
A SURVEY ON LARGE LANGUAGE MODELS FOR MATHEMATICAL REASONING

A PREPRINT

Peng-Yuan Wang^{1,2*}, Tian-Shuo Liu^{1,2*}, Chenyang Wang^{1,2}, Yi-Di Wang^{1,2}, Shu Yan¹, Cheng-Xing Jia^{1,2}, Xu-Hui Liu^{1,2}, Xin-Wei Chen³, Jia-Cheng Xu^{4,5}, Ziniu Li⁶, Yang Yu^{1,2,◊}

¹ National Key Laboratory for Novel Software Technology, Nanjing University, China

² School of Artificial Intelligence, Nanjing University, China

³ Polixir.ai

⁴ Nanyang Technological University, Singapore

⁵ Skywork AI, Singapore

⁶ The Chinese University of Hong Kong, Shenzhen

* Equal contribution

◊ Corresponding: yuy@nju.edu.cn

ABSTRACT

Mathematical reasoning has long represented one of the most fundamental and challenging frontiers in artificial intelligence research. In recent years, large language models (LLMs) have achieved significant advances in this area. This survey examines the development of mathematical reasoning abilities in LLMs through two high-level cognitive phases: comprehension, where models gain mathematical understanding via diverse pretraining strategies, and answer generation, which has progressed from direct prediction to step-by-step Chain-of-Thought (CoT) reasoning. We review methods for enhancing mathematical reasoning, ranging from training-free prompting to fine-tuning approaches such as supervised fine-tuning and reinforcement learning, and discuss recent work on extended CoT and “test-time scaling”. Despite notable progress, fundamental challenges remain in terms of capacity, efficiency, and generalization. To address these issues, we highlight promising research directions, including advanced pretraining and knowledge augmentation techniques, formal reasoning frameworks, and meta-generalization through principled learning paradigms. This survey tries to provide some insights for researchers interested in enhancing reasoning capabilities of LLMs and for those seeking to apply these techniques to other domains.

1 Introduction

“Can machines think?” This profound question, posed by Alan Turing in the 1950s [Turing, 1950], established the philosophical foundation for modern artificial intelligence. Since then, enabling machines to reason has remained a central and enduring objective in AI research. Among the most rigorous and illuminating assessments of machine reasoning is mathematical problem-solving, a domain that demands not only the manipulation of symbols, but also the representation and understanding of abstract concepts, the construction of formal arguments, and the transfer of principles to new and varied contexts. As such, mathematical reasoning provides a precise and demanding lens through which to evaluate and advance machine intelligence. Beginning in the 1960s, AI researchers sought to endow machines with mathematical reasoning abilities by developing systems capable of representing and manipulating formal knowledge. Early work centered on symbolic rule-based systems [Feigenbaum et al., 1963, Bobrow et al., 1964], which depended on handcrafted rules and pattern matching [Slagle, 1965, Fletcher, 1985]. While pioneering, these approaches were restricted to narrow domains and lacked generalization. Subsequent efforts, such as

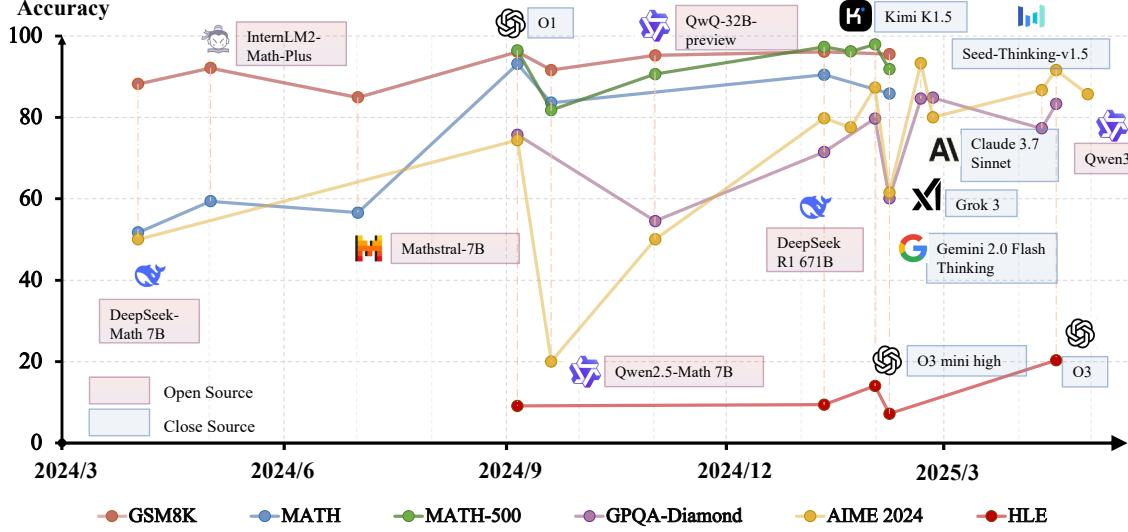


Figure 1: The rapid advancements in mathematical capabilities in recent years.

semantic parsing methods [Kwiatkowski et al., 2013, Goldwasser and Roth, 2014], focused on mapping problem text to structured logical forms, but continued to rely heavily on human engineering and struggled to scale to the full diversity of mathematical tasks.

