminteri. Studying and improving reasoning in humans and machines. *Communications Psychology*, 2(1):51, 2024.

[674] Guanghao Ye, Khiem Duc Pham, Xinzhi Zhang, Sivakanth Gopi, Baolin Peng, Beibin Li, Janardhan Kulkarni, and Huseyin A Inan. On the emergence of thinking in llms i: Searching for the right intuition. *arXiv preprint arXiv:2502.06773*, 2025.

[675] Tian Ye, Zicheng Xu, Yuanzhi Li, and Zeyuan Allen-Zhu. Physics of language models: Part 2.2, how to learn from mistakes on grade-school math problems. In *The Thirteenth International Conference on Learning Representations*, January 2025. URL https://openreview.net/forum?id=zpDGwcmMV4.

[676] Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*, 2025.

[677] Zihuiwen Ye, Fraser Greenlee-Scott, Max Bartolo, Phil Blunsom, Jon Ander Campos, and Matthias Gallé. Improving reward models with synthetic critiques. *arXiv preprint arXiv:2405.20850*, 2024.

[678] Zihuiwen Ye, Luckeciano Carvalho Melo, Younesse Kaddar, Phil Blunsom, Sam Staton, and Yarin Gal. Uncertainty-aware step-wise verification with generative reward models. *arXiv preprint arXiv:2502.11250*, 2025.

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

[680] Hao Yi, Qingyang Li, Yulan Hu, Fuzheng Zhang, Di Zhang, and Yong Liu. Sppd: Self-training with process preference learning using dynamic value margin. *arXiv preprint arXiv:2502.13516*, 2025.

[681] Zhangyue Yin, Qiushi Sun, Qipeng Guo, Zhiyuan Zeng, Xiaonan Li, Junqi Dai, Qinyuan Cheng, Xuanjing Huang, and Xipeng Qiu. Reasoning in flux: Enhancing large language models reasoning through uncertainty-aware adaptive guidance. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2401–2416, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.131. URL https://aclanthology.org/2024.acl-long.131/.

[682] Huaiyuan Ying, Shuo Zhang, Linyang Li, Zhejian Zhou, Yunfan Shao, Zhaoye Fei, Yichuan Ma, Jiawei Hong, Kuikun Liu, Ziyi Wang, et al. Internlm-math: Open math large language models toward verifiable reasoning. *arXiv preprint arXiv:2402.06332*, 2024.

[683] Eunseop Yoon, Hee Suk Yoon, SooHwan Eom, Gunsoo Han, Daniel Wontae Nam, Daejin Jo, Kyoung-Woon On, Mark A Hasegawa-Johnson, Sungwoong Kim, and Chang D Yoo. Tlcr: Token-level continuous reward for fine-grained reinforcement learning from human feedback. *arXiv preprint arXiv:2407.16574*, 2024.

[684] Dian Yu, Baolin Peng, Ye Tian, Linfeng Song, Haitao Mi, and Dong Yu. Siam: Self-improving code-assisted mathematical reasoning of large language models. *arXiv preprint arXiv:2408.15565*, 2024.

[685] Fei Yu, Anningzhe Gao, and Benyou Wang. OVM, outcome-supervised value models for planning in mathematical reasoning. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 858–875, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.55. URL https://aclanthology.org/2024.findings-naacl.55/.

[686] Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. Natural language reasoning, a survey. *ACM Comput. Surv.*, 56(12), October 2024. ISSN 0360-0300. doi: 10.1145/3664194. URL https://doi.org/10.1145/3664194.

[687] Fei Yu, Yingru Li, and Benyou Wang. Uncertainty-aware search and value models: Mitigating search scaling flaws in llms. *arXiv preprint arXiv:2502.11155*, 2025.

[688] Longhui Yu, Weisen Jiang, Han Shi, Jincheng YU, Zhengying Liu, Yu Zhang, James Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. In *The Twelfth International Conference on Learning Representations*, January 2024. URL https://openreview.net/forum?id=N8N0hgNDRt.

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

[690] Xiao Yu, Baolin Peng, Vineeth Vajipey, Hao Cheng, Michel Galley, Jianfeng Gao, and Zhou Yu. ExACT: Teaching AI agents to explore with reflective-MCTS and exploratory learning. In *The Thirteenth International Conference on Learning Representations*, January 2025. URL https://openreview.net/forum?id=GBIUbwW9D8.

