From System 1 to System 2: A Survey of Reasoning Large Language Models

Zhong-Zhi Li[*], Duzhen Zhang[*], Ming-Liang Zhang[§], Jiaxin Zhang[§], Zengyan Liu[§], Yuxuan Yao[§], Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhiwei Li, Bao-long Bi, Ling-rui Mei, Jun-Feng Fang, Xiao Liang
Zhijiang Guo[†], Le Song[†], Cheng-Lin Liu[†], *Fellow, IEEE*

Abstract—Achieving human-level intelligence requires refining the transition from the fast, intuitive System 1 to the slower, more deliberate System 2 reasoning. While System 1 excels in quick, heuristic decisions, System 2 relies on logical reasoning for more accurate judgments and reduced biases. Foundational Large Language Models (LLMs) excel at fast decision-making but lack the depth for complex reasoning, as they have not yet fully embraced the step-by-step analysis characteristic of true System 2 thinking. Recently, reasoning LLMs like OpenAI's o1/o3 and DeepSeek's R1 have demonstrated expert-level performance in fields such as mathematics and coding, closely mimicking the deliberate reasoning of System 2 and showcasing human-like cognitive abilities. This survey begins with a brief overview of the progress in foundational LLMs and the early development of System 2 technologies, exploring how their combination has paved the way for reasoning LLMs. Next, we discuss how to construct reasoning LLMs, analyzing their features, the core methods enabling advanced reasoning, and the evolution of various reasoning LLMs. Additionally, we provide an overview of reasoning benchmarks, offering an in-depth comparison of the performance of representative reasoning LLMs. Finally, we explore promising directions for advancing reasoning LLMs and maintain a real-time GitHub Repository to track the latest developments. We hope this survey will serve as a valuable resource to inspire innovation and drive progress in this rapidly evolving field.

Index Terms—Slow-thinking, Large Language Models, Human-like Reasoning, Decision Making in AI, AGI

1 Introduction

> *"Don't teach. Incentivize."*
>
> —*Hyung Won Chung, OpenAI*

Achieving human-level intelligence requires refining the transition from *System 1* to *System 2* reasoning [1]–[5]. Dual-system theory suggests that human cognition operates through two modes: *System 1*, which is fast, automatic, and intuitive, enabling quick decisions with minimal effort, and *System 2*, which is slower, more analytical, and deliberate [6], [7]. While *System 1* is efficient for routine tasks, it is prone to cognitive biases, especially in complex or uncertain situations, leading to judgment errors. In contrast, *System 2* relies on logical reasoning and systematic thinking, resulting in more accurate and rational decisions [8]–[11]. By mitigating the biases of *System 1*, *System 2* provides a more refined approach to problem-solving [12]–[15].

The development of foundational Large Language Models (LLMs)[1] has marked a major milestone in Artificial Intelligence (AI). Models such as GPT-4o [16] and DeepSeek-v3 [17] have demonstrated impressive capabilities in text generation, language translation, and a variety of perception tasks [18]–[28]. These models, trained on extensive datasets and utilizing advanced algorithms, excel in understanding and generating human-like responses. However, despite their impressive achievements, foundational LLMs operate in a manner similar to *System 1* reasoning, relying on fast, heuristic-driven decision-making. While they perform ex-

Version: v1 (major update on February 24, 2025)
[*] *Core contribution.* [§] *Significant contribution.* [†] *Corresponding author.*

Duzhen Zhang, Jiahua Dong, and Le Song are with the Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE (E-mail: bladedancer957@gmail.com; dongjiahua1995@gmail.com; le.song@mbzuai.ac.ae).
Zhong-Zhi Li, Pei-Jie Wang, Bao-Long Bi, Ling-Rui Mei, Xiuyi Chen, Fei Yin, and Cheng-Lin Liu are with the Institute of Automation, Chinese Academy of Sciences, Beijing, China (E-mail: lizhongzhi2022@ia.ac.cn; wangpeijie2023@ia.ac.cn, bibaolong23z@ict.ac.cn, meilingrui25b@ict.ac.cn hugheren.chan@gmail.com; fyin@nlpr.ia.ac.cn; liucl@nlpr.ia.ac.cn).
Ming-Liang Zhang is with the Alibaba Group, Beijing, China (E-mail: luxing.zml@taobao.com).
Bao-Long Bi, Ling-Rui Mei are with the Chinese Academy of Sciences, Beijing, China (E-mail: bibaolong23z@ict.ac.cn, meilingrui25b@ict.ac.cn).
Xiao Liang is with the University of California,Los Angles, USA (E-mail: luxing.zml@taobao.com).
Zengyan Liu, Yuxuan Yao, Zhiwei Li, and Zhijiang Guo are with City University of Hong Kong and the Hong Kong University of Science and Technology (Guangzhou), China (E-mail: zengyaliu2-c@my.cityu.edu.hk; yuxuanyao3-c@my.cityu.edu.hk; zhiweil466@gmail.com; zhijiangguo@hkust-gz.edu.cn).
Jiaxin Zhang is with the University of Strathclyde, Glasgow, UK (E-mail: jiaxin.zhang@strath.ac.uk).
Haotian Xu is with the Xiaohongshu Inc, Beijing, China (E-mail: xuhaotian@xiaohongshu.com).
Yingying Zhang is with the East China Normal University, Shanghai, China (E-mail: yyzhang@fem.ecnu.edu.cn).
Junfeng Fang is with the Nanyang Technological University, Singapore (E-mail: junhaozheng47@outlook.com).
Junhao Zheng is with the South China University of Technology, Guangzhou, China (E-mail: junhaozheng47@outlook.com).

[1] In this paper, "reasoning" refers to answering questions involving complex, multi-step processes with intermediate steps. **Foundational LLMs:** LLMs with basic reasoning abilities, handling simple or single-step tasks. **Reasoning LLMs:** LLMs that excel in complex tasks like coding and mathematical proofs, incorporating a "thinking" process–tasks that foundational LLMs struggle with.

ceptionally well in providing rapid responses, they often fall short in scenarios requiring deep, logical analysis and precision in complex reasoning tasks. This limitation becomes especially clear in situations involving intricate problem-solving, logical analysis, or nuanced understanding, where these models do not yet match human cognitive abilities.

In contrast, reasoning LLMs represent a significant advancement in the evolution of language models. Models like OpenAI's o1/o3 [29], [30] and DeepSeek's R1 [31] are designed to emulate the slower, more deliberate reasoning associated with *System 2* thinking. Unlike foundational LLMs, reasoning LLMs are equipped with mechanisms for processing information step-by-step, allowing them to make more accurate and rational decisions. This shift from fast-thinking, intuitive processes to more methodical, reasoning-driven models enables reasoning LLMs to tackle complex tasks, such as advanced mathematics [32]–[37], logical reasoning [38]–[44], and multimodal reasoning [45]–[47], with expert-level performance, exhibiting human-like cognitive abilities. As a result, reasoning LLMs are increasingly seen as capable of achieving the deep, logical thinking needed for tasks that were once considered beyond AI's reach. The recent timeline of reasoning LLMs is presented in Figure 1.

1.1 Structure of the Survey

This survey offers a comprehensive overview of the key concepts, methods, and challenges involved in the development of reasoning LLMs. As illustrated in Figure 2, this survey is organized as follows:

1) Section 2 offers a concise overview of the progress in foundational LLMs (Section 2.1) and the early development of key *System 2* technologies, including symbolic logic systems (Section 2.2), Monte Carlo Tree Search (MCTS) (Section 2.3), and Reinforcement Learning (RL) (Section 2.4), highlighting how their combination has paved the way for reasoning LLMs.
2) Section 3 introduces reasoning LLMs and outlines their construction process. Specifically, Section 3.1 presents the characteristics of reasoning LLMs from two perspectives: output behavior (Section 3.1.1) and training dynamics (Section 3.1.2), emphasizing their differences from foundational LLMs. Section 3.2 identifies the core methods necessary for achieving advanced reasoning capabilities, focusing on five aspects: Structure Search (Section 3.2.1), Reward Modeling (Section 3.2.2), Self Improvement (Section 3.2.3), Macro Action (Section 3.2.4), and Reinforcement Fine-Tuning (Section 3.2.5). Each section delves into the specific characteristics of these methods and introduces representative reasoning LLMs for each approach. Section 3.3 traces the evolutionary stages of reasoning LLMs.
3) Section 4 evaluates representative reasoning LLMs. Specifically, Section 4.1 reviews current mainstream reasoning benchmarks, covering both plain text and multimodal benchmarks across various task types. Section 4.2 outlines the current evaluation metrics, while Section 4.3 analyzes and compares the performance of mainstream reasoning LLMs with their foundational counterparts based on these benchmarks.
4) Section 5 summarizes several recent technical areas related to Reasoning LLMs. Section 5.1 summarizes technologies related to large-scale RL training. Section 5.2 discusses several safety issues related to LRM. Section 5.3 summarizes the field of integrating Reasoning LLMs with Agents. Section 5.4 summarizes the technologies of Adaptive Reasoning LLMs and Efficient Reasoning LLMs.
5) Section 6 highlights the limitations of existing reasoning LLMs and outlines several promising future development directions for these models.
6) Finally, we conclude the paper in Section 7 and provide a real-time tracking GitHub Repository to monitor the latest developments in the field.

We hope this survey serves as a valuable resource, fostering innovation and progress in this rapidly evolving domain.

1.2 Contribution of the Survey

Recently, several analyses and replications of specific technical approaches have been conducted [48]–[55], yet there remains a lack of systematic analysis and organization. Research [56] has focused only on slow-thinking methods during testing. Meanwhile, studies [57]–[59] have primarily concentrated on training or achieving reasoning LLMs, often from the perspective of RL.

Our survey distinguishes itself from and contributes to the existing literature in the following ways:

1) Rather than focusing on a single technical approach, we offer a comprehensive overview of the key concepts, methods, and challenges involved in reasoning LLMs.
2) We summarize the key advancements of early *System 2* and how they have paved the way for reasoning LLMs, specifically in combination with foundational LLMs–a crucial aspect often overlooked in previous works.
3) We present a more thorough and inclusive summary of the core methods necessary for constructing reasoning LLMs, including but not limited to RL.

2 Foundations of Reasoning LLMs

In this section, we provide a concise overview of the progress in foundational LLMs and the early development of key *System 2* technologies, highlighting critical advancements that, when combined with foundational LLMs, have paved the way for reasoning LLMs. These advancements include symbolic logic systems, MCTS, and RL.

2.1 Foundational LLMs

The development of foundational LLMs saw significant advancements with the introduction of pretrained Transformers [18] in 2018-2019, notably through BERT [19] and GPT [21]. These models leveraged unsupervised pretraining on vast text corpora, followed by fine-tuning for task-specific applications. This approach enabled them to develop a broad language understanding before specializing in tasks such as sentiment analysis, entity recognition, and question answering. BERT's bidirectional context processing improved word understanding, while GPT excelled in text generation with its unidirectional design.

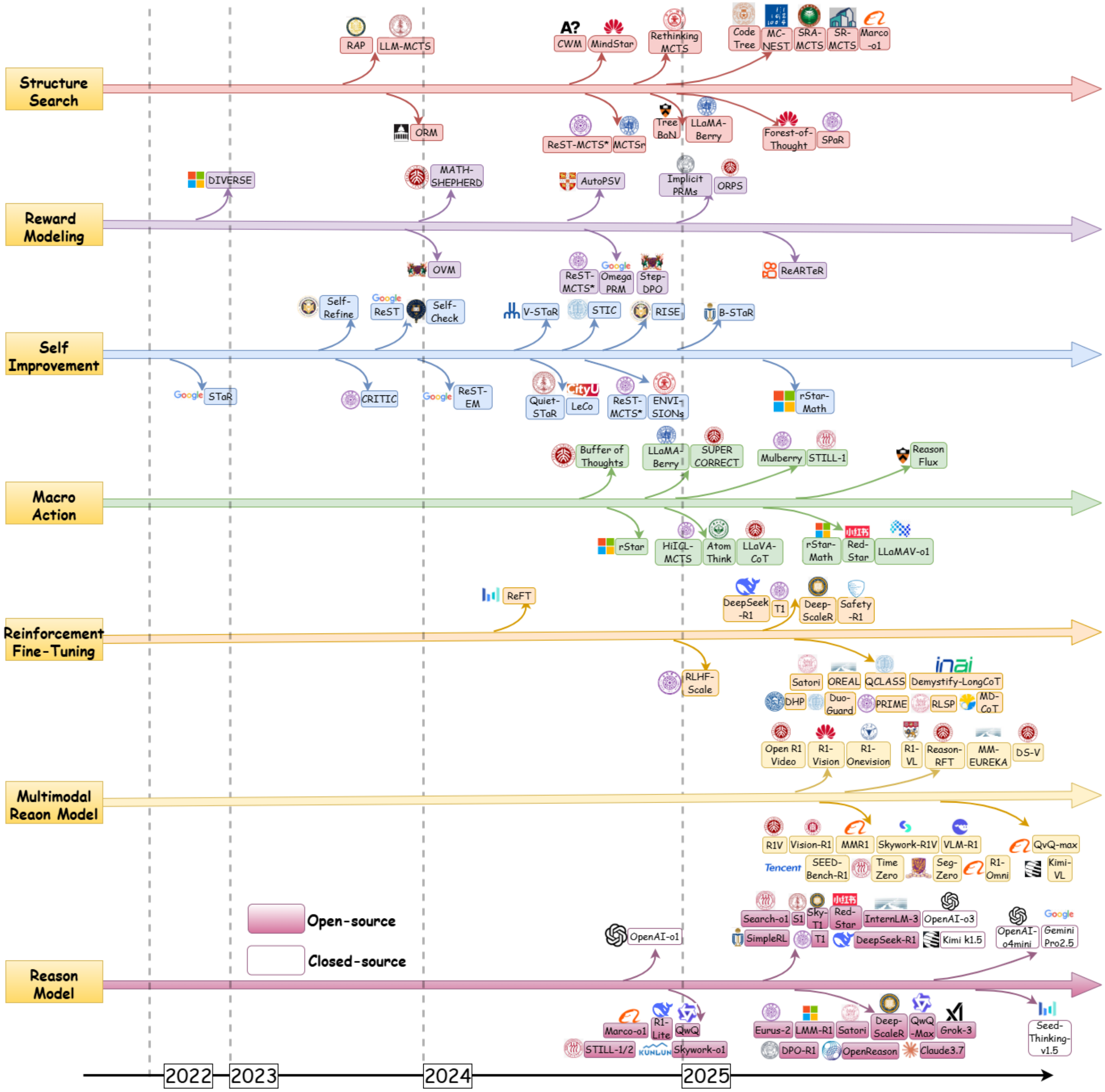

Fig. 1. The recent timeline of reasoning LLMs, covering core methods and the release of open-source and closed-source reproduction projects.

The release of GPT-2 [22] in 2019, with 1.5 billion parameters, marked a significant leap in generative performance, though it also raised ethical concerns. GPT-3 [23], with 175 billion parameters, further demonstrated the power of unsupervised pretraining, excelling in few-shot learning and performing well across a wide range of NLP tasks. In subsequent years, multimodal models like CLIP [60] and DALL-E [61] emerged, integrating text and visual inputs. These models enabled new tasks, such as generating images from text, and enhanced human-computer interaction.

By 2023-2024, models such as GPT-4/4o [16], [62], LLaMA [25], and LLaVA [27] demonstrated advanced capabilities in reasoning, contextual understanding, and multimodal reasoning, processing both text and images [63]–[65]. DeepSeek-V3 [17], featuring a 671B Mixture-of-Expert architecture [66]–[68], outperforms several other LLMs on key benchmarks while offering significant improvements in efficiency and processing speed. The evolution of foundational LLMs has revolutionized AI, enabling more sophisticated applications in language comprehension, problem-solving, and human-machine collaboration.

Summary: The development of foundational LLMs has progressed from pretrained transformers like BERT to multimodal models such as GPT-4, enhancing language understanding, text generation, and image processing. This advancement has led to significant breakthroughs in AI, improving language comprehension, problem-solving, and human-computer interaction. Building on deep learning advancements [18], [69]–[83], foundational LLMs can learn extensive world knowledge and semantic relationships from

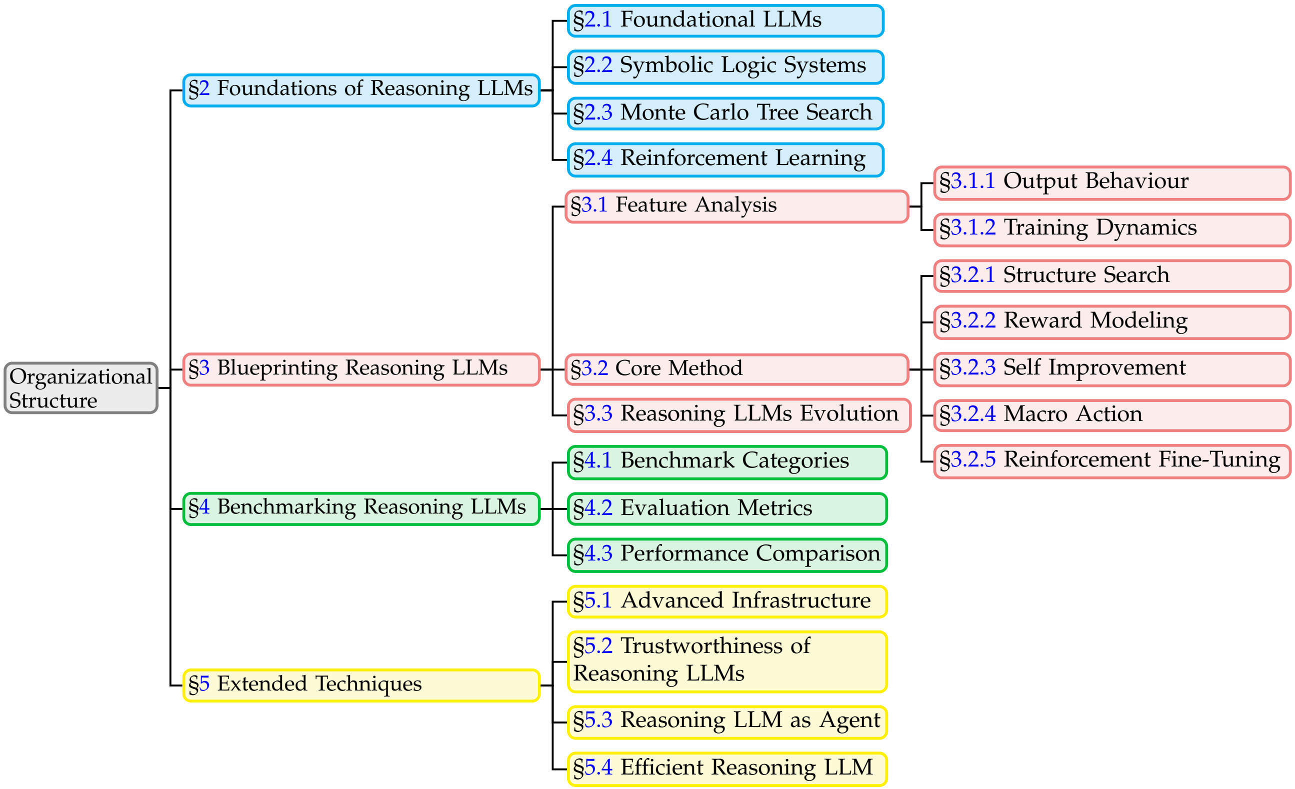
Organizational Structure
§2 Foundations of Reasoning LLMs
§2.1 Foundational LLMs
§2.2 Symbolic Logic Systems
§2.3 Monte Carlo Tree Search
§2.4 Reinforcement Learning
§3 Blueprinting Reasoning LLMs
§3.1 Feature Analysis
§3.1.1 Output Behaviour
§3.1.2 Training Dynamics
§3.2 Core Method
§3.2.1 Structure Search
§3.2.2 Reward Modeling
§3.2.3 Self Improvement
§3.2.4 Macro Action
§3.2.5 Reinforcement Fine-Tuning
§3.3 Reasoning LLMs Evolution
§4 Benchmarking Reasoning LLMs
§4.1 Benchmark Categories
§4.2 Evaluation Metrics
§4.3 Performance Comparison
§5 Extended Techniques
§5.1 Advanced Infrastructure
§5.2 Trustworthiness of Reasoning LLMs
§5.3 Reasoning LLM as Agent
§5.4 Efficient Reasoning LLM

Fig. 2. The primary organizational structure of the survey.

vast textual or multimodal data. This enables them to exhibit emergent capabilities such as In-Context Learning (ICL) [84], [85], prompt engineering [86], [87], and Chain-of-Thought (CoT) reasoning [2], significantly enhancing their adaptability and creative problem-solving abilities.

Despite this progress, foundational LLMs operate similarly to *System 1* reasoning, relying on fast, heuristic-driven decision-making and lacking the step-by-step analysis characteristic of *System 2*. However, their developments lay a solid foundation for future reasoning LLMs–especially when integrated with the following early *System 2* technologies. This combination paves the way for more versatile, flexible, and human-like reasoning models.

2.2 Symbolic Logic Systems

Symbolic logic systems mark the earliest phase of AI, utilizing rules and logical principles to represent knowledge and draw conclusions [88], [89]. They are particularly effective in structured domains, where formal logic ensures precision.

Prolog, a logic programming language based on first-order logic, allows users to define facts, rules, and reason through queries. It has been pivotal in symbolic reasoning systems, especially in NLP and expert systems [90]–[92]. Logic-based systems like Prolog employ propositional and predicate logic for formal reasoning [93], [94]. From the 1960s to the early 1980s, this approach dominated AI, with systems like IBM's LISP [95] for symbolic computation and Resolution Theorem Provers [96] for automated reasoning. In the 1970s, Marvin Minsky introduced Frames, which organized knowledge into structured frameworks, influencing both expert systems and cognitive science [97].

Summary: Symbolic logic systems were pivotal milestones in early AI development. Based on formal logic, they excelled in well-defined problems, particularly in structured environments. However, they also exposed the limitations of rigid, rule-based systems. Despite these constraints, symbolic logic remains foundational to the progress of AI.

Recent advancements in reasoning LLMs have greatly enhanced the emulation of human-like *System 2* cognitive processes through sophisticated thought architectures, known as Macro Action frameworks (Section 3.2.4). By combining symbolic templates or rules with foundational LLMs, macro actions have significantly improved their reasoning capabilities. Integrating macro actions into foundational LLMs has transformed their ability to handle complex reasoning tasks, as hierarchical planning allows models to make high-level decisions before delving into specific problem details, mirroring symbolic logic's structured approach.

2.3 Monte Carlo Tree Search

MCTS is a simulation-based search algorithm for decision-making and planning [98]. It constructs a search tree through four steps: *Selection*, which chooses the child node with the highest priority using the UCB1 formula:

$$\text{UCB1} = \frac{w_i}{n_i} + c\sqrt{\frac{\ln N}{n_i}}, \quad (1)$$

where w_i is the total reward of node i, n_i is its visit count, N is the parent node's visit count, and c balances exploration and exploitation. *Expansion* adds new nodes, *Simulation* performs random rollouts to evaluate them, and *Backpropagation* updates node statistics. MCTS has been widely used in tasks

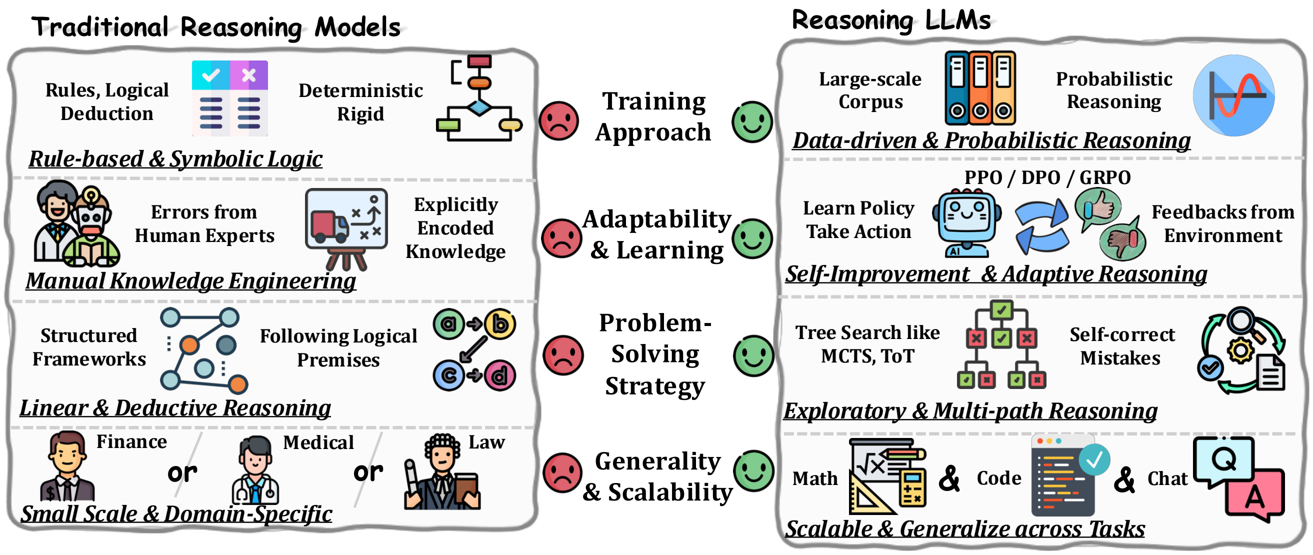
Traditional Reasoning Models
Rules, Logical Deduction
Deterministic Rigid
Rule-based & Symbolic Logic
Errors from Human Experts
Explicitly Encoded Knowledge
Manual Knowledge Engineering
Structured Frameworks
Following Logical Premises
Linear & Deductive Reasoning
Finance
or
Medical
or
Law
Small Scale & Domain-Specific
Training Approach
Adaptability & Learning
Problem-Solving Strategy
Generality & Scalability
Reasoning LLMs
Large-scale Corpus
Probabilistic Reasoning
Data-driven & Probabilistic Reasoning
PPO / DPO / GRPO
Learn Policy Take Action
AI
Feedbacks from Environment
Self-Improvement & Adaptive Reasoning
Tree Search like MCTS, ToT
Self-correct Mistakes
Exploratory & Multi-path Reasoning
Math
&
Code
&
Chat
Q
A
Scalable & Generalize across Tasks

Fig. 3. A comprehensive comparison of traditional reasoning models and reasoning LLMs. Reasoning LLMs offer significant advantages over traditional models in areas such as training approaches, adaptability and learning, problem-solving strategies, and generality and scalability.

such as optimizing strategies in board games like Go [99] and in robotic path planning, where it helps robots navigate dynamic environments effectively [100].

Summary: MCTS has played a crucial role in the development of reasoning LLMs, particularly in Structural Search (Section 3.2.1). By simulating potential future reasoning paths and backpropagating estimated rewards, MCTS helps foundational LLMs efficiently identify the most promising, high-reward paths. This process mirrors human-like planning, where future consequences of decisions are considered before taking action. By dynamically exploring multiple reasoning trajectories, MCTS enables models to avoid getting stuck in suboptimal paths, making it easier to navigate complex decision spaces. This integration has significantly enhanced the ability of LLMs to handle intricate and dynamic reasoning problems, such as those requiring long-term planning or multi-step logical inferences. It has allowed LLMs to make more strategic and informed decisions, improving their overall performance in tasks that involve nuanced reasoning and strategic exploration.

2.4 Reinforcement Learning

RL is a type of machine learning where an agent learns to make decisions by interacting with an environment and receiving feedback in the form of rewards, aiming to maximize cumulative rewards over time [101]. Early breakthroughs in RL, such as Q-learning [102] and DQNs [103], revolutionized the field by enabling the handling of complex state spaces using Deep Neural Networks (DNNs) [104]. These methods paved the way for scaling RL to real-world tasks, where traditional tabular approaches fell short. The advent of deep RL marked a significant step forward, combining the power of deep learning with RL to process high-dimensional inputs, such as images and unstructured data.

A landmark achievement in deep RL was AlphaGo, which demonstrated RL's potential by defeating a world champion in the complex game of Go through self-play [105]. This success highlighted deep RL's ability to thrive in environments with large, continuous action spaces and uncertainty. Building on this, AlphaZero advanced the approach by mastering multiple board games—chess, Go, and Shogi—using self-play, MCTS, and DNNs [106]. AlphaZero's ability to learn entirely from scratch, without prior human knowledge, showcased RL's power in environments requiring long-term strategy and planning.

AlphaStar further expanded the boundaries of deep RL by excelling in the real-time strategy game StarCraft II. Unlike board games, StarCraft II presents dynamic, partially observable environments and demands multi-step, real-time decision-making [107]. AlphaStar's success in this domain demonstrated deep RL's capacity to adapt to complex decision-making scenarios that require both strategic planning and tactical execution. These advancements in RL and deep RL have greatly expanded AI's potential, transitioning from well-defined, static environments to dynamic, complex settings that demand continuous learning and adaptation.

Summary: Deep RL has proven highly effective in solving complex decision-making tasks. AlphaGo exemplifies this by learning strategies through self-play and defeating the world champion in Go. This self-play concept laid the foundation for Self Improvement technology (Section 3.2.3) in reasoning LLMs, both relying on continuous feedback and adjustments to optimize strategies.

In RL, reward shaping has been crucial, especially for multi-step reasoning tasks [108]. By adjusting the reward signal to provide more granular feedback during intermediate steps, it helps agents navigate complex decision-making paths. This concept inspired the development of Reward Modeling (Section 3.2.2), particularly the process reward model, in reasoning LLMs. This model offers step-by-step supervision to identify and correct errors in the reasoning process. By mimicking human reasoning, the process reward model ensures more robust and interpretable results, especially in tasks like mathematical problem-solving and code generation, where step-by-step evaluation is critical.

Moreover, RL itself is a powerful tool for reasoning LLMs (Section 3.2.5). With a reward mechanism, RL guides foundational LLMs to find optimal solutions, especially in dynamic reasoning problems. Its simplicity and efficiency make RL invaluable for training and optimizing reasoning LLMs, enhancing the intelligence and self-evolution of AI models. The integration of RL has led to significant advancements in reasoning LLMs, as demonstrated by DeepSeek-R1 [31], offering more flexible and efficient solutions.

3 Blueprinting Reasoning LLMs

In this section, we first analyze the features of reasoning LLMs from both output behavior and training dynamics perspectives. We then provide a detailed overview of the core methods that enable their advanced reasoning capabilities. Finally, we summarize the evolution of reasoning LLMs. A comprehensive comparison of traditional reasoning models and reasoning LLMs is shown in Figure 3.

3.1 Analysis of the Features of Reasoning LLMs

3.1.1 Output Behaviour Perspective

Explore and Planning Structure: Recent empirical studies have revealed that reasoning LLMs demonstrate a strong tendency for exploratory behavior in their output structures, especially when compared to models such as WizardMath [109] and DeepSeekMath [110], which primarily rely on conventional CoT reasoning approaches. This exploratory behavior is evident in the formulation of novel hypotheses and the pursuit of alternative solution paths. Research by [49] suggests that slow-thinking models engage in a latent generative process, particularly noticeable during the prediction of subsequent tokens. This claim is supported by [31], which observes that similar behaviors naturally arise during RL scale training. Furthermore, the Quiet-STaR framework [111] introduces an auxiliary pre-training phase focused on next-token prediction, highlighting the critical role of internal deliberation and exploratory mechanisms prior to content generation. Collectively, these findings underscore the complex and dynamic nature of reasoning processes in advanced LLMs, emphasizing the interaction between exploration and structured reasoning within their operational frameworks.