Recent advances in large language models (LLMs) have profoundly reshaped the field of natural language representation and understanding. The introduction of transformer-based architectures and instruction-tuned models such as ChatGPT has led to remarkable progress in natural language problem-solving. In this context, mathematical reasoning, a long-standing and rigorous benchmark for AI systems, has become a focal point for evaluating LLM capabilities. As illustrated in Figure 1, state-of-the-art models now exhibit significantly enhanced performance across a wide range of complex mathematical benchmarks. For example, Grok 3 Beta achieved an impressive score of 83.9% on the AIME 2024 [aim, 2024], a prestigious Olympiad-level contest. This result places the model within the top 2.5% of all participants nationwide¹, as shown in Figure 3. As LLMs continue to excel in mathematical problem-solving, their impact is becoming increasingly significant across various domains. In Finance, they can leverage complex mathematical reasoning capabilities to process both structured tables and unstructured text, enabling Financial Document Question Answering, leading to handling complex mathematical scenarios [Srivastava et al., 2024]. In the Medical field, advanced LLMs can work as a medical diagnostic assistant, achieving high accuracy in answering specialized healthcare-related questions [Wu et al., 2024a]. Driven by the rapid advancements and increasing applicability of mathematical reasoning in LLMs, this work seeks to consolidate the fragmented developments into a coherent framework.

The ability of LLMs to perform mathematical reasoning can be grounded roughly in two elements: comprehension of mathematical concepts and the generation of solutions through step-by-step deduction. Comprehension involves acquiring a broad and flexible understanding of diverse mathematical concepts, problem formats, and difficulty levels to enable effective reasoning and problem-solving. Recent research has demonstrated that large-scale pre-training on mathematical corpora, such as textbooks, academic publications, and problem datasets, enables LLMs to internalize domain-specific knowledge, terminology, and the contextual reasoning patterns characteristic of mathematical discourse [Shao et al., 2024, Yang et al., 2024b]. This data-driven paradigm marks a significant shift from earlier approaches based on rule-matching [Slagle, 1965, Fletcher, 1985] or semantic parsing [Kwiatkowski et al., 2013, Goldwasser and Roth, 2014], moving toward a more holistic and contextualized understanding of mathematics. The second element, solution generation, focuses on producing logically coherent intermediate steps during problem-solving. A key technique for achieving this in LLMs is chain-of-thought (CoT) prompting [Wei et al., 2022b], which encourages models to emulate deductive reasoning by generating step-by-step explanations. This approach has proven effective in enabling LLMs to tackle more complex, abstract, and real-world mathematical challenges, as it structures the reasoning process into interpretable and verifiable steps.

¹Human performance is based on data from MAA: <https://maa.edvistas.com/eduview>.

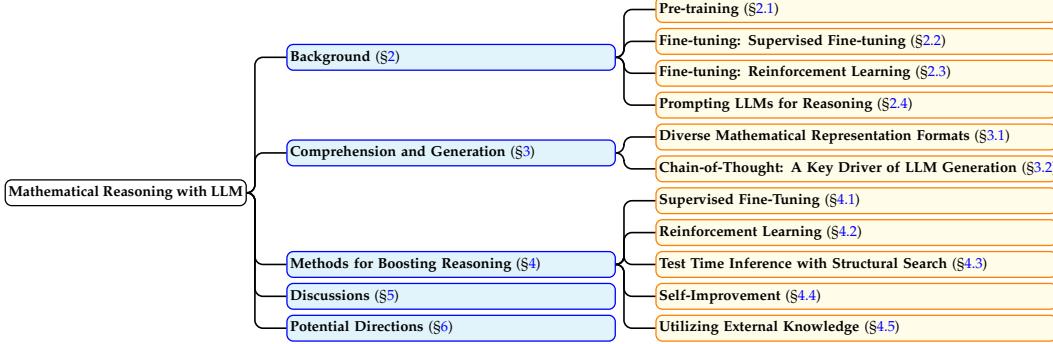


Figure 2: The organization of this survey.