[691] Yiyao Yu, Yuxiang Zhang, Dongdong Zhang, Xiao Liang, Hengyuan Zhang, Xingxing Zhang, Ziyi Yang, Mahmoud Khademi, Hany Awadalla, Junjie Wang, et al. Chain-of-reasoning: Towards unified mathematical reasoning in large language models via a multi-paradigm perspective. *arXiv preprint arXiv:2501.11110*, 2025.

[692] Yue Yu, Zhengxing Chen, Aston Zhang, Liang Tan, Chenguang Zhu, Richard Yuanzhe Pang, Yundi Qian, Xuewei Wang, Suchin Gururangan, Chao Zhang, et al. Self-generated critiques boost reward modeling for language models. *arXiv preprint arXiv:2411.16646*, 2024.

[693] Zeping Yu, Yonatan Belinkov, and Sophia Ananiadou. Back attention: Understanding and enhancing multi-hop reasoning in large language models. *arXiv preprint arXiv:2502.10835*, 2025.

[694] Zhaojian Yu, Yilun Zhao, Arman Cohan, and Xiao-Ping Zhang. Humaneval pro and mbpp pro: Evaluating large language models on self-invoking code generation. *arXiv preprint arXiv:2412.21199*, 2024.

[695] Zhouliang Yu, Yuhuan Yuan, Tim Z Xiao, Fuxiang Frank Xia, Jie Fu, Ge Zhang, Ge Lin, and Weiyang Liu. Generating symbolic world models via test-time scaling of large language models. *arXiv preprint arXiv:2502.04728*, 2025.

[696] Zhuohao Yu, Weizheng Gu, Yidong Wang, Zhengran Zeng, Jindong Wang, Wei Ye, and Shikun Zhang. Outcome-refining process supervision for code generation. *arXiv preprint arXiv:2412.15118*, 2024.

[697] Zishun Yu, Tengyu Xu, Di Jin, Karthik Abinav Sankararaman, Yun He, Wenxuan Zhou, Zhouhao Zeng, Eryk Helenowski, Chen Zhu, Sinong Wang, et al. Think smarter not harder: Adaptive reasoning with inference aware optimization. *arXiv preprint arXiv:2501.17974*, 2025.

[698] Jiahao Yuan, Dehui Du, Hao Zhang, Zixiang Di, and Usman Naseem. Reversal of thought: Enhancing large language models with preference-guided reverse reasoning warm-up. *arXiv preprint arXiv:2410.12323*, 2024.

[699] Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. Free process rewards without process labels. *arXiv preprint arXiv:2412.01981*, 2024.

[700] Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Boji Shan, Zeyuan Liu, Jia Deng, Huimin Chen, Ruobing Xie, Yankai Lin, Zhenghao Liu, Bowen Zhou, Hao Peng, Zhiyuan Liu, and Maosong Sun. Advancing LLM reasoning generalists with preference trees. In *The Thirteenth International Conference on Learning Representations*, January 2025. URL https://openreview.net/forum?id=2ea5TNVR0c.

[701] Michelle Yuan, Elman Mansimov, Katerina Margatina, Anurag Pratik, Daniele Bonadiman, Monica Sunkara, Yi Zhang, Yassine Benajiba, et al. A study on leveraging search and self-feedback for agent reasoning. *arXiv preprint arXiv:2502.12094*, 2025.

[702] Siyu Yuan, Zehui Chen, Zhiheng Xi, Junjie Ye, Zhengyin Du, and Jiecao Chen. Agent-r: Training language model agents to reflect via iterative self-training. *arXiv preprint arXiv:2501.11425*, 2025.

[703] Weizhe Yuan, Jane Yu, Song Jiang, Karthik Padthe, Yang Li, Dong Wang, Ilia Kulikov, Kyunghyun Cho, Yuandong Tian, Jason E Weston, and Xian Li. Naturalreasoning: Reasoning in the wild with 2.8m challenging questions, 2025. URL https://arxiv.org/abs/2502.13124.

[704] Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*, 2023.