Verification and Check Structure: Analysis of OpenAI's o1 [29] and o3 [30] models indicates that their reasoning frameworks incorporate both macro-level actions for long-term strategic planning and micro-level actions, including "*Wait*", "*Hold on*", "*Alternatively*", and "*Let's pause*". These micro actions facilitate meticulous verification and iterative checking processes, ensuring precision in task execution. Such a dual-layered approach underscores the models' capacity to balance overarching goals with granular, detail-oriented operations, thereby enhancing their overall functionality and reliability. To emulate this characteristic, Marco-o1 [112], during the MCTS process for constructing Long-CoT, assigns each tree node the state of "*Wait! Maybe I made some mistakes! I need to rethink from scratch*", thereby facilitating the reflective nature of Long-CoT. Huatuo-o1 [113] employs a multi-agent framework to address the issue of incorrect CoT generation during validation. This is achieved by incorporating a prompt with "*Backtracking*" and "*Correction*" functionalities, which enables the correction process.

Longer Inference Length & Time: Recent research [49]–[52], [114], [115] indicates that reasoning LLMs often generate outputs exceeding 2000 tokens to tackle complex problems in coding and mathematics. However, this extended output length can sometimes lead to overthinking, where the model spends excessive time on a problem without necessarily improving the solution. Studies [49] highlight that while autoregressive generation and Classic CoT can effectively solve simpler problems, they struggle with more complex tasks. Research [116], [117] shows that in multimodal domains, many problems demand careful observation, comparison, and deliberation. Additionally, Search-o1 [118] suggests that slow-thinking mechanisms are particularly beneficial in areas requiring external knowledge or where potential knowledge conflicts arise. In medical scenarios [119], complex problems, such as those requiring test-time scaling techniques, demonstrate significant improvements [52].

Overly Cautious & Simple Problem Trap: Currently, reasoning LLMs have demonstrated strong performance in domains such as competitive-level mathematics [31], [54], [120], [121], complex coding [122], medical question answering [52], [113], and multilingual translation [112], [123]. These scenarios require the model to perform fine-grained analysis of the problem and execute careful logical reasoning based on the given conditions. Interestingly, even for straightforward problems like "*2+3=?*", reasoning LLMs can exhibit overconfidence or uncertainty. Recent research [124] notes that o1-like models tend to generate multiple solution rounds for easier math problems, often exploring unnecessary paths. This behavior contrasts with the lack of diverse exploratory actions for simpler questions, indicating a potential inefficiency in the model's reasoning process.

3.1.2 Training Dynamic Perspective

Amazing Data Efficiency: Unlike traditional approaches that focus on expanding instruction sets with uniformly distributed difficulty levels, Studies [52], [54] suggest that constructing Slow-thinking CoT datasets with a focus on hard samples leads to better generalization in fields like medicine and mathematics. This approach diverges from the conventional practice of collecting diverse and evenly distributed instruction datasets.

Sparse Training Method: Contrary to conventional wisdom, the development of effective reasoning LLMs does not require extensive datasets or dense reward signals. For example, STILL2 [51] demonstrated impressive performance using only 5,000 distilled samples, while Sky-T1 [121] achieved performance parity with QwQ [120] using just 17,000 Long-CoT samples. Similarly, RedStar [54] achieved exceptional results across both textual and multimodal tasks with only 4,000 core LongCoT samples. In comparison to simple CoT, Slow-thinking Supervised Fine-Tuning (SFT) data exhibits remarkable sample efficiency, often delivering comparable results with just 1/100th of the sample size. Additionally, research [125] emphasizes the significant training potential of online RL scaling algorithms, suggesting that non-dense RL supervision and even rule-based reward structures are sufficient for achieving high performance.

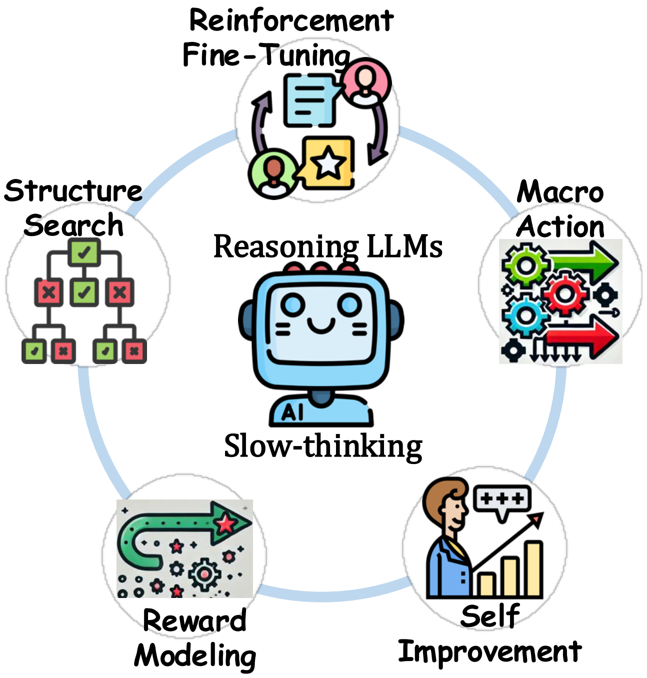

Fig. 4. The core methods enabling reasoning LLMs.

Parameter Characteristic: Training LLMs for slow-thinking, as characterized by the LongCoT approach, results in relatively uniform gradient norms across different layers. In contrast, fast-thinking, exemplified by the simplified CoT method, generates larger gradient magnitudes in the earlier layers, along with significant variability in gradient norms across layers. Empirical evidence suggests that larger models, particularly those exceeding 30 billion parameters, are more compatible with reasoning LLMs training due to their enhanced capacity for complex reasoning. Additionally, experiments conducted by RedStar [54] show that the benefits of data scaling vary across model sizes, with scaling effects being more pronounced and effective in larger models. This finding is supported by Deepseek-R1's research [31], which demonstrates that a 670-billion-parameter model achieves performance metrics closely approximating those of the o1 benchmark, highlighting the scalability advantages of larger architectures in advanced reasoning tasks.

3.2 Core Method

In this section, we provide an overview of the core methods that drive the advanced reasoning capabilities of reasoning LLMs, as shown in Figure 4. These include Structure Search, Reward Modeling, Self Improvement, Macro Action, and Reinforcement Fine-Tuning. We also highlight representative reasoning LLMs for each method.

3.2.1 Structure Search

Reasoning LLMs aim to achieve high accuracy and depth in solving complex problems by emulating the deliberate nature of human reasoning. However, despite recent advancements, current foundational LLMs face inherent limitations when addressing intricate reasoning tasks. These limitations arise from their lack of an internal world model to simulate environmental states, their inability to predict the long-term outcomes of reasoning paths, and their failure to iteratively refine reasoning steps based on future states or rewards [8]. As a result, these shortcomings hinder foundational LLMs from effectively balancing exploration and exploitation in vast reasoning spaces, creating challenges in tasks that require multi-step reasoning, such as complex mathematics, logical inference, or strategic decision-making [149].

MCTS, a powerful search and optimization algorithm, effectively addresses these challenges by providing a structured framework to explore and evaluate reasoning paths systematically. It operates by constructing a reasoning tree, where each node represents a reasoning state, and actions expand the tree by considering potential next steps. Through the simulation of future states and the iterative backpropagation of estimated rewards, MCTS allows foundational LLMs to efficiently identify high-reward reasoning paths, mirroring human planning processes. This approach aligns with the core principles of reasoning LLMs, where thorough analysis and deliberate exploration are essential for generating well-reasoned outputs. Recent methods, such as RAP [14], enhance foundational LLMs by integrating MCTS with a world model, enabling the system to iteratively refine intermediate reasoning steps and improve future predictions. Similarly, Forest-of-Thought [127] utilizes MCTS to dynamically explore multiple reasoning trajectories, revisiting flawed paths and refining outcomes. We list the recent methods in Table 1.

The application of MCTS in reasoning tasks extends beyond traditional problem-solving to highly specialized domains. For example, frameworks like SRA-MCTS [135] and MC-NEST [134] showcase the utility of MCTS in tackling technical challenges such as code generation and mathematical reasoning, where intermediate steps are iteratively evaluated and refined. In fields like instructional alignment, frameworks such as SPaR [136] and Marco-o1 [112] leverage MCTS to refine responses and align reasoning trajectories with human preferences or desired outcomes. Additionally, task-specific implementations like HuatuoGPT-o1 [113] underscore MCTS's crucial role in navigating highly specialized domains, such as medical reasoning, where accuracy and robustness are paramount.

MCTS also enables models to go beyond single-pass reasoning methods, such as CoT or Tree-of-Thought, by incorporating mechanisms to revisit, critique, and refine reasoning steps dynamically [132], [150]. This iterative capability is essential for tackling tasks with vast decision spaces or those requiring long-term planning, where earlier decisions can significantly impact final outcomes. By allowing LLMs to simulate, evaluate, and refine multiple reasoning paths, MCTS introduces a level of adaptability and strategic exploration that traditional approaches lack. As shown by AlphaZero-like tree-search [126] and Search-o1 [118], MCTS enables reasoning LLMs to not only achieve better performance on specific tasks but also exhibit enhanced generalization capabilities across diverse domains.

The integration of MCTS into LLMs depends on defining actions and rewards to guide reasoning path exploration and assess quality. We classify the actions in prior work into four categories:

1) **Reasoning Steps as Nodes:** Actions represent intermediate reasoning steps or decisions, such as select-

TABLE 1
Summary of tree search-based methods. **Tag:** N=Natural language, C=Code, E=Math Expression, or A=Action.

Method	Node		Evaluate	Rollout	Tasks	LLM Models
	Format	Partial				
RAP [14]	N	✓	LLM Self-correction	LLM-based Prediction	Planning, Reasoning	LLaMA-33B
ORM [126]	N	✓	Value/Reward Function	N/A	Multiple Tasks	LLaMA-2-7B, GPT-2-small
Forest-of-Thought [127]	N	✓	LLM Self-correction	Self-refinement & Iterative Improvement	Planning, Reasoning	LLaMA-3, Mistral7B, GLM-4-9B
CodeTree [128]	N C	✓	Execution Accuracy + LLM Self-correction	Code Execution	Code Generation	GPT-4
TreeBoN [129]	N	✓	Value/Reward Function	Speculative & Dynamic Strategies	Planning, Reasoning	LLaMA-3-8B
CWM [130]	C	✓	Compare with Golden Data	Code Execution	Alignment Task	N/A
LLM-MCTS [131]	N	✗	LLM Self-correction + Policy	LLM-Based Prediction	Household Environments	GPT-2, GPT-3.5
RethinkMCTS [132]	C	✗	Execution Accuracy	Code Execution	Code Generation	GPT-3.5-turbo, GPT-4o-mini
MCTSr [133]	N	✗	LLM Self-correction	Self-Refinement & Iterative Improvement	Mathematical Reasoning	LLaMA-3-8B
MC-NEST [134]	N	✗	LLM Self-correction	Reasoning Path Generation	Mathematical Reasoning	GPT-4o, Phi-3-mini
SRA-MCTS [135]	N	✓	Execution Accuracy	Reasoning Path Generation	Code Generation	LLaMA-3-70B-Instruct
SPaR [136]	N	✓	LLM Self-correction	Self-Refinement & Iterative Improvement	Instruction Following	LLaMA-3-8B
MindStar [137]	N	✓	Value/Reward Function	Reasoning Path Generation	Mathematical Reasoning	LLaMA-2-13B, Mistral-7B
SR-MCTS [138]	E	✓	Compare with Golden Data	Math Expression Generation	Financial Fraud Detection	GPT
LLaMA-Berry [139]	N	✗	Compare with Other Solutions	Math Expression Generation	Mathematical Reasoning	LLaMA-3.1-8B
Macro-o1 [112]	N A	✗	LLM's Output Probabilities	Reasoning Path Generation	Multiple Tasks	Qwen2-7B-Instruct
ReST-MCTS* [140]	N	✓	Probability to Correct Answer	Reasoning Path Generation	Mathematical Reasoning	LLaMA-3, Mistral-7B, SciGLM-6B
CoMCTS [141]	N	✓	Compare with Other Solutions	Reasoning Path Generation	Multiple Tasks	GPT-4o, Qwen2-VL-7B, LLaMA-3.2-11B
C-MCTS [142]	N	✓	Compare with Golden Data	Math Expression Generation	Mathematical Reasoning	Qwen2.5
rStar-Math [143]	E	✓	Compare with Other Solutions	Math Expression Generation	Mathematical Reasoning	Phi3-mini-Instruct, Qwen2.5-Math-1.5B, Qwen2.5-Math-7B
AStar [144]	N	✓	Compare with Other Solutions	Reasoning Path Generation	Multiple Tasks	Qwen2.5-7B, Qwen2-VL-2B, Qwen2-VL-7B
DeepSolution [145]	N A	✗	Compare with Golden Data	Reasoning Path Generation	Multiple Tasks	Qwen2.5-7B-Instruct
VisuoThink [146]	N	✓	Compare with Golden Data	Reasoning Path Generation	Multiple Tasks	GPT-4o, Qwen2-VL-72B-Instruct, Claude-3.5-sonnet
TongGeometry [147]	E	✓	Compare with Golden Data	Math Expression Generation	Mathematical Reasoning	GPT-4
PPO-MCTS [148]	N	✓	Compare with Other Solutions	Reasoning Path Generation	Alignment Task	LLaMA-7B

ing rules, applying transformations, or generating subquestions [14], [126], [127], [149].

2) **Token-level Decisions:** Actions involve generating tokens or sequences (*e.g.*, the next word, phrase, or code snippet) [128], [129], [136], [151].
3) **Task-specific Structures:** Actions are domain-specific, such as moving blocks in blocksworld, constructing geometry in geometry problem-solving, or modifying workflows in task planning [130], [131], [152].
4) **Self-correction and Exploration:** Actions focus on revisiting, refining, or backtracking to improve previous reasoning steps [132], [133], [153].

Additionally, we classify the reward design into five categories:

1) **Outcome-based Rewards:** Rewards focus on the correctness or validity of the final outcome or solution, including the validation of reasoning paths or task success [134], [149], [152].
2) **Stepwise Evaluations:** Rewards are assigned at intermediate steps based on the quality of each step or its contribution toward the final outcome [14], [126], [135].
3) **Self-evaluation Mechanisms:** Rewards rely on the model's own confidence or self-assessment (*e.g.*, likelihood, next-word probability, or confidence scores) [129], [136], [137].
4) **Domain-specific Criteria:** Rewards are tailored to specific tasks, such as symmetry and complexity in geometry or alignment with human preferences in text generation [131], [138], [152].
5) **Iterative Preference Learning:** Rewards are derived from comparing multiple solutions or reasoning paths, guiding learning dynamically [112], [139], [140].

Summary: Despite its advantages, structure search-based (*i.e.*, MCTS) reasoning LLMs often suffer from substantial computational overhead due to the large number of simulations required. This makes them less suitable for tasks that demand real-time decision-making or operate under resource constraints [154]. Additionally, the effectiveness of MCTS is highly dependent on well-designed reward mechanisms and action definitions, which can vary significantly across different domains, thus posing challenges to its generalizability [155].

3.2.2 *Reward Modeling*

Two primary training paradigms are used to tackle multi-step reasoning tasks: outcome supervision and process supervision. Outcome supervision emphasizes the correctness of the final answer at a higher level of granularity, and the resulting model is referred to as the Outcome Reward Model (ORM) [32], [181], [182]. In contrast, process supervision provides step-by-step labels for the solution trajectory, evaluating the quality of each reasoning step. The resulting model is known as the Process Reward Model (PRM) [37], [156], [183].

PRM offers significant advantages [160], [184] in complex reasoning tasks for several key reasons. First, it provides fine-grained, step-wise supervision, allowing for the identification of specific errors within a solution path. This feature is especially valuable for RL and automated error correction. Second, PRM closely mirrors human reasoning behavior, which relies on accurate intermediate steps to reach correct conclusions. Unlike ORM, PRM avoids situations where incorrect reasoning can still lead to a correct final answer, thus ensuring more robust and interpretable reasoning. While PRM has primarily been applied to complex mathematical problems, its benefits have recently driven applications in other fields. For instance, ORPS [163] utilizes PRM to address complex code generation challenges, while Step-DPO [179] combines process supervision with the Direct Preference Optimization (DPO) algorithm [185] to improve long-chain mathematical reasoning. The framework has also shown promise in multimodal settings, with M-STAR [174] demonstrating how self-evolving training can optimize PRM for vision-language tasks, and Visual-PRM [175] establishing effective process reward modeling for multimodal reasoning through its Best-of-N evaluation approach. A summary of Reward Modeling method is presented in Table 2.

Summary: Despite the advantages of PRMs, they present several challenges. The primary difficulty is obtaining process supervision-labeled data, which is often both costly and time-consuming. To address concerns related to scalability,

TABLE 2
Summary of Reward Modeling method.

Category		Methods	Data Source	Model Refinement		Applications	Characteristic
				Strategy	Learning		
ORM		Verifier Training [32]	Existing Data	Verification	SL	Math Reasoning	GSM8K, ORM
		ORM PRM Comparison [156]	Human Annotation	Feedback-guided	SFT & RL	Math Reasoning	PRM ORM Analysis
		OVM [157]	Sampling	Feedback-guided	SFT	Math Reasoning	Guided Decoding
		ENCORE [158]	Existing Data	Entropy-guided	Training-free	Safety Tasks	Entropy of Safety Attribute
PRM	Outcome-based	DIVERSE [159]	Prompting	Fine-tuning	SFT	Multiple Reasoning Tasks	Weighted Voting Verifier
		MATH-SHEPHERD [160]	Sampling	Feedback-guided	SFT & RL	Math Reasoning	Correctness Score Assignment
		AutoPSV [161]	Prompting	Feedback-guided	SFT	Math / Commonsense Reasoning	Automated Process Supervision
		Implicit PRMs [162]	Sampling	Fine-tuning	SFT & RL	Math Reasoning	Obtaining PRM from ORM
		ORPS [163]	Sampling	Feedback-guided	SFT	Code Generation	Supervising Outcome Refinement
		R-PRM [164]	Sampling	Feedback-guided	SFT & RL	Math Reasoning	Reasoning-Driven Supervision
		BiRM [165]	Prompting	Feedback-guided	SFT	Math Reasoning	Bidirectional Reward Signals
		DeepSeek-GRM [166]	Sampling	Feedback-guided	SFT & RL	Multiple Reasoning Tasks	Inference-Time Scalability
		RewardAgent [167]	Existing Data	Feedback-guided	SFT & RL	NLP Tasks	Human & Verifiable Signals
		PAR [168]	Sampling	Feedback-guided	SFT & RL	NLP Tasks	Centered Reward Shaping
		SCIR [169]	Sampling	Feedback-guided	SFT & RL	NLP Tasks	Self-Consistency Enforcement
	MCTS	ReST-MCTS* [170]	Sampling	Self-training	SFT & RL	Multiple Reasoning Tasks	MCTS and Self-training
		OmegaPRM [171]	MCTS with Binary Search	Feedback-guided	SFT	Math Reasoning	Divide-and-Conquer MCTS
		Consensus Filtering [172]	MCTS Data Construction	Feedback-guided	SFT	Math Reasoning	Consensus Filtering Mechanism
		ReARTeR [173]	Sampling	Feedback-guided	SFT & RL	QA	Retrieval-Augmented Generation
	Multimodal	MSTaR [174]	Sampling	Self-training	SFT	Multiple Reasoning Tasks	Adaptive Temperature Adjustment
		VisualPRM [175]	Sampling	Fine-tuning	SFT	Math Reasoning	Multimoda PRM, BoN
		UnifiedReward [176]	Sampling	Feedback-guided	SFT & RL	Multimodal Tasks	Unified Multimodal Reward Modeling
	Others	Pro-Out Feedback [177]	Existing Data & Annotation	Feedback-guided	SFT & RL	Math Reasoning	Process & Outcome Supervision
		Verify Step-by-Step [178]	Human Annotation	Feedback-guided	SFT	Math Reasoning	Process Reward Annotation
		Step-DPO [179]	Sampling	Feedback-guided	SFT & RL	Math Reasoning	Step-wise Preference Pairs
		AdaptiveStep [180]	Response Dividing	Feedback-guided	SFT	Math Reasoning, Code Generation	Dividing Reasoning Steps

efficiency, and accuracy, researchers have explored various automated annotation methods. For example, MATH-SHEPHERD [160] utilizes the correctness of the final answer to define the quality of intermediate steps based on their potential to lead to the correct outcome, automating the step-wise data collection process. ReST-MCTS* [170] combines process reward guidance with MCTS to generate higher-quality reasoning traces through extensive rollouts. Similarly, OmegaPRM [171] employs the MCTS framework while introducing a divide-and-conquer algorithm for automated process supervision data generation. Another novel approach involves using ORM to train a PRM. Yuan et al. [162] propose training a PRM implicitly by leveraging ORM training on cheaper datasets, under mild reward parameterization assumptions. They also provide theoretical guarantees for the performance of this implicit PRM, demonstrating its practicality and cost-effectiveness.

In addition to data collection, PRMs face challenges related to trustworthiness [173], categorized as follows:

1) **Lack of Explanations:** Current PRMs often generate scores for reasoning steps without sufficient explanations, limiting interpretability and hindering their usefulness in refining reasoning during test-time.
2) **Bias in Training Data:** Data collection methods, such as MCTS, tend to introduce distributional biases, assigning disproportionately higher scores to the majority of questions. As a result, PRMs struggle to effectively identify erroneous reasoning steps.
3) **Early-Step Bias:** PRMs show lower accuracy in predicting rewards for earlier reasoning steps compared to those closer to the final answer. This issue stems from the increased randomness and uncertainty associated with the initial steps in the reasoning process.

3.2.3 Self Improvement

Reasoning LLMs exemplify a progression from weak to strong supervision, while traditional CoT fine-tuning faces challenges in scaling effectively. Self improvement, using the model's exploration capabilities for self-supervision, gradually enhances LLMs performance [201] in tasks such as translation [189], mathematics [186], [190], and multimodal perception [193]. This approach fosters exploration and application within reasoning LLMs [143], [216]–[218]. A summary of Self Improvement method is presented in Table 3.

Training-based self improvement in LLMs can be categorized based on exploration and improvement strategies. The exploration phase focuses on data collection to facilitate subsequent training improvements, with notable variations in approach. STaR [186] uses few-shot examples for data gathering, while ReST [189], ReST-EM [190], and ENVISIONS [191] rely on multiple samplings of complete trajectories. Quiet-STaR [111] explores at the token level, introducing concepts like meta-tokens and non-myopic loss to enhance supervision. Additionally, ReST-MCTS* [170] and rStar-Math [143] generate training data through MCTS.

Improvement strategies also exhibit significant diversity. For instance, STaR and its derivatives, such as V-STaR [**?**] and B-STaR [188], combine filtering with SFT. ReST and its variants typically introduce innovative reward calculation methods to enhance RL training for policy models. RISE [192] incorporates external feedback, recording rewards and refining responses through distillation during the improvement process. Notably, rStar-Math [143] demonstrates that small models have achieved *System 2* reflective capabilities through self-evolving training approaches.

Test-time self improvement leverages the consistency of a model's internal knowledge to correct hallucinations during inference. These approaches can be categorized into three main types: methods that refine answers using

TABLE 3
Summary of Self Improvement method.

Stage	Methods	Data Source	Model Refinement		Application
			Feedback	Strategy	
Training	STaR [186]	Few-shot	Language Model	SFT	QA, Arithmetic Reasoning
	Quiet-STaR [111]	Token-level Exploration	Language Model	RL	QA, Arithmetic Reasoning
	V-STaR [187]	Sampling	Verifier	SFT	Arithmetic Reasoning, Code Generation
	B-STaR [188]	Sampling	Reward Model	SFT	Arithmetic Reasoning, Code Generation
	rStar-Math [143]	MCTS Data Construction	Reward Model	SFT	Arithmetic Reasoning
	ReST [189]	Sampling	Reward Model	RL	Machine Translation
	ReST-EM [190]	Sampling	Language Model	EM for RL	Arithmetic Reasoning, Code Generation
	ReST-MCTS* [170]	Sampling	Reward Model	SFT, RL	Reasoning
	ENVISIONS [191]	Sampling	Environment Guided	SFT	Web Agents, Reasoning
	RISE [192]	Sampling	Reward Function	Weighted SFT	Arithmetic Reasoning
	STIC [193]	Few-shot	Language Model	SFT	Vision Language Model Tasks
	SIRLC [194]	Question Answeing	Language Model	RL	Reasoning, Translation, Summary
	AlpacaFarm [195]	Existing Data	Language Model	SFT	None (Intrinsic Evaluation)
	PSRLM [196]	Sampling	Language Model	RL	Reasoning
	SCRIT [197]	Sampling	Language Model	SFT	Arithmetic Reasoning
	S^2R [198]	Sampling	Language Model	SFT & RL	Arithmetic Reasoning
	Self-Training [199]	Sampling	Language Model	SFT	Arithmetic Reasoning
	STL [200]	Sampling	Language Model	SFT & RL	State-Value Estimation
	Genius [201]	Sampling	Language Model	RL	Arithmetic Reasoning
	START [202]	Sampling	Language Model	SFT & RL	Arithmetic Reasoning, Code Generation
	AlphaMath [203]	Sampling	Language Model	RL	Arithmetic Reasoning
	HS-STAR [204]	Sampling	Language Model	SFT & RL	Arithmetic Reasoning
Inference	Self-Refine [205]	Independent of Training Data	Language Model	Few-shot Demonstration	Code Generation, Sentiment Reversal, Acronym Generation
	Self-Check [206]	Independent of Training Data	Language Model	Step Check	QA, Arithmetic Reasoning
	CRITIC [207]	Independent of Training Data	Language Model	External Tools	QA, Arithmetic Reasoning, Detoxification
	EffiLearner [208]	Independent of Training Data	Language Model	External Tools	Code Generation
	ROSE [209]	Independent of Training Data	Language Model	Distributed Prompt	Safety, Knowledge
	Self-Verification [210]	Independent of Training Data	Language Model	Re-Ranking	Arithmetic Reasoning
	SelfEval-Decoding [211]	Independent of Training Data	Language Model	Beam Search	Aritnmetic/Symbolic Reasoning
	IPS [212]	Independent of Training Data	Language Model	Constrained Decoding	Dialogue
	Control-DAG [213]	Independent of Training Data	Language Model	Constrained Decoding	Dialogue, Open-domain Generation
	Look-Back [214]	Independent of Training Data	Language Model	Contrastive Decoding	Alleviating Repetitions
	LeCo [196]	Independent of Training Data	Language Model	Constrained Decoding	QA, Reasoning
	ProgCo [215]	Independent of Training Data	Language Model	Program-driven Verification	Instruction-following, Arithmetic Reasoning

prompts [205], [206], approaches that utilize external tools [207], and techniques that leverage logits without the need for external tools or prompts [196], [214].

3.2.4 Macro Action

Recent advancements in LLMs have driven progress in emulating human-like *System 2* cognitive processes via sophisticated thought architectures, often referred to as macro action frameworks. These structured reasoning systems go beyond traditional token-level autoregressive generation by introducing hierarchical cognitive phases, such as strategic planning, introspective verification, and iterative refinement. This approach not only enhances the depth of reasoning but also broadens the solution space, enabling more robust and diverse problem-solving pathways. A summary of Macro Action method is presented in Table 4.

We classify the progress of macro action into two aspects:

1) **Test-time Scaling through Macro Action Operationalization:** Recent research identifies two key methodologies for improving reasoning performance during inference and test-time scaling. HiICL-MCTS [222] employs a deliberate search through seed data to generate action-chain templates consisting of macro actions, thereby facilitating an action-chain-guided approach to test-time reasoning. ReasonFlux [224] utilizes an iterative test-time scaling framework, harnessing external high-level thought templates to iteratively refine and update the current CoT.
2) **Macro Action-Enhanced Data Synthesis Paradigms:** A key application of macro actions in complex reasoning is in the synthesis of reasoning data [237]. In data synthesis and training frameworks, macro action architectures enhance reasoning diversity and generalization. Recent research has shown that integrating or synthesizing a CoT process with macro actions within the reasoning sequence can significantly improve the data efficiency of the reasoning chain. For instance, LLaVA-CoT [229] enhances CoT data synthesis by externalizing intermediate reasoning steps across multiple modalities. AtomThink [231] generates the AMATH-SFT dataset using a structured g1 prompt [238], achieving superior performance on long-horizon reasoning tasks compared to traditional CoT approaches. CoAct [239] introduces a dual-agent collaborative reasoning framework, where a global planning agent executes overarching macro-actions, while a local execution agent carries out specific sub-actions within those broader actions.

Macro actions also play a crucial role in enhancing self improvement frameworks. rStar-Math [143] utilizes high-level deliberate search through Code-augmented CoT, generating diverse and reliable solutions while achieving proactive search capabilities. Satori [240] integrates CoT with RL, incorporating "*<reflect>*"-style macro actions to diversify exploration and alleviate policy saturation in online RL environments. Huatuo-o1 [113] combines hierarchical planning with domain-specific knowledge bases to improve medical reasoning. Additionally, ReasonFlux [224] dynamically reconfigures reasoning templates (*e.g.*, breaking down calculus problems into symbolic and numeric phases) to align with the problem structure.

TABLE 4
Summary of Macro Action method. **Tag:**I= Image,T= Text,V= Video. Action Category: AD: Analysis and Decomposition, IPR: Information Processing and Reasoning, VC: Verification and Correction, GO: Generation and Optimization, EB: Exploration and Backtracking.

Methods	**Usage**	**Action Attribute**					**Main Action Category**
		Action Source	**Action Number**	**Learning**	**Reflection**	**Modality**	
Self-Check [219]	Verification	Human-Designed	4	ICL	✓	T	AD, VC
LeMa [220]	Synthetic Data	Human-Designed	3	ICL & SFT	✓	T	VC, IPR
REFINER [221]	Verification/Exploration	Human-Designed	2	ICL & SFT	✓	T	VC, AD
HiICL-MCTS [222]	Exploration	Human-Designed	5	ICL	✓	T	VC, EB, AD
SUPERCORRECT [223]	Distill	In-Context Learning	Dynamic	SFT & RL	✗	T	AD, IPR
ReasonFlux [224]	Synthetic Data/Exploration	Human-Designed	~500	ICL & SFT & RL	✗	T	AD, IPR
rStar [225]	Exploration	Human-Designed	5	ICL & RL	✓	T	VC, GO, EB
LLaMA-Berry [226]	Exploration	Human-Designed	2	ICL & RL	✓	T	VC, EB
Huatuo-o1 [113]	Synthetic Data	Human-Designed	4	ICL & SFT	✓	T	VC
Marco-o1 [112]	Verification	Human-Designed	1	ICL & SFT	✓	T	VC
BoT [227]	Exploration	In-Context Learning	Dynamic	ICL	✗	T	AD, IPR
rStar-Math [143]	Exploration	In-Context Learning	1	ICL & RL	✓	T	AD, IPR
Mulberry [228]	Synthetic Data	In-Context Learning	1	ICL & SFT	✓	T	VC, EB
LLaVA-CoT [229]	Synthetic Data/Exploration	Human-Designed	4	SFT	✗	I T	AD, IPR
LLaMAV-o1 [230]	Verification/Exploration	Human-Designed	4173	Curriculum Learning	✓	I T	AD, IPR
AtomThink [231]	Synthetic Data/Exploration	In-Context Learning	>100	SFT & RL	✓	I T	AD, IPR, EB
RedStar [54]	Distill	Human-Designed	2	SFT	✓	I T	AD, VC
Auto-CoT [13]	Exploration	In-Context Learning	2	ICL	✗	T	AD, IPR, GO
PoT [232]	Verification	In-Context Learning	1	ICL	✗	T	AD, IPR, GO
PAL [233]	Verification	In-Context Learning	1	ICL	✗	T	AD, IPR, GO
Decomposed Prompt [234]	Exploration	Human-Designed	3	ICL	✗	T	AD, IPR
Least-to-Most [235]	Exploration	Human-Designed	2	ICL	✗	T	AD, IPR
CoR-Math [236]	Synthetic Data	Human-Designed	3	SFT	✓	T	AS, SR, NLR

3.2.5 *Reinforcement Fine-Tuning*

Reinforcement Fine-Tuning (RFT) [241] is an innovative technique recently introduced by OpenAI, designed to enable developers and engineers to fine-tune existing models for specific domains or complex tasks. Unlike general SFT, RFT focuses on optimizing the model's reasoning process by using a reward mechanism to guide the model's evolution, thereby enhancing its reasoning capabilities and accuracy. The core of RFT lies in improving the model's performance in a specific domain with minimal high-quality training data [242], an appropriate reward model [243], and a stable optimization process in long-context [244]–[247]. A summary of RFT method is presented in Table 5.