Despite their impressive language understanding abilities, pre-trained LLMs often struggle to produce contextually appropriate and relevant responses when applied directly to downstream tasks, frequently resulting in repetitive or irrelevant output. To mitigate these shortcomings, prompt engineering, particularly approaches like In-Context Learning (ICL) [Jie and Lu, 2023, Zhou et al., 2022], has emerged as a vital strategy for enhancing the reasoning capabilities of LLMs through careful design of input prompts. In addition to prompt engineering, fine-tuning² strategies have been shown to further boost model performance. For example, supervised fine-tuning (SFT) on high-quality demonstrations adapts pre-trained models to specific tasks or domains [Yang et al., 2024b, Ho et al., 2023, Magister et al., 2023]. This process helps align models with instruction-following objectives, although it introduces challenges such as overfitting and limitations on exploration [Li et al., 2025b, Wang et al., 2024a, Zeng et al., 2025]. Reinforcement learning (RL) methods [Guo et al., 2025] further empower models to improve their problem-solving abilities through trial-and-error exploration. Recent research has also shown that extending CoT reasoning by generating longer token sequences during inference [OpenAI, 2024, Snell et al., 2024] can significantly enhance the reasoning capabilities of LLMs, resulting in more structured and accurate solutions. Techniques such as search-based methods [Yao et al., 2023, Feng et al., 2024b] and reinforcement fine-tuning [OpenAI, 2024, Trung et al., 2024] can be incorporated to further refine the model’s reasoning abilities, especially when working with extended CoT reasoning.

Recently, a wide range of methods and models have been developed to advance mathematical reasoning. Several surveys have reviewed aspects of this field. For example, Zhang et al. [2019] and Meadows and Freitas [2022] focus on traditional approaches to mathematical problem-solving. Other works, such as Ahn et al. [2024] and Lu et al. [2023], primarily discuss datasets and fine-tuning methods. The survey by Yan et al. [2024] summarizes progress in multi-modal mathematical reasoning based on LLMs. In the broader context of reasoning, Li et al. [2025c] offers a systematic analysis of advancements in System 2 thinking, while Xu et al. [2025] introduces the application of reinforcement learning to reasoning tasks. Additionally, Zhou et al. [2025] reviews RL-based multi-modal reasoning. This survey tries to complement these works by offering a comprehensive and up-to-date overview of recent developments in LLM-based mathematical reasoning, helping to synthesize insights across this rapidly evolving field.

In this paper, we review the research in the field of mathematical reasoning as illustrated in Figure 2. First, in Section 2, we give a brief introduction to the background of LLM. In Section 3, we summarize recent innovations and identify some essential elements that enable mathematical reasoning. Section 4 reviews related methods, which can boost reasoning in LLM, including prompting, fine-tuning, test-time scaling, and self-improvement. Finally, Sections 5 and 6 discuss current limitations and potential directions for future research.

2 Background

2.1 Pre-Training

Large language models (LLMs) acquire knowledge and develop general language understanding across various domains through pre-training on extensive corpora [Zhao et al., 2023]. Given a dataset \mathcal{D}_{pre} =

²Here, we refer supervised fine-tuning, reinforcement learning, and any other stages of guided learning after pre-training of LLMs as “fine-tuning” instead of “post-training”, for exactness of verbal expression.

$\{x^{(i)}\}_{i=1,\dots,N}$, where $x^{(i)}$ represents the i -th sequence of tokens in the dataset, the pre-training optimization objective is typically a form of language modeling, such as predicting the next token:

$$\max_{\theta} \sum_{x \in \mathcal{D}_{pre}} \sum_{t=1}^{|x|} \log (\pi_{\theta}(x_t | x_{<t})) \quad (1)$$

where π_{θ} is the model parameterized by θ , x_t is the token in sequence x at position t , and $x_{<t}$ are the preceding tokens.

During the pre-training phase, access to high-quality and highly diverse data is fundamental in shaping the reasoning, comprehension, and generalization abilities of LLMs [Yang et al., 2024a, Tovron et al., 2023]. Vast amounts of text corpora, including books, papers, websites, and code, can be leveraged to improve the performance of LLMs. However, raw data often contains noise and inconsistencies, making effective data selection and filtering techniques [Shao et al., 2024, Joulin, 2016, Zhang et al., 2024d, Lin et al., 2024b] important for curating high-quality information. Beyond relying solely on raw data, recent developments in LLM training have incorporated the use of synthetic data [Wang and Lu, 2023, Ying et al., 2024b, Zhou et al., 2024, Yang et al., 2024b]. By combining high-quality raw data with synthetically generated examples, LLMs can potentially achieve enhanced capabilities in areas such as reasoning.

2.2 Fine-Tuning: Supervised Fine-Tuning

To adapt LLMs to specific downstream tasks or to align their behavior with human instructions, a common stage is supervised fine-tuning (SFT). Given a dataset $\mathcal{D}_{sft} = \{(x^{(i)}, y^{(i)})\}_{i=1,\dots,M}$, where x_i is an instructional prompt or input sequence and y_i is the desired output sequence, SFT refines the model by optimizing a similar language modeling objective:

$$\max_{\theta} \sum_{(x,y) \in \mathcal{D}_{sft}} \sum_{t=1}^{|y|} \log (\pi_{\theta}(y_t | x, y_{<t})) \quad (2)$$

This process trains the model to mimic the “golden” responses provided in the SFT dataset.