[705] Xiang Yue, Tianyu Zheng, Ge Zhang, and Wenhu Chen. Mammoth2: Scaling instructions from the web. *Advances in Neural Information Processing Systems*, 37:90629–90660, 2025.

[706] Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo Ma, Haodong Duan, Wenwei Zhang, et al. Internlm-xcomposer2.5-reward: A simple yet effective multi-modal reward model. *arXiv preprint arXiv:2501.12368*, 2025.

[707] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, November 2022.

[708] Eric Zelikman, Georges Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah D Goodman. Quiet-star: Language models can teach themselves to think before speaking. *arXiv preprint arXiv:2403.09629*, 2024.

[709] Huaye Zeng, Dongfu Jiang, Haozhe Wang, Ping Nie, Xiaotong Chen, and Wenhu Chen. Acecoder: Acing coder rl via automated test-case synthesis. *arXiv preprint arXiv:2502.01718*, 2025.

[710] Thomas Zeng, Shuibai Zhang, Shutong Wu, Christian Classen, Daewon Chae, Ethan Ewer, Minjae Lee, Heeju Kim, Wonjun Kang, Jackson Kunde, et al. Versaprm: Multi-domain process reward model via synthetic reasoning data. *arXiv preprint arXiv:2502.06737*, 2025.

[711] Weihao Zeng, Yuzhen Huang, Lulu Zhao, Yijun Wang, Zifei Shan, and Junxian He. B-star: Monitoring and balancing exploration and exploitation in self-taught reasoners. *arXiv preprint arXiv:2412.17256*, 2024.

[712] Weihao Zeng, Yuzhen Huang, Wei Liu, Keqing He, Qian Liu, Zejun Ma, and Junxian He. 7b model and 8k examples: Emerging reasoning with reinforcement learning is both effective and efficient. https://hkust-nlp.notion.site/simplerl-reason, January 2025. Notion Blog.

[713] Yongcheng Zeng, Xinyu Cui, Xuanfa Jin, Guoqing Liu, Zexu Sun, Quan He, Dong Li, Ning Yang, Jianye Hao, Haifeng Zhang, et al. Aries: Stimulating self-refinement of large language models by iterative preference optimization. *arXiv preprint arXiv:2502.05605*, 2025.

[714] Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Bo Wang, Shimin Li, Yunhua Zhou, Qipeng Guo, Xuanjing Huang, and Xipeng Qiu. Scaling of search and learning: A roadmap to reproduce o1 from reinforcement learning perspective. *arXiv preprint arXiv:2412.14135*, 2024.

[715] Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Yunhua Zhou, and Xipeng Qiu. Revisiting the test-time scaling of o1-like models: Do they truly possess test-time scaling capabilities? *arXiv preprint arXiv:2502.12215*, 2025.

[716] Zhongshen Zeng, Yinhong Liu, Yingjia Wan, Jingyao Li, Pengguang Chen, Jianbo Dai, Yuxuan Yao, Rongwu Xu, Zehan Qi, Wanru Zhao, Linling Shen, Jianqiao Lu, Haochen Tan, Yukang Chen, Hao Zhang, Zhan Shi, Bailin Wang, Zhijiang Guo, and Jiaya Jia. MR-ben: A meta-reasoning benchmark for evaluating system-2 thinking in LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, June 2024. URL https://openreview.net/forum?id=GN2qbxZlni.

[717] Zihao Zeng, Xuyao Huang, Boxiu Li, and Zhijie Deng. Sift: Grounding llm reasoning in contexts via stickers. *arXiv preprint arXiv:2502.14922*, 2025.

[718] Yuexiang Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Shengbang Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, and Sergey Levine. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, September 2024. URL https://openreview.net/forum?id=nBjmMF2IZU.

[719] Zaifu Zhan, Shuang Zhou, Huixue Zhou, Jiawen Deng, Yu Hou, Jeremy Yeung, and Rui Zhang. An evaluation of deepseek models in biomedical natural language processing. *arXiv preprint arXiv:2503.00624*, 2025.

[720] Alexander Zhang, Marcus Dong, Jiaheng Liu, Wei Zhang, Yejie Wang, Jian Yang, Ge Zhang, Tianyu Liu, Zhongyuan Peng, Yingshui Tan, et al. Codecriticbench: A holistic code critique benchmark for large language models. *arXiv preprint arXiv:2502.16614*, 2025.