DeepSeek-R1 [31], which employs a verifier reward-based strategy, has shown significant performance improvements compared to traditional methods like SoS [248]. Key advantages include:

1) **Simplified Training Pipeline:** RL supervision streamlines data construction and training processes, eliminating the need for complex stepwise search mechanisms.
2) **Enhanced Scalability:** Online RL training facilitates efficient scaling on large datasets, particularly for complex reasoning tasks.
3) **Emergent Properties:** DeepSeek-R1 [31] demonstrates unique emergent capabilities, such as Long-CoT reasoning, which are difficult to achieve through SFT alone.

The success of R1 in unimodal reasoning has spurred its adaptation to MLLM, achieving SOTA performance across diverse domains (math [249], [250], medical imaging [251], segmentation [252]). Key strategies include rule-based reward systems to incentivize structured reasoning (*e.g.*, stepwise validity/accuracy rewards in R1-VL [253]), modality-agnostic training frameworks (*e.g.*, Vision-R1's [254] cold-start CoT data generation), and efficient parameter utilization (*e.g.*, MMR1 [249]'s 7B models rivaling larger proprietary counterparts). The work emphasizes transparency through explicit reasoning paths (*e.g.*, MedVLM-R1's [251] interpretable medical analysis) and open-source contributions (code [255], data [250], benchmarks [256], [257]), fostering reproducibility and MLLM-community improvement.

Despite the strengths of RFT, it still faces the following challenges:

1) **Unclear Mechanism behind Reasoning:** The underlying mechanisms driving the reasoning improvements in DeepSeek-R1 remain poorly understood. For example, while DeepSeek-R1 exhibits emergent properties (*e.g.*, "Emergent Length Increasing", "Aha moments"), studies such as [294] suggest that capabilities like Long-CoT might already exist in the base model, rather than solely emerging from RL training. Furthermore, performance gains observed in smaller models (*e.g.*, Qwen-Math-2B/7B [295]) occur without noticeable "Aha moments", complicating causal interpretations.
2) **Reward Model Saturation:** Many existing RL algorithms face reward model saturation, typically manifested as exploration collapse after around 100 training steps. Although DeepSeek-R1 alleviates this issue through specialized reward formatting, methods like ReFT [243] and Satori [240] propose alternating sampling and SFT distillation to combat reward hacking and exploration collapse.
3) **Unstable Long-CoT Generation:** Long reasoning chains generated by RFT are prone to instability, including context overflow, failure to return final answers, and sensitivity to reward shaping [124]. For instance, methods like [290] inadvertently introduce cosine reward functions, which degrade performance with increased iterations. O1-Prune [296] uses post-hoc length pruning techniques [258] (via RL/SFT) to stabilize outputs. Recent VL models such as Open-R1-Video [280] and Seg-Zero [252] show instability in chain generation when handling long videos or fine-grained segmentation reasoning.

Reinforcement Learning without External Reward. Reinforcement learning without explicit answer-based reward signals has recently been shown to be effective even in the absence of external rewards. INTUITOR [297]

TABLE 5
Summary of RFT method. **Tag**:I= Image,T= Text,V= Video.

Category	Methods	Model Attribute		Incentivize Attribute				Application & Benchmark
		Foundational LLMs	Feedback Modality	Reward Type	Algorithm	Learning	Incentivize Sample	
Reason RFT Project								
LLM	DeepSeek-R1-Zero [31]	DeepSeek-V3	T	Rule-Outcome-Reward	GPRO	RL	800K	Multiple Tasks
	DeepSeek-R1 [31]	DeepSeek-V3	T	Rule-Outcome-Reward	GPRO	RL & SFT	800K	Multiple Tasks
	Kimi k1.5 [258]	–	I T	Rule-Outcome-Reward	PPO*	RL & SFT	–	Multiple Tasks
	ReFT [243]	Galactica, CodeLLama	T	Rule-Outcome-Reward	PPO*	RL & SFT	3k/7k/8k/15k	GSM8k/SVAMP/MathQA
	RFTT [259]	LLaMA-3-3/8B-Instruct, Qwen-2.5-7B-Instruct	T	Rule-Outcome-Reward	Reinforce++	RL & SFT	1.2K	Multiple Math Task
	Satori [240]	Qwen-2.5-Math-7B	T	Rule-Outcome-Reward	PPO	RL & SFT	66K	Multiple Math Task
	QCLASS [260]	Llama-2-7B-Chat	T	Process-Reward	QNet	RL & SFT	1.9K/1.5K/3.3K	WebShop, ALFWorld, SciWorld
	PRIME [261]	Qwen2.5-Math-7B	T	Rule-Process-Outcome-Reward	PPO	RL & SFT	150K	Math, Code Tasks
	DeepScaleR [262]	DeepSeek-R1-Distill-Qwen-1.5B	T	Rule-Outcome-Reward	Iteratively GPRO	RL	40K	Multiple Math Task
	PURE [263]	Qwen2.5-Math-7B	T	Rule-Process-Outcome-Reward	PPO+RLOO	RL	8K	Multiple Math Task
	SimpleRL [125]	Qwen2.5-Math-7B	T	Rule-Outcome-Reward	PPO	RL	8K	Multiple Math Task
	Open-R1 [264]	Qwen2.5-1.5B-Instruct	T	Rule-Outcome-Reward	GPRO	RL & SFT	8K	Multiple Math, Code Task
	TinyZero [265]	Qwen2.5-0.5B/3B	T	Rule-Outcome-Reward	GPRO	RL	–	CountDown Task
	Ota-Zero [266]	Qwen-2.5-Series, DeepSeek-Series, Rho, Llama-3.x	T	Rule-Outcome-Reward	GRPO	RL	0.5K	CountDown Task
	Ota [267]	RHO-1b/Qwen2.5-3B	T	Rule-Outcome-Reward	GPRO/PPO	RL	7.5K	GSM8K
	LIMR [268]	Qwen-Math-7B	T	Rule-Outcome-Reward	PPO	RL	1.3K	Multiple Math Task
	Critic-RL [269]	Qwen2.5-Coder-32B	T	Rule-Outcome-Reward	GPRO*	RL & SFT	18.8K	Multiple Code Task
	Logic-R1 [270]	Qwen2.5-7B-Instruct-1M	T	Rule-Outcome-Reward	REINFORCE++*	RL	5K	Multiple Math, Logic Task
	Online-DPO-R1 [271]	Qwen2.5-MATH-7B	T	Rule-Outcome-Reward	DPO	RL& SFT	207.5K	Multiple Math Task
	OpenReason-Zero [272]	Qwen-2.5-7B/32B	T	Rule-Outcome-Reward	PPO	RL	57K	Multiple Math Task, GPQA, MMLU
	RLAIF [273]	PaLM 2 Extra-Small	T	Rule-Outcome-Reward	RLAIF	RL	–	Summary and Conversation Generation
MLLM	RLHF-V [274]	OmniLMM-12B	I T	Process-Reward	DDPO	RL	1.4K	Multiple Tasks
	MM-RLHF [275]	LLaVA-onevision-7B	I T V	Process-Reward	MM-DPO	RL	120K	MM-RLHF-RewardBench/SafetyBench
	Align-DS-V [276]	LLaVA-v1.5-7B,Qwen2-VL	I T V	Process-Reward	PPO, DPO	RL & SFT	200K	Align-Anything, Eval-Anything
	R1V [277]	Qwen2-VL,Qwen2.5-VL	I T	Rule-Outcome-Reward	GRPO	RL	70K/70K/8K	Multiple Tasks
	VLM-R1 [278]	Qwen2.5-VL	I T	Rule-Outcome-Reward	GRPO	RL	120K	Multiple Tasks
	LMM-R1 [279]	Qwen2.5-VL	I T	Rule-Outcome-Reward	PPO/RLOO	RL	8K	Multiple Tasks
	Open-R1-Video [280]	Qwen2-VL-7B	I T V	Rule-Outcome-Reward	GRPO	RL	4K	Multiple Tasks
	Easy-R1 [255]	Qwen2.5-VL	I T	Rule-Outcome-Reward	GRPO	RL	3K	Multiple Tasks
	Efficient-R1-VLLM [281]	DeepSeek-VL2 MoE	I T	Rule-Outcome-Reward	GRPO	RL	–	Multiple Tasks
	MMR1 [249]	Qwen2.5-VL-7B	I T	Rule-Outcome-Reward	GRPO	RL	6K	Multiple Math Tasks
	MedVLM-R1 [251]	Qwen2-VL-2B	I T	Rule-Outcome-Reward	GRPO	RL	17.3K	Medical VQA
	MM-EUREKA [250]	InternVL2.5-8B/38B	I T	Rule-Outcome-Reward	RLOO	RL	54K	Multiple Math Tasks
	R1-Omni [282]	HumanOmni-0.5B	I T V	Rule-Outcome-Reward	GRPO	RL & SFT	0.58K/15.3K	Multimodal Emotion Recognition
	R1-Onevision [256]	Qwen2.5-VL-3B/7B	I T	Rule-Outcome-Reward	GRPO	RL & SFT	155K	Multiple Tasks
	R1-VL [253]	Qwen2-VL-2B/7B	I T	Rule-Outcome-Reward	stepGRPO	RL & SFT	260K	Multiple Tasks
	VisualThinker R1 Zero [283]	Qwen2-VL-2B, Qwen-2-VL-2B-Instruct	I T	Rule-Outcome-Reward	GRPO	RL	218K	CVBench
	LALM AQA [284]	Qwen2-Audio-7B-Instruc	I T V	Rule-Outcome-Reward	GRPO	RL	38k	MMAU
	Seg-Zero [252]	Qwen2.5-VL-3B	I T	Rule-Outcome-Reward	GRPO	RL	9K	RefCOCO, ReasonSeg
	Skywork R1V [285]	InternViT-6B, DeepSeek-R1-Distill-Qwen-32B	I T	Rule-Outcome-Reward	GRPO	RL & SFT	–	Multiple Tasks
	TimeZero [286]	Qwen2.5-VL-7B	I T V	Rule-Outcome-Reward	GRPO	RL	37.4K	Charades-STA, ActivityNet
	Vision-R1 [254]	Qwen-2.5-VL-7B-Instruct	I T	Rule-Outcome-Reward	GRPO	RL & SFT	10K/200K	Multiple Tasks
	Visual-RFT [287]	Qwen2-VL-2B/7B	I T	Rule-Outcome-Reward	GRPO	RL	–	Visual Perception Tasks
	Reason-RFT [288]	Qwen-2-VL-2B/7B	I T	Rule-Outcome-Reward	GRPO	RL	–	Multiple Tasks
	SEED-Bench-R1 [287]	Qwen2-VL-Instruct-7B	I T V	Rule-Outcome-Reward	GRPO	RL	50.2K	Video Understanding
	STAR-R1 [289]	Qwen2.5-VL-7B	I T V	Rule-Outcome-Reward	GRPO	RL & SFT	9K	Visual Reasoning
Analysis RFT Project								
LLM	Demystify-LongCoT [290]	Llama-3.1-8B, Qwen2.5 -7B-Math	T	Rule-Outcome-Reward	PPO/Reinforce++	RL & SFT	7.5K	Multiple Math, MMLU
	RLHF-Scale [291]	GLM4-9B	T	Process-Reward	PPO	RL	11K	Multiple Tasks
	MD-CoT [292]	–	–	–	–	–	–	–
	GraidientUnified [293]	–	–	–	–	–	–	–

and RENT [298] propose leveraging the model's confidence—specifically, the entropy of its self-generated responses—as an intrinsic reward signal during RL training, thereby utilizing internal feedback rather than external supervision. Genius [299] demonstrates that using uncertainty from future sampling steps as a reward can further enhance model reasoning. EMPO [300] utilizes semantic clusters [301] to normalize rewards within groups of responses, thereby minimizing predictive entropy. From another perspective, self-consistency-based methods [298], [302] find that using self-consistent answers as pseudo-gold labels in RL enables self-evolution, and [302] further demonstrates the effectiveness of self-consistency-based RL training for test-time scaling. Similarly, MM-UPT [303] introduces a voting-based reward for unsupervised RL training of MLLMs. More surprisingly, [304] shows that a model's reasoning can be substantially improved using "spurious rewards," such as random or even completely incorrect rewards. However, these alternative approaches still yield inferior results compared to reward signals from golden answers and exhibit limited compatibility with answer-based rewards, indicating that further exploration is needed.

Entropy in RL Exploration. Since the exploration plays a crucial role in RL [101], adequate exploration is necessary to achieve optimal performance. DAPO [305] finds that increasing the clip-high ratio in GRPO helps prevent entropy collapse during training. Skywork-OR1 [306] demonstrates that strategically tuning sampling-related parameters and introducing an entropy loss help maintain diversity during long-term RL training. [307] states that entropy minimization (EM) shares the same objective as RLVR in unlocking the pretrained model's latent potential, and shows that a single unsupervised sample in EM training can elicit reasoning in LLMs. [308] presents an empirical transformation between the entropy and model performance during RLVR training and proposes a covariance-based entropy control strategy. Nevertheless, research on the role of entropy during RL training remains empirical, and deeper theoretical analysis is urgently needed.

Future directions for RFT may include several exciting and innovative advancements, such as:

1) **Efficient and Stable RL Frameworks:** There is a need to develop more robust RL algorithms that prevent reward saturation and exploration collapse. [290] reveals that REINFORCE++ [309] underperforms when combined with KL divergence regularization, suggesting the need for alternative methods. Future work should revisit classic RL algorithms in the context of modern LLMs training to optimize both stability and efficiency.
2) **Scaling RFT:** Current RL-Supervise models rely on curated, verifiable prompts selected from large-scale datasets. Future research should focus on synthesizing high-quality, diverse prompts to improve generalization. [291] shows that merely scaling policy/reward models or increasing sample sizes results in diminishing returns, while expanding the scope of PRM and R1 training data holds greater promise. Hybrid approaches, such as combining RL with SFT or curriculum learning, should be explored to enhance scalability.
3) **Controlling Long-CoT Stability:** Adaptive reward

shaping mechanisms are needed to balance reasoning length, coherence, and answer correctness. Techniques such as O1-Prune [296] demonstrate the value of post-hoc length regularization, but dynamic in-training controls are necessary. Hierarchical RL frameworks should be investigated to decompose long reasoning chains into manageable sub-tasks, reducing instability.

4) **Theoretical and Empirical Analysis:** It is essential to clarify the relationship between RL training and the capabilities of the base model. For instance, it should be determined whether emergent properties (*e.g.*, Long-CoT) arise from RL optimization or are latent traits of the base model. However, this phenomenon of Long-CoT is often difficult to appear in multimodal situations (*e.g.*, [255], [277], [279], [310]). Systematic studies on reward design principles (*e.g.*, sparse vs. dense rewards, multi-objective balancing) should be conducted to avoid unintended behaviors such as reward hacking.

Summary: RFT presents a promising direction for advancing LLMs reasoning, as evidenced by DeepSeek-R1 [31]. However, challenges such as reward saturation, unstable long reasoning chains, and unclear emergent mechanisms require urgent attention. Future efforts should prioritize algorithmic innovation, scalable prompt synthesis, and theoretical grounding to fully unlock the potential of RL-driven reasoning LLMs.

3.3 Evolutionary of Reasoning LLMs

The evolution of reasoning LLMs has progressed by several distinct stages, with various strategies developed to overcome the limitations of direct autoregressive inference and build more advanced slow-thinking reasoning architectures.

In the early stages, reasoning LLMs primarily focused on enhancing pre-trained LLMs with external reasoning algorithms, without altering the underlying model parameters. Approaches such as Tree of Thoughts [394], [395] and Reasoning via Planning [14] utilized LLMs-driven Breadth-First Search, Depth-First Search, and MCTS [98], [127], [129], [396] to simulate human-like reasoning processes. These methods represented reasoning as tree or graph traversals, where intermediate reasoning states were depicted as nodes, and various reasoning strategies produced distinct reasoning paths. The final decision was made through additional voting mechanisms [3] or Monte Carlo-based value estimation to identify the optimal path.

However, these externalized slow-reasoning approaches introduced several challenges:

1) **Limited Exploration Space:** The search-based methods required predefined constraints on the breadth, depth, and granularity of the search space, which often restricted the LLM's exploration to a narrow reasoning space. Furthermore, the reasoning strategies across different child nodes of the same parent node frequently lacked sufficient diversity, further limiting exploration.
2) **Limited Experience Sharing:** Exploration experiences and reasoning information across different paths could only be assessed based on reward models or self-consistency among outcomes. Additionally, search-based methods significantly increased computational overhead, relying on reward models such as PRM/ORM for tree pruning or speculative decoding techniques to accelerate inference.

To overcome these limitations, subsequent models such as rSTaR [225], LLaMAV-o1 [230], HiICL-MCTS [222], Mulberry [228], g1 [238], and Thinking-Claude [397] introduced richer action spaces. These enhanced action spaces offered high-level planning cues, broadening the model's exploration scope and enabling more comprehensive structured search processes. However, this approach necessitated careful design of the action spaces to ensure their effectiveness.

With the introduction of models like o1 [29] and QwQ [120], external reasoning paradigms were internalized within the LLM's context. These models initially performed exploratory macro-planning to generate an initial reasoning path, followed by contextual exploration of alternative paths. Through mechanisms like "Rethink" and "Verification", these models produced extended reasoning chains. To replicate this internalized capability, STILL-1 [396] linearized tree search outputs into long reasoning chains with attributes such as "Rethink", "Wait", and "Explore New Path". Similarly, STILL-2 [53] and sky-T1 [121] synthesized long reasoning chains using distillation techniques. Approaches like Virgo [398] have attempted to distill text-based slow-thinking reasoning into multimodal LLMs; their performance improvements in tasks such as MathVision [365], which demand detailed visual understanding, have been marginal. However, the linearized reasoning chains derived from search-based methods struggled to match the quality of those produced by distillation approaches. Additionally, extending slow-thinking reasoning capabilities from text-based domains to multimodal contexts remains a significant challenge, especially in tasks requiring fine-grained perception [117], [399]–[401].

Recent advancements, including DeepSeek-R1 [31] and Kimi-k1.5 [258], have demonstrated the potential of RL to enhance models like DeepSeek-V3 [17], resulting in the emergence of complex behaviors such as long reasoning chains, reflective reasoning, and advanced planning capabilities. Remarkably, these sophisticated behaviors were achieved through simple RL scaling. SimpleRL [125] sought to replicate these capabilities using a streamlined pipeline and minimal codebase, while R1V [277] explored the development of multimodal reasoning models based on multimodal foundation architectures. However, though R1 has been proven to significantly enhance reasoning abilities in the LLMs field, there remain many challenges to explore in the MLLMs domain. These include difficulties in maintaining a consistent slow thinking process when handling complex visual inputs and achieving training benefits comparable to those acquired through unimodal RL.

Summary: The evolution of reasoning LLMs has shifted from externally augmented reasoning to internally embedded reasoning. Recent developments emphasize the potential of RL-based scaling to unlock advanced capabilities.

4 Benchmarking Reasoning LLMs

The development of a robust benchmark is crucial for documenting the advancements in reasoning LLMs capabilities and for identifying promising research directions for future progress. Here, we review the benchmarks from three key

TABLE 6
Statistics of benchmarks for reasoning LLMs.

Domain	Benchmark	Question Type	Venue	Language	Size	Level
Math	AIME 2024 [311]	Open-End	-	English	30	Competition
	BBH [312]	Hybrid	Findings ACL 2023	English	23	Challenging
	MATH-500 [37]	Open-End	ICLR 2024	English	500	Competition
	AMC 2023 [313]	Open-End	–	English	30	Competition
	Olympiad Bench [314]	Open-End	ACL 2024	English/Chinese	8,476	Competition
	Putnam-AXIOM [315]	Open-End	NeurIPS 2024	English	236	Competition
	PRM800K [178]	Open-End	ICLR 2024	English	800,000	Hybrid
	FrontierMath [316]	Open-End	ArXiv 2024	English	-	Expert
	ProcessBench [317]	Open-End	ArXiv 2024	English	3400	Competition
	LiveBench [318]	Open-End	ICLR 2025	English	Frequently Updated	Expert
	AIME 2025	Open-End	Hugging Face	English	13	Competition
	ThinkBench [319]	Hybrid	ArXiv 2025	English	2,912	Expert
	MATH-Perturb [320]	Open-End	ArXiv 2025	English	279	Competition
	ZebraLogic [321]	Open-End	ArXiv 2025	English	1,000	Hybrid
	QuestBench [322]	Choice	ArXiv 2025	English	38,882	Hybrid
	Math-RoB [323]	Open-End	ArXiv 2025	English	-	High School
	GSM-Ranges [324]	Open-End	ArXiv 2025	English	-	Middle School
Code	Codeforces	Open-End	-	English	-	Expert
	CodeContests [325]	Open-End	Science 2022	English	13,610	Competition
	SWE-bench [326]	Open-End	ICLR 2024	English	2,294	Expert
	LiveCodeBench [327]	Open-End	ArXiv 2024	English	-	Expert
	CodeCriticBench [328]	Hybrid	ArXiv 2025	English	-	Expert
Science	GPQA Diamond [329]	Choice	COLM 2024	English	448	University
	MR-Ben [330]	Hybrid	NeurIPS 2024	English	5,975	Hybrid
	MMLU-Pro [331]	Choice	NeurIPS 2024	English	12,032	Hybrid
	MHPP [332]	Open-End	ArXiv 2024	English	210	Expert
	RewardBench [333]	Hybrid	ArXiv 2024	English	-	Hybrid
	MR-Ben [334]	Open-End	NeurIPS 2024	English	5,975	Hybrid
	ReaLMistake [335]	Open-End	COLM 2024	English	-	Expert
	CriticBench [336]	Open-End	ACL 2024	English	-	Hybrid
	JudgeBench [337]	Open-End	ArXiv 2024	English	-	Hybrid
	TPBench [338]	Open-End	ArXiv 2025	English	57	University
	ProBench [339]	Open-End	ArXiv 2025	English/Chinese	790	Competition
	EquiBench [340]	Open-End	ArXiv 2025	English	2,400	Hybrid
	SuperGPQA [341]	Choice	ArXiv 2025	English	26,529	University
	Sys2Bench [342]	Open-End	ArXiv 2025	English	-	Hybrid
	PRMBench [343]	Open-End	ArXiv 2025	English	6,216	Expert
	DeltaBench [344]	Open-End	ArXiv 2025	English	-	Expert
	FINEREASON [345]	Open-End	ArXiv 2025	English	-	Expert
Agent	ARC [346]	Open-End	ArXiv 2019	Symbolic	1,000	Expert
	WebShop [347]	Open-End	NeurIPS 2022	English	1,600	Hybrid
	SciWorld [348]	Open-End	EMNLP 2022	English	7,200	Hybrid
	WebArena [349]	Open-End	ICLR 2024	English	812	Hybrid
	TextCraft [350]	Open-End	NAACL 2024	English	200	Hybrid
	Osworld [351]	Open-End	NeurIPS 2024	English	369	Hybrid
	GAMABench [352]	Open-End	ArXiv 2024	English	-	Hybrid
	Mle-bench [353]	Open-End	ArXiv 2025	English	-	Competition
	ToolComp [354]	Open-End	ArXiv 2025	English	-	Hybrid
	Mobile-Agent-E [355]	Open-End	ArXiv 2025	English	25	Hybrid
	Text2World [356]	Open-End	ArXiv 2025	English	-	Hybrid
	WebGames [357]	Open-End	ArXiv 2025	English	50	Hybrid
	Ui-r1 [358]	Open-End	ArXiv 2025	English	136	Hybrid
Medicine	JAMA Clinical [359]	Choice	NAACL 2025	English	1,524	Expert
	Medbullets [359]	Choice	NAACL 2025	English	308	Expert
	MedQA [360]	Choice	ArXiv 2020	English/Chinese	61,097	Expert
	MEDEC [361]	Open-End	ArXiv 2024	English	3,848	Expert
	MedXpertQA [362]	Choice	ArXiv 2025	English	4,460	Expert

aspects: categories, evaluation metrics, and performance comparisons, while offering our reflections and insights.

4.1 Benchmark Categories

We categorize reasoning benchmarks by task type, which can be broadly divided into math, code, scientific, agent [402], medical, and multimodal reasoning. The detailed statistics for these benchmarks are presented in Table 6.

4.1.1 Benchmark Introduction

1) **Math Problems:** We document the current popular competition-level mathematical benchmarks to showcase the capabilities of reasoning LLMs, including AIME 2024 [311], MATH-500 [37], AMC 2023 [313], and Olympiad Bench [314].
2) **Code Problems:** Code problems [403], [404] requires solid foundation and high logical thinking to evaluate the reasoning ability of reasoning LLMs such as Codeforces, SWE-bench [326], and LiveCodeBench [327].
3) **Scientific Problems:** Scientific benchmarks, *i.e.*, GPQA Diamond [329] and MMLU-Pro [331], involve multi-domains reasoning about chemistry, biology, and physics, which requires extensive knowledge accumulation and integrated reasoning.

TABLE 7
Statistics of benchmarks for reasoning MLLMs.

Domain	Benchmark	Question Type	Venue	Language	Size	Level
Multimodality	MMMU [363]	Hybrid	CVPR 2024	English	11,500	Hybrid
	MathVista [364]	Hybrid	ICLR 2024	English	6,141	Middle School
	MathVision [365]	Hybrid	NeurIPS 2024	English	3,040	Middle/High School
	CMMaTH [366]	Hybrid	COLING 2025	English/Chinese	23,856	Middle/High School
	PGPS9K [367]	Hybrid	IJCAI 2023	English	9,023	Middle School
	ZeroBench [368]	Open-End	ArXiv 2025	English	100/334	Impossible
	MME-CoT [369]	Hybrid	ArXiv 2025	English	1,130	Hybrid
	MM-IQ [370]	Choice	ArXiv 2025	English/Chinese	2,710	Hybrid
	Multimodal RewardBench [371]	Hybrid	ArXiv 2025	English	5,211	Hybrid
	GRAB [372]	Open-End	ArXiv 2024	English	2,710	Hybrid
	SciFIBench [373]	Choice	ArXiv 2024	English	2,000	Challenging
	MV-MATH [374]	Hybrid	ArXiv 2025	English	2,009	Middle/High School
	ScienceQA [375]	Choice	NeurIPS 2022	English	21,000	Hybrid
	Plot2Code [376]	Open-End	ArXiv 2024	English	132	Hybrid
	M3CoT [377]	Open-End	ArXiv 2024	English	-	Hybrid
	PUZZLEVQA [378]	Hybrid	ACL 2024	English	-	Hybrid
	MolPuzzle [379]	Hybrid	NeurIPS 2024	English	23,000	Expert
	HumanEval-V [380]	Open-End	ArXiv 2024	English	108	Hybrid
	CoMT [381]	Choice	AAAI 2025	English	3,853	Hybrid
	ChartMimic [382]	Open-End	ArXiv 2024	English	4,800	Expert
	OlympicArena [383]	Hybrid	NeurIPS 2024	English/Chinese	11,163	Competition
	CVQA&CPVQA [384]	Open-End	ArXiv 2025	English	1664	Hybrid
	ENIGMAEVAL [385]	Open-End	ArXiv 2025	English	1184	Challenging
	CODE-VISION [386]	Open-End	ArXiv 2025	English	438	Competition
	MKRC [387]	Open-End	ArXiv 2025	English	7010	Hybrid
	MMSciBench [388]	Hybrid	ArXiv 2025	English/Chinese	4,482	High School
	LEGO-Puzzles [389]	Hybrid	ArXiv 2025	English	1,100	Expert
	JustLogic [390]	Open-End	ArXiv 2025	English	7,000	Expert
	HUMANITY'S LAST EXAM [391]	Hybrid	ArXiv 2025	English	2,700	Challenging
	DivIL [392]	Hybrid	TMLR	English	-	Hybrid
	ErrorRadar [393]	Open-End	ArXiv 2024	English	2,500	Middle/High School

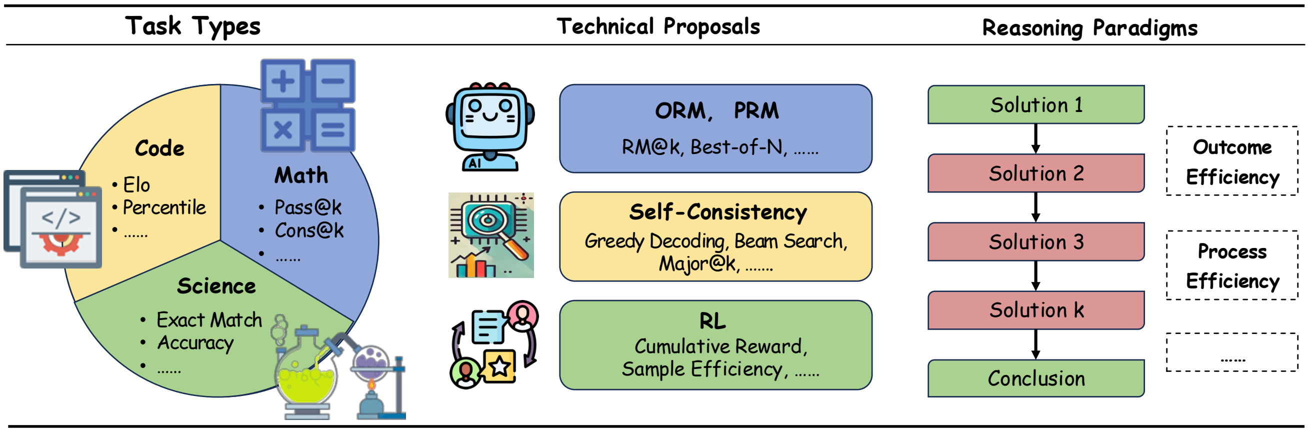

Fig. 5. Various evaluation metrics of reasoning LLMs divided by task types, technical proposals, and reasoning paradigms.

4) **Agent Reasoning:** Realistic tasks often involve complex planning and tool usage, leading to the creation of agent reasoning benchmarks [405]. For example, WebShop [347] and WebArena [349] focus on web operations, while SciWorld [353] and TextCraft [350] are centered around scientific research.
5) **Medical Reasoning:** Medicine fundamentally involves complex reasoning, spanning tasks from diagnostic decision making to treatment planning. Benchmarks of JAMA Clinical Challenge [359], Medbullets [359], and MedQA [360] offer model measurements that mimic the doctor's disease diagnosis.
6) **Multimodal Reasoning:** Multimodal reasoning, such as benchmarks of MMMU [363] , MathVista [364], MME-CoT [369] and MM-IQ [370], requires cross-modal thinking in combination with text and images. Especially for those visual-centered problems, in benchmarks MathVision [365], MathVerse [406], CMMaTH [366] , PGPS9K [367], SciFIBench [373] and GRAB [372], put forward higher requirements for reasoning MMs.

4.1.2 Summary

The field of LLMs has advanced rapidly in recent years, with benchmark performance consistently improving. Simple reasoning benchmarks, such as GSM8K [32], MATH-500 [37], and ScienceQA [375], have approached performance saturation. Recent studies on reasoning LLMs [54], [143] show that models designed for long reasoning chains do not significantly outperform those designed for shorter chains on these benchmarks. This highlights the urgent need to

TABLE 8
Performance of Different Models, including Basic LLMs and Reasoning LLMs, on Plain Text Benchmarks. The **red** denotes the highest result, and the **blue** denotes the second highest result.