The choice and quality of the task-specific dataset play a significant role in the effectiveness of SFT. Methods such as data augmentation and data synthesis are commonly used to enhance dataset quality and diversity. To further enhance model performance, fine-tuning can incorporate knowledge distillation [Hinton, 2015], where a student model learns from the outputs of a more capable teacher model. The teacher’s outputs can provide additional supervisory signals beyond ground truth labels, potentially improving fluency, consistency, and efficiency.

2.3 Fine-Tuning: Reinforcement Learning

While SFT helps LLMs learn to follow instructions and generate relevant text, it may not always be sufficient to align model outputs with complex human preferences, ensure factual accuracy, or reduce undesirable behaviors like generating harmful or biased content [OpenAI, 2023, Team and Google, 2023]. Reinforcement learning (RL) offers a framework to further refine LLM behavior based on broader notions of quality, often captured by a reward signal [Ziegler et al., 2019]. A prominent application of RL in this domain is reinforcement learning from human feedback (RLHF), where human preferences are used to train a reward model, which then guides the LLM’s fine-tuning process [Christiano et al., 2017, Bai et al., 2022, Ouyang et al., 2022]. This approach allows the LLM to learn from scalar feedback signals that can represent nuanced aspects of response quality, such as helpfulness, harmlessness, and honesty, which can be difficult to specify directly in an SFT objective [Askell et al., 2021, Menick et al., 2022].

The RL framework is typically formalized as a Markov Decision Process (MDP). An MDP is formally represented as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma, \rho_0, T)$, where:

- \mathcal{A} represents the action space. For an LLM, an action a_t is typically the selection of the next token from the vocabulary [Ouyang et al., 2022].
- \mathcal{S} represents the state space. In the context of LLMs, a state s_t at time step t can be defined as the sequence of tokens generated so far, often including the initial prompt: $s_t = (x, a_1, a_2, \dots, a_{t-1})$ [Ouyang et al., 2022].

- $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ represents the state transition probability function, $P(s_{t+1}|s_t, a_t)$. In LLM generation, the transition is often deterministic: generating token a_t in state s_t leads to a unique next state $s_{t+1} = (s_t, a_t)$. Thus, $P(s_{t+1}|s_t, a_t) = 1$ if s_{t+1} is the concatenation of s_t and a_t , and 0 otherwise.
- $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ represents the reward function. This function provides a scalar feedback signal. In RLHF, this reward is often given by a separate reward model trained on human preference data, which evaluates the quality of generated sequences [Christiano et al., 2017, Ouyang et al., 2022]. Rewards can be sparse (e.g., given only at the end of a sequence) or dense (e.g., per token).
- $\gamma \in [0, 1]$ is the discount factor, which balances the importance of immediate versus future rewards.
- $\rho_0 : \mathcal{S} \rightarrow [0, 1]$ represents the initial state distribution. For LLMs, this is typically determined by the distribution of input prompts x .
- T denotes the horizon or maximum episode length (e.g., maximum sequence length).

The agent, in this case the LLM (also referred to as the policy $\pi_\theta(a_t|s_t)$), aims to learn a policy that maximizes the expected cumulative discounted reward:

$$\max_{\theta} \mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_{t=0}^{T-1} \gamma^t r(s_t, a_t) \right] \quad (3)$$

where $\tau = (s_0, a_0, s_1, a_1, \dots, s_{T-1}, a_{T-1}, s_T)$ is a trajectory (a sequence of states and actions), $s_0 \sim \rho_0$, $a_t \sim \pi_\theta(\cdot|s_t)$, and $s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)$.

The reward function $r(s_t, a_t)$ is crucial. It can be categorized into two types:

- **Outcome-based rewards:** These are assigned based on the final result (e.g., evaluating the complete generated sequence for helpfulness or correctness) [Stiennon et al., 2020].
- **Process-based rewards:** These evaluate intermediate steps or the reasoning process leading to the outcome, which can provide more granular feedback [Wang and Zhou, 2024].

Common RL algorithms like Proximal Policy Optimization (PPO) were often used to optimize the LLM policy based on the rewards from the reward model [Schulman et al., 2017, Ouyang et al., 2022, Bai et al., 2022]. The overall goal is to steer the LLM towards generating outputs that are more aligned with desired characteristics that might be underspecified or difficult to learn through SFT alone [OpenAI, 2023, Team and Google, 2023]. It is worth noticing that, once the reward model is given, RLHF can train LLMs without response data, significantly improving the generalization ability of LLMs.

2.4 Prompting LLMs for Reasoning

Prompting has emerged as a simple yet effective way for eliciting and enhancing the reasoning capabilities of large language models (LLMs). The initial prompting paradigm, exemplified by models such as GPT-2 [Radford et al., 2019], was characterized as the zero-shot setting. In this paradigm, models were prompted with only task instructions, allowing them to function as multi-task systems without task-specific examples. Building upon this foundation, researchers introduced few-shot prompting, wherein carefully designed, high-quality examples included in the prompt enable LLMs to infer reasoning strategies from the provided context [Zhou et al., 2022, Jie and Lu, 2023]. This technique has demonstrated improved performance over zero-shot prompting for certain tasks. However, crafting high-quality demonstrations for a diverse array of reasoning tasks can present practical challenges, and complex few-shot prompts may significantly increase the computational cost during inference.