[721] Beichen Zhang, Yuhong Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Haodong Duan, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Booststep: Boosting mathematical capability of large language models via improved single-step reasoning. *arXiv preprint arXiv:2501.03226*, 2025.

[722] Bohan Zhang, Xiaokang Zhang, Jing Zhang, Jifan Yu, Sijia Luo, and Jie Tang. Cot-based synthesizer: Enhancing llm performance through answer synthesis. *arXiv preprint arXiv:2501.01668*, 2025.

[723] Che Zhang, Zhenyang Xiao, Chengcheng Han, Yixin Lian, and Yuejian Fang. Learning to check: Unleashing potentials for self-correction in large language models. *arXiv preprint arXiv:2402.13035*, 2024.

[724] Chi Zhang, Jiajun Song, Siyu Li, Yitao Liang, Yuxi Ma, Wei Wang, Yixin Zhu, and Song-Chun Zhu. Proposing and solving olympiad geometry with guided tree search. *arXiv preprint arXiv:2412.10673*, 2024.

[725] Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. ReST-MCTS*: LLM self-training via process reward guided tree search. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, September 2024. URL https://openreview.net/forum?id=8rcFOqEud5.

[726] Di Zhang, Xiaoshui Huang, Dongzhan Zhou, Yuqiang Li, and Wanli Ouyang. Accessing gpt-4 level mathematical olympiad solutions via monte carlo tree self-refine with llama-3 8b. *arXiv preprint arXiv:2406.07394*, 2024.

[727] Di Zhang, Jianbo Wu, Jingdi Lei, Tong Che, Jiatong Li, Tong Xie, Xiaoshui Huang, Shufei Zhang, Marco Pavone, Yuqiang Li, et al. Llama-berry: Pairwise optimization for o1-like olympiad-level mathematical reasoning. *arXiv preprint arXiv:2410.02884*, 2024.

[728] Fengji Zhang, Linquan Wu, Huiyu Bai, Guancheng Lin, Xiao Li, Xiao Yu, Yue Wang, Bei Chen, and Jacky Keung. Humaneval-v: Evaluating visual understanding and reasoning abilities of large multimodal models through coding tasks. *arXiv preprint arXiv:2410.12381*, 2024.

[729] Hanning Zhang, Pengcheng Wang, Shizhe Diao, Yong Lin, Rui Pan, Hanze Dong, Dylan Zhang, Pavlo Molchanov, and Tong Zhang. Entropy-regularized process reward model. *arXiv preprint arXiv:2412.11006*, 2024.

[730] Hongbo Zhang, Han Cui, Guangsheng Bao, Linyi Yang, Jun Wang, and Yue Zhang. Direct value optimization: Improving chain-of-thought reasoning in llms with refined values. *arXiv preprint arXiv:2502.13723*, 2025.

[731] Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xionghui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, et al. Aflow: Automating agentic workflow generation. *arXiv preprint arXiv:2410.10762*, 2024.

[732] Jintian Zhang, Yuqi Zhu, Mengshu Sun, Yujie Luo, Shuofei Qiao, Lun Du, Da Zheng, Huajun Chen, and Ningyu Zhang. Lightthinker: Thinking step-by-step compression. *arXiv preprint arXiv:2502.15589*, 2025.

[733] Kechi Zhang, Ge Li, Jia Li, Yihong Dong, and Zhi Jin. Focused-dpo: Enhancing code generation through focused preference optimization on error-prone points. *arXiv preprint arXiv:2502.11475*, 2025.

[734] Kexun Zhang, Shang Zhou, Danqing Wang, William Yang Wang, and Lei Li. Scaling llm inference with optimized sample compute allocation. *arXiv preprint arXiv:2410.22480*, 2024.

[735] Kongcheng Zhang, Qi Yao, Baisheng Lai, Jiaxing Huang, Wenkai Fang, Dacheng Tao, Mingli Song, and Shunyu Liu. Reasoning with reinforced functional token tuning. *arXiv preprint arXiv:2502.13389*, 2025.

[736] Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. *arXiv preprint arXiv:2408.15240*, 2024.