	Model	Math		Code			General	
		AIME 2024 *(Pass@1)*	MATH-500 *(Pass@1)*	LiveCodeBench *(Pass@1-CoT)*	Codeforces *(Percentile)*	SWE Verified *(Resolved)*	MMLU *(Pass@1)*	GPQA-Diamond *(Pass@1)*
Basic LLMs	Gemini-2.5-Pro [408]	**92.0**	-	70.4	-	**63.8**	81.7	**84.0**
	Gemini-2.0-Pro [409]	-	91.8	36.0	-	-	86.5	64.7
	GPT-4o [16]	9.3	74.6	34.2	23.6	38.8	87.2	49.9
	Claude-3.5-Sonnet [410]	16.0	78.3	33.8	20.3	50.8	88.3	65.0
	Deepseek-V3 [17]	39.2	90.2	36.2	58.7	42.0	88.5	59.1
	Claude 3.7 Sonnet [411]	-	82.2	-	-	**62.3**	-	68.0
Reasoning LLMs	Eurus-2-7B-PRIME [261]	26.7	79.2	-	-	-	-	-
	InternLM3-8B-Instruct [412]	20.0	83.0	17.8	-	-	76.6	37.4
	rStar-Math-7B [143]	46.7	81.6	-	-	-	82.7	54.9
	STILL-2-32B [53]	46.7	90.2	-	-	-	-	-
	Redstar-code-math [54]	53.3	91.2	-	-	-	-	-
	Search-o1 [118]	56.7	86.4	33.0	-	-	-	63.6
	QwQ [120]	50.0	90.6	41.9	62.0	-	-	54.5
	s1-32B [413]	56.7	93.0	-	-	-	-	59.6
	OpenAI o1-mini [414]	63.6	90.0	53.8	93.4	41.6	85.2	60.0
	LIMO-32B [415]	57.1	94.8	-	-	-	-	66.7
	Kimi k1.5 long-CoT [258]	77.5	96.2	62.5	94.0	-	-	-
	OpenAI o3-mini [30]	**87.3**	**97.9**	**84.6**	-	49.3	86.9	**79.7**
	Seed-Thinking-v1.5 [416]	86.7	-	64.9	55.0	47.0	87.0	77.3
	DeepSeek-R1-Distill-Qwen-1.5B [417]	28.9	83.9	16.9	-	-	-	33.8
	DeepSeek-R1-Distill-Qwen-7B [417]	55.5	92.8	37.6	-	-	-	49.1
	DeepSeek-R1-Distill-Qwen-14B [417]	69.7	93.9	53.1	-	-	-	59.1
	DeepSeek-R1-Distill-Qwen-32B [417]	72.6	94.3	57.2	-	-	-	62.1
	DeepSeek-R1-Distill-Llama-8B [417]	50.4	89.1	39.6	-	-	-	49.0
	DeepSeek-R1-Distill-Llama-70B [417]	70.0	94.5	57.5	-	-	-	65.2
	DeepSeek-R1 [31]	79.8	**97.3**	**65.9**	**96.3**	49.2	**90.8**	71.5
	OpenAI-o1 [29]	79.2	96.4	63.4	**96.6**	48.9	**91.8**	75.7

establish new benchmarks that more effectively assess the reasoning capabilities of reasoning LLMs. Moreover, current benchmarks are limited, focusing mainly on solid reasoning tasks. Soft reasoning benchmarks, lacking explicitly defined correct answers, offer a more nuanced evaluation, better capturing the complexities and subtleties of human-like reasoning. Furthermore, it is essential to address the issue of data leakage in evaluation processes [407]. Ensuring the confidentiality and neutrality of evaluation data is critical to preserving the integrity and reliability of benchmark results.

4.2 Evaluation Metrics

Depending on task types, technical proposals, and reasoning paradigms, various evaluation metrics have been introduced for reasoning LLMs as shown in Figure 5. These metrics are designed to more accurately assess the model's performance in handling complex reasoning tasks, ensuring that both the quality and coherence of the generated solutions are effectively measured.

4.2.1 Task Types

In terms of benchmark categories, mathematical reasoning typically uses two main metrics: *Pass@k* and *Cons@k*. The *Pass@k* metric evaluates the model's ability to generate a correct solution within k attempts, measuring the likelihood of success within a limited number of tries. On the other hand, *Cons@k* assesses whether the model consistently produces correct or logically coherent solutions, highlighting the stability and reliability of its reasoning capabilities. For code tasks, the key metrics are *Elo* and *Percentile*, both of which measure the relative skill in generating correct code compared to other models or human programmers. In scientific tasks, evaluation generally employs *Exact Match (EM)* and *Accuracy* for fill-in-the-blank and multiple-choice questions, respectively. The *EM* metric judges whether the model's output exactly matches the expected solution, while *Accuracy* measures the proportion of correct answers out of the total number of questions.

4.2.2 Technical Proposals

Based on technical routes, the schemes with ORM or PRM often leverage *RM@k* and *Best-of-N* two evaluation indicators. *RM@k* measures whether the reward model can rank the good answer higher in the top k candidates according to reward score, and *Best-of-N* chooses the solution with highest score from N generated reasoning trajectories. Methods for self-consistency are evaluated using *Greedy Decoding*, *Beam Search*, and *Major@k*. *Greedy Decoding* and *Beam Search* control the randomness of the inference process by limiting the sampling range. *Major@k* selects the solution with the most consistent results from k candidate solutions. In RL, metrics reflect both performance in achieving desired outcomes and the efficiency of the learning process. For example, *Cumulative Reward* measures the total reward received by the agent over time, while *Sample Efficiency* assesses the efficiency of the agent's sample usage during learning.

4.2.3 Reasoning Paradigms

For reasoning paradigm of the multi-turn solution generation in reasoning LLMs, *Outcome Efficiency* and *Process Efficiency* [124] are proposed recently to evaluate the efficiency of long thinking specifically. *Outcome Efficiency* metric empirically evaluates how effectively later solutions contribute to accuracy improvements, formulating as the ratio of efficient tokens that contribute to reaching the correct answer, to all output tokens. *Process Efficiency* metric evaluates the contribution of later solutions to solution diversity empirically, concretely representing as the ratio of tokens of distinct solutions to all solution tokens. These two indicators reveal to the overthinking issue of existing reasoning LLMs to simple problems certainly.

4.2.4 Summary

Most of the existing evaluation metrics are judged according to the final answer. It is imperative to develop a comprehensive assessment framework that considers various aspects of the reasoning process in view of the large inference computation consumption. Current popular evaluation frameworks, such as LMMs-Eval [418], OpenCompass [419], and PRMBench [343], lack efficiency and their metrics do not adequately account for the computational and temporal efficiency of the reasoning process. To address these shortcomings, we highly recommend exploring more efficient proxy tasks as potential solutions. By identifying and utilizing tasks that better capture the nuances of long reasoning chains, we can develop more robust and effective evaluation metrics to enhance the overall assessment framework, ensuring that it not only measures the accuracy of the final output but also evaluates the efficiency and coherence of the reasoning process throughout.

4.3 Performance Comparison

In this section, we compare the performance of different reasoning LLMs and their corresponding foundational LLMs on plain text benchmarks, such as math and code problems, as well as on multimodal benchmarks. The comprehensive real-time leaderboard is available on Arena.

4.3.1 Performance on Plain Text Benchmarks

As shown in Table 8, reasoning LLMs, such as DeepSeek-R1 [31] and OpenAI-o1/o3 [29], [30], demonstrate exceptional performance across a wide range of tasks, including math, coding, and other general tasks. These models achieve high scores on multiple plain-text benchmarks, such as AIME 2024, MATH-500, and LiveCodeBench, showcasing their robust text-based reasoning abilities. In contrast, foundational LLMs, like GPT-4o [62], Claude-3.5-Sonnet [410], and DeepSeek-V3 [17], generally perform less effectively than reasoning LLMs, particularly in math and coding tasks (*e.g.*, AIME 2024 and Codeforces). For example, OpenAI-o1 outperforms GPT-4o by 69.9% and 73% on these tasks, respectively. Moreover, DeepSeek-R1, based on the DeepSeek-V3 architecture, surpasses its predecessor on all benchmarks, further highlighting the advantages of the reasoning LLMs.

4.3.2 Performance on Multimodal Benchmarks

As shown in Table 9, reasoning LLMs continue to excel in multimodal tasks. OpenAI-o1 [29] performs strongly in vision tasks, achieving the highest score of 78.2% on MMMU and outperforming its corresponding foundational LLM, GPT-4o [62], by 7.2% on MathVista. However, the performance improvement in multimodal tasks is less pronounced compared to text-only tasks. This can be attributed in part to the limitations of current multimodal reasoning LLM techniques, as well as the lack of sufficient datasets to fully assess the multimodal capabilities of reasoning LLMs.

TABLE 9
Performance of Models, including Basic LLMs and Reasoning LLMs, on Multimodal Benchmarks. The **red** denotes the highest result, and the **blue** denotes the second highest result.

	Model	MMMU	Mathvista	Mathvision	Olympiadbench
Basic LLMs	Claude-3.5-Sonnet [410]	68.3	65.3	38.0	-
	GPT-4o [16]	69.1	63.8	30.4	**25.9**
	Gemini 2.0 Pro [409]	72.7	-	-	-
	Llama 4 Maverick [420]	73.4	73.7	-	-
	Claude-3.7-Sonnet [411]	75.0	-	-	-
	Gemini 2.5 Pro [408]	**81.7**	-	-	-
Reasoning LLMs	Skywork-R1V-38B [285]	69.0	67.5	-	-
	LLaVA-CoT [229]	-	54.8	-	-
	Kimi k1.5 long-CoT [258]	70.0	74.9	-	-
	Qwen2.5-VL-72B [421]	70.2	74.8	**38.1**	-
	QvQ-72B-preview [422]	70.3	71.4	35.9	20.4
	Kimi k1.6 preview	-	**80.0**	**53.3**	-
	MMR1-Math-v0-7B [249]	-	71.0	30.2	-
	MM-EUREKA [250]	-	64.2	26.6	**37.3**
	R1-Onevision-7B [256]	-	64.1	29.9	-
	Doubao-1.5-pro [423]	73.8	78.8	48.6	48.5
	OpenAI-o1 [29]	78.2	71.0	-	-
	OpenAI o4-mini [424]	**81.6**	**84.3**	-	-

4.3.3 Summary

In summary, reasoning LLMs show strong performance across both plain text and multimodal benchmarks, particularly excelling in math and coding tasks, where they outperform foundational LLMs by a large margin. Although the improvement in multimodal tasks is not as pronounced as in text-only tasks, reasoning LLMs still surpass their counterparts, highlighting their potential for processing both image and text data. These results emphasize the versatility and effectiveness of reasoning LLMs across a broad spectrum of reasoning tasks, with potential for further advancements in multimodal reasoning techniques.

5 Extended Techniques

5.1 Advanced Infrastructure

In the realm of System-2 paradigms, reinforcement learning stands as a pivotal methodology to elevate the performance of reasoning models, finding ubiquitous employment in LLM, MLLM, and agentic applications such as Agent-R1 [425] and UI-R1 [426]. Recent advancements have seen state-of-the-art RL frameworks like VeRL [427], OpenRLHF [428], and ReaLHF [429] emerge as dominant tools in this domain. Despite their divergent architectural designs, these frameworks adhere to a canonical RLHF workflow comprising three core stages:

- **Generative Rollout Phase**: The actor network generates outputs in response to a batch of input prompts. In agentic contexts, this process transcends mere static response generation, necessitating dynamic interactions with complex environments to facilitate adaptive decision-making.
- **Preprocessing Pipeline**: Post-actor inference, the framework undertakes a sophisticated derivation of auxiliary information critical for RL training. This includes computing log probabilities from the reference policy to quantify behavioral consistency, estimating value functions via the critic model to assess state desirability, and calculating advantage estimates to prioritize high-impact learning signals.
- **Optimization Stage**: With the curated dataset from the generative and preprocessing stages—encompassing

actor outputs, critic evaluations, and policy gradients—the system initiates iterative parameter updates. This stage harmonizes actor-critic dynamics to refine decision-making strategies, ensuring alignment with task objectives and environmental feedback.

This structured workflow underscores the synergy between generative modeling, value estimation, and policy optimization, cementing RLHF as a cornerstone for developing robust, adaptive intelligent systems.

In the Generative Rollout Phase, contemporary inference engines such as vLLM [430] and SGLang [431] are leveraged to facilitate high-throughput deployment, ensuring efficient realization of generative workflows.

For the Optimization Stage, diverse parallelism paradigms—including data parallelism, pipeline parallelism, and tensor parallelism—are systematically implemented. This is achieved through state-of-the-art distributed training frameworks like Megatron-LM [432] and MegaScale [433], which incorporate sophisticated 3D parallelism. These architectural designs enable seamless orchestration of computational resources, enhancing both scalability and training efficiency across distributed systems.

5.1.1 Taxonomy of RL Infrastructure

While distinct RLHF frameworks broadly adhere to an analogous workflow, notable discrepancies emerge in their architectural configurations. The most pronounced differentiator resides in whether they employ a **Single-Controller** or **Multi-Controller** paradigm, a distinction that fundamentally shapes their operational dynamics and scalability profiles.

Within the **Single-Controller** paradigm, a centralized command hub governs the holistic execution trajectory of distributed programs, enabling users to craft core dataflow functionalities as an integrated workflow. However, the propagation of coordination messages from this central hub to all worker nodes engenders significant dispatch overhead during the execution of large-scale dataflow graphs across sprawling clusters. VeRL and ReaLHF exemplify this architectural model with notable prominence, illustrating both its cohesive design advantages and scalability challenges in high-scale distributed environments.

In the domain of **Multi-Controller** architectures, each device—hereafter referred to as a worker—is equipped with its own bespoke controller. Cutting-edge distributed systems for LLM training and inference demonstrate a pronounced preference for the multi-controller paradigm. This inclination is primarily attributable to its inherent scalability and the minimal scheduling overhead it incurs.

In the context of implementing the RLHF workflow within a multi-controller architectural framework, practitioners are required to meticulously choreograph the integration of codebase governing inter-device collective communication protocols, computational logic frameworks, and point-to-point data transfer mechanisms across each device's operational runtime environment. This imperative orchestration gives rise to profoundly nested code architectures, wherein computational routines and data transmission operations become inextricably interwoven. The resultant architectural intricacy poses formidable challenges for development, maintenance, and performance optimization—given that the inter-dependencies between functional components necessitate elaborate debugging methodologies and precision tuning to ensure seamless cross-component synchronization and optimal resource allocation. OpenRLHF stands as a paradigmatic exemplar of such systems, embodying the complex interplay between modular device autonomy and centralized coordination required to sustain efficient RLHF pipeline execution.

5.1.2 Prospective Challenges of RL Infrastructure

1) **Long-Text-Generation**. The prioritization of long-text generation in reasoning models presents substantial challenges for the generation stage. During Long-CoT generation, the substantial variability in response lengths across different prompts leads to significant GPU under-utilization due to prolonged idle periods. In the context of agentic reinforcement learning (Agentic-RL), the generation stage must not only efficiently produce lengthy responses but also interact dynamically with environments. As environments scale in complexity and diversity, this interaction imposes escalating challenges on generative efficiency and adaptability. For long-text generation, numerous studies have introduced sophisticated rollout strategies to address these issues. For example, Kimi-1.5 [258] employs a partial-rollout framework, eschewing the need for monolithic response generation. Instead, it processes and stores textual segments incrementally, enabling the creation of significantly longer outputs while maintaining rapid iteration cycles. During training, selective exclusion of non-critical segments from loss computation optimizes the learning process, enhancing both operational efficiency and system scalability. Concurrently, Seed-Thinking-1.5 [434] introduces a Streaming Rollout System that decouples model evolution from runtime execution, facilitating dynamic adjustment of on/off-policy sampling. This architectural innovation ensures that generative processes remain responsive to environmental feedback without compromising computational efficiency, thus advancing the state-of-the-art in scalable long-text generation frameworks.
2) **Scalable Design for Agent-Environment Interaction**. In the realm of agent-environment interaction scenarios [435], the dynamic interplay between agents and their surroundings poses a formidable challenge, primarily stemming from the heterogeneous latencies inherent across diverse environments [426], [436]. To address this, we posit that leveraging an asynchronous message-queue architecture such as the producer-consumer paradigm becomes imperative for effectively managing the temporal discrepancies in agent-environment exchanges. This approach mitigates the adverse effects of variable delays, ensuring robust coordination even in systems characterized by unpredictable response times. Regarding environmental stability, our analysis, rooted in scaling-law principles, underscores the critical role of scalable environmental design in advancing LLM/MLLM-based agentic applications (e.g., DeepResearch [437], DeepSearch [438], and autonomous tooling frameworks like Open-

Manus [439]). Unlike traditional LLM/MLLM training pipelines, where data resides statically within file systems, agentic environments typically consist of distributed assemblages of HTTP servers. This architectural distinction introduces a nontrivial stability challenge for large-scale Agentic-RL training, as the reliability of such systems hinges on the seamless integration and resilience of networked components. Scaling environments to accommodate growing computational demands while maintaining operational consistency thus emerges as a foundational requirement for realizing the full potential of next-generation agentic systems.

3) **Protocol Design and Security Considerations**. Developing interfaces between disparate HTTP servers and LLMs can impose substantial overhead, as each integration often requires bespoke engineering efforts. A standardized protocol—such as the Model-Computer Protocol (MCP) [440]—facilitates seamless interaction between LLMs/MLMs and tools without requiring intimate knowledge of tool-specific implementations. This framework streamlines development by mitigating the need for redundant interface design, thereby enhancing interoperability and reducing engineering latency. However, this architectural elegance introduces a critical security imperative: third-party MCP servers may harbor malicious code that could compromise the integrity of the training ecosystem. Such risks necessitate robust safeguards to prevent unauthorized access, code injection, or systemic sabotage, ensuring that the benefits of protocol standardization are not overshadowed by potential vulnerabilities. Proactive measures—including rigorous code audits [441], secure authentication mechanisms [442], and runtime monitoring [443] are essential to maintain the reliability and safety of the integrated system, striking a balance between operational efficiency and defensive resilience.

5.2 Trustworthiness of Reasoning LLMs

As RLLMs grow more capable, their trustworthiness becomes increasingly critical—particularly in domains requiring transparency, factual grounding, and safety. Unlike general LLMs, RLLMs introduce unique trust challenges due to their explicit multi-step reasoning, reliance on external tools, and complex decision traces. Recent work has begun addressing these issues by enhancing interpretability, managing knowledge conflicts, improving robustness, grounding reasoning in evidence, and enabling controllability.

1) **Interpretability of Reasoning Processes:** RLLMs often produce complex reasoning traces, making it hard to assess whether outputs reflect genuine internal decision-making. Faithfulness metrics have been introduced to evaluate the alignment between explanations and model behavior [444]. Mechanistic analyses explore how reasoning steps are encoded in neurons or attention heads [445], [446]. Other studies examine how gradient dynamics differ when training models for slow, deliberate reasoning [447]. These efforts enhance transparency while preserving performance.
2) **Knowledge Conflicts** Knowledge inconsistencies arise when internal model memory conflicts with external context, undermining reasoning. Recent surveys categorize such conflicts and highlight their impact on reasoning accuracy [448], [449]. Methods have been proposed to detect and resolve these conflicts using special prompt [450], decoding [451], [452] and alignment [453], [454] technologies, while advanced editing approaches [451], [455], [456] allow updated facts to be incorporated into reasoning traces without full model retraining.
3) **Safety & Robustness:** Reasoning enables models to reflect on safety but also opens new vulnerabilities. Alignment procedures often reduce reasoning performance, a trade-off termed "safety tax" [457]. Chain-of-thought attacks, such as prompt hijacking and input slowdown, expose structural vulnerabilities in reasoning flows [458], [459]. Backdoor-based jailbreaks further demonstrate that specific reasoning patterns can trigger harmful behavior [460], [461]. To counter this, researchers propose deliberative safety alignment [462], reasoning-based safeguard modules [463], and dedicated safe-CoT benchmarks [464].
4) **External Grounding and Truthfulness:** RLLMs benefit from grounding reasoning in external evidence to mitigate hallucination. Retrieval-augmented frameworks dynamically adapt context during reasoning [465], [466]. Other prompting techniques first extract chains of evidence from the input before generating an answer [467]. Reinforcement methods reward reasoning paths that include faithful, verifiable external content [468]. Symbolic-verification modules also check reasoning consistency against retrieved data [173].
5) **Consistency and Controllability:** To improve reliability, RLLMs must generate consistent reasoning across runs and support external control. Dual-model self-reflection and critique techniques help detect and revise flawed reasoning [469]. Prompting strategies that separate high-level plans from step-wise solutions improve coherence [470]–[472]. Studies also show that excessively long reasoning chains reduce accuracy, prompting step-length control mechanisms [473].

Across interpretability, safety, grounding, and control, recent research demonstrates growing attention to the trustworthiness of RLLMs. Still, many challenges remain, particularly in balancing reasoning ability with safety, ensuring consistency under open-domain inputs, and aligning long reasoning chains with verified knowledge. Future efforts must continue integrating reasoning supervision, verification, and robust alignment into unified RLLM frameworks.

5.3 Reasoning LLM as Agent

LLMs are increasingly employed as autonomous agents capable of complex reasoning and interaction with external environments. This involves enabling them to actively seek information, utilize diverse tools, and dynamically refine their reasoning. Recent advancements span enhanced search augmentation, broader tool integration, novel training methodologies, and evolving evaluation frameworks.

5.3.1 Search-Augmented Reasoning

Rather than relying solely on internal knowledge, these models are being trained to interact with external search engines to access up-to-date information. Reinforcement learning emerged as a prominent technique for training LLMs to autonomously generate relevant search queries as an integral part of their reasoning chains [474], [475], [476]. Frameworks were developed to incentivize models to invoke search proactively during reasoning, optimizing the interaction based on outcomes rather than predefined processes or distillation [477]. Furthermore, research explored the synergistic potential of combining open-source reasoning LLMs with specialized search tools, demonstrating that such combinations can achieve performance competitive with larger, proprietary systems [478]. The complexity of information retrieval was also addressed through collaborative approaches, where multiple LLM agents work together, potentially using sophisticated search strategies like MCTS, to tackle complex search-based reasoning tasks [479].

5.3.2 Tool Integration Beyond Search

Beyond web search, the scope of tool integration for LLM agents expanded considerably. Efforts focused on enabling LLMs to leverage a wider array of external tools, such as code interpreters and specialized solvers. Self-learning frameworks were introduced, allowing models to acquire tool-using skills through methods like hint-based learning and rejection sampling fine-tuning, effectively teaching them when and how to employ external functionalities [480]. This capability proved valuable in specialized domains; for instance, systems were developed to automate the complex process of modeling and solving Operations Research (OR) problems by translating natural language descriptions into formal mathematical models and generating executable code for solvers [481]. More comprehensive agentic reasoning frameworks emerged, integrating multiple capabilities—including web search [482], [483], code execution [484], and structured reasoning-context memory [485]–[487] (e.g., using a mind map structure)—to support more robust and multifaceted problem-solving.

5.3.3 Training Methodologies

Innovations in training methodologies were crucial for developing these sophisticated agentic capabilities. Research demonstrated the effectiveness of reinforcement learning techniques for enhancing reasoning abilities, even in the context of smaller LLMs, suggesting that advanced reasoning is achievable without massive parameter counts [488]. Further emphasizing self-improvement through reinforcement learning, some approaches employ a two-stage paradigm involving initial format tuning to instill structured reasoning patterns incorporating meta-actions, followed by large-scale RL optimization enabling autoregressive search and self-correction with minimal initial supervision [489]. RL-based training was successfully scaled to complex, specialized domains like software engineering, utilizing lightweight, rule-based reward signals to guide the learning process effectively [490]. Advanced RL strategies, such as explicit policy optimization over multiple interaction turns, were explored to improve strategic reasoning performance without requiring initial supervised fine-tuning phases [491]. Additionally, novel training paradigms were developed to cultivate complex reasoning patterns, such as long chains of thought. Bootstrapping methods, for example, enabled the development of these capabilities without relying heavily on advanced teacher models or extensive human annotations, offering a more scalable approach to training deep reasoning [492].

5.3.4 Evaluation Benchmarks and Metrics

As LLM agent capabilities advanced, evaluation methodologies evolved in tandem. New benchmarks were introduced specifically to assess the complex, multi-step reasoning and tool-use abilities required in real-world applications [493]. Interactive platforms and "arenas" gained prominence, providing environments for systematically comparing the performance of different LLM agents on user-defined tasks, often incorporating community voting and feedback mechanisms for a more holistic evaluation [494], [435], [495], [496]. Alongside these new developments, established benchmarks measuring core reasoning abilities (e.g., AIME [311], GPQA [329]) and code generation/execution (e.g., LiveCodeBench [327], SWE-Bench [326]) continued to serve as important yardsticks for tracking progress in foundational capabilities.

5.4 Efficient Reasoning LLM

Despite LongCoT demonstrating strong generalization and reasoning abilities, the autoregressive paradigm of language models imposes a significant reasoning burden, limiting the application of these models in agent-based or edge scenarios [497]. Additionally, numerous studies have shown that language models often exhibit excessive reasoning, with a lot of redundant inference in the thought chains. As a result, various technical approaches have emerged to improve the reasoning efficiency of Reason LLMs. We categorize these approaches into the following categories based on their strategies:

1) **Building reasoning-budget-sensitive LLMs:** CoD [498] and TALE-EP [499] attempt to impose reasoning budget constraints on Reason LLMs to control the overall reasoning cost. Models like L1 [500], TOPS [501], O1-Pruner [296], and Kimi-k1.5 [258] add length penalties.
2) **Building diverse reasoning-length data for post-training:** CoT-Valve [502] synthesizes diverse-length thought chain data through data interpolation. TOPS samples corresponding versions using a budget-sensitive data model, and C3oT [503] compresses the original LLM output to obtain shorter Short-CoT and trains them jointly.
3) **Using external models or switching mechanisms [504] to intervene in the reasoning scope:** Routellm [505] introduces multiple routers to find the most suitable reasoning model for a given problem, while Self-REF [506] uses confidence to route the model's reasoning difficulty. Methods like JudgeLRM [507]–[509] and DeepSeek-GRM [166] attempt to construct generalized Reward Models for efficient test-time scaling, eliminating irrelevant reasoning paths.
4) **Using efficient representations to execute reasoning:** TokenSkip selects data based on the importance of tokens and executes reasoning using compressed, more concise thought chain representations. COCONUT [510] attempts to execute more efficient reasoning in the latent space. ICoT-KD [511] and CCoT [511] try to build more efficient reasoning strategies in the hidden space, while Token Assorted combines hidden space reasoning with text-based reasoning to

balance interpretability and efficiency. Heima aims to build hidden-space reasoning for multimodal models.

5) **Non-autoregressive reasoning models:** Diffusion-LM avoids a large number of autoregressive prediction steps due to its non-autoregressive nature, enabling efficient reasoning. LLaDA [512] constructs a diffusion-based LM with an enormous parameter size, achieving scalability based on a non-autoregressive language model architecture. Diffusion-of-Thoughts [513] uses post-training strategies to transform an LLM into a denoising process for token prediction, significantly improving reasoning efficiency.
6) **New architectures for efficient reasoning:** One of the key bottlenecks in long thought-chain prediction is the computational cost of ultra-long context. Mamba [514] and RWKV [515], among others, use linear attention or state space techniques to effectively enhance the model's reasoning efficiency.

6 Challenges & Future Directions

Despite the rapid advancements in reasoning LLMs, several challenges persist, limiting their generalizability and practical applicability. This section outlines these challenges and highlights potential research directions to address them.

6.1 Efficient Reasoning LLMs

While reasoning LLMs excel at solving complex problems via extended inference, their reliance on long autoregressive reasoning within large-scale architectures presents significant efficiency challenges [516], [517]. For example, many problems on platforms like Codeforces require over 10,000 tokens of reasoning, resulting in high latency. As noted in [124], even when a reasoning LLM identifies the correct solution early, it often spends considerable time verifying its reasoning. Recent reports, such as Deepseek-R1 [31], suggest that self-improvement via RL is more effective in larger models, while smaller-scale large language models (SLMs) (*e.g.*, 3B and 7B models as explored by [125] and [265], [290]) struggle to match performance in slow-thinking reasoning tasks.

Future research should focus on two key areas: (1) integrating external reasoning tools to enable early stopping and verification mechanisms, thus improving the efficiency of long inference chains, and (2) exploring strategies to implement slow-thinking reasoning capabilities in SLMs without sacrificing performance.

6.2 Collaborative Slow & Fast-thinking Systems

A key challenge in reasoning LLMs is the loss of fast-thinking capabilities, which results in inefficiencies when simple tasks require unnecessary deep reasoning. Unlike humans, who fluidly switch between fast (*System 1*) and slow (*System 2*) thinking, current reasoning LLMs struggle to maintain this balance. While reasoning LLMs ensure deliberate and thorough reasoning, fast-thinking systems rely on prior knowledge for quick responses. Despite efforts such as the *System 1-2* switcher [116], speculative decoding [518]–[520], and interactive continual learning [521]–[523], integrating both modes of thinking remains challenging. This often leads to inefficiencies in domain-specific tasks and underutilized strengths in more complex scenarios.

Future research should focus on developing adaptive switching mechanisms, joint training frameworks, and co-evolution strategies to harmonize the efficiency of fast-thinking systems with the precision of reasoning LLMs. Achieving this balance is crucial for advancing the field and creating more versatile AI systems.

6.3 Reasoning LLMs For Science

Reasoning LLMs play a crucial role in scientific research [524], [525], enabling deep, structured analysis that goes beyond the heuristic-based fast-thinking models. Their value becomes especially clear in fields that demand complex reasoning, such as medicine and mathematics. In medicine, particularly in differential diagnosis and treatment planning, reasoning LLMs (*e.g.*, inference-time scaling) enhance AI's step-by-step reasoning, improving diagnostic accuracy where traditional scaling methods fall short [52]. In mathematics, approaches like FunSearch [526] incorporate slow-thinking principles to push beyond previous discoveries, showcasing the potential of AI-human collaboration.

Beyond these fields, reasoning LLMs can foster advancements in physics, engineering, and computational biology by refining model formulation and hypothesis testing. Investing in reasoning LLMs research not only bridges the gap between AI's computational power and human-like analytical depth but also paves the way for more reliable, interpretable, and groundbreaking scientific discoveries.

6.4 Deep Integration of Neural and Symbolic Systems

Despite significant advancements in reasoning LLMs, their limited transparency and interpretability restrict their performance in more complex real-world reasoning tasks. The reliance on large-scale data patterns and lack of clear reasoning pathways makes it challenging to handle intricate or ambiguous problems effectively. Early symbolic logic systems, while less adaptable, offered better explainability and clearer reasoning steps, leading to more reliable performance in such cases.

A promising future direction is the deep integration of neural and symbolic systems. Google's AlphaGeometry [527] and AlphaGeometry2 [528] combine reasoning LLMs with symbolic engines, achieving breakthroughs in the International Olympiad in Mathematics (IMO). In particular, AlphaGeometry2 utilizes the Gemini-based model [409], [529], [530] and a more efficient symbolic engine, improving performance by reducing rule sets and enhancing key concept handling [531]–[533]. The system now covers a broader range of geometric concepts, including locus theorems and linear equations. A new search algorithm and knowledge-sharing mechanism accelerate the process. This system solved 84% of IMO geometry problems (2000-2024), surpassing gold medalists' averages. In contrast, reasoning LLMs like OpenAI-o1 [29] failed to solve any problems. The integration of neural and symbolic systems offers a balanced approach, improving both adaptability and interpretability, with vast potential for complex real-world reasoning tasks beyond mathematical geometry problems.

6.5 Multilingual Reasoning LLMs

Current reasoning LLMs perform well in high-resource languages like English and Chinese, demonstrating strong capabilities in tasks such as translation and various reasoning tasks [112], [123], [534]. These models excel in environments where large-scale data and diverse linguistic resources are available. However, their performance in low-resource languages remains limited [535], facing challenges related to data sparsity, stability, safety, and overall performance. These issues hinder the effectiveness of reasoning LLMs in languages that lack substantial linguistic datasets and resources.