To further augment the reasoning performance achievable with zero-shot prompting, researchers introduced the influential technique of CoT prompting [Wang et al., 2024b, Zhong et al., 2024, Wang et al., 2023b, Imani et al., 2023, Wei et al., 2022b, Kojima et al., 2022, Chen et al., 2024b, Yao et al., 2023, Zhang et al., 2023b, Ghosh et al., 2024, Yin et al., 2024]. Simple instructive phrases, such as “Let’s think step by step”, guide the model to generate intermediate reasoning steps, which has been shown to improve its problem-solving accuracy. This approach has been subsequently extended by methods like self-consistency [Huang et al., 2023], which involves generating multiple reasoning paths and selecting the most consistent outcome, thereby enhancing the reliability of the final answer. Nevertheless, the manual design of effective CoT prompts for a wide range of problems can be a labor-intensive process.

Beyond linear reasoning pathways, more advanced prompting strategies have been developed to explore structured reasoning. The tree-of-thoughts (ToT) approach [Yao et al., 2023] models reasoning as a tree-like

structure of thought trajectories and employs search algorithms to navigate this structure and explore various potential solution paths. Extending this concept, graph-of-thoughts (GoT) [Besta et al., 2024a] generalizes this structure to a graph, potentially offering more powerful reasoning capabilities and flexible backtracking mechanisms. In recent years, prompting strategies have evolved from straightforward knowledge elicitation to supporting complex, multi-stage reasoning processes, which may or may not incorporate external support.

3 LLMs' Mathematical Reasoning from Comprehension and Generation

Developing large language models (LLMs) capable of sophisticated mathematical reasoning presents a significant challenge. Such models must effectively master abstract symbols, complex notations, and advanced mathematical concepts to solve problems. This endeavor is inherently complex, yet recent years have witnessed considerable advancements in this domain through various techniques and approaches. To navigate this evolving landscape, we propose a structured framework to organize recent innovations. By identifying the core components that underpin mathematical reasoning capabilities, this framework provides a foundation for our subsequent discussion of performance-enhancing techniques.

When addressing mathematical problems, these models first comprehend the relevant knowledge [Allen-Zhu and Li, 2024] and then proceed to decompose and solve the problem. Inspired by the mechanisms of the human brain [Collins and Koechlin, 2012, Friederici, 2011], a structured framework can be conceptualized around two high-level cognitive components: comprehension and answer generation. Comprehension requires LLMs to parse and contextualize mathematical structures and concepts—ranging from arithmetic operations to geometric relationships and theorem formalisms—utilizing both textual and visual representations. Answer generation has evolved from direct prediction to more elaborate step-by-step CoT reasoning.

Responding to mathematical problems via LLMs involves a prompt-guided process that necessitates a thorough understanding of the task and the ability to interpret formal representations. Once the task is comprehended, the model can apply reasoning to derive a solution. For arithmetic problems, this involves performing calculations and manipulating numbers according to established rules. For word problems, it requires extracting mathematical relationships from textual descriptions to formulate and solve mathematical expressions. In geometry, the model must be able to visualize and analyze spatial relationships involving shapes, sizes, angles, and their relative positions. Theorem proving entails leveraging a rich knowledge base, combining formal logical analysis with an understanding of mathematical axioms, theorems, and proof strategies to establish rigorous and sound conclusions.

3.1 Diverse Mathematical Representation Formats for LLMs to Comprehend

Developing mathematical comprehension in LLMs involves training them on a wide array of mathematical tasks and objectives, designed to address challenges across different domains and levels of difficulty. These tasks span a broad spectrum, from basic arithmetic and algebra typical of primary and middle school curricula, to more advanced topics such as geometry, calculus, and Olympiad-level problems. Furthermore, mathematical problems are presented in multiple formats, including textual descriptions, formal mathematical notation, and visual representations like graphs and geometric figures. By incorporating these diverse problem types into training datasets, LLMs are encouraged to develop a flexible understanding that enables them to process and reason about mathematical concepts across various formats and complexities.

In the early stages of artificial intelligence research [Feigenbaum et al., 1963, Bobrow et al., 1964], initial attempts to solve mathematical problems involved designing solver systems reliant on manually written rules and pattern matching. However, these solvers depended heavily on human intervention and could only handle a limited set of predefined scenarios [Slagle, 1965, Fletcher, 1985]. Subsequently, semantic parsing-based approaches were introduced [Kwiatkowski et al., 2013, Goldwasser and Roth, 2014], which aimed to transform problem statements into structured logical representations, akin to syntax trees. These methods, however, sought to explicitly encode human-derived mathematical understanding into models, thereby restricting their applicability to a narrow range of predefined mathematical problems.