[737] Ming-Liang Zhang, Fei yin, and Cheng-Lin Liu. A multi-modal neural geometric solver with textual clauses parsed from diagram. In Edith Elkind, editor, *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 3374–3382. International Joint Conferences on Artificial Intelligence Organization, 8 2023. doi: 10.24963/ijcai.2023/376. URL https://doi.org/10.24963/ijcai.2023/376. Main Track.

[738] Qingjie Zhang, Han Qiu, Di Wang, Haoting Qian, Yiming Li, Tianwei Zhang, and Min-lie Huang. Understanding the dark side of llms' intrinsic self-correction. *arXiv preprint arXiv:2412.14959*, 2024.

[739] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer, October 2024.

[740] Shaowei Zhang and Deyi Xiong. BackMATH: Towards backward reasoning for solving math problems step by step. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, Steven Schockaert, Kareem Darwish, and Apoorv Agarwal, editors, *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 466–482, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-industry.40/.

[741] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023.

[742] Shimao Zhang, Xiao Liu, Xin Zhang, Junxiao Liu, Zheheng Luo, Shujian Huang, and Yeyun Gong. Process-based self-rewarding language models. *arXiv preprint arXiv:2503.03746*, 2025.

[743] Wenjing Zhang, Xuejiao Lei, Zhaoxiang Liu, Ning Wang, Zhenhong Long, Peijun Yang, Jiaojiao Zhao, Minjie Hua, Chaoyang Ma, Kai Wang, et al. Safety evaluation of deepseek models in chinese contexts. *arXiv preprint arXiv:2502.11137*, 2025.

[744] Wenqi Zhang, Yongliang Shen, Linjuan Wu, Qiuying Peng, Jun Wang, Yueting Zhuang, and Weiming Lu. Self-contrast: Better reflection through inconsistent solving perspectives. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3602–3622, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10. 18653/v1/2024.acl-long.197. URL https://aclanthology.org/2024.acl-long.197/.

[745] Xinyu Zhang, Yuxuan Dong, Yanrui Wu, Jiaxing Huang, Chengyou Jia, Basura Fernando, Mike Zheng Shou, Lingling Zhang, and Jun Liu. Physreason: A comprehensive benchmark towards physics-based reasoning. *arXiv preprint arXiv:2502.12054*, 2025.

[746] Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. Chain of preference optimization: Improving chain-of-thought reasoning in llms. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 333–356. Curran Associates, Inc., September 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/00d80722b756de0166523a87805dd00f-Paper-Conference.pdf.

[747] Yifan Zhang, Wenyu Du, Dongming Jin, Jie Fu, and Zhi Jin. Finite state automata inside transformers with chain-of-thought: A mechanistic study on state tracking. *arXiv preprint arXiv:2502.20129*, 2025.

[748] Yong Zhang, Bingyuan Zhang, Zhitao Li, Ming Li, Ning Cheng, Minchuan Chen, Tao Wei, Jun Ma, Shaojun Wang, and Jing Xiao. Self-enhanced reasoning training: Activating latent reasoning in small models for enhanced reasoning distillation. *arXiv preprint arXiv:2502.12744*, 2025.

[749] Yongheng Zhang, Qiguang Chen, Min Li, Wanxiang Che, and Libo Qin. AutoCAP: Towards automatic cross-lingual alignment planning for zero-shot chain-of-thought. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9191–9200, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.546. URL https://aclanthology.org/2024.findings-acl.546/.

[750] Yongheng Zhang, Qiguang Chen, Jingxuan Zhou, Peng Wang, Jiasheng Si, Jin Wang, Wenpeng Lu, and Libo Qin. Wrong-of-thought: An integrated reasoning framework with multi-perspective verification and wrong information. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6644–6653, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.388. URL https://aclanthology.org/2024.findings-emnlp.388/.

[751] Yudi Zhang, Lu Wang, Meng Fang, Yali Du, Chenghua Huang, Jun Wang, Qingwei Lin, Mykola Pechenizkiy, Dongmei Zhang, Saravan Rajmohan, et al. Distill not only data but also rewards: Can smaller language models surpass larger ones? *arXiv preprint arXiv:2502.19557*, 2025.

[752] Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Jaekyeom Kim, Moontae Lee, Honglak Lee, and Lu Wang. Small language models need strong verifiers to self-correct reasoning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15637–15653, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.924. URL https://aclanthology.org/2024.findings-acl.924/.