Future research should prioritize overcoming the challenges posed by data scarcity and cultural biases in low-resource languages. Innovations such as parameter sharing across reasoning LLMs and the incremental injection of domain-specific knowledge could help mitigate these challenges, enabling faster adaptation of slow-thinking capabilities to a broader range of languages. This would not only enhance the effectiveness of reasoning LLMs in these languages but also ensure more equitable access to advanced AI technologies.

6.6 Safe Reasoning LLMs

The rapid development of reasoning LLMs like OpenAI-o1 [29] and DeepSeek-R1 [31] has led to the rise of superintelligent models capable of continuous self-evolution. However, this progress brings challenges in safety and control [536]–[538]. RL, a key training method, introduces risks such as reward hacking, generalization failures, and language mixing, which can lead to harmful outcomes. Ensuring the safety of such systems like DeepSeek-R1 is urgent. While RL enhances reasoning, its uncontrollable nature raises concerns about safely guiding these models. SFT addresses some issues but is not a complete solution. A hybrid approach combining RL and SFT is needed to reduce harmful outputs while maintaining model effectiveness [539], [540].

As these models surpass human cognitive capabilities, ensuring their safe, responsible, and transparent use is crucial. This requires ongoing research to develop methods for controlling and guiding their actions, thereby balancing AI power with ethical decision-making.

7 Conclusion

This paper presents a comprehensive survey that advances research on reasoning LLMs. We begin with an overview of the progress in foundational LLMs and key early *System 2* technologies, including symbolic logic, MCTS, and RL, exploring how each, when combined with foundational LLMs, has paved the way for reasoning LLMs. We then provide a detailed feature analysis of the latest reasoning LLMs, examining the core methods that enable their advanced reasoning capabilities and highlighting representative models. Through a review of mainstream reasoning benchmarks and performance comparisons, we offer valuable insights into the current state of the field. Looking ahead, we identify promising research directions and continue to track developments via our real-time GitHub Repository. This survey aims to inspire innovation and foster progress in the rapidly evolving field of reasoning LLMs.

References

[1] W. Hua and Y. Zhang, “System 1+ system 2= better world: Neural-symbolic chain of logic reasoning,” in *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022, pp. 601–612.

[2] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[3] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou, “Self-Consistency Improves Chain of Thought Reasoning in Language Models,” in *The Eleventh International Conference on Learning Representations*, 2023.

[4] D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. V. Le *et al.*, “Least-to-Most Prompting Enables Complex Reasoning in Large Language Models,” in *The Eleventh International Conference on Learning Representations*, 2023.

[5] E. Zelikman, Y. Wu, J. Mu, and N. D. Goodman, “STaR: Self-taught reasoner bootstrapping reasoning with reasoning,” in *Proc. the 36th International Conference on Neural Information Processing Systems*, vol. 1126, 2024.

[6] J. S. B. Evans, “Heuristic and analytic processes in reasoning,” *British Journal of Psychology*, vol. 75, no. 4, pp. 451–468, 1984.

[7] D. Kahneman, “Maps of bounded rationality: Psychology for behavioral economics,” *American economic review*, vol. 93, no. 5, pp. 1449–1475, 2003.

[8] J. Huang and K. C.-C. Chang, “Towards Reasoning in Large Language Models: A Survey,” in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 1049–1065.

[9] S. Qiao, Y. Ou, N. Zhang, X. Chen, Y. Yao, S. Deng, C. Tan, F. Huang, and H. Chen, “Reasoning with Language Model Prompting: A Survey,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 5368–5393.

[10] B. Wang, S. Min, X. Deng, J. Shen, Y. Wu, L. Zettlemoyer, and H. Sun, “Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 2717–2739.

[11] O. Shaikh, H. Zhang, W. Held, M. Bernstein, and D. Yang, “On Second Thought, Let’s Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 4454–4470.

[12] H. Shao, S. Qian, H. Xiao, G. Song, Z. Zong, L. Wang, Y. Liu, and H. Li, “Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning,” in *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[13] Z. Zhang, A. Zhang, M. Li, and A. Smola, “Automatic Chain of Thought Prompting in Large Language Models,” in *The Eleventh International Conference on Learning Representations*, 2023.

[14] S. Hao, Y. Gu, H. Ma, J. Hong, Z. Wang, D. Wang, and Z. Hu, “Reasoning with Language Model is Planning with World Model,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 8154–8173.

[15] Y. Zhang, “Meta prompting for agi systems,” *arXiv preprint arXiv:2311.11482*, 2023.

[16] OpenAI, “Hello GPT-4o,” May 2024. [Online]. Available: https://openai.com/index/hello-gpt-4o/

[17] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan *et al.*, “Deepseek-v3 technical report,” *arXiv preprint arXiv:2412.19437*, 2024.

[18] A. Vaswani, “Attention is all you need,” *Advances in Neural Information Processing Systems*, 2017.

[19] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[20] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” *CoRR*, vol. abs/1907.11692, 2019.

[21] A. Radford, "Improving language understanding by generative pre-training," 2018.

[22] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[23] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[24] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[25] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[26] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[27] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[28] D. Zhang, Y. Yu, J. Dong, C. Li, D. Su, C. Chu, and D. Yu, "MM-LLMs: Recent Advances in MultiModal Large Language Models," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*. Association for Computational Linguistics, 2024, pp. 12 401–12 430.

[29] OpenAI, "Learning to reason with LLMs," September 2024. [Online]. Available: https://openai.com/index/learning-to-reason-with-llms/

[30] ——, "OpenAI o3-mini," January 2025. [Online]. Available: https://openai.com/index/openai-o3-mini/

[31] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning," *arXiv preprint arXiv:2501.12948*, 2025.

[32] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano *et al.*, "Training verifiers to solve math word problems," *arXiv preprint arXiv:2110.14168*, 2021.

[33] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in neural information processing systems*, vol. 35, pp. 22 199–22 213, 2022.

[34] Y. Liu, A. Singh, C. D. Freeman, J. D. Co-Reyes, and P. J. Liu, "Improving large language model fine-tuning for solving math problems," *arXiv preprint arXiv:2310.10047*, 2023.

[35] X. Zhu, J. Wang, L. Zhang, Y. Zhang, Y. Huang, R. Gan, J. Zhang, and Y. Yang, "Solving Math Word Problems via Cooperative Reasoning induced Language Models," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 4471–4485.

[36] P. Lu, L. Qiu, K.-W. Chang, Y. N. Wu, S.-C. Zhu, T. Rajpurohit, P. Clark, and A. Kalyan, "Dynamic Prompt Learning via Policy Gradient for Semi-structured Mathematical Reasoning," in *The Eleventh International Conference on Learning Representations*, 2023.

[37] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe, "Let's Verify Step by Step," in *The Twelfth International Conference on Learning Representations*, 2024.

[38] F. Yao, C. Tian, J. Liu, Z. Zhang, Q. Liu, L. Jin, S. Li, X. Li, and X. Sun, "Thinking like an expert: Multimodal hypergraph-of-thought (hot) reasoning to boost foundation modals," *arXiv preprint arXiv:2308.06207*, 2023.

[39] Y. Yao, Z. Li, and H. Zhao, "Beyond Chain-of-Thought, Effective Graph-of-Thought Reasoning in Language Models," *arXiv preprint arXiv:2305.16582*, 2023.

[40] Y. Wen, Z. Wang, and J. Sun, "Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models," *arXiv preprint arXiv:2308.09729*, 2023.

[41] B. Lei, C. Liao, C. Ding *et al.*, "Boosting logical reasoning in large language models through a new framework: The graph of thought," *arXiv preprint arXiv:2308.08614*, 2023.

[42] M. Jin, Q. Yu, D. Shu, H. Zhao, W. Hua, Y. Meng, Y. Zhang, and M. Du, "The impact of reasoning step length on large language models," *arXiv preprint arXiv:2401.04925*, 2024.

[43] M. Besta, N. Blach, A. Kubicek, R. Gerstenberger, M. Podstawski, L. Gianinazzi, J. Gajda, T. Lehmann, H. Niewiadomski, P. Nyczyk *et al.*, "Graph of thoughts: Solving elaborate problems with large language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, 2024, pp. 17 682–17 690.

[44] P. Cheng, T. Hu, H. Xu, Z. Zhang, Y. Dai, L. Han, and N. Du, "Self-playing Adversarial Language Game Enhances LLM Reasoning," *arXiv preprint arXiv:2404.10642*, 2024.

[45] H. You, R. Sun, Z. Wang, L. Chen, G. Wang, H. Ayyubi, K.-W. Chang, and S.-F. Chang, "IdealGPT: Iteratively Decomposing Vision and Language Reasoning via Large Language Models," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 11 289–11 303.

[46] P. Wu and S. Xie, "V?: Guided Visual Search as a Core Mechanism in Multimodal LLMs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 084–13 094.

[47] Z. Chen, R. Sun, W. Liu, Y. Hong, and C. Gan, "GENOME: Generative Neuro-Symbolic Visual Reasoning by Growing and Reusing Modules," in *International Conference on Learning Representations*, 2024.

[48] S. Wu, Z. Peng, X. Du, T. Zheng, M. Liu, J. Wu, J. Ma, Y. Li, J. Yang, W. Zhou *et al.*, "A Comparative Study on Reasoning Patterns of OpenAI's o1 Model," *arXiv preprint arXiv:2410.13639*, 2024.

[49] V. Xiang, C. Snell, K. Gandhi, A. Albalak, A. Singh, C. Blagden, D. Phung, R. Rafailov, N. Lile, D. Mahan *et al.*, "Towards System 2 Reasoning in LLMs: Learning How to Think With Meta Chain-of-Though," *arXiv preprint arXiv:2501.04682*, 2025.

[50] Y. Qin, X. Li, H. Zou, Y. Liu, S. Xia, Z. Huang, Y. Ye, W. Yuan, H. Liu, Y. Li *et al.*, "O1 Replication Journey: A Strategic Progress Report–Part 1," *arXiv preprint arXiv:2410.18982*, 2024.

[51] Z. Huang, H. Zou, X. Li, Y. Liu, Y. Zheng, E. Chern, S. Xia, Y. Qin, W. Yuan, and P. Liu, "O1 Replication Journey–Part 2: Surpassing O1-preview through Simple Distillation, Big Progress or Bitter Lesson?" *arXiv preprint arXiv:2411.16489*, 2024.

[52] Z. Huang, G. Geng, S. Hua, Z. Huang, H. Zou, S. Zhang, P. Liu, and X. Zhang, "O1 Replication Journey–Part 3: Inference-time Scaling for Medical Reasoning," *arXiv preprint arXiv:2501.06458*, 2025.

[53] Y. Min, Z. Chen, J. Jiang, J. Chen, J. Deng, Y. Hu, Y. Tang, J. Wang, X. Cheng, H. Song, W. X. Zhao, Z. Liu, Z. Wang, and J.-R. Wen, "Imitate, Explore, and Self-Improve: A Reproduction Report on Slow-thinking Reasoning Systems," *arXiv preprint arXiv:2412.09413*, 2024.

[54] H. Xu, X. Wu, W. Wang, Z. Li, D. Zheng, B. Chen, Y. Hu, S. Kang, J. Ji, Y. Zhang *et al.*, "RedStar: Does Scaling Long-CoT Data Unlock Better Slow-Reasoning Systems?" *arXiv preprint arXiv:2501.11284*, 2025.

[55] Z. Zeng, Q. Cheng, Z. Yin, B. Wang, S. Li, Y. Zhou, Q. Guo, X. Huang, and X. Qiu, "Scaling of Search and Learning: A Roadmap to Reproduce o1 from Reinforcement Learning Perspective," *arXiv preprint arXiv:2412.14135*, 2024.

[56] Y. Ji, J. Li, H. Ye, K. Wu, J. Xu, L. Mo, and M. Zhang, "Test-time Computing: from System-1 Thinking to System-2 Thinking," *arXiv preprint arXiv:2501.02497*, 2025.

[57] M. Besta, J. Barth, E. Schreiber, A. Kubicek, A. Catarino, R. Gerstenberger, P. Nyczyk, P. Iff, Y. Li, S. Houliston *et al.*, "Reasoning Language Models: A Blueprint," *arXiv preprint arXiv:2501.11223*, 2025.

[58] Y. Zhang, S. Mao, T. Ge, X. Wang, A. de Wynter, Y. Xia, W. Wu, T. Song, M. Lan, and F. Wei, "LLM as a Mastermind: A Survey of Strategic Reasoning with Large Language Models," *arXiv preprint arXiv:2404.01230*, 2024.

[59] F. Xu, Q. Hao, Z. Zong, J. Wang, Y. Zhang, J. Wang, X. Lan, J. Gong, T. Ouyang, F. Meng *et al.*, "Towards Large Reasoning Models: A Survey of Reinforced Reasoning with Large Language Models," *arXiv preprint arXiv:2501.09686*, 2025.

[60] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[61] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation,"

in *International conference on machine learning*. Pmlr, 2021, pp. 8821–8831.

[62] OpenAI, "GPT-4 Technical Report," 2023.

[63] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 716–23 736, 2022.

[64] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," in *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, 2023, pp. 19 730–19 742.

[65] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. C. H. Hoi, "InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[66] J. He, J. Qiu, A. Zeng, Z. Yang, J. Zhai, and J. Tang, "Fastmoe: A fast mixture-of-expert training system," *arXiv preprint arXiv:2103.13262*, 2021.

[67] N. Du, Y. Huang, A. M. Dai, S. Tong, D. Lepikhin, Y. Xu, M. Krikun, Y. Zhou, A. W. Yu, O. Firat *et al.*, "Glam: Efficient scaling of language models with mixture-of-experts," in *International conference on machine learning*. PMLR, 2022, pp. 5547–5569.

[68] D. Dai, C. Deng, C. Zhao, R. Xu, H. Gao, D. Chen, J. Li, W. Zeng, X. Yu, Y. Wu *et al.*, "DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 1280–1297.

[69] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[70] Y. LeCun, Y. Bengio *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[71] S. Hochreiter, "Long Short-term Memory," *Neural Computation MIT-Press*, 1997.

[72] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[73] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[74] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[75] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[76] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, 2014, pp. 1724–1734.

[77] I. Sutskever, "Sequence to Sequence Learning with Neural Networks," *arXiv preprint arXiv:1409.3215*, 2014.

[78] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[79] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[80] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[81] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[82] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[83] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[84] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer, "Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?" in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 11 048–11 064.

[85] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, B. Chang *et al.*, "A survey on in-context learning," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 1107–1128.

[86] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt, "A prompt pattern catalog to enhance prompt engineering with chatgpt," *arXiv preprint arXiv:2302.11382*, 2023.

[87] B. Lester, R. Al-Rfou, and N. Constant, "The Power of Scale for Parameter-Efficient Prompt Tuning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 3045–3059.

[88] C. I. Lewis, C. H. Langford, and P. Lamprecht, *Symbolic logic*. Dover publications New York, 1959, vol. 170.

[89] R. Carnap, *Introduction to symbolic logic and its applications*. Courier Corporation, 2012.

[90] A. Colmerauer, "An introduction to Prolog III," *Communications of the ACM*, vol. 33, no. 7, pp. 69–90, 1990.

[91] W. F. Clocksin and C. S. Mellish, *Programming in PROLOG*. Springer Science & Business Media, 2003.

[92] K. R. Apt *et al.*, *From logic programming to Prolog*. Prentice Hall London, 1997, vol. 362.

[93] M. P. Singh, A. S. Rao, and M. P. Georgeff, *Formal methods in DAI: Logic-based representation and reasoning*. MIT Press Cambridge, 1999, vol. 8.

[94] R. G. Jeroslow, "Computation-oriented reductions of predicate to propositional logic," *Decision Support Systems*, vol. 4, no. 2, pp. 183–197, 1988.

[95] J. McCarthy, "History of LISP," in *History of programming languages*, 1978, pp. 173–185.

[96] L. Bachmair and H. Ganzinger, "Resolution Theorem Proving." *Handbook of automated reasoning*, vol. 1, no. 02, 2001.

[97] M. Minsky *et al.*, "A framework for representing knowledge," 1974.

[98] C. B. Browne, E. Powley, D. Whitehouse, S. M. Lucas, P. I. Cowling, P. Rohlfshagen, S. Tavener, D. Perez, S. Samothrakis, and S. Colton, "A survey of monte carlo tree search methods," *IEEE Transactions on Computational Intelligence and AI in games*, vol. 4, no. 1, pp. 1–43, 2012.

[99] S. Gelly and D. Silver, "Monte-Carlo tree search and rapid action value estimation in computer Go," *Artificial Intelligence*, vol. 175, no. 11, pp. 1856–1875, 2011.

[100] M. Świechowski, K. Godlewski, B. Sawicki, and J. Mańdziuk, "Monte Carlo tree search: A review of recent modifications and applications," *Artificial Intelligence Review*, vol. 56, no. 3, pp. 2497–2562, 2023.

[101] R. S. Sutton, A. G. Barto *et al.*, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998, vol. 1, no. 1.

[102] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, pp. 279–292, 1992.

[103] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[104] R. R. Torrado, P. Bontrager, J. Togelius, J. Liu, and D. Perez-Liebana, "Deep reinforcement learning for general video game ai," in *2018 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE, 2018, pp. 1–8.

[105] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of Go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[106] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton *et al.*, "Mastering the game of go without human knowledge," *nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[107] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev *et al.*, "Grandmaster level in StarCraft II using multi-agent reinforcement learning," *nature*, vol. 575, no. 7782, pp. 350–354, 2019.

[108] A. Y. Ng, D. Harada, and S. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in *Icml*, vol. 99. Citeseer, 1999, pp. 278–287.

[109] H. Luo, Q. Sun, C. Xu, P. Zhao, J. Lou, C. Tao, X. Geng, Q. Lin, S. Chen, and D. Zhang, "Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct," *arXiv preprint arXiv:2308.09583*, 2023.

[110] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu *et al.*, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," *arXiv preprint arXiv:2402.03300*, 2024.

[111] E. Zelikman, G. Harik, Y. Shao, V. Jayasiri, N. Haber, and N. D. Goodman, "Quiet-star: Language models can teach themselves to think before speaking," *arXiv preprint arXiv:2403.09629*, 2024.

[112] Y. Zhao, H. Yin, B. Zeng, H. Wang, T. Shi, C. Lyu, L. Wang, W. Luo, and K. Zhang, "Marco-o1: Towards open reasoning models for open-ended solutions," *arXiv preprint arXiv:2411.14405*, 2024.

[113] J. Chen, Z. Cai, K. Ji, X. Wang, W. Liu, R. Wang, J. Hou, and B. Wang, "Huatuogpt-o1, towards medical complex reasoning with llms," *arXiv preprint arXiv:2412.18925*, 2024.

[114] Z. Tang, Z. Chen, L. Li, X. Song, Y. Deng, Y. Shen, G. Chen, P. Spirtes, and K. Zhang, "Reflection-window decoding: Text generation with selective refinement," *arXiv preprint arXiv:2502.03678*, 2025.

[115] F. Teng, Z. Yu, Q. Shi, J. Zhang, C. Wu, and Y. Luo, "Atom of thoughts for markov llm test-time scaling," *arXiv preprint arXiv:2502.12018*, 2025.

[116] G. Sun, M. Jin, Z. Wang, C.-L. Wang, S. Ma, Q. Wang, Y. N. Wu, Y. Zhang, and D. Liu, "Visual agents as fast and slow thinkers," *arXiv preprint arXiv:2408.08862*, 2024.

[117] H. Wei, Y. Yin, Y. Li, J. Wang, L. Zhao, J. Sun, Z. Ge, and X. Zhang, "Slow Perception: Let's Perceive Geometric Figures Step-by-step," *arXiv preprint arXiv:2412.20631*, 2024.

[118] X. Li, G. Dong, J. Jin, Y. Zhang, Y. Zhou, Y. Zhu, P. Zhang, and Z. Dou, "Search-o1: Agentic search-enhanced large reasoning models," *arXiv preprint arXiv:2501.05366*, 2025.

[119] X. Tang, D. Shao, J. Sohn, J. Chen, J. Zhang, J. Xiang, F. Wu, Y. Zhao, C. Wu, W. Shi, A. Cohan, and M. Gerstein, "Medagentsbench: Benchmarking thinking models and agent frameworks for complex medical reasoning," 2025. [Online]. Available: https://arxiv.org/abs/2503.07459

[120] Q. Team, "QwQ: Reflect Deeply on the Boundaries of the Unknown," November 2024. [Online]. Available: https://qwenlm.github.io/blog/qwq-32b-preview/

[121] N. Team, "Sky-T1: Train your own O1 preview model within $450," 2025, accessed: 2025-01-09. [Online]. Available: https://novasky-ai.github.io/posts/sky-t1

[122] Y. Zhang, S. Wu, Y. Yang, J. Shu, J. Xiao, C. Kong, and J. Sang, "o1-coder: an o1 replication for coding," *arXiv preprint arXiv:2412.00154*, 2024.

[123] J. Wang, F. Meng, Y. Liang, and J. Zhou, "DRT-o1: Optimized Deep Reasoning Translation via Long Chain-of-Thought," *arXiv preprint arXiv:2412.17498*, 2024.

[124] X. Chen, J. Xu, T. Liang, Z. He, J. Pang, D. Yu, L. Song, Q. Liu, M. Zhou, Z. Zhang *et al.*, "Do NOT Think That Much for 2 + 3=? On the Overthinking of o1-Like LLMs," *arXiv preprint arXiv:2412.21187*, 2024.

[125] W. Zeng, Y. Huang, W. Liu, K. He, Q. Liu, Z. Ma, and J. He, "7B Model and 8K Examples: Emerging Reasoning with Reinforcement Learning is Both Effective and Efficient," 2025, notion Blog. [Online]. Available: https://hkust-nlp.notion.site/simplerl-reason

[126] Z. Wan, X. Feng, M. Wen, S. M. McAleer, Y. Wen, W. Zhang, and J. Wang, "Alphazero-like tree-search can guide large language model decoding and training," in *Forty-first International Conference on Machine Learning*, 2024.

[127] Z. Bi, K. Han, C. Liu, Y. Tang, and Y. Wang, "Forest-of-Thought: Scaling Test-Time Compute for Enhancing LLM Reasoning," *CoRR*, vol. abs/2412.09078, 2024.

[128] J. Li, H. Le, Y. Zhou, C. Xiong, S. Savarese, and D. Sahoo, "CodeTree: Agent-guided Tree Search for Code Generation with Large Language Models," *CoRR*, vol. abs/2411.04329, 2024.

[129] J. Qiu, Y. Lu, Y. Zeng, J. Guo, J. Geng, H. Wang, K. Huang, Y. Wu, and M. Wang, "TreeBoN: Enhancing Inference-Time Alignment with Speculative Tree-Search and Best-of-N Sampling," *CoRR*, vol. abs/2410.16033, 2024.

[130] N. Dainese, M. Merler, M. Alakuijala, and P. Marttinen, "Generating Code World Models with Large Language Models Guided by Monte Carlo Tree Search," *CoRR*, vol. abs/2405.15383, 2024.

[131] Z. Zhao, W. S. Lee, and D. Hsu, "Large Language Models as Commonsense Knowledge for Large-Scale Task Planning," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[132] Q. Li, W. Xia, K. Du, X. Dai, R. Tang, Y. Wang, Y. Yu, and W. Zhang, "RethinkMCTS: Refining Erroneous Thoughts in Monte Carlo Tree Search for Code Generation," *CoRR*, vol. abs/2409.09584, 2024.

[133] D. Zhang, X. Huang, D. Zhou, Y. Li, and W. Ouyang, "Accessing GPT-4 level Mathematical Olympiad Solutions via Monte Carlo Tree Self-refine with LLaMa-3 8B," *CoRR*, vol. abs/2406.07394, 2024.

[134] G. Rabby, F. Keya, P. Zamil, and S. Auer, "MC-NEST - Enhancing Mathematical Reasoning in Large Language Models with a Monte Carlo Nash Equilibrium Self-Refine Tree," *CoRR*, vol. abs/2411.15645, 2024.

[135] B. Xu, Y. Lin, Y. Li, and Y. Gao, "SRA-MCTS: Self-driven Reasoning Augmentation with Monte Carlo Tree Search for Code Generation," *CoRR*, vol. abs/2411.11053, 2024.

[136] J. Cheng, X. Liu, C. Wang, X. Gu, Y. Lu, D. Zhang, Y. Dong, J. Tang, H. Wang, and M. Huang, "SPaR: Self-Play with Tree-Search Refinement to Improve Instruction-Following in Large Language Models," *CoRR*, vol. abs/2412.11605, 2024.

[137] J. Kang, X. Z. Li, X. Chen, A. Kazemi, and B. Chen, "MindStar: Enhancing Math Reasoning in Pre-trained LLMs at Inference Time," *CoRR*, vol. abs/2405.16265, 2024.

[138] P. Kadam, "GPT-Guided Monte Carlo Tree Search for Symbolic Regression in Financial Fraud Detection," *CoRR*, vol. abs/2411.04459, 2024.

[139] D. Zhang, J. Wu, J. Lei, T. Che, J. Li, T. Xie, X. Huang, S. Zhang, M. Pavone, Y. Li, W. Ouyang, and D. Zhou, "LLaMA-Berry: Pairwise Optimization for O1-like Olympiad-Level Mathematical Reasoning," *CoRR*, vol. abs/2410.02884, 2024.

[140] D. Zhang, S. Zhoubian, Y. Yue, Y. Dong, and J. Tang, "ReST-MCTS*: LLM Self-Training via Process Reward Guided Tree Search," *CoRR*, vol. abs/2406.03816, 2024.

[141] H. Yao, J. Huang, W. Wu, J. Zhang, Y. Wang, S. Liu, Y. Wang, Y. Song, H. Feng, L. Shen, and D. Tao, "Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search," 2024. [Online]. Available: https://arxiv.org/abs/2412.18319

[142] Q. Lin, B. Xu, Z. Li, Z. Hao, K. Zhang, and R. Cai, "Leveraging constrained monte carlo tree search to generate reliable long chain-of-thought for mathematical reasoning," 2025. [Online]. Available: https://arxiv.org/abs/2502.11169

[143] X. Guan, L. L. Zhang, Y. Liu, N. Shang, Y. Sun, Y. Zhu, F. Yang, and M. Yang, "rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking," 2025. [Online]. Available: https://arxiv.org/abs/2501.04519

[144] J. Wu, M. Feng, S. Zhang, R. Jin, F. Che, Z. Wen, and J. Tao, "Boosting multimodal reasoning with mcts-automated structured thinking," 2025. [Online]. Available: https://arxiv.org/abs/2502.02339

[145] Z. Li, H. Yu, X. Chen, H. Lin, Y. Lu, F. Huang, X. Han, Y. Li, and L. Sun, "Deepsolution: Boosting complex engineering solution design via tree-based exploration and bi-point thinking," 2025. [Online]. Available: https://arxiv.org/abs/2502.20730

[146] Y. Wang, S. Wang, Q. Cheng, Z. Fei, L. Ding, Q. Guo, D. Tao, and X. Qiu, "Visuothink: Empowering lvlm reasoning with multimodal tree search," 2025. [Online]. Available: https://arxiv.org/abs/2504.09130

[147] C. Zhang, J. Song, S. Li, Y. Liang, Y. Ma, W. Wang, Y. Zhu, and S.-C. Zhu, "Proposing and solving olympiad geometry with guided tree search," 2024. [Online]. Available: https://arxiv.org/abs/2412.10673

[148] J. Liu, A. Cohen, R. Pasunuru, Y. Choi, H. Hajishirzi, and A. 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*, 2024. [Online]. Available: https://openreview.net/forum?id=kh9Zt2Ldmn

[149] Y. Xie, A. Goyal, W. Zheng, M. Kan, T. P. Lillicrap, K. Kawaguchi, and M. Shieh, “Monte Carlo Tree Search Boosts Reasoning via Iterative Preference Learning,” *CoRR*, vol. abs/2405.00451, 2024.

[150] J. Y. Koh, S. McAleer, D. Fried, and R. Salakhutdinov, “Tree Search for Language Model Agents,” *CoRR*, vol. abs/2407.01476, 2024.

[151] J. Liu, A. Cohen, R. Pasunuru, Y. Choi, H. Hajishirzi, and A. Celikyilmaz, “Don’t throw away your value model! Generating more preferable text with Value-Guided Monte-Carlo Tree Search decoding,” 2024. [Online]. Available: https://arxiv.org/abs/2309.15028

[152] C. Zhang, J. Song, S. Li, Y. Liang, Y. Ma, W. Wang, Y. Zhu, and S. Zhu, “Proposing and solving olympiad geometry with guided tree search,” *CoRR*, vol. abs/2412.10673, 2024.

[153] H. Jiang, Y. Ma, C. Ding, K. Luan, and X. Di, “Towards Intrinsic Self-Correction Enhancement in Monte Carlo Tree Search Boosted Reasoning via Iterative Preference Learning,” 2024. [Online]. Available: https://arxiv.org/abs/2412.17397

[154] H. Xu, “No Train Still Gain. Unleash Mathematical Reasoning of Large Language Models with Monte Carlo Tree Search Guided by Energy Function,” *CoRR*, vol. abs/2309.03224, 2023.

[155] M. Kemmerling, D. Lütticke, and R. H. Schmitt, “Beyond games: a systematic review of neural Monte Carlo tree search applications,” *Appl. Intell.*, vol. 54, no. 11-12, pp. 1020–1046, 2024.

[156] J. Uesato, N. Kushman, R. Kumar, F. Song, N. Siegel, L. Wang, A. Creswell, G. Irving, and I. Higgins, “Solving math word problems with process-and outcome-based feedback,” *arXiv preprint arXiv:2211.14275*, 2022.

[157] F. Yu, A. Gao, and B. Wang, “OVM, Outcome-supervised Value Models for Planning in Mathematical Reasoning,” in *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024, pp. 858–875.

[158] X. Li, X. Chen, J. Fan, E. H. Jiang, and M. Gao, “Multi-head reward aggregation guided by entropy,” 2025. [Online]. Available: https://arxiv.org/abs/2503.20995

[159] Y. Li, Z. Lin, S. Zhang, Q. Fu, B. Chen, J.-G. Lou, and W. Chen, “Making large language models better reasoners with step-aware verifier,” *arXiv preprint arXiv:2206.02336*, 2022.

[160] P. Wang, L. Li, Z. Shao, R. Xu, D. Dai, Y. Li, D. Chen, Y. Wu, and Z. Sui, “Math-shepherd: Verify and reinforce llms step-by-step without human annotations,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 9426–9439.

[161] J. Lu, Z. Dou, W. Hongru, Z. Cao, J. Dai, Y. Feng, and Z. Guo, “Autopsv: Automated process-supervised verifier,” in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[162] L. Yuan, W. Li, H. Chen, G. Cui, N. Ding, K. Zhang, B. Zhou, Z. Liu, and H. Peng, “Free process rewards without process labels,” *arXiv preprint arXiv:2412.01981*, 2024.

[163] Z. Yu, W. Gu, Y. Wang, Z. Zeng, J. Wang, W. Ye, and S. Zhang, “Outcome-Refining Process Supervision for Code Generation,” *arXiv preprint arXiv:2412.15118*, 2024.