With the advent of models like ChatGPT, researchers have observed that formulating models as sophisticated LLMs and scaling up both model and data size can significantly enhance their capabilities [Wei et al., 2022a]. For example, Figure 3 illustrates the impact of model scale. Inspired by this observation, training on larger-scale mathematical datasets has been shown to improve a model's performance on mathematical tasks, leading to enhanced comprehension and generalization [Yang et al., 2024b, Touvron et al., 2023]. A

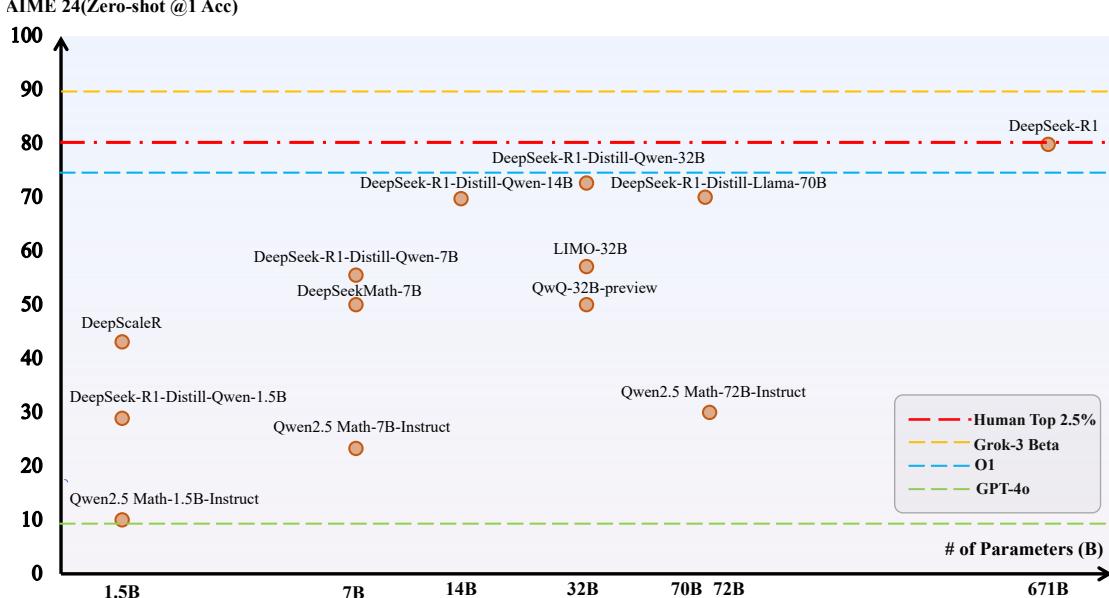


Figure 3: Scaling up LLMs leads to surpassing the top 2.5% of human participants on AIME 2024

well-curated dataset is crucial for exposing the model to diverse mathematical contexts and problem-solving patterns, which aids in the development of deeper and more robust mathematical comprehension. For instance, OpenWebMath [Paster et al., 2024] comprises 14.7B tokens sourced from Common Crawl, offering a broad range of extracted text or L^AT_EX content that includes core mathematical materials such as theorems, definitions, proofs, questions and answers, and formal mathematics, as well as interdisciplinary documents. MathPile [Wang et al., 2024e] is a high-quality, diverse math-focused corpus containing approximately 9.5B tokens. For a comprehensive list of datasets, please refer to Table 1. Adjusting the data distribution within the pretraining corpus, for example by incorporating error-correction data, has been demonstrated to enhance the higher-level reasoning abilities of LLMs [Ye et al., 2024].

Furthermore, mathematical reasoning problems frequently involve diverse inputs that extend beyond traditional text-only formats. Over the past year, multimodal mathematical reasoning has emerged as a significant research focus for multimodal large language models (MLLMs). By introducing additional modalities, such as images, audio, and video, multimodal inputs can reduce the reliance on verbose textual descriptions and provide essential information for reasoning. However, multimodal data inherently exhibit heterogeneity [Liang et al., 2024]; specifically, information from different modalities cannot always be directly mapped into a shared latent space. Moreover, compared to text, multimodal inputs often contain a greater amount of noise, such as irrelevant details in images, which imposes additional comprehension challenges [Qu et al., 2025]. Therefore, effective multimodal reasoning requires models not only to perceive and understand objects within each modality but also to perform reasoning based on key information embedded in complex multimodal contexts [Qu et al., 2025, Zheng et al., 2023]. As multimodal learning is not the primary focus of this survey, interested readers are referred to [Yan et al., 2024, Zhou et al., 2025] for a more comprehensive overview.