[753] Yuxiang Zhang, Shangxi Wu, Yuqi Yang, Jiangming Shu, Jinlin Xiao, Chao Kong, and Jitao Sang. o1-coder: an o1 replication for coding. *arXiv preprint arXiv:2412.00154*, 2024.

[754] Yuxiang Zhang, Yuqi Yang, Jiangming Shu, Yuhang Wang, Jinlin Xiao, and Jitao Sang. Openrft: Adapting reasoning foundation model for domain-specific tasks with reinforcement fine-tuning. *arXiv preprint arXiv:2412.16849*, 2024.

[755] Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning. *arXiv preprint arXiv:2501.07301*, 2025.

[756] Zhongwang Zhang, Pengxiao Lin, Zhiwei Wang, Yaoyu Zhang, and Zhi-Qin John Xu. Complexity control facilitates reasoning-based compositional generalization in transformers. *arXiv preprint arXiv:2501.08537*, 2025.

[757] Zhuosheng Zhang, Aston Zhang, Mu Li, hai zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, June 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=y1pPWFVfvR.

[758] Eric Zhao, Pranjal Awasthi, and Sreenivas Gollapudi. Sample, scrutinize and scale: Effective inference-time search by scaling verification. *arXiv preprint arXiv:2502.01839*, 2025.

[759] Jun Zhao, Jingqi Tong, Yurong Mou, Ming Zhang, Qi Zhang, and Xuanjing Huang. Exploring the compositional deficiency of large language models in mathematical reasoning. *arXiv preprint arXiv:2405.06680*, 2024.

[760] Lili Zhao, Yang Wang, Qi Liu, Mengyun Wang, Wei Chen, Zhichao Sheng, and Shijin Wang. Evaluating large language models through role-guide and self-reflection: A comparative study. In *The Thirteenth International Conference on Learning Representations*, January 2025. URL https://openreview.net/forum?id=E36NHwe7Zc.

[761] Xueliang Zhao, Wei Wu, Jian Guan, and Lingpeng Kong. Promptcot: Synthesizing olympiad-level problems for mathematical reasoning in large language models. *arXiv preprint arXiv:2503.02324*, 2025.

[762] Xufeng Zhao, Mengdi Li, Wenhao Lu, Cornelius Weber, Jae Hee Lee, Kun Chu, and Stefan Wermter. Enhancing zero-shot chain-of-thought reasoning in large language models through logic. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6144–6166, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.543/.

[763] Yachao Zhao, Bo Wang, and Yan Wang. Explicit vs. implicit: Investigating social bias in large language models through self-reflection. *arXiv preprint arXiv:2501.02295*, 2025.

[764] Yichong Zhao and Susumu Goto. Can frontier llms replace annotators in biomedical text mining? analyzing challenges and exploring solutions. *arXiv preprint arXiv:2503.03261*, 2025.

[765] Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. Marco-o1: Towards open reasoning models for open-ended solutions. *arXiv preprint arXiv:2411.14405*, 2024.

[766] Zilong Zhao, Yao Rong, Dongyang Guo, Emek Gözlüklü, Emir Gülboy, and Enkelejda Kasneci. Stepwise self-consistent mathematical reasoning with large language models. *arXiv preprint arXiv:2402.17786*, 2024.

[767] Zirui Zhao, Wee Sun Lee, and David Hsu. Large language models as commonsense knowledge for large-scale task planning. *Advances in Neural Information Processing Systems*, 36:31967–31987, December 2023.

[768] Chuanyang Zheng, Zhengying Liu, Enze Xie, Zhenguo Li, and Yu Li. Progressive-hint prompting improves reasoning in large language models. In *AI for Math Workshop @ ICML 2024*, June 2024. URL https://openreview.net/forum?id=UkFEs3ciz8.

[769] Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Processbench: Identifying process errors in mathematical reasoning. *arXiv preprint arXiv:2412.06559*, 2024.

[770] Jiani Zheng, Lu Wang, Fangkai Yang, Chaoyun Zhang, Lingrui Mei, Wenjie Yin, Qingwei Lin, Dongmei Zhang, Saravan Rajmohan, and Qi Zhang. Vem: Environment-free exploration for training gui agent with value environment model. *arXiv preprint arXiv:2502.18906*, 2025.