[164] S. She, J. Liu, Y. Liu, J. Chen, X. Huang, and S. Huang, “R-prm: Reasoning-driven process reward modeling,” 2025. [Online]. Available: https://arxiv.org/abs/2503.21295

[165] W. Chen, W. He, Z. Xi, H. Guo, B. Hong, J. Zhang, R. Zheng, N. Li, T. Gui, Y. Li, Q. Zhang, and X. Huang, “Better process supervision with bi-directional rewarding signals,” 2025. [Online]. Available: https://arxiv.org/abs/2503.04618

[166] Z. Liu, P. Wang, R. Xu, S. Ma, C. Ruan, P. Li, Y. Liu, and Y. Wu, “Inference-time scaling for generalist reward modeling,” 2025. [Online]. Available: https://arxiv.org/abs/2504.02495

[167] H. Peng, Y. Qi, X. Wang, Z. Yao, B. Xu, L. Hou, and J. Li, “Agentic reward modeling: Integrating human preferences with verifiable correctness signals for reliable reward systems,” 2025. [Online]. Available: https://arxiv.org/abs/2502.19328

[168] J. Fu, X. Zhao, C. Yao, H. Wang, Q. Han, and Y. Xiao, “Reward shaping to mitigate reward hacking in rlhf,” 2025. [Online]. Available: https://arxiv.org/abs/2502.18770

[169] X. Zhou, Y. Guo, R. Ma, T. Gui, Q. Zhang, and X. Huang, “Self-consistency of the internal reward models improves self-rewarding language models,” 2025. [Online]. Available: https://arxiv.org/abs/2502.08922

[170] D. Zhang, S. Zhoubian, Z. Hu, Y. Yue, Y. Dong, and J. Tang, “Rest-mcts*: Llm self-training via process reward guided tree search,” *arXiv preprint arXiv:2406.03816*, 2024.

[171] L. Luo, Y. Liu, R. Liu, S. Phatale, H. Lara, Y. Li, L. Shu, Y. Zhu, L. Meng, J. Sun *et al.*, “Improve Mathematical Reasoning in Language Models by Automated Process Supervision,” *arXiv preprint arXiv:2406.06592*, 2024.

[172] Z. Zhang, C. Zheng, Y. Wu, B. Zhang, R. Lin, B. Yu, D. Liu, J. Zhou, and J. Lin, “The lessons of developing process reward models in mathematical reasoning,” *arXiv preprint arXiv:2501.07301*, 2025.

[173] Z. Sun, Q. Wang, W. Yu, X. Zang, K. Zheng, J. Xu, X. Zhang, S. Yang, and H. Li, “ReARTeR: Retrieval-Augmented Reasoning with Trustworthy Process Rewarding,” *arXiv preprint arXiv:2501.07861*, 2025.

[174] W. Liu, J. Li, X. Zhang, F. Zhou, Y. Cheng, and J. He, “Diving into self-evolving training for multimodal reasoning,” *arXiv preprint arXiv:2412.17451*, 2024.

[175] W. Wang, Z. Gao, L. Chen, Z. Chen, J. Zhu, X. Zhao, Y. Liu, Y. Cao, S. Ye, X. Zhu *et al.*, “Visualprm: An effective process reward model for multimodal reasoning,” *arXiv preprint arXiv:2503.10291*, 2025.

[176] Y. Wang, Y. Zang, H. Li, C. Jin, and J. Wang, “Unified reward model for multimodal understanding and generation,” 2025. [Online]. Available: https://arxiv.org/abs/2503.05236

[177] J. Uesato, N. Kushman, R. Kumar, F. Song, N. Siegel, L. Wang, A. Creswell, G. Irving, and I. Higgins, “Solving math word problems with process- and outcome-based feedback,” 2022. [Online]. Available: https://arxiv.org/abs/2211.14275

[178] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe, “Let’s verify step by step,” in *The Twelfth International Conference on Learning Representations*, 2023.

[179] X. Lai, Z. Tian, Y. Chen, S. Yang, X. Peng, and J. Jia, “Step-dpo: Step-wise preference optimization for long-chain reasoning of llms,” *arXiv preprint arXiv:2406.18629*, 2024.

[180] Y. Liu, J. Lu, Z. Chen, C. Qu, J. K. Liu, C. Liu, Z. Cai, Y. Xia, L. Zhao, J. Bian *et al.*, “AdaptiveStep: Automatically Dividing Reasoning Step through Model Confidence,” *arXiv preprint arXiv:2502.13943*, 2025.

[181] F. Yu, A. Gao, and B. Wang, “Outcome-supervised verifiers for planning in mathematical reasoning,” *arXiv preprint arXiv:2311.09724*, 2023.

[182] N. Chen, Z. Hu, Q. Zou, J. Wu, Q. Wang, B. Hooi, and B. He, “Judgelrm: Large reasoning models as a judge,” *arXiv preprint arXiv:2504.00050*, 2025.

[183] Y. Li, Z. Lin, S. Zhang, Q. Fu, B. Chen, J.-G. Lou, and W. Chen, “Making language models better reasoners with step-aware verifier,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 5315–5333.

[184] Z. Wu, Y. Hu, W. Shi, N. Dziri, A. Suhr, P. Ammanabrolu, N. A. Smith, M. Ostendorf, and H. Hajishirzi, “Fine-grained human feedback gives better rewards for language model training,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 59 008–59 033, 2023.

[185] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, “Direct preference optimization: Your language model is secretly a reward model,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[186] E. Zelikman, Y. Wu, J. Mu, and N. D. Goodman, “STaR: Bootstrapping Reasoning With Reasoning,” in *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[187] A. Hosseini, X. Yuan, N. Malkin, A. Courville, A. Sordoni, and R. Agarwal, “V-STaR: Training Verifiers for Self-Taught Reasoners,” 2024. [Online]. Available: https://arxiv.org/abs/2402.06457

[188] W. Zeng, Y. Huang, L. Zhao, Y. Wang, Z. Shan, and J. He, “B-STaR: Monitoring and Balancing Exploration and Exploitation in Self-Taught Reasoners,” 2024. [Online]. Available: https://arxiv.org/abs/2412.17256

[189] Ç. Gülçehre, T. L. Paine, S. Srinivasan, K. Konyushkova, L. Weerts, A. Sharma, A. Siddhant, A. Ahern, M. Wang, C. Gu,

W. Macherey, A. Doucet, O. Firat, and N. de Freitas, “Reinforced Self-Training (ReST) for Language Modeling,” *CoRR*, vol. abs/2308.08998, 2023.

[190] A. Singh, J. D. Co-Reyes, R. Agarwal, A. Anand, P. Patil, X. Garcia, P. J. Liu, J. Harrison, J. Lee, K. Xu, A. Parisi, A. Kumar, A. Alemi, A. Rizkowsky, A. Nova, B. Adlam, B. Bohnet, G. Elsayed, H. Sedghi, I. Mordatch, I. Simpson, I. Gur, J. Snoek, J. Pennington, J. Hron, K. Kenealy, K. Swersky, K. Mahajan, L. Culp, L. Xiao, M. L. Bileschi, N. Constant, R. Novak, R. Liu, T. Warkentin, Y. Qian, Y. Bansal, E. Dyer, B. Neyshabur, J. Sohl-Dickstein, and N. Fiedel, “Beyond Human Data: Scaling Self-Training for Problem-Solving with Language Models,” 2024. [Online]. Available: https://arxiv.org/abs/2312.06585

[191] F. Xu, Q. Sun, K. Cheng, J. Liu, Y. Qiao, and Z. Wu, “Interactive Evolution: A Neural-Symbolic Self-Training Framework For Large Language Models,” *CoRR*, vol. abs/2406.11736, 2024.

[192] Y. Qu, T. Zhang, N. Garg, and A. Kumar, “Recursive Introspection: Teaching Language Model Agents How to Self-Improve,” 2024. [Online]. Available: https://arxiv.org/abs/2407.18219

[193] Y. Deng, P. Lu, F. Yin, Z. Hu, S. Shen, Q. Gu, J. Zou, K.-W. Chang, and W. Wang, “Enhancing Large Vision Language Models with Self-Training on Image Comprehension,” 2024. [Online]. Available: https://arxiv.org/abs/2405.19716

[194] J. Pang, P. Wang, K. Li, X. Chen, J. Xu, Z. Zhang, and Y. Yu, “Language Model Self-improvement by Reinforcement Learning Contemplation,” in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[195] Y. Dubois, C. X. Li, R. Taori, T. Zhang, I. Gulrajani, J. Ba, C. Guestrin, P. Liang, and T. B. Hashimoto, “AlpacaFarm: A Simulation Framework for Methods that Learn from Human Feedback,” in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[196] Y. Yao, H. Wu, Z. Guo, B. Zhou, J. Gao, S. Luo, H. Hou, X. Fu, and L. Song, “Learning From Correctness Without Prompting Makes LLM Efficient Reasoner,” *CoRR*, vol. abs/2403.19094, 2024.

[197] Z. Tang, Z. Li, Z. Xiao, T. Ding, R. Sun, B. Wang, D. Liu, F. Huang, T. Liu, B. Yu, and J. Lin, “Enabling scalable oversight via self-evolving critic,” 2025. [Online]. Available: https://arxiv.org/abs/2501.05727

[198] R. Ma, P. Wang, C. Liu, X. Liu, J. Chen, B. Zhang, X. Zhou, N. Du, and J. Li, “S^2r: Teaching llms to self-verify and self-correct via reinforcement learning,” 2025. [Online]. Available: https://arxiv.org/abs/2502.12853

[199] T. Munkhbat, N. Ho, S. H. Kim, Y. Yang, Y. Kim, and S.-Y. Yun, “Self-training elicits concise reasoning in large language models,” 2025. [Online]. Available: https://arxiv.org/abs/2502.20122

[200] E. Mendes and A. Ritter, “Language models can self-improve at state-value estimation for better search,” 2025. [Online]. Available: https://arxiv.org/abs/2503.02878

[201] F. Xu, H. Yan, C. Ma, H. Zhao, Q. Sun, K. Cheng, J. He, J. Liu, and Z. Wu, “Genius: A generalizable and purely unsupervised self-training framework for advanced reasoning,” 2025. [Online]. Available: https://arxiv.org/abs/2504.08672

[202] C. Li, M. Xue, Z. Zhang, J. Yang, B. Zhang, X. Wang, B. Yu, B. Hui, J. Lin, and D. Liu, “Start: Self-taught reasoner with tools,” 2025. [Online]. Available: https://arxiv.org/abs/2503.04625

[203] G. Chen, M. Liao, C. Li, and K. Fan, “Alphamath almost zero: Process supervision without process,” 2024. [Online]. Available: https://arxiv.org/abs/2405.03553

[204] F. Xiong, H. Xu, Y. Wang, R. Cheng, Y. Wang, and X. Chu, “Hs-star: Hierarchical sampling for self-taught reasoners via difficulty estimation and budget reallocation,” 2025. [Online]. Available: https://arxiv.org/abs/2505.19866

[205] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Gupta, B. P. Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, and P. Clark, “Self-Refine: Iterative Refinement with Self-Feedback,” in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[206] N. Miao, Y. W. Teh, and T. Rainforth, “SelfCheck: Using LLMs to Zero-Shot Check Their Own Step-by-Step Reasoning,” in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[207] Z. Gou, Z. Shao, Y. Gong, Y. Shen, Y. Yang, N. Duan, and W. Chen, “CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing,” in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[208] D. Huang, J. Dai, H. Weng, P. Wu, Y. Qing, H. Cui, Z. Guo, and J. Zhang, “Effilearner: Enhancing efficiency of generated code via self-optimization,” in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/99c66755871ae101a4cef87c67fb29e8-Abstract-Conference.html

[209] Q. Zhong, L. Ding, J. Liu, B. Du, and D. Tao, “ROSE Doesn’t Do That: Boosting the Safety of Instruction-Tuned Large Language Models with Reverse Prompt Contrastive Decoding,” 2024. [Online]. Available: https://arxiv.org/abs/2402.11889

[210] Y. Weng, M. Zhu, F. Xia, B. Li, S. He, S. Liu, B. Sun, K. Liu, and J. Zhao, “Large Language Models are Better Reasoners with Self-Verification,” in *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*. Association for Computational Linguistics, 2023, pp. 2550–2575.

[211] Y. Xie, K. Kawaguchi, Y. Zhao, X. Zhao, M.-Y. Kan, J. He, and Q. Xie, “Self-Evaluation Guided Beam Search for Reasoning,” 2023. [Online]. Available: https://arxiv.org/abs/2305.00633

[212] Y. Yao, H. Wu, Q. Xu, and L. Song, “Fine-grained Conversational Decoding via Isotropic and Proximal Search,” 2023. [Online]. Available: https://arxiv.org/abs/2310.08130

[213] J. Chen, W. Lin, J. Mei, and B. Byrne, “Control-DAG: Constrained Decoding for Non-Autoregressive Directed Acyclic T5 using Weighted Finite State Automata,” 2024. [Online]. Available: https://arxiv.org/abs/2404.06854

[214] N. Xu, C. Zhou, A. Celikyilmaz, and X. Ma, “Look-back Decoding for Open-Ended Text Generation,” 2023. [Online]. Available: https://arxiv.org/abs/2305.13477

[215] X. Song, Y. Wu, W. Wang, J. Liu, W. Su, and B. Zheng, “Progco: Program helps self-correction of large language models,” 2025. [Online]. Available: https://arxiv.org/abs/2501.01264

[216] T. Anthony, Z. Tian, and D. Barber, “Thinking Fast and Slow with Deep Learning and Tree Search,” 2017. [Online]. Available: https://arxiv.org/abs/1705.08439

[217] Y. Tong, D. Li, S. Wang, Y. Wang, F. Teng, and J. Shang, “Can LLMs Learn from Previous Mistakes? Investigating LLMs’ Errors to Boost for Reasoning,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 3065–3080.

[218] Y. Tong, S. Wang, D. Li, Y. Wang, S. Han, Z. Lin, C. Huang, J. Huang, and J. Shang, “Optimizing Language Model’s Reasoning Abilities with Weak Supervision,” *arXiv preprint arXiv:2405.04086*, 2024.

[219] N. Miao, Y. W. Teh, and T. Rainforth, “Selfcheck: Using llms to zero-shot check their own step-by-step reasoning,” *arXiv preprint arXiv:2308.00436*, 2023.

[220] S. An, Z. Ma, Z. Lin, N. Zheng, J.-G. Lou, and W. Chen, “Learning from mistakes makes llm better reasoner,” *arXiv preprint arXiv:2310.20689*, 2023.

[221] Z. Li, X. Hu, A. Liu, K. Zheng, S. Huang, and H. Xiong, “Refiner: Restructure retrieval content efficiently to advance question-answering capabilities,” *arXiv preprint arXiv:2406.11357*, 2024.

[222] J. Wu, M. Feng, S. Zhang, F. Che, Z. Wen, and J. Tao, “Beyond examples: High-level automated reasoning paradigm in in-context learning via mcts,” *arXiv preprint arXiv:2411.18478*, 2024.

[223] L. Yang, Z. Yu, T. Zhang, M. Xu, J. E. Gonzalez, B. Cui, and S. Yan, “Supercorrect: Supervising and correcting language models with error-driven insights,” *arXiv preprint arXiv:2410.09008*, 2024.

[224] L. Yang, Z. Yu, B. Cui, and M. Wang, “ReasonFlux: Hierarchical LLM Reasoning via Scaling Thought Templates,” 2025. [Online]. Available: https://arxiv.org/abs/2502.06772

[225] Z. Qi, M. Ma, J. Xu, L. L. Zhang, F. Yang, and M. Yang, “Mutual reasoning makes smaller llms stronger problem-solvers,” *arXiv preprint arXiv:2408.06195*, 2024.

[226] D. Zhang, J. Wu, J. Lei, T. Che, J. Li, T. Xie, X. Huang, S. Zhang, M. Pavone, Y. Li *et al.*, “Llama-berry: Pairwise optimization for

o1-like olympiad-level mathematical reasoning," *arXiv preprint arXiv:2410.02884*, 2024.

[227] L. Yang, Z. Yu, T. Zhang, S. Cao, M. Xu, W. Zhang, J. E. Gonzalez, and B. Cui, "Buffer of Thoughts: Thought-Augmented Reasoning with Large Language Models," *arXiv preprint arXiv:2406.04271*, 2024.

[228] H. Yao, J. Huang, W. Wu, J. Zhang, Y. Wang, S. Liu, Y. Wang, Y. Song, H. Feng, L. Shen *et al.*, "Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search," *arXiv preprint arXiv:2412.18319*, 2024.

[229] G. Xu, P. Jin, L. Hao, Y. Song, L. Sun, and L. Yuan, "LLaVA-o1: Let Vision Language Models Reason Step-by-Step," *arXiv preprint arXiv:2411.10440*, 2024.

[230] O. Thawakar, D. Dissanayake, K. More, R. Thawkar, A. Heakl, N. Ahsan, Y. Li, M. Zumri, J. Lahoud, R. M. Anwer *et al.*, "LlamaV-o1: Rethinking Step-by-step Visual Reasoning in LLMs," *arXiv preprint arXiv:2501.06186*, 2025.

[231] K. Xiang, Z. Liu, Z. Jiang, Y. Nie, R. Huang, H. Fan, H. Li, W. Huang, Y. Zeng, J. Han *et al.*, "AtomThink: A Slow Thinking Framework for Multimodal Mathematical Reasoning," *arXiv preprint arXiv:2411.11930*, 2024.

[232] W. Chen, X. Ma, X. Wang, and W. W. Cohen, "Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks," *arXiv preprint arXiv:2211.12588*, 2022.

[233] L. Gao, A. Madaan, S. Zhou, U. Alon, P. Liu, Y. Yang, J. Callan, and G. Neubig, "Pal: Program-aided language models," in *International Conference on Machine Learning*. PMLR, 2023, pp. 10 764–10 799.

[234] T. Khot, H. Trivedi, M. Finlayson, Y. Fu, K. Richardson, P. Clark, and A. Sabharwal, "Decomposed prompting: A modular approach for solving complex tasks," *arXiv preprint arXiv:2210.02406*, 2022.

[235] D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. Le *et al.*, "Least-to-most prompting enables complex reasoning in large language models," *arXiv preprint arXiv:2205.10625*, 2022.

[236] Y. Yu, Y. Zhang, D. Zhang, X. Liang, H. Zhang, X. Zhang, Z. Yang, M. Khademi, H. Awadalla, J. Wang, Y. Yang, and F. Wei, "Chain-of-reasoning: Towards unified mathematical reasoning in large language models via a multi-paradigm perspective," 2025. [Online]. Available: https://arxiv.org/abs/2501.11110

[237] Z. Tan, D. Li, S. Wang, A. Beigi, B. Jiang, A. Bhattacharjee, M. Karami, J. Li, L. Cheng, and H. Liu, "Large language models for data annotation and synthesis: A survey," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 930–957.

[238] B. Klieger *et al.*, "g1: Using Llama-3.1 70b on Groq to create o1-like reasoning chains," 2024. [Online]. Available: https://github.com/bklieger-groq/g1

[239] X. Hou, M. Yang, W. Jiao, X. Wang, Z. Tu, and W. X. Zhao, "CoAct: A Global-Local Hierarchy for Autonomous Agent Collaboration," *arXiv preprint arXiv:2406.13381*, 2024.

[240] M. Shen, G. Zeng, Z. Qi, Z.-W. Hong, Z. Chen, W. Lu, G. Wornell, S. Das, D. Cox, and C. Gan, "Satori: Reinforcement Learning with Chain-of-Action-Thought Enhances LLM Reasoning via Autoregressive Search," *arXiv preprint arXiv:2502.02508*, 2025.

[241] OpenAI, "Reinforcement fine-tuning," 2024.

[242] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang *et al.*, "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–45, 2024.

[243] L. Trung, X. Zhang, Z. Jie, P. Sun, X. Jin, and H. Li, "Reft: Reasoning with reinforced fine-tuning," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 7601–7614.

[244] interconnects.ai, "Blob reinforcement fin-tuning," 2024.

[245] H. Yu, X. Wu, W. Yin, D. Zhang, and S. Hu, "Codepmp: Scalable preference model pretraining for large language model reasoning," *arXiv preprint arXiv:2410.02229*, 2024.

[246] C. Gao, X. Wu, Z. Lin, D. Zhang, and S. Hu, "Nextlong: Toward effective long-context training without long documents," *arXiv preprint arXiv:2501.12766*, 2025.

[247] C. Gao, X. Wu, Q. Fu, and S. Hu, "Quest: Query-centric data synthesis approach for long-context scaling of large language model," *arXiv preprint arXiv:2405.19846*, 2024.

[248] K. Gandhi, D. Lee, G. Grand, M. Liu, W. Cheng, A. Sharma, and N. D. Goodman, "Stream of search (sos): Learning to search in language, 2024," *URL https://arxiv. org/abs/2404.03683*, 2024.

[249] S. Leng, J. Wang, J. Li, H. Zhang, Z. Hu, B. Zhang, H. Zhang, Y. Jiang, X. Li, D. Zhao, F. Wang, Y. Rong, A. Sun, and S. Lu, "Mmr1: Advancing the frontiers of multimodal reasoning," https://github.com/LengSicong/MMR1, 2025.

[250] F. Meng, L. Du, Z. Liu, Z. Zhou, Q. Lu, D. Fu, B. Shi, W. Wang, J. He, K. Zhang *et al.*, "Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning," *arXiv preprint arXiv:2503.07365*, 2025.

[251] J. Pan, C. Liu, J. Wu, F. Liu, J. Zhu, H. B. Li, C. Chen, C. Ouyang, and D. Rueckert, "Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning," *arXiv preprint arXiv:2502.19634*, 2025.

[252] Y. Liu, B. Peng, Z. Zhong, Z. Yue, F. Lu, B. Yu, and J. Jia, "Seg-Zero: Reasoning-Chain Guided Segmentation via Cognitive Reinforcement," Mar. 2025, arXiv:2503.06520 [cs]. [Online]. Available: http://arxiv.org/abs/2503.06520

[253] J. Zhang, J. Huang, H. Yao, S. Liu, X. Zhang, S. Lu, and D. Tao, "R1-VL: Learning to Reason with Multimodal Large Language Models via Step-wise Group Relative Policy Optimization," Mar. 2025, arXiv:2503.12937 [cs]. [Online]. Available: http://arxiv.org/abs/2503.12937

[254] W. Huang, B. Jia, Z. Zhai, S. Cao, Z. Ye, F. Zhao, Z. Xu, Y. Hu, and S. Lin, "Vision-R1: Incentivizing Reasoning Capability in Multimodal Large Language Models," Mar. 2025, arXiv:2503.06749 [cs]. [Online]. Available: http://arxiv.org/abs/2503.06749

[255] Y. Zheng, J. Lu, S. Wang, and Y. Xiong, "EasyR1: An Efficient, Scalable, Multi-Modality RL Training Framework," 2025. [Online]. Available: https://github.com/hiyouga/EasyR1

[256] Y. Yang, X. He, H. Pan, X. Jiang, Y. Deng, X. Yang, H. Lu, D. Yin, F. Rao, M. Zhu, B. Zhang, and W. Chen, "R1-Onevision: Advancing Generalized Multimodal Reasoning through Cross-Modal Formalization," Mar. 2025, arXiv:2503.10615 [cs]. [Online]. Available: http://arxiv.org/abs/2503.10615

[257] Y. Chen, Y. Ge, R. Wang, Y. Ge, L. Qiu, Y. Shan, and X. Liu, "Exploring the Effect of Reinforcement Learning on Video Understanding: Insights from SEED-Bench-R1," Mar. 2025, arXiv:2503.24376 [cs] version: 1. [Online]. Available: http://arxiv.org/abs/2503.24376

[258] K. Team, A. Du, B. Gao, B. Xing, C. Jiang, C. Chen, C. Li, C. Xiao, C. Du, C. Liao *et al.*, "Kimi k1. 5: Scaling Reinforcement Learning with LLMs," *arXiv preprint arXiv:2501.12599*, 2025.

[259] K. Zhang, Q. Yao, B. Lai, J. Huang, W. Fang, D. Tao, M. Song, and S. Liu, "Reasoning with reinforced functional token tuning," *arXiv preprint arXiv:2502.13389*, 2025.

[260] Z. Lin, Y. Tang, X. Yao, D. Yin, Z. Hu, Y. Sun, and K.-W. Chang, "QLASS: Boosting Language Agent Inference via Q-Guided Stepwise Search," 2025. [Online]. Available: https://arxiv.org/abs/2502.02584

[261] G. Cui, L. Yuan, Z. Wang, H. Wang, W. Li, B. He, Y. Fan, T. Yu, Q. Xu, W. Chen *et al.*, "Process Reinforcement through Implicit Rewards," *arXiv preprint arXiv:2502.01456*, 2025.

[262] M. Luo, S. Tan, J. Wong, X. Shi, W. Tang, M. Roongta, C. Cai, J. Luo, T. Zhang, E. Li, R. A. Popa, and I. Stoica, "DeepScaleR: Surpassing O1-Preview with a 1.5B Model by Scaling RL," 2025, notion Blog.

[263] J. Cheng, L. Li, G. Xiong, J. Shao, and Y. Lv, "Stop gamma decay: Min-form credit assignment is all process reward model needs for reasoning," 2025, notion Blog.

[264] H. Team, "Open r1: A fully open reproduction of deepseek-r1." 2025, github Project. [Online]. Available: https://github.com/huggingface/open-r1

[265] J. Pan, J. Zhang, X. Wang, L. Yuan, H. Peng, and A. Suhr, "TinyZero," 2025, accessed: 2025-01-24. [Online]. Available: https://github.com/Jiayi-Pan/TinyZero

[266] Z. Liu, C. Chen, W. Li, T. Pang, C. Du, and M. Lin, "There may not be aha moment in r1-zero-like training — a pilot study," 2025, notion Blog. [Online]. Available: https://oatllm.notion.site/oat-zero

[267] Z. Liu, C. Chen, C. Du, W. S. Lee, and M. Lin, "Oat: A research-friendly framework for llm online alignment," 2025. [Online]. Available: https://github.com/sail-sg/oat

[268] X. Li, H. Zou, and P. Liu, "Limr: Less is more for rl scaling," *arXiv preprint arXiv:2502.11886*, 2025.

[269] Z. Xie, L. Chen, W. Mao, J. Xu, L. Kong *et al.*, "Teaching language models to critique via reinforcement learning," *arXiv preprint arXiv:2502.03492*, 2025.

[270] T. Xie, Z. Gao, Q. Ren, H. Luo, Y. Hong, B. Dai, J. Zhou, K. Qiu, Z. Wu, and C. Luo, "Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning," 2025. [Online]. Available: https://arxiv.org/abs/2502.14768

[271] H. Zhang, J. Yao, C. Ye, W. Xiong, and T. Zhang, "Online-dpo-r1: Unlocking effective reasoning without the ppo overhead," 2025, notion Blog.

[272] J. Hu, Y. Zhang, Q. Han, D. Jiang, and H.-Y. S. Xiangyu Zhang, "Open-reasoner-zero: An open source approach to scaling reinforcement learning on the base model," 2025. [Online]. Available: https://github.com/Open-Reasoner-Zero/Open-Reasoner-Zero

[273] H. Lee, S. Phatale, H. Mansoor, K. R. Lu, T. Mesnard, J. Ferret, C. Bishop, E. Hall, V. Carbune, and A. Rastogi, "Rlaif: Scaling reinforcement learning from human feedback with ai feedback," 2023.

[274] T. Yu, Y. Yao, H. Zhang, T. He, Y. Han, G. Cui, J. Hu, Z. Liu, H.-T. Zheng, M. Sun *et al.*, "Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 807–13 816.

[275] Y.-F. Zhang, T. Yu, H. Tian, C. Fu, P. Li, J. Zeng, W. Xie, Y. Shi, H. Zhang, J. Wu *et al.*, "Mm-rlhf: The next step forward in multimodal llm alignment," *arXiv preprint arXiv:2502.10391*, 2025.

[276] J. Ji, J. Zhou, H. Lou, B. Chen, D. Hong, X. Wang, W. Chen, K. Wang, R. Pan, J. Li, M. Wang, J. Dai, T. Qiu, H. Xu, D. Li, W. Chen, J. Song, B. Zheng, and Y. Yang, "Align anything: Training all-modality models to follow instructions with language feedback," 2024. [Online]. Available: https://arxiv.org/abs/2412.15838

[277] L. Chen, L. Li, H. Zhao, Y. Song, and Vinci, "R1-V: Reinforcing Super Generalization Ability in Vision-Language Models with Less Than $3," 2025, accessed: 2025-02-02. [Online]. Available: https://github.com/Deep-Agent/R1-V

[278] H. Shen, Z. Zhang, Q. Zhang, R. Xu, and T. Zhao, "Vlm-r1: A stable and generalizable r1-style large vision-language model," 2025, accessed: 2025-02-15. [Online]. Available: https://github.com/om-ai-lab/VLM-R1

[279] Y. Peng, G. Zhang, X. Geng, and X. Yang, "Lmm-r1," 2025, accessed: 2025-02-13. [Online]. Available: https://github.com/TideDra/lmm-r1

[280] X. Wang and P. Peng, "Open-r1-video," 2025. [Online]. Available: https://github.com/Wang-Xiaodong1899/Open-R1-Video

[281] B. Bizhe, W. Shao, and Q. Zhang, "Efficient-r1-vllm: Efficient rl-tuned moe vision-language model for reasoning," 2025. [Online]. Available: https://github.com/baibizhe/Efficient-R1-VLLM

[282] J. Zhao, X. Wei, and L. Bo, "R1-omni: Explainable omni-multimodal emotion recognition with reinforcing learning," *arXiv preprint arXiv:2503.05379*, 2025.

[283] H. Zhou, X. Li, R. Wang, M. Cheng, T. Zhou, and C.-J. Hsieh, "R1-Zero's "Aha Moment" in Visual Reasoning on a 2B Non-SFT Model," Mar. 2025, arXiv:2503.05132 [cs]. [Online]. Available: http://arxiv.org/abs/2503.05132

[284] G. Li, J. Liu, H. Dinkel, Y. Niu, J. Zhang, and J. Luan, "Reinforcement Learning Outperforms Supervised Fine-Tuning: A Case Study on Audio Question Answering," Mar. 2025, arXiv:2503.11197 [cs]. [Online]. Available: http://arxiv.org/abs/2503.11197

[285] Y. Peng, Chris, X. Wang, Y. Wei, J. Pei, W. Qiu, A. Jian, Y. Hao, J. Pan, T. Xie, L. Ge, R. Zhuang, X. Song, Y. Liu, and Y. Zhou, "Skywork r1v: Pioneering multimodal reasoning with chain-of-thought," 2025. [Online]. Available: https://huggingface.co/Skywork/Skywork-R1V-38B

[286] Y. Wang, B. Xu, Z. Yue, Z. Xiao, Z. Wang, L. Zhang, D. Yang, W. Wang, and Q. Jin, "TimeZero: Temporal Video Grounding with Reasoning-Guided LVLM," Mar. 2025, arXiv:2503.13377 [cs]. [Online]. Available: http://arxiv.org/abs/2503.13377

[287] Z. Liu, Z. Sun, Y. Zang, X. Dong, Y. Cao, H. Duan, D. Lin, and J. Wang, "Visual-RFT: Visual Reinforcement Fine-Tuning," Mar. 2025, arXiv:2503.01785 [cs]. [Online]. Available: http://arxiv.org/abs/2503.01785

[288] H. Tan, Y. Ji, X. Hao, M. Lin, P. Wang, Z. Wang, and S. Zhang, "Reason-RFT: Reinforcement Fine-Tuning for Visual Reasoning," Mar. 2025, arXiv:2503.20752 [cs]. [Online]. Available: http://arxiv.org/abs/2503.20752

[289] Z. Li, Z. Ma, M. Li, S. Li, Y. Rong, T. Xu, Z. Zhang, D. Zhao, and W. Huang, "Star-r1: Spatial transformation reasoning by reinforcing multimodal llms," 2025. [Online]. Available: https://arxiv.org/abs/2505.15804

[290] E. Yeo, Y. Tong, M. Niu, G. Neubig, and X. Yue, "Demystifying Long Chain-of-Thought Reasoning in LLMs," *arXiv preprint arXiv:2502.03373*, 2025.

[291] Z. Hou, P. Du, Y. Niu, Z. Du, A. Zeng, X. Liu, M. Huang, H. Wang, J. Tang, and Y. Dong, "Does RLHF Scale? Exploring the Impacts From Data, Model, and Method," *arXiv preprint arXiv:2412.06000*, 2024.