Training LLMs on large-scale corpora has become the mainstream approach. Many studies have investigated the internal comprehension mechanisms of LLMs and found evidence of forward planning, where models engage their knowledge and consider multiple possibilities before reacting, and memory shortcuts, where they bypass standard reasoning paths [Lindsey et al., 2025]. By defining model features as human-interpretable concepts, ranging from low-level elements (e.g., specific words or phrases) to high-level abstractions (e.g., sentiment, plans, reasoning steps), research has shown that key concepts in a prompt are actively represented and activated inside LLMs [Templeton et al., 2024]. Furthermore, LLMs have demonstrated the ability to identify the underlying structure behind surface-level problems and invoke distilled skills to solve associated tasks [Guo et al., 2024, Didolkar et al., 2024]. Also, semantic embeddings in large language models have been shown to exhibit linear structure, enabling concept relationships to be captured through vector arithmetic [Arora et al., 2019]. Collectively, these findings suggest that LLMs go beyond surface-level

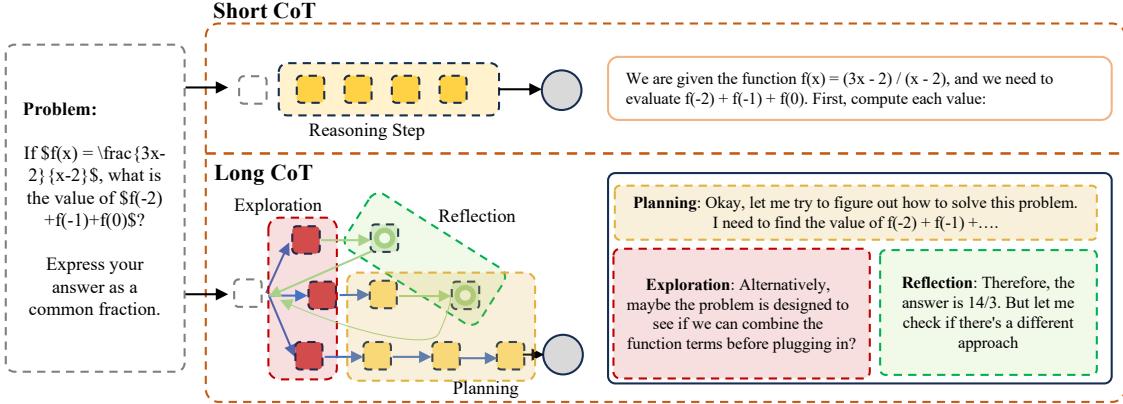


Figure 4: An illustration of short and long CoT.

pattern matching; they internalize conceptual structures and flexibly integrate them across diverse reasoning scenarios.

3.2 Chain-of-Thought: A Key Driver of LLM Generation

LLMs encode task instructions into latent representations through transformer architectures, subsequently generating responses via autoregressive sequence completion. However, direct generation methods frequently fail to produce accurate solutions for tasks requiring complex mathematical reasoning. CoT prompting addresses this limitation by guiding LLMs to generate intermediate reasoning steps before producing final answers, thereby enhancing performance on multi-step problems. This section presents a characterization of CoT and examines the underlying mechanisms that account for its effectiveness.

3.2.1 Formulation of CoT

CoT prompting induces LLMs to produce intermediate reasoning steps prior to generating final answers. This approach facilitates progressive complexity reduction, information augmentation, and irrelevant information filtering throughout the reasoning process. The demonstrated effectiveness of CoT has catalyzed the development of advanced techniques for enhancing LLM reasoning capabilities. Wei et al. [2022b] demonstrated that generating sequences of intermediate reasoning steps significantly improves LLM performance on complex reasoning tasks, particularly those involving mathematical or logical operations [Sprague et al., 2024a]. Furthermore, Kojima et al. [2022] demonstrated that LLMs can perform zero-shot CoT reasoning through the simple addition of reasoning-trigger phrases (e.g., “Let’s think step by step”) to prompts, enabling reasoning capabilities without explicit step-by-step demonstrations.

Initial CoT implementations employ shallow, linear reasoning processes characterized by sequential answer derivation and limited intermediate steps [Mirzadeh et al., 2024]. Recent advances, including OpenAI O1 [OpenAI, 2024] and DeepSeek R1 [Guo et al., 2025], introduce extended sequential CoT reasoning through test-time scaling, enabling more comprehensive and structured reasoning processes. This extended approach, termed *long CoT* [Chen et al., 2025b; Li, 2025], incorporates iterative exploration and self-reflection within problem spaces. Figure 4 illustrates the distinction between long CoT and traditional short CoT approaches. Test-time scaling enables models to identify inconsistencies in intermediate steps and implement corrective measures to maintain coherence and accuracy. Additionally, models may explore multiple solution paths and backtrack when specific approaches prove incorrect, yielding more robust and reliable inference outcomes.

3.2.2 Unveiling the Mechanism behind CoT

While CoT demonstrably enhances LLM performance in mathematical domains [Wei et al., 2022b, Kojima et al., 2022], the underlying mechanisms remain incompletely understood. A comprehensive investigation of these fundamental factors is essential for maximizing CoT effectiveness. This review examines mechanisms from both theoretical and empirical perspectives.

Table 1: A summary of pre-train math datasets. **Synth.** indicates that the dataset contains synthetic data.