[771] Kunhao Zheng, Juliette Decugis, Jonas Gehring, Taco Cohen, benjamin negrevergne, and Gabriel Synnaeve. What makes large language models reason in (multi-turn) code generation? In *The Thirteenth International Conference on Learning Representations*, January 2025. URL https://openreview.net/forum?id=Zk9guOl9NS.

[772] Xin Zheng, Jie Lou, Boxi Cao, Xueru Wen, Yuqiu Ji, Hongyu Lin, Yaojie Lu, Xianpei Han, Debing Zhang, and Le Sun. Critic-cot: Boosting the reasoning abilities of large language model via chain-of-thoughts critic. *arXiv preprint arXiv:2408.16326*, 2024.

[773] Zhi Zheng, Zhuoliang Xie, Zhenkun Wang, and Bryan Hooi. Monte carlo tree search for comprehensive exploration in llm-based automatic heuristic design. *arXiv preprint arXiv:2501.08603*, 2025.

[774] Jianyuan Zhong, Zeju Li, Zhijian Xu, Xiangyu Wen, and Qiang Xu. Dyve: Thinking fast and slow for dynamic process verification. *arXiv preprint arXiv:2502.11157*, 2025.

[775] Qihuang Zhong, Kang Wang, Ziyang Xu, Juhua Liu, Liang Ding, and Bo Du. Achieving> 97% on gsm8k: Deeply understanding the problems makes llms better solvers for math word problems. *arXiv preprint arXiv:2404.14963*, 2024.

[776] Tianyang Zhong, Zhengliang Liu, Yi Pan, Yutong Zhang, Yifan Zhou, Shizhe Liang, Zihao Wu, Yanjun Lyu, Peng Shu, Xiaowei Yu, et al. Evaluation of openai o1: Opportunities and challenges of agi. *arXiv preprint arXiv:2409.18486*, 2024.

[777] Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning, acting, and planning in language models. In *Forty-first International Conference on Machine Learning*, May 2024. URL https://openreview.net/forum?id=njwv9BsGHF.

[778] Aojun Zhou, Ke Wang, Zimu Lu, Weikang Shi, Sichun Luo, Zipeng Qin, Shaoqing Lu, Anya Jia, Linqi Song, Mingjie Zhan, and Hongsheng Li. Solving challenging math word problems using GPT-4 code interpreter with code-based self-verification. In *The Twelfth International Conference on Learning Representations*, January 2024. URL https://openreview.net/forum?id=c8McWs4Av0.

[779] Changzhi Zhou, Xinyu Zhang, Dandan Song, Xiancai Chen, Wanli Gu, Huipeng Ma, Yuhang Tian, Mengdi Zhang, and Linmei Hu. Refinecoder: Iterative improving of large language models via adaptive critique refinement for code generation. *arXiv preprint arXiv:2502.09183*, 2025.

[780] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*, February 2023. URL https://openreview.net/forum?id=WZH7099tgfM.

[781] Fan Zhou, Haoyu Dong, Qian Liu, Zhoujun Cheng, Shi Han, and Dongmei Zhang. Reflection of thought: Inversely eliciting numerical reasoning in language models via solving linear systems. *arXiv preprint arXiv:2210.05075*, 2022.

[782] 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.

[783] Jin Peng Zhou, Charles E Staats, Wenda Li, Christian Szegedy, Kilian Q Weinberger, and Yuhuai Wu. Don't trust: Verify – grounding LLM quantitative reasoning with autoformalization. In *The Twelfth International Conference on Learning Representations*, January 2024. URL https://openreview.net/forum?id=V5tdi14ple.

[784] Jin Peng Zhou, Kaiwen Wang, Jonathan Chang, Zhaolin Gao, Nathan Kallus, Kilian Q Weinberger, Kianté Brantley, and Wen Sun. $q\sharp$: Provably optimal distributional rl for llm post-training. *arXiv preprint arXiv:2502.20548*, 2025.

[785] Kaiwen Zhou, Chengzhi Liu, Xuandong Zhao, Shreedhar Jangam, Jayanth Srinivasa, Gaowen Liu, Dawn Song, and Xin Eric Wang. The hidden risks of large reasoning models: A safety assessment of r1. *arXiv preprint arXiv:2502.12659*, 2025.