[292] J. Kim, D. Wu, J. Lee, and T. Suzuki, "Metastable Dynamics of Chain-of-Thought Reasoning: Provable Benefits of Search, RL and Distillation," 2025. [Online]. Available: https://arxiv.org/abs/2502.01694

[293] M. Li, Y. Li, Z. Li, and T. Zhou, "How instruction and reasoning data shape post-training: Data quality through the lens of layer-wise gradients," 2025. [Online]. Available: https://arxiv.org/abs/2504.10766

[294] Z. Liu, C. Chen, W. Li, T. Pang, C. Du, and M. Lin, "There May Not be Aha Moment in R1-Zero-like Training — A Pilot Study," 2025, notion Blog. [Online]. Available: https://oatllm.notion.site/oat-zero

[295] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei *et al.*, "Qwen2. 5 technical report," *arXiv preprint arXiv:2412.15115*, 2024.

[296] H. Luo, L. Shen, H. He, Y. Wang, S. Liu, W. Li, N. Tan, X. Cao, and D. Tao, "O1-Pruner: Length-Harmonizing Fine-Tuning for O1-Like Reasoning Pruning," *arXiv preprint arXiv:2501.12570*, 2025.

[297] X. Zhao, Z. Kang, A. Feng, S. Levine, and D. Song, "Learning to reason without external rewards," *arXiv preprint arXiv:2505.19590*, 2025.

[298] S. Shafayat, F. Tajwar, R. Salakhutdinov, J. Schneider, and A. Zanette, "Can large reasoning models self-train?" *arXiv preprint arXiv:2505.21444*, 2025.

[299] F. Xu, H. Yan, C. Ma, H. Zhao, Q. Sun, K. Cheng, J. He, J. Liu, and Z. Wu, "Genius: A generalizable and purely unsupervised self-training framework for advanced reasoning," *arXiv preprint arXiv:2504.08672*, 2025.

[300] Q. Zhang, H. Wu, C. Zhang, P. Zhao, and Y. Bian, "Right question is already half the answer: Fully unsupervised llm reasoning incentivization," *arXiv preprint arXiv:2504.05812*, 2025.

[301] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, "Detecting hallucinations in large language models using semantic entropy," *Nature*, vol. 630, no. 8017, pp. 625–630, 2024.

[302] Y. Zuo, K. Zhang, S. Qu, L. Sheng, X. Zhu, B. Qi, Y. Sun, G. Cui, N. Ding, and B. Zhou, "Ttrl: Test-time reinforcement learning," *arXiv preprint arXiv:2504.16084*, 2025.

[303] L. Wei, Y. Li, C. Wang, Y. Wang, L. Kong, W. Huang, and L. Sun, "Unsupervised post-training for multi-modal llm reasoning via grpo," *arXiv preprint arXiv:2505.22453*, 2025.

[304] R. Shao, S. S. Li, R. Xin, S. Geng, Y. Wang, S. Oh, S. S. Du, N. Lambert, S. Min, R. Krishna, Y. Tsvetkov, H. Hajishirzi, P. W. Koh, and L. Zettlemoyer, "Spurious rewards: Rethinking training signals in rlvr," https://rethink-rlvr.notion.site/Spurious-Rewards-Rethinking-Training-Signals-in-RLVR-1f4df34dac18809488 2025, notion Blog.

[305] Q. Yu, Z. Zhang, R. Zhu, Y. Yuan, X. Zuo, Y. Yue, T. Fan, G. Liu, L. Liu, X. Liu *et al.*, "Dapo: An open-source llm reinforcement learning system at scale," *arXiv preprint arXiv:2503.14476*, 2025.

[306] J. He, J. Liu, C. Y. Liu, R. Yan, C. Wang, P. Cheng, X. Zhang, F. Zhang, J. Xu, W. Shen *et al.*, "Skywork open reasoner 1 technical report," *arXiv preprint arXiv:2505.22312*, 2025.

[307] Z. Gao, L. Chen, J. Zhou, and B. Dai, "One-shot entropy minimization," *arXiv preprint arXiv:2505.20282*, 2025.

[308] G. Cui, Y. Zhang, J. Chen, L. Yuan, Z. Wang, Y. Zuo, H. Li, Y. Fan, H. Chen, W. Chen *et al.*, "The entropy mechanism of reinforcement learning for reasoning language models," *arXiv preprint arXiv:2505.22617*, 2025.

[309] J. Hu, "REINFORCE++: A Simple and Efficient Approach for Aligning Large Language Models," *arXiv preprint arXiv:2501.03262*, 2025.

[310] H. Zhang, C. Li, W. Wu, S. Mao, Y. xia, I. Vulić, Z. Zhang, L. Wang, T. Tan, and F. Wei, "A call for new recipes to enhance spatial reasoning in mllms," 2025. [Online]. Available: https://arxiv.org/abs/2504.15037

[311] AI-MO, "Aime 2024," 2024. [Online]. Available: https://huggingface.co/datasets/AI-MO/aimo-validation-aime

[312] M. Suzgun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung, A. Chowdhery, Q. V. Le, E. H. Chi, D. Zhou *et al.*, "Challenging big-bench tasks and whether chain-of-thought can solve them," *arXiv preprint arXiv:2210.09261*, 2022.

[313] AI-MO, "Amc 2023," 2024. [Online]. Available: https://huggingface.co/datasets/AI-MO/aimo-validation-amc

[314] C. He, R. Luo, Y. Bai, S. Hu, Z. L. Thai, J. Shen, J. Hu, X. Han, Y. Huang, Y. Zhang *et al.*, "Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems," *arXiv preprint arXiv:2402.14008*, 2024.

[315] A. Gulati, B. Miranda, E. Chen, E. Xia, K. Fronsdal, B. de Moraes Dumont, and S. 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.

[316] E. Glazer, E. Erdil, T. Besiroglu, D. Chicharro, E. Chen, A. Gunning, C. F. Olsson, J.-S. Denain, A. Ho, E. d. O. Santos *et al.*, "Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai," *arXiv preprint arXiv:2411.04872*, 2024.

[317] C. Zheng, Z. Zhang, B. Zhang, R. Lin, K. Lu, B. Yu, D. Liu, J. Zhou, and J. Lin, "Processbench: Identifying process errors in mathematical reasoning," *arXiv preprint arXiv:2412.06559*, 2024.

[318] C. White, S. Dooley, M. Roberts, A. Pal, B. Feuer, S. Jain, R. Shwartz-Ziv, N. Jain, K. Saifullah, S. Dey, Shubh-Agrawal, S. S. Sandha, S. V. Naidu, C. Hegde, Y. LeCun, T. Goldstein, W. Neiswanger, and M. Goldblum, "Livebench: A challenging, contamination-limited LLM benchmark," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=sKYHBTAxVa

[319] S. Huang, L. Yang, Y. Song, S. Chen, L. Cui, Z. Wan, Q. Zeng, Y. Wen, K. Shao, W. Zhang *et al.*, "Thinkbench: Dynamic out-of-distribution evaluation for robust llm reasoning," *arXiv preprint arXiv:2502.16268*, 2025.

[320] K. Huang, J. Guo, Z. Li, X. Ji, J. Ge, W. Li, Y. Guo, T. Cai, H. Yuan, R. Wang *et al.*, "Math-perturb: Benchmarking llms' math reasoning abilities against hard perturbations," *arXiv preprint arXiv:2502.06453*, 2025.

[321] B. Y. Lin, R. L. Bras, K. Richardson, A. Sabharwal, R. Poovendran, P. Clark, and Y. Choi, "Zebralogic: On the scaling limits of llms for logical reasoning," *arXiv preprint arXiv:2502.01100*, 2025.

[322] B. Z. Li, B. Kim, and Z. Wang, "Questbench: Can llms ask the right question to acquire information in reasoning tasks?" *arXiv preprint arXiv:2503.22674*, 2025.

[323] T. Yu, Y. Jing, X. Zhang, W. Jiang, W. Wu, Y. Wang, W. Hu, B. Du, and D. Tao, "Benchmarking reasoning robustness in large language models," *arXiv preprint arXiv:2503.04550*, 2025.

[324] S. Shrestha, M. Kim, and K. Ross, "Mathematical reasoning in large language models: Assessing logical and arithmetic errors across wide numerical ranges," *arXiv preprint arXiv:2502.08680*, 2025.

[325] Y. Li, D. Choi, J. Chung, N. Kushman, J. Schrittwieser, R. Leblond, T. Eccles, J. Keeling, F. Gimeno, A. Dal Lago *et al.*, "Competition-level code generation with alphacode," *Science*, vol. 378, no. 6624, pp. 1092–1097, 2022.

[326] C. E. Jimenez, J. Yang, A. Wettig, S. Yao, K. Pei, O. Press, and K. R. Narasimhan, "SWE-bench: Can Language Models Resolve Real-world Github Issues?" in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=VTF8yNQM66

[327] N. Jain, K. Han, A. Gu, W.-D. Li, F. Yan, T. Zhang, S. Wang, A. Solar-Lezama, K. Sen, and I. Stoica, "Livecodebench: Holistic and contamination free evaluation of large language models for code," *arXiv preprint arXiv:2403.07974*, 2024.

[328] A. Zhang, M. Dong, J. Liu, W. Zhang, Y. Wang, J. Yang, G. Zhang, T. Liu, Z. Peng, Y. Tan *et al.*, "Codecriticbench: A holistic code critique benchmark for large language models," *arXiv preprint arXiv:2502.16614*, 2025.

[329] D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman, "GPQA: A Graduate-Level Google-Proof Q&A Benchmark," in *First Conference on Language Modeling*, 2024. [Online]. Available: https://openreview.net/forum?id=Ti67584b98

[330] Z. Zeng, Y. Liu, Y. Wan, J. Li, P. Chen, J. Dai, Y. Yao, R. Xu, Z. Qi, W. Zhao, L. Shen, J. Lu, H. Tan, Y. Chen, H. Zhang, Z. Shi, B. Wang, Z. Guo, and J. Jia, "Mr-ben: A meta-reasoning benchmark for evaluating system-2 thinking in llms," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/d81cb1f4dc6e13aeb45553f80b3d6837-Abstract-Conference.html

[331] Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang *et al.*, "Mmlu-pro: A more robust and challenging multi-task language understanding benchmark," *arXiv preprint arXiv:2406.01574*, 2024.

[332] J. Dai, J. Lu, Y. Feng, D. Huang, G. Zeng, R. Ruan, M. Cheng, H. Tan, and Z. Guo, "Mhpp: Exploring the capabilities and limitations of language models beyond basic code generation," *arXiv preprint arXiv:2405.11430*, 2024.

[333] N. Lambert, V. Pyatkin, J. Morrison, L. Miranda, B. Y. Lin, K. Chandu, N. Dziri, S. Kumar, T. Zick, Y. Choi *et al.*, "Rewardbench: Evaluating reward models for language modeling," *arXiv preprint arXiv:2403.13787*, 2024.

[334] Z. Zeng, Y. Liu, Y. Wan, J. Li, P. Chen, J. Dai, Y. Yao, R. Xu, Z. Qi, W. Zhao *et al.*, "Mr-ben: A meta-reasoning benchmark for evaluating system-2 thinking in llms," *arXiv preprint arXiv:2406.13975*, 2024.

[335] R. Kamoi, S. S. S. Das, R. Lou, J. J. Ahn, Y. Zhao, X. Lu, N. Zhang, Y. Zhang, R. H. Zhang, S. R. Vummanthala *et al.*, "Evaluating llms at detecting errors in llm responses," *arXiv preprint arXiv:2404.03602*, 2024.

[336] Z. Lin, Z. Gou, T. Liang, R. Luo, H. Liu, and Y. Yang, "CriticBench: Benchmarking LLMs for critique-correct reasoning," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 1552–1587. [Online]. Available: https://aclanthology.org/2024.findings-acl.91/

[337] S. Tan, S. Zhuang, K. Montgomery, W. Y. Tang, A. Cuadron, C. Wang, R. A. Popa, and I. Stoica, "Judgebench: A benchmark for evaluating llm-based judges," *arXiv preprint arXiv:2410.12784*, 2024.

[338] D. J. Chung, Z. Gao, Y. Kvasiuk, T. Li, M. Münchmeyer, M. Rudolph, F. Sala, and S. C. Tadepalli, "Theoretical physics benchmark (tpbench)–a dataset and study of ai reasoning capabilities in theoretical physics," *arXiv preprint arXiv:2502.15815*, 2025.

[339] L. Yang, R. Jin, L. Shi, J. Peng, Y. Chen, and D. Xiong, "Probench: Benchmarking large language models in competitive programming," *arXiv preprint arXiv:2502.20868*, 2025.

[340] A. Wei, J. Cao, R. Li, H. Chen, Y. Zhang, Z. Wang, Y. Sun, Y. Liu, T. S. Teixeira, D. Yang *et al.*, "Equibench: Benchmarking code reasoning capabilities of large language models via equivalence checking," *arXiv preprint arXiv:2502.12466*, 2025.

[341] X. Du, Y. Yao, K. Ma, B. Wang, T. Zheng, K. Zhu, M. Liu, Y. Liang, X. Jin, Z. Wei *et al.*, "Supergpqa: Scaling llm evaluation across 285 graduate disciplines," *arXiv preprint arXiv:2502.14739*, 2025.

[342] S. Parashar, B. Olson, S. Khurana, E. Li, H. Ling, J. Caverlee, and S. Ji, "Inference-time computations for llm reasoning and planning: A benchmark and insights," *arXiv preprint arXiv:2502.12521*, 2025.

[343] M. Song, Z. Su, X. Qu, J. Zhou, and Y. Cheng, "Prmbench: A fine-grained and challenging benchmark for process-level reward models," *arXiv preprint arXiv:2501.03124*, 2025.

[344] Y. He, S. Li, J. Liu, W. Wang, X. Bu, G. Zhang, Z. Peng, Z. Zhang, Z. Zheng, W. Su *et al.*, "Can large language models detect errors in long chain-of-thought reasoning?" *arXiv preprint arXiv:2502.19361*, 2025.

[345] G. Chen, W. Xu, H. Zhang, H. P. Chan, C. Liu, L. Bing, D. Zhao, A. T. Luu, and Y. Rong, "Finereason: Evaluating and improving llms' deliberate reasoning through reflective puzzle solving," *arXiv preprint arXiv:2502.20238*, 2025.

[346] F. Chollet, "On the measure of intelligence," *arXiv preprint arXiv:1911.01547*, 2019.

[347] S. Yao, H. Chen, J. Yang, and K. Narasimhan, "Webshop: Towards scalable real-world web interaction with grounded language agents," *Advances in Neural Information Processing Systems*, vol. 35, pp. 20 744–20 757, 2022.

[348] J. S. Chan, N. Chowdhury, O. Jaffe, J. Aung, D. Sherburn, E. Mays, G. Starace, K. Liu, L. Maksin, T. Patwardhan, L. Weng,

and A. Mądry, "Mle-bench: Evaluating machine learning agents on machine learning engineering," 2025. [Online]. Available: https://arxiv.org/abs/2410.07095

[349] S. Zhou, F. F. Xu, H. Zhu, X. Zhou, R. Lo, A. Sridhar, X. Cheng, Y. Bisk, D. Fried, U. Alon *et al.*, "WebArena: A Realistic Web Environment for Building Autonomous Agents," *arXiv preprint arXiv:2307.13854*, 2023. [Online]. Available: https://webarena.dev

[350] A. Prasad, A. Koller, M. Hartmann, P. Clark, A. Sabharwal, M. Bansal, and T. Khot, "ADaPT: As-Needed Decomposition and Planning with Language Models," in *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024, pp. 4226–4252.

[351] T. Xie, D. Zhang, J. Chen, X. Li, S. Zhao, R. Cao, T. J. Hua, Z. Cheng, D. Shin, F. Lei *et al.*, "Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments," *Advances in Neural Information Processing Systems*, vol. 37, pp. 52 040–52 094, 2024.

[352] J.-t. Huang, E. J. Li, M. H. Lam, T. Liang, W. Wang, Y. Yuan, W. Jiao, X. Wang, Z. Tu, and M. 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.

[353] R. Wang, P. Jansen, M.-A. Côté, and P. Ammanabrolu, "ScienceWorld: Is your Agent Smarter than a 5th Grader?" 2022. [Online]. Available: https://arxiv.org/abs/2203.07540

[354] V. Nath, P. Raja, C. Yoon, and S. Hendryx, "Toolcomp: A multi-tool reasoning & process supervision benchmark," *arXiv preprint arXiv:2501.01290*, 2025.

[355] Z. Wang, H. Xu, J. Wang, X. Zhang, M. Yan, J. Zhang, F. Huang, and H. Ji, "Mobile-agent-e: Self-evolving mobile assistant for complex tasks," *arXiv preprint arXiv:2501.11733*, 2025.

[356] X. Zhang, Y. Dong, Y. Wu, J. Huang, C. Jia, B. Fernando, M. Z. Shou, L. Zhang, and J. Liu, "Physreason: A comprehensive benchmark towards physics-based reasoning," *arXiv preprint arXiv:2502.12054*, 2025.

[357] G. Thomas, A. J. Chan, J. Kang, W. Wu, F. Christianos, F. Greenlee, A. Toulis, and M. Purtorab, "Webgames: Challenging general-purpose web-browsing ai agents," *arXiv preprint arXiv:2502.18356*, 2025.

[358] Z. Lu, Y. Chai, Y. Guo, X. Yin, L. Liu, H. Wang, G. Xiong, and H. Li, "Ui-r1: Enhancing action prediction of gui agents by reinforcement learning," *arXiv preprint arXiv:2503.21620*, 2025.

[359] H. Chen, Z. Fang, Y. Singla, and M. Dredze, "Benchmarking Large Language Models on Answering and Explaining Challenging Medical Questions," *arXiv preprint arXiv:2402.18060*, 2024.

[360] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, "What disease does this patient have? a large-scale open domain question answering dataset from medical exams," *Applied Sciences*, vol. 11, no. 14, p. 6421, 2021.

[361] A. B. Abacha, W.-w. Yim, Y. Fu, Z. Sun, M. Yetisgen, F. Xia, and T. Lin, "Medec: A benchmark for medical error detection and correction in clinical notes," *arXiv preprint arXiv:2412.19260*, 2024.

[362] Y. Zuo, S. Qu, Y. Li, Z. Chen, X. Zhu, E. Hua, K. Zhang, N. Ding, and B. Zhou, "Medxpertqa: Benchmarking expert-level medical reasoning and understanding," *arXiv preprint arXiv:2501.18362*, 2025.

[363] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun *et al.*, "Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9556–9567.

[364] P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, and J. Gao, "MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts," in *International Conference on Learning Representations (ICLR)*, 2024.

[365] K. Wang, J. Pan, W. Shi, Z. Lu, M. Zhan, and H. Li, "Measuring multimodal mathematical reasoning with math-vision dataset," *arXiv preprint arXiv:2402.14804*, 2024.

[366] Z.-Z. Li, M.-L. Zhang, F. Yin, Z.-L. Ji, J.-F. Bai, Z.-R. Pan, F.-H. Zeng, J. Xu, J.-X. Zhang, and C.-L. Liu, "Cmmath: A chinese multi-modal math skill evaluation benchmark for foundation models," *arXiv preprint arXiv:2407.12023*, 2024.

[367] M.-L. Zhang, F. Yin, and C.-L. Liu, "A Multi-Modal Neural Geometric Solver with Textual Clauses Parsed from Diagram," in *IJCAI*, 2023.

[368] J. Roberts, M. R. Taesiri, A. Sharma, A. Gupta, S. Roberts, I. Croitoru, S.-V. Bogolin, J. Tang, F. Langer, V. Raina *et al.*, "Zerobench: An impossible visual benchmark for contemporary large multimodal models," *arXiv preprint arXiv:2502.09696*, 2025.

[369] D. Jiang, R. Zhang, Z. Guo, Y. Li, Y. Qi, X. Chen, L. Wang, J. Jin, C. Guo, S. Yan *et al.*, "Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency," *arXiv preprint arXiv:2502.09621*, 2025.

[370] H. Cai, Y. Yang, and W. Hu, "Mm-iq: Benchmarking human-like abstraction and reasoning in multimodal models," *arXiv preprint arXiv:2502.00698*, 2025.

[371] M. Yasunaga, L. Zettlemoyer, and M. Ghazvininejad, "Multimodal rewardbench: Holistic evaluation of reward models for vision language models," *arXiv preprint arXiv:2502.14191*, 2025.

[372] J. Roberts, K. Han, and S. Albanie, "Grab: A challenging graph analysis benchmark for large multimodal models," *arXiv preprint arXiv:2408.11817*, 2024.

[373] J. Roberts, K. Han, N. Houlsby, and S. Albanie, "Scifibench: Benchmarking large multimodal models for scientific figure interpretation," *arXiv preprint arXiv:2405.08807*, 2024.

[374] P. Wang, Z. Li, F. Yin, D. Ran, and C. Liu, "MV-MATH: evaluating multimodal math reasoning in multi-visual contexts," *CoRR*, vol. abs/2502.20808, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2502.20808

[375] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan, "Learn to explain: Multimodal reasoning via thought chains for science question answering," *Advances in Neural Information Processing Systems*, vol. 35, pp. 2507–2521, 2022.

[376] C. Wu, Y. Ge, Q. Guo, J. Wang, Z. Liang, Z. Lu, Y. Shan, and P. Luo, "Plot2code: A comprehensive benchmark for evaluating multi-modal large language models in code generation from scientific plots," *arXiv preprint arXiv:2405.07990*, 2024.

[377] Q. Chen, L. Qin, J. Zhang, Z. Chen, X. Xu, and W. Che, "M3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought," *arXiv preprint arXiv:2405.16473*, 2024.

[378] Y. K. Chia, V. T. Y. Han, D. Ghosal, L. Bing, and S. Poria, "Puzzlevqa: Diagnosing multimodal reasoning challenges of language models with abstract visual patterns," *arXiv preprint arXiv:2403.13315*, 2024.

[379] K. Guo, B. Nan, Y. Zhou, T. Guo, Z. Guo, M. Surve, Z. Liang, N. Chawla, O. Wiest, and X. Zhang, "Can llms solve molecule puzzles? a multimodal benchmark for molecular structure elucidation," *Advances in Neural Information Processing Systems*, vol. 37, pp. 134 721–134 746, 2024.

[380] F. Zhang, L. Wu, H. Bai, G. Lin, X. Li, X. Yu, Y. Wang, B. Chen, and J. Keung, "Humaneval-v: Evaluating visual understanding and reasoning abilities of large multimodal models through coding tasks," *arXiv preprint arXiv:2410.12381*, 2024.

[381] Z. Cheng, Q. Chen, J. Zhang, H. Fei, X. Feng, W. Che, M. Li, and L. Qin, "Comt: A novel benchmark for chain of multi-modal thought on large vision-language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 22, 2025, pp. 23 678–23 686.

[382] C. Yang, C. Shi, Y. Liu, B. Shui, J. Wang, M. Jing, L. Xu, X. Zhu, S. Li, Y. Zhang *et al.*, "Chartmimic: Evaluating lmm's cross-modal reasoning capability via chart-to-code generation," *arXiv preprint arXiv:2406.09961*, 2024.

[383] Z. Huang, Z. Wang, S. Xia, X. Li, H. Zou, R. Xu, R.-Z. Fan, L. Ye, E. Chern, Y. Ye *et al.*, "Olympicarena: Benchmarking multi-discipline cognitive reasoning for superintelligent ai," *Advances in Neural Information Processing Systems*, vol. 37, pp. 19 209–19 253, 2024.

[384] C. Wang, L. Zhang, Z. Wang, and Y. 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.

[385] C. J. Wang, D. Lee, C. Menghini, J. Mols, J. Doughty, A. Khoja, J. Lynch, S. Hendryx, S. Yue, and D. Hendrycks, "Enigmaeval: A benchmark of long multimodal reasoning challenges," *arXiv preprint arXiv:2502.08859*, 2025.

[386] H. Wang, X. Zhou, Z. Xu, K. Cheng, Y. Zuo, K. Tian, J. Song, J. Lu, W. Hu, and X. Liu, "Code-vision: Evaluating multimodal llms logic understanding and code generation capabilities," *arXiv preprint arXiv:2502.11829*, 2025.

[387] B. Jia, J. Zhang, H. Zhang, and X. Wan, "Exploring and evaluating multimodal knowledge reasoning consistency of multimodal large language models," *arXiv preprint arXiv:2503.04801*, 2025.

[388] X. Ye, C. Li, S. Chen, X. Tang, and W. Wei, "Mmscibench: Bench-

marking language models on multimodal scientific problems," *arXiv preprint arXiv:2503.01891*, 2025.

[389] K. Tang, J. Gao, Y. Zeng, H. Duan, Y. Sun, Z. Xing, W. Liu, K. Lyu, and K. Chen, "Lego-puzzles: How good are mllms at multi-step spatial reasoning?" *arXiv preprint arXiv:2503.19990*, 2025.

[390] M. K. Chen, X. Zhang, and D. Tao, "Justlogic: A comprehensive benchmark for evaluating deductive reasoning in large language models," *arXiv preprint arXiv:2501.14851*, 2025.

[391] L. Phan, A. Gatti, Z. Han, N. Li, J. Hu, H. Zhang, C. B. C. Zhang, M. Shaaban, J. Ling, S. Shi *et al.*, "Humanity's last exam," *arXiv preprint arXiv:2501.14249*, 2025.

[392] W. Jiaqi, Y. Zhou, Z. Zhang, Q. Chen, Y. Chen, and J. Cheng, "Divil: Unveiling and addressing over-invariance for out-of-distribution generalization," *Transactions on Machine Learning Research.*

[393] Y. Yan, S. Wang, J. Huo, H. Li, B. Li, J. Su, X. Gao, Y.-F. Zhang, T. Xu, Z. Chu *et al.*, "Errorradar: Benchmarking complex mathematical reasoning of multimodal large language models via error detection," *arXiv preprint arXiv:2410.04509*, 2024.

[394] J. Muralidharan and T. Thomas, "Deliberate Problem-solving with a Large Language Model as a Brainstorm Aid Using a Checklist for Prompt Generation," *The Journal of the Association of Physicians of India*, vol. 72, no. 5, pp. 89–90, 2024.

[395] J. Yu, R. He, and R. Ying, "Thought propagation: An analogical approach to complex reasoning with large language models," *arXiv preprint arXiv:2310.03965*, 2023.

[396] J. Jiang, Z. Chen, Y. Min, J. Chen, X. Cheng, J. Wang, Y. Tang, H. Sun, J. Deng, W. X. Zhao *et al.*, "Technical Report: Enhancing LLM Reasoning with Reward-guided Tree Search," *arXiv preprint arXiv:2411.11694*, 2024.

[397] F. Lyu *et al.*, "Thinking Claude," 2024. [Online]. Available: https://github.com/richards199999/Thinking-Claude

[398] Y. Du, Z. Liu, Y. Li, W. X. Zhao, Y. Huo, B. Wang, W. Chen, Z. Liu, Z. Wang, and J.-R. Wen, "Virgo: A preliminary exploration on reproducing o1-like mllm," *arXiv preprint arXiv:2501.01904*, 2025.

[399] Y. Li, Z. Lai, W. Bao, Z. Tan, A. Dao, K. Sui, J. Shen, D. Liu, H. Liu, and Y. Kong, "Visual large language models for generalized and specialized applications," *arXiv preprint arXiv:2501.02765*, 2025.

[400] Z.-Z. Li, M.-L. Zhang, F. Yin, and C.-L. Liu, "Lans: A layout-aware neural solver for plane geometry problem," *arXiv preprint arXiv:2311.16476*, 2023.

[401] K. Cheng, W. Song, J. Fan, Z. Ma, Q. Sun, F. Xu, C. Yan, N. Chen, J. Zhang, and J. Chen, "Caparena: Benchmarking and analyzing detailed image captioning in the llm era," 2025. [Online]. Available: https://arxiv.org/abs/2503.12329

[402] B. Liu, X. Li, J. Zhang, J. Wang, T. He, S. Hong, H. Liu, S. Zhang, K. Song, K. Zhu *et al.*, "Advances and challenges in foundation agents: From brain-inspired intelligence to evolutionary, collaborative, and safe systems," *arXiv preprint arXiv:2504.01990*, 2025.

[403] Q. Sun, Z. Yin, X. Li, Z. Wu, X. Qiu, and L. Kong, "Corex: Pushing the boundaries of complex reasoning through multi-model collaboration," *arXiv preprint arXiv:2310.00280*, 2023.

[404] Q. Sun, Z. Chen, F. Xu, K. Cheng, C. Ma, Z. Yin, J. Wang, C. Han, R. Zhu, S. Yuan *et al.*, "A survey of neural code intelligence: Paradigms, advances and beyond," *arXiv preprint arXiv:2403.14734*, 2024.

[405] Z. Xi, Y. Ding, W. Chen, B. Hong, H. Guo, J. Wang, D. Yang, C. Liao, X. Guo, W. He, S. Gao, L. Chen, R. Zheng, Y. Zou, T. Gui, Q. Zhang, X. Qiu, X. Huang, Z. Wu, and Y.-G. Jiang, "AgentGym: Evolving Large Language Model-based Agents across Diverse Environments," 2024.

[406] R. Zhang, D. Jiang, Y. Zhang, H. Lin, Z. Guo, P. Qiu, A. Zhou, P. Lu, K.-W. Chang, P. Gao *et al.*, "MathVerse: Does Your Multi-modal LLM Truly See the Diagrams in Visual Math Problems?" *arXiv preprint arXiv:2403.14624*, 2024.

[407] Y. Li, Y. Guo, F. Guerin, and C. Lin, "An open-source data contamination report for large language models," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 528–541.

[408] G. DeepMind, "Gemini 2.5 Pro," March 2025. [Online]. Available: https://deepmind.google/technologies/gemini/pro/

[409] ——, "Gemini 2.0 Pro," October 2024. [Online]. Available: https://deepmind.google/technologies/gemini/pro/

[410] Claude, "Claude 3.5 Sonnet," June 2024. [Online]. Available: https://www.anthropic.com/news/claude-3-5-sonnet

[411] ——, "Claude 3.7 Sonnet," February 2025. [Online]. Available: https://www.anthropic.com/news/claude-3-7-sonnet

[412] I. Team, "InternLM2 Technical Report," 2024.

[413] N. Muennighoff, Z. Yang, W. Shi, X. L. Li, L. Fei-Fei, H. Hajishirzi, L. Zettlemoyer, P. Liang, E. Candès, and T. Hashimoto, "s1: Simple test-time scaling," *arXiv preprint arXiv:2501.19393*, 2025.