[786] Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and José Hernández-Orallo. Larger and more instructable language models become less reliable. *Nature*, 634(8032):61–68, 2024.

[787] Li Zhou, Ruijie Zhang, Xunlian Dai, Daniel Hershcovich, and Haizhou Li. Large language models penetration in scholarly writing and peer review. *arXiv preprint arXiv:2502.11193*, 2025.

[788] Ruochen Zhou, Minrui Xu, Shiqi Chen, Junteng Liu, Yunqi Li, LIN Xinxin, Zhengyu Chen, and Junxian He. AI for math or math for AI? on the generalization of learning mathematical problem solving. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*, 2024. URL https://openreview.net/forum?id=xlnvZ85CSo.

[789] Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations*, January 2024. URL https://openreview.net/forum?id=oKn9c6ytLx.

[790] Xin Zhou, Yiwen Guo, Ruotian Ma, Tao Gui, Qi Zhang, and Xuanjing Huang. Self-consistency of the internal reward models improves self-rewarding language models. *arXiv preprint arXiv:2502.08922*, 2025.

[791] Yang Zhou, Hongyi Liu, Zhuoming Chen, Yuandong Tian, and Beidi Chen. Gsm-infinite: How do your llms behave over infinitely increasing context length and reasoning complexity? *arXiv preprint arXiv:2502.05252*, 2025.

[792] Dawei Zhu, Xiyu Wei, Guangxiang Zhao, Wenhao Wu, Haosheng Zou, Junfeng Ran, Xun Wang, Lin Sun, Xiangzheng Zhang, and Sujian Li. Chain-of-thought matters: Improving long-context language models with reasoning path supervision. *arXiv preprint arXiv:2502.20790*, 2025.

[793] Junda Zhu, Lingyong Yan, Shuaiqiang Wang, Dawei Yin, and Lei Sha. Reasoning-to-defend: Safety-aware reasoning can defend large language models from jailbreaking. *arXiv preprint arXiv:2502.12970*, 2025.

[794] Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, et al. Multiagentbench: Evaluating the collaboration and competition of llm agents. *arXiv preprint arXiv:2503.01935*, 2025.

[795] Tinghui Zhu, Kai Zhang, Jian Xie, and Yu Su. Deductive beam search: Decoding deducible rationale for chain-of-thought reasoning. In *First Conference on Language Modeling*, July 2024. URL https://openreview.net/forum?id=S1XnUsqwr7.

[796] Xinyu Zhu, Junjie Wang, Lin Zhang, Yuxiang Zhang, Yongfeng Huang, Ruyi Gan, Jiaxing Zhang, and Yujiu Yang. Solving math word problems via cooperative reasoning induced language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4471–4485, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.245. URL https://aclanthology.org/2023.acl-long.245/.

[797] Zihao Zhu, Hongbao Zhang, Mingda Zhang, Ruotong Wang, Guanzong Wu, Ke Xu, and Baoyuan Wu. Bot: Breaking long thought processes of o1-like large language models through backdoor attack. *arXiv preprint arXiv:2502.12202*, 2025.

[798] Ziyu Zhuang, Qiguang Chen, Longxuan Ma, Mingda Li, Yi Han, Yushan Qian, Haopeng Bai, Weinan Zhang, and Liu Ting. Through the lens of core competency: Survey on evaluation of large language models. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum)*, pages 88–109, Harbin, China, August 2023. Chinese Information Processing Society of China. URL https://aclanthology.org/2023.ccl-2.8/.

[799] Alireza S Ziabari, Nona Ghazizadeh, Zhivar Sourati, Farzan Karimi-Malekabadi, Payam Piray, and Morteza Dehghani. Reasoning on a spectrum: Aligning llms to system 1 and system 2 thinking. *arXiv preprint arXiv:2502.12470*, 2025.

[800] Henry Peng Zou, Zhengyao Gu, Yue Zhou, Yankai Chen, Weizhi Zhang, Liancheng Fang, Yibo Wang, Yangning Li, Kay Liu, and Philip S Yu. Testnuc: Enhancing test-time computing approaches through neighboring unlabeled data consistency. *arXiv preprint arXiv:2502.19163*, 2025.

[801] Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025.