[414] OpenAI, "OpenAI o1-mini," September 2024. [Online]. Available: https://openai.com/index/openai-o1-mini-advancing-cost-efficient-reasoning/

[415] Y. Ye, Z. Huang, Y. Xiao, E. Chern, S. Xia, and P. Liu, "LIMO: Less is More for Reasoning," 2025. [Online]. Available: https://arxiv.org/abs/2502.03387

[416] B. Seed, :, Y. Yuan, Y. Yue, M. Wang, X. Zuo, J. Chen, L. Yan, W. Xu, C. Zhang, X. Liu, C. Wang, T. Fan, L. Liu, Q. Yu, X. Wei, Z. Lin, R. Zhu, Q. Yang, C. Wei, J. He, G. Liu, Z. Wu, X. Yu, Z. Liu, J. Xu, J. Chen, H. Pan, S. Hu, Z. Du, W. Wang, Z. Sun, C. Lou, B. Ma, Z. Wang, M. Zhang, W. Zhang, G. Liu, K. Jiang, H. Lin, R. Zhang, J. Liu, L. Han, J. Chi, W. Zhang, J. Xu, J. Yuan, Z. Xiao, Y. Xian, J. Wu, K. Hua, N. Zhou, J. Duan, H. Lu, C. Wang, J. Ou, S. Wang, X. Jin, X. Yao, C. Xu, W. Ma, Z. An, R. Pang, X. Xiao, J. Su, Y. Zhang, T. Sun, K. Liu, Y. Sun, K. Shen, S. Zhang, Y. Ma, X. Bin, J. Li, Y. Luo, D. Liu, S. Zhan, Y. Li, Y. Yang, D. Zhu, K. Shen, C. Li, X. Zhou, L. Xiang, and Y. Wu, "Seed-thinking-v1.5: Advancing superb reasoning models with reinforcement learning," 2025. [Online]. Available: https://arxiv.org/abs/2504.13914

[417] DeepSeek-AI, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," 2025. [Online]. Available: https://arxiv.org/abs/2501.12948

[418] K. Zhang, B. Li, P. Zhang, F. Pu, J. A. Cahyono, K. Hu, S. Liu, Y. Zhang, J. Yang, C. Li, and Z. Liu, "LMMs-Eval: Reality Check on the Evaluation of Large Multimodal Models," 2024. [Online]. Available: https://arxiv.org/abs/2407.12772

[419] O. Contributors, "OpenCompass: A Universal Evaluation Platform for Foundation Models," 2023. [Online]. Available: https://github.com/open-compass/opencompass

[420] "The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation." [Online]. Available: https://ai.meta.com/blog/llama-4-multimodal-intelligence/

[421] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, "Qwen2. 5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[422] Q. Team, "QVQ: To See the World with Wisdom," December 2024. [Online]. Available: https://qwenlm.github.io/blog/qvq-72b-preview/

[423] Doubao, "Doubao-1.5-pro," January 2025. [Online]. Available: https://team.doubao.com/en/special/doubao_1_5_pro

[424] OpenAI, "OpenAI o4-mini," April 2025. [Online]. Available: https://openai.com/index/openai-o3-mini/

[425] J. Ouyang, R. Yan, Y. Luo, M. Cheng, Q. Liu, Z. Liu, S. Yu, and D. Wang, "Training powerful llm agents with end-to-end reinforcement learning," GitHub, 2025. [Online]. Available: https://github.com/0russwest0/Agent-R1

[426] Z. Lu, Y. Chai, Y. Guo, X. Yin, L. Liu, H. Wang, G. Xiong, and H. Li, "Ui-r1: Enhancing action prediction of gui agents by reinforcement learning," *arXiv preprint arXiv:2503.21620*, 2025.

[427] Volcengine, "Verl: Volcano engine reinforcement learning," n.d., accessed: 2025-04-21. [Online]. Available: https://github.com/volcengine/verl

[428] J. Hu, X. Wu, Z. Zhu, Xianyu, W. Wang, D. Zhang, and Y. Cao, "Openrlhf: An easy-to-use, scalable and high-performance rlhf framework," 2024. [Online]. Available: https://arxiv.org/abs/2405.11143

[429] Z. Mei, W. Fu, K. Li, G. Wang, H. Zhang, and Y. Wu, "Realhf: Optimized rlhf training for large language models through parameter reallocation," 2024. [Online]. Available: https://arxiv.org/abs/2406.14088

[430] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica, "Efficient memory management for large language model serving with pagedattention," 2023. [Online]. Available: https://arxiv.org/abs/2309.06180

[431] L. Zheng, L. Yin, Z. Xie, C. Sun, J. Huang, C. H. Yu, S. Cao, C. Kozyrakis, I. Stoica, J. E. Gonzalez, C. Barrett, and Y. Sheng, "Sglang: Efficient execution of structured language model programs," 2024. [Online]. Available: https://arxiv.org/abs/2312.07104

[432] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro, "Megatron-lm: Training multi-billion parameter

language models using model parallelism," 2020. [Online]. Available: https://arxiv.org/abs/1909.08053

[433] Z. Jiang, H. Lin, Y. Zhong, Q. Huang, Y. Chen, Z. Zhang, Y. Peng, X. Li, C. Xie, S. Nong, Y. Jia, S. He, H. Chen, Z. Bai, Q. Hou, S. Yan, D. Zhou, Y. Sheng, Z. Jiang, H. Xu, H. Wei, Z. Zhang, P. Nie, L. Zou, S. Zhao, L. Xiang, Z. Liu, Z. Li, X. Jia, J. Ye, X. Jin, and X. Liu, "Megascale: Scaling large language model training to more than 10,000 gpus," 2024. [Online]. Available: https://arxiv.org/abs/2402.15627

[434] B. Seed, "Seed-thinking v1.5," 2023, accessed: 2025-04-21. [Online]. Available: https://github.com/ByteDance-Seed/Seed-Thinking-v1.5

[435] N. Yekollu, A. Bohra, A. Chirumamilla, K. Wen, S. K. W.-L. Chiang, A. Angelopoulos, J. E. Gonzalez, I. Stoica, and S. G. Patil, "Agent arena," 2024.

[436] X. Xia and R. Luo, "Gui-r1: A generalist r1-style vision-language action model for gui agents," *arXiv preprint arXiv:2504.10458*, 2025.

[437] J. Wu, J. Zhu, and Y. Liu, "Agentic reasoning: Reasoning llms with tools for the deep research," 2025. [Online]. Available: https://arxiv.org/abs/2502.04644

[438] S. Alzubi, C. Brooks, P. Chiniya, E. Contente, C. von Gerlach, L. Irwin, Y. Jiang, A. Kaz, W. Nguyen, S. Oh, H. Tyagi, and P. Viswanath, "Open deep search: Democratizing search with open-source reasoning agents," 2025. [Online]. Available: https://arxiv.org/abs/2503.20201

[439] X. Liang, J. Xiang, Z. Yu, J. Zhang, S. Hong, S. Fan, and X. Tang, "Openmanus: An open-source framework for building general ai agents," 2025. [Online]. Available: https://doi.org/10.5281/zenodo.15186407

[440] X. Hou, Y. Zhao, S. Wang, and H. Wang, "Model context protocol (mcp): Landscape, security threats, and future research directions," 2025. [Online]. Available: https://arxiv.org/abs/2503.23278

[441] C. Wang, Z. Chen, T. Li, Y. Zhao, and Y. Liu, "Towards trustworthy llms for code: A data-centric synergistic auditing framework," *arXiv preprint arXiv:2410.09048*, 2024.

[442] T. Nandy, M. Y. I. B. Idris, R. M. Noor, L. M. Kiah, L. S. Lun, N. B. A. Juma'at, I. Ahmedy, N. A. Ghani, and S. Bhattacharyya, "Review on security of internet of things authentication mechanism," *IEEE Access*, vol. 7, pp. 151 054–151 089, 2019.

[443] H. Liu, S. Dass, R. Martín-Martín, and Y. Zhu, "Model-based runtime monitoring with interactive imitation learning," 2023. [Online]. Available: https://arxiv.org/abs/2310.17552

[444] W. J. Yeo, R. Satapathy, R. Goh, and E. Cambria, "How interpretable are reasoning explanations from prompting large language models?" in *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024, pp. 2148–2164.

[445] S. Dutta, J. Singh, S. Chakrabarti, and T. Chakraborty, "How to think step-by-step: A mechanistic understanding of chain-of-thought reasoning," *arXiv preprint arXiv:2402.18312*, 2024.

[446] M. Li, Y. Li, and T. Zhou, "What happened in llms layers when trained for fast vs. slow thinking: A gradient perspective," *arXiv preprint arXiv:2410.23743*, 2024.

[447] ——, "What happened in llms layers when trained for fast vs. slow thinking: A gradient perspective," *arXiv preprint arXiv:2410.23743*, 2024.

[448] R. Xu, Z. Qi, Z. Guo, C. Wang, H. Wang, Y. Zhang, and W. Xu, "Knowledge conflicts for LLMs: A survey," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024, pp. 8541–8565.

[449] B. Bi, S. Liu, Y. Wang, L. Mei, J. Fang, H. Gao, S. Ni, and X. Cheng, "Is factuality enhancement a free lunch for llms? better factuality can lead to worse context-faithfulness," *arXiv preprint arXiv:2404.00216*, 2024.

[450] Y. Wang, S. Feng, H. Wang, W. Shi, V. Balachandran, T. He, and Y. Tsvetkov, "Resolving knowledge conflicts in large language models," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[451] B. Bi, S. Liu, L. Mei, Y. Wang, P. Ji, and X. Cheng, "Decoding by contrasting knowledge: Enhancing llms' confidence on edited facts," *arXiv preprint arXiv:2405.11613*, 2024.

[452] Y.-S. Chuang, Y. Xie, H. Luo, Y. Kim, J. Glass, and P. He, "Dola: Decoding by contrasting layers improves factuality in large language models," *arXiv preprint arXiv:2309.03883*, 2023.

[453] S.-C. Lin, L. Gao, B. Oguz, W. Xiong, J. Lin, S. Yih, and X. Chen, "Flame: Factuality-aware alignment for large language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 115 588–115 614, 2024.

[454] B. Bi, S. Huang, Y. Wang, T. Yang, Z. Zhang, H. Huang, L. Mei, J. Fang, Z. Li, F. Wei *et al.*, "Context-dpo: Aligning language models for context-faithfulness," *arXiv preprint arXiv:2412.15280*, 2024.

[455] Z. Li, H. Jiang, H. Chen, B. Bi, Z. Zhou, F. Sun, J. Fang, and X. Wang, "Reinforced lifelong editing for language models," *arXiv preprint arXiv:2502.05759*, 2025.

[456] B. Bi, S. Liu, Y. Wang, L. Mei, H. Gao, Y. Xu, and X. Cheng, "Adaptive token biaser: Knowledge editing via biasing key entities," *arXiv preprint arXiv:2406.12468*, 2024.

[457] T. Huang, S. Hu, F. Ilhan, S. F. Tekin, Z. Yahn, Y. Xu, and L. Liu, "Safety tax: Safety alignment makes your large reasoning models less reasonable," *arXiv preprint arXiv:2503.00555*, 2025.

[458] M. Y. Guan, M. Joglekar, E. Wallace, S. Jain, B. Barak, A. Helyar, R. Dias, A. Vallone, H. Ren, J. Wei, H. W. Chung, S. Toyer, J. Heidecke, A. Beutel, and A. Glaese, "Deliberative alignment: Reasoning enables safer language models," *arXiv preprint arXiv:2412.16339*, 2024.

[459] A. Kumar, J. Roh, A. Naseh, M. Karpinska, M. Iyyer, A. Houmansadr, and E. Bagdasarian, "Overthink: Slowdown attacks on reasoning llms," *arXiv preprint arXiv:2502.02542*, 2025.

[460] M. Kuo, J. Zhang, A. Ding, Q. Wang, L. DiValentin, Y. Bao, W. Wei, D.-C. Juan, H. Li, and Y. 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.

[461] Z. Zhu, H. Zhang, M. Zhang, R. Wang, G. Wu, K. Xu, and B. Wu, "Bot: Breaking long thought processes of o1-like large language models through backdoor attack," *arXiv preprint arXiv:2502.12202*, 2025.

[462] Y. Liu, H. Gao, S. Zhai, J. Xia, T. Wu, Z. Xue, Y. Chen, K. Kawaguchi, J. Zhang, and B. Hooi, "Guardreasoner: Towards reasoning-based llm safeguards," *arXiv preprint arXiv:2501.18492*, 2025.

[463] X. Wen, W. Zhou, W. J. Mo, and M. Chen, "Thinkguard: Deliberative slow thinking leads to cautious guardrails," *arXiv preprint arXiv:2502.13458*, 2025.

[464] F. Jiang, Z. Xu, Y. Li, L. Niu, Z. Xiang, B. Li, B. Y. Lin, and R. Poovendran, "Safechain: Safety of language models with long chain-of-thought reasoning capabilities," *arXiv preprint arXiv:2502.12025*, 2025.

[465] X. Li, R. Zhao, Y. K. Chia, B. Ding, S. Joty, S. Poria, and L. Bing, "Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources," in *International Conference on Learning Representations (ICLR)*, 2024.

[466] B. Bi, S. Liu, Y. Wang, L. Mei, H. Gao, J. Fang, and X. Cheng, "Struedit: Structured outputs enable the fast and accurate knowledge editing for large language models," *arXiv preprint arXiv:2409.10132*, 2024.

[467] H. Tran, Z. Yao, J. Wang, Y. Zhang, Z. Yang, and H. Yu, "Rare: Retrieval-augmented reasoning enhancement for large language models," *arXiv preprint arXiv:2412.02830*, 2024.

[468] M. R. Parvez, "Chain of evidences and evidence to generate: Prompting for context grounded and retrieval augmented reasoning," *arXiv preprint arXiv:2401.05787*, 2024.

[469] J. Li, Y. Zhou, J. Lu, G. Tyen, L. Gui, C. Aloisi, and Y. He, "Two heads are better than one: Dual-model verbal reflection at inference-time," *arXiv preprint arXiv:2502.19230*, 2025.

[470] Y. Wang, S. Zhao, Z. Wang, H. Huang, M. Fan, Y. Zhang, Z. Wang, H. Wang, and T. Liu, "Strategic chain-of-thought: Guiding accurate reasoning in llms through strategy elicitation," *arXiv preprint arXiv:2409.03271*, 2024.

[471] Y. Ge, S. Liu, Y. Wang, L. Mei, L. Chen, B. Bi, and X. Cheng, "Innate reasoning is not enough: In-context learning enhances reasoning large language models with less overthinking," *arXiv preprint arXiv:2503.19602*, 2025.

[472] S. Yang, J. Wu, W. Ding, N. Wu, S. Liang, M. Gong, H. Zhang, and D. Zhang, "Quantifying the robustness of retrieval-augmented language models against spurious features in grounding data," *arXiv preprint arXiv:2503.05587*, 2025.

[473] Y. Wu, Y. Wang, T. Du, S. Jegelka, and Y. Wang, "When more is less: Understanding chain-of-thought length in llms," *arXiv preprint arXiv:2502.07266*, 2025.

[474] B. Jin, H. Zeng, Z. Yue, D. Wang, H. Zamani, and J. Han, "Search-r1: Training llms to reason and leverage search

engines with reinforcement learning," 2025. [Online]. Available: https://arxiv.org/abs/2503.09516

[475] M. Chen, T. Li, H. Sun, Y. Zhou, C. Zhu, H. Wang, J. Z. Pan, W. Zhang, H. Chen, F. Yang, Z. Zhou, and W. Chen, "Research: Learning to reason with search for llms via reinforcement learning," 2025. [Online]. Available: https://arxiv.org/abs/2503.19470

[476] X. Li, G. Dong, J. Jin, Y. Zhang, Y. Zhou, Y. Zhu, P. Zhang, and Z. Dou, "Search-o1: Agentic search-enhanced large reasoning models," 2025. [Online]. Available: https://arxiv.org/abs/2501.05366

[477] H. Song, J. Jiang, Y. Min, J. Chen, Z. Chen, W. X. Zhao, L. Fang, and J.-R. Wen, "R1-searcher: Incentivizing the search capability in llms via reinforcement learning," 2025. [Online]. Available: https://arxiv.org/abs/2503.05592

[478] S. Alzubi, C. Brooks, P. Chiniya, E. Contente, C. von Gerlach, L. Irwin, Y. Jiang, A. Kaz, W. Nguyen, S. Oh, H. Tyagi, and P. Viswanath, "Open deep search: Democratizing search with open-source reasoning agents," 2025. [Online]. Available: https://arxiv.org/abs/2503.20201

[479] S. Yang, Y. Li, W. Lam, and Y. Cheng, "Multi-llm collaborative search for complex problem solving," 2025. [Online]. Available: https://arxiv.org/abs/2502.18873

[480] C. Li, M. Xue, Z. Zhang, J. Yang, B. Zhang, X. Wang, B. Yu, B. Hui, J. Lin, and D. Liu, "Start: Self-taught reasoner with tools," 2025. [Online]. Available: https://arxiv.org/abs/2503.04625

[481] B. Zhang and P. Luo, "Or-llm-agent: Automating modeling and solving of operations research optimization problem with reasoning large language model," 2025. [Online]. Available: https://arxiv.org/abs/2503.10009

[482] A. Fourney, G. Bansal, H. Mozannar, C. Tan, E. Salinas, Erkang, Zhu, F. Niedtner, G. Proebsting, G. Bassman, J. Gerrits, J. Alber, P. Chang, R. Loynd, R. West, V. Dibia, A. Awadallah, E. Kamar, R. Hosn, and S. Amershi, "Magentic-one: A generalist multi-agent system for solving complex tasks," 2024. [Online]. Available: https://arxiv.org/abs/2411.04468

[483] Q. Sun, K. Cheng, Z. Ding, C. Jin, Y. Wang, F. Xu, Z. Wu, C. Jia, L. Chen, Z. Liu *et al.*, "Os-genesis: Automating gui agent trajectory construction via reverse task synthesis," *arXiv preprint arXiv:2412.19723*, 2024.

[484] T. Zheng, G. Zhang, T. Shen, X. Liu, B. Y. Lin, J. Fu, W. Chen, and X. Yue, "Opencodeinterpreter: Integrating code generation with execution and refinement," 2025. [Online]. Available: https://arxiv.org/abs/2402.14658

[485] C. Packer, S. Wooders, K. Lin, V. Fang, S. G. Patil, I. Stoica, and J. E. Gonzalez, "Memgpt: Towards llms as operating systems," 2024. [Online]. Available: https://arxiv.org/abs/2310.08560

[486] J. Wu, J. Zhu, and Y. Liu, "Agentic reasoning: Reasoning llms with tools for the deep research," 2025. [Online]. Available: https://arxiv.org/abs/2502.04644

[487] F. Lei, J. Chen, Y. Ye, R. Cao, D. Shin, H. Su, Z. Suo, H. Gao, W. Hu, P. Yin, V. Zhong, C. Xiong, R. Sun, Q. Liu, S. Wang, and T. Yu, "Spider 2.0: Evaluating language models on real-world enterprise text-to-sql workflows," 2025. [Online]. Available: https://arxiv.org/abs/2411.07763

[488] Q.-A. Dang and C. Ngo, "Reinforcement learning for reasoning in small llms: What works and what doesn't," 2025. [Online]. Available: https://arxiv.org/abs/2503.16219

[489] M. Shen, G. Zeng, Z. Qi, Z.-W. Hong, Z. Chen, W. Lu, G. Wornell, S. Das, D. Cox, and C. Gan, "Satori: Reinforcement learning with chain-of-action-thought enhances llm reasoning via autoregressive search," 2025. [Online]. Available: https://arxiv.org/abs/2502.02508

[490] Y. Wei, O. Duchenne, J. Copet, Q. Carbonneaux, L. Zhang, D. Fried, G. Synnaeve, R. Singh, and S. I. Wang, "Swe-rl: Advancing llm reasoning via reinforcement learning on open software evolution," 2025. [Online]. Available: https://arxiv.org/abs/2502.18449

[491] X. Liu, K. Wang, Y. Li, Y. Wu, W. Ma, A. Kong, F. Huang, J. Jiao, and J. Zhang, "Epo: Explicit policy optimization for strategic reasoning in llms via reinforcement learning," 2025. [Online]. Available: https://arxiv.org/abs/2502.12486

[492] B. Pang, H. Dong, J. Xu, S. Savarese, Y. Zhou, and C. Xiong, "Bolt: Bootstrap long chain-of-thought in language models without distillation," 2025. [Online]. Available: https://arxiv.org/abs/2502.03860

[493] D. Zhang, S. Zhoubian, M. Cai, F. Li, L. Yang, W. Wang, T. Dong, Z. Hu, J. Tang, and Y. Yue, "Datascibench: An llm agent benchmark for data science," 2025. [Online]. Available: https://arxiv.org/abs/2502.13897

[494] Y. Chai, H. Li, J. Zhang, L. Liu, G. Liu, G. Wang, S. Ren, S. Huang, and H. Li, "A3: Android agent arena for mobile gui agents," 2025. [Online]. Available: https://arxiv.org/abs/2501.01149

[495] R. Bonatti, D. Zhao, F. Bonacci, D. Dupont, S. Abdali, Y. Li, Y. Lu, J. Wagle, K. Koishida, A. Bucker, L. Jang, and Z. Hui, "Windows agent arena: Evaluating multi-modal os agents at scale," 2024. [Online]. Available: https://arxiv.org/abs/2409.08264

[496] S. Zhou, F. F. Xu, H. Zhu, X. Zhou, R. Lo, A. Sridhar, X. Cheng, T. Ou, Y. Bisk, D. Fried, U. Alon, and G. Neubig, "Webarena: A realistic web environment for building autonomous agents," 2024. [Online]. Available: https://arxiv.org/abs/2307.13854

[497] Z. Wu, Z. Wu, F. Xu, Y. Wang, Q. Sun, C. Jia, K. Cheng, Z. Ding, L. Chen, P. P. Liang, and Y. Qiao, "Os-atlas: A foundation action model for generalist gui agents," 2024. [Online]. Available: https://arxiv.org/abs/2410.23218

[498] S. Xu, W. Xie, L. Zhao, and P. He, "Chain of draft: Thinking faster by writing less," *arXiv preprint arXiv:2502.18600*, 2025.

[499] T. Han, Z. Wang, C. Fang, S. Zhao, S. Ma, and Z. Chen, "Token-budget-aware llm reasoning," *arXiv preprint arXiv:2412.18547*, 2024.

[500] P. Aggarwal and S. Welleck, "L1: Controlling how long a reasoning model thinks with reinforcement learning," *arXiv preprint arXiv:2503.04697*, 2025.

[501] W. Yang, S. Ma, Y. Lin, and F. Wei, "Towards thinking-optimal scaling of test-time compute for llm reasoning," *arXiv preprint arXiv:2502.18080*, 2025.

[502] X. Ma, G. Wan, R. Yu, G. Fang, and X. Wang, "Cot-valve: Length-compressible chain-of-thought tuning," 2025. [Online]. Available: https://arxiv.org/abs/2502.09601

[503] Y. Kang, X. Sun, L. Chen, and W. Zou, "C3ot: Generating shorter chain-of-thought without compromising effectiveness," *arXiv preprint arXiv:2412.11664*, 2024.

[504] Z. Yin, Q. Sun, Q. Guo, Z. Zeng, X. Li, J. Dai, Q. Cheng, X. Huang, and X. Qiu, "Reasoning in flux: Enhancing large language models reasoning through uncertainty-aware adaptive guidance," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 2401–2416. [Online]. Available: https://aclanthology.org/2024.acl-long.131/

[505] I. Ong, A. Almahairi, V. Wu, W.-L. Chiang, T. Wu, J. E. Gonzalez, M. W. Kadous, and I. Stoica, "Routellm: Learning to route llms with preference data," 2025. [Online]. Available: https://arxiv.org/abs/2406.18665

[506] Y.-N. Chuang, H. Zhou, P. Sarma, P. Gopalan, J. Boccio, S. Bolouki, and X. Hu, "Learning to route llms with confidence tokens," *arXiv preprint arXiv:2410.13284*, 2025.

[507] N. Chen, Z. Hu, Q. Zou, J. Wu, Q. Wang, B. Hooi, and B. He, "Judgelrm: Large reasoning models as a judge," *arXiv preprint arXiv:2504.00050*, 2025.

[508] D. Li, B. Jiang, L. Huang, A. Beigi, C. Zhao, Z. Tan, A. Bhattacharjee, Y. Jiang, C. Chen, T. Wu *et al.*, "From generation to judgment: Opportunities and challenges of llm-as-a-judge," *arXiv preprint arXiv:2411.16594*, 2024.

[509] D. Li, R. Sun, Y. Huang, M. Zhong, B. Jiang, J. Han, X. Zhang, W. Wang, and H. Liu, "Preference leakage: A contamination problem in llm-as-a-judge," *arXiv preprint arXiv:2502.01534*, 2025.

[510] S. Hao, S. Sukhbaatar, D. Su, X. Li, Z. Hu, J. Weston, and Y. Tian, "Training large language models to reason in a continuous latent space," *arXiv preprint arXiv:2412.06769*, 2024.

[511] J. Cheng and B. Van Durme, "Compressed chain of thought: Efficient reasoning through dense representations," *arXiv preprint arXiv:2412.13171*, 2024.

[512] S. Nie, F. Zhu, Z. You, X. Zhang, J. Ou, J. Hu, J. Zhou, Y. Lin, J.-R. Wen, and C. Li, "Large language diffusion models," 2025. [Online]. Available: https://arxiv.org/abs/2502.09992

[513] J. Ye, S. Gong, L. Chen, L. Zheng, J. Gao, H. Shi, C. Wu, X. Jiang, Z. Li, W. Bi, and L. Kong, "Diffusion of thoughts: Chain-of-thought reasoning in diffusion language models," 2024. [Online]. Available: https://arxiv.org/abs/2402.07754

[514] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[515] B. Peng, E. Alcaide, Q. Anthony, A. Albalak, S. Arcadinho, S. Biderman, H. Cao, X. Cheng, M. Chung, M. Grella, K. K. GV, X. He, H. Hou, J. Lin, P. Kazienko, J. Kocon, J. Kong, B. Koptyra, H. Lau, K. S. I. Mantri, F. Mom, A. Saito, G. Song, X. Tang, B. Wang, J. S. Wind, S. Wozniak, R. Zhang, Z. Zhang, Q. Zhao, P. Zhou, Q. Zhou, J. Zhu, and R.-J. Zhu, “Rwkv: Reinventing rnns for the transformer era,” 2023. [Online]. Available: https://arxiv.org/abs/2305.13048

[516] Y. Liu, J. Wu, Y. He, H. Gao, H. Chen, B. Bi, J. Zhang, Z. Huang, and B. Hooi, “Efficient inference for large reasoning models: A survey,” 2025. [Online]. Available: https://arxiv.org/abs/2503.23077

[517] F. Xu, H. Yan, C. Ma, H. Zhao, J. Liu, Q. Lin, and Z. Wu, “ϕ-decoding: Adaptive foresight sampling for balanced inference-time exploration and exploitation,” 2025. [Online]. Available: https://arxiv.org/abs/2503.13288

[518] Y. Leviathan, M. Kalman, and Y. Matias, “Fast inference from transformers via speculative decoding,” in *International Conference on Machine Learning*, 2023, pp. 19 274–19 286.

[519] X. Ning, Z. Lin, Z. Zhou, Z. Wang, H. Yang, and Y. Wang, “Skeleton-of-thought: Large language models can do parallel decoding,” *Proceedings ENLSP-III*, 2023.

[520] X. Miao, G. Oliaro, Z. Zhang, X. Cheng, Z. Wang, Z. Zhang, R. Y. Y. Wong, A. Zhu, L. Yang, X. Shi *et al.*, “SpecInfer: Accelerating Generative Large Language Model Serving with Tree-based Speculative Inference and Verification,” *arXiv preprint arXiv:2305.09781*, 2023.

[521] B. Qi, X. Chen, J. Gao, D. Li, J. Liu, L. Wu, and B. Zhou, “Interactive continual learning: Fast and slow thinking,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12 882–12 892.

[522] A. Wang, L. Song, Y. Tian, B. Peng, D. Yu, H. Mi, J. Su, and D. Yu, “Litesearch: Efficacious tree search for llm,” *arXiv preprint arXiv:2407.00320*, 2024.

[523] P. Hu, J. Qi, X. Li, H. Li, X. Wang, B. Quan, R. Wang, and Y. Zhou, “Tree-of-mixed-thought: Combining fast and slow thinking for multi-hop visual reasoning,” *arXiv preprint arXiv:2308.09658*, 2023.

[524] Y. Zheng, S. Sun, L. Qiu, D. Ru, C. Jiayang, X. Li, J. Lin, B. Wang, Y. Luo, R. Pan *et al.*, “OpenResearcher: Unleashing AI for Accelerated Scientific Research,” *arXiv preprint arXiv:2408.06941*, 2024.

[525] Q. Sun, Z. Liu, C. Ma, Z. Ding, F. Xu, Z. Yin, H. Zhao, Z. Wu, K. Cheng, Z. Liu, J. Wang, Q. Li, X. Tang, T. Xie, X. Feng, X. Li, B. Kao, W. Wang, B. Qi, L. Kong, and Z. Wu, “Scienceboard: Evaluating multimodal autonomous agents in realistic scientific workflows,” 2025. [Online]. Available: https://arxiv.org/abs/2505.19897

[526] B. Romera-Paredes, M. Barekatain, A. Novikov, M. Balog, M. P. Kumar, E. Dupont, F. J. Ruiz, J. S. Ellenberg, P. Wang, O. Fawzi *et al.*, “Mathematical discoveries from program search with large language models,” *Nature*, vol. 625, no. 7995, pp. 468–475, 2024.

[527] T. H. Trinh, Y. Wu, Q. V. Le, H. He, and T. Luong, “Solving olympiad geometry without human demonstrations,” *Nature*, vol. 625, no. 7995, pp. 476–482, 2024.

[528] Y. Chervonyi, T. H. Trinh, M. Olšák, X. Yang, H. Nguyen, M. Menegali, J. Jung, V. Verma, Q. V. Le, and T. Luong, “Gold-medalist Performance in Solving Olympiad Geometry with AlphaGeometry2,” *arXiv preprint arXiv:2502.03544*, 2025.

[529] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican *et al.*, “Gemini: a family of highly capable multimodal models,” *arXiv preprint arXiv:2312.11805*, 2023.

[530] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang *et al.*, “Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context,” *arXiv preprint arXiv:2403.05530*, 2024.

[531] H. Wang, H. Xin, C. Zheng, L. Li, Z. Liu, Q. Cao, Y. Huang, J. Xiong, H. Shi, E. Xie, J. Yin, Z. Li, H. Liao, and X. Liang, “Lego-prover: Neural theorem proving with growing libraries,” 2023. [Online]. Available: https://arxiv.org/abs/2310.00656

[532] J. Xiong, J. Shen, Y. Yuan, H. Wang, Y. Yin, Z. Liu, L. Li, Z. Guo, Q. Cao, Y. Huang, C. Zheng, X. Liang, M. Zhang, and Q. Liu, “Trigo: Benchmarking formal mathematical proof reduction for generative language models,” 2023. [Online]. Available: https://arxiv.org/abs/2310.10180

[533] H. Wang, Y. Yuan, Z. Liu, J. Shen, Y. Yin, J. Xiong, E. Xie, H. Shi, Y. Li, L. Li *et al.*, “Dt-solver: Automated theorem proving with dynamic-tree sampling guided by proof-level value function,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 12 632–12 646.

[534] H. Zhang, C. Shang, S. Wang, D. Zhang, F. Yao, R. Sun, Y. Yu, Y. Yang, and F. Wei, “Shifcon: Enhancing non-dominant language capabilities with a shift-based contrastive framework,” *arXiv preprint arXiv:2410.19453*, 2024.

[535] N. Chen, Z. Zheng, N. Wu, L. Shou, M. Gong, Y. Song, D. Zhang, and J. Li, “Breaking language barriers in multilingual mathematical reasoning: Insights and observations,” *arXiv preprint arXiv:2310.20246*, 2023.

[536] L. Mei, S. Liu, Y. Wang, B. Bi, and X. Chen, “Slang: New concept comprehension of large language models,” *arXiv preprint arXiv:2401.12585*, 2024.

[537] L. Mei, S. Liu, Y. Wang, B. Bi, R. Yuan, and X. Cheng, “Hiddenguard: Fine-grained safe generation with specialized representation router,” *arXiv preprint arXiv:2410.02684*, 2024.

[538] L. Mei, S. Liu, Y. Wang, B. Bi, J. Mao, and X. Cheng, “" not aligned" is not" malicious": Being careful about hallucinations of large language models’ jailbreak,” *arXiv preprint arXiv:2406.11668*, 2024.

[539] M. Parmar and Y. Govindarajulu, “Challenges in Ensuring AI Safety in DeepSeek-R1 Models: The Shortcomings of Reinforcement Learning Strategies,” *arXiv preprint arXiv:2501.17030*, 2025.

[540] Y. Liu, H. Gao, S. Zhai, J. Xia, T. Wu, Z. Xue, Y. Chen, K. Kawaguchi, J. Zhang, and B. Hooi, “Guardreasoner: Towards reasoning-based llm safeguards,” 2025. [Online]. Available: https://arxiv.org/abs/2501.18492