Dataset	Target Domain	Synth.	Size(Tokens/Pairs)	Release Time
AMPS [Hendrycks et al., 2021]	Math Competition	✓	0.7B	Mar-2021
ProofPile [Zhangir Azerbayev, 2023]	General Math	✓	8.3B	Nov-2022
ProofPile2 [Azerbayev et al., 2024]	General Math	✓	55B	Oct-2023
OpenWebMath [Paster et al., 2024]	General Math	✗	14.7B	Oct-2023
MathPile [Wang et al., 2024e]	General Math	✓	9.5B	Dec-2023
AutoMathText [Zhang et al., 2024d]	General Math	✗	-	Feb-2024

From theoretical perspective, Feng et al. [2023] employed circuit complexity theory [Arora and Barak, 2009] to demonstrate that while bounded-depth Transformers require super-polynomial size for direct solution of basic arithmetic and equation tasks, autoregressive Transformers achieve successful solutions through CoT derivation generation with constant-size architectures. These models generalize effectively to longer input sequences, indicating that CoT enables internalization of reasoning processes rather than mere memorization of input-output mappings. From an expressiveness perspective, CoT enhances transformer capacity by enabling intermediate token generation before final answer production. This methodology provides greater expressive power for sequential reasoning tasks, as intermediate tokens function as recurrent states [Merrill and Sabharwal, 2024]. Liu et al. [2024b] formulated and analyze the hypothesis that CoT enables serial computations beyond the capabilities of vanilla transformers. Prystawski et al. [2023] applied a Bayesian framework to elucidate how intermediate steps contribute to enhanced reasoning performance, while Tutunov et al. [2023] developed a two-level hierarchical graphical model characterizing LLM reasoning sequence generation. From an in-context learning perspective, CoT improves performance by structuring compositional function learning into two phases, reducing sample complexity and enabling complex function learning beyond non-CoT method capabilities [Li et al., 2023].

From the empirical perspective, Wu et al. [2023a] demonstrated that CoT enables LLMs to maintain robust attention on semantically relevant prompt tokens. Wang et al. [2023a] identified two critical factors affecting CoT effectiveness: semantic relevance of demonstration examples and correct ordering of reasoning steps. Jin et al. [2024] established a task-dependent relationship for optimal CoT length, where simpler tasks benefit from shorter sequences while complex tasks require extended reasoning steps. Despite evidence supporting CoT superiority over direct answer generation, empirical observations reveal that language models frequently produce correct answers despite errors in intermediate reasoning. Competition-level tasks exhibit error rates within generated processes reaching 51.8% [Zheng et al., 2024a]. Sprague et al. [2024b] demonstrated that CoT responses may be suboptimal for non-symbolic reasoning tasks. Recent investigations indicate that LLM-generated reasoning steps often lack reliability and fail to accurately reflect step-by-step logical processes [Arcuschin et al., 2025, Chen et al., 2025c]. Current LLM-generated CoT does not constitute rigorous chains of logically entailed steps but rather resembles heuristic processes that leverage recursive CoT structure to generate sufficiently rich intermediate content, increasing effective model depth and supporting final answers despite intermediate step errors.

4 Methods for Boosting Reasoning

During the pre-training phase, LLMs are trained on vast amounts of data to enhance the comprehension of mathematical problems. However, pre-trained models often struggle with producing contextually appropriate responses. Therefore, improving model performance is crucial for effective mathematical reasoning. Fine-tuning plays a crucial role in enhancing a model’s instruction-following and generation capabilities.

4.1 Supervised Fine-Tuning

Supervised Fine-Tuning (SFT) aims to align a pretrained language model with high-quality, human-crafted by supervised learning. These datasets follow clear formatting patterns, which provide structural priors that help constrain and guide the model’s generation space. This structure plays a crucial role in enabling effective exploration during subsequent RL. The efficacy of SFT hinges on the construction of high-quality supervised demonstrations.

Table 2: A summary of SFT math datasets. **Synth.** indicates that the dataset contains synthetic data.

Dataset	Target Domain	Synth.	Size(Tokens/Pairs)	Release Time
NaturalProofs [Welleck et al., 2021]	Theorem Proving	✗	48K	Mar-2021
Lean Workbook [Ying et al., 2024a]	Theorem Proving	✗	57K	Jun-2024
NuminaMath [Li et al., 2024c]	Math Competition	✓	860K	Jul-2024
CARP [Zhang et al., 2023a]	General Math (zh)	✗	10M	Jun-2023
MetaMathQA [Yu et al., 2024]	General Math	✓	395K	Sep-2023
MathInstruct [Yue et al., 2024a]	General Math	✓	260K	Sep-2023
MMIQC [Liu et al., 2024a]	General Math	✓	1.57M	Jan-2024
OpenMathInstruct-1 [Toshniwal et al., 2024b]	General Math	✓	1.8M	Feb-2024
MathScaleQA [Tang et al., 2024]	General Math	✓	2M	Mar-2024
WebInstruct [Yue et al., 2024b]	General Math	✗	10M	May-2024
OpenMathInstruct-2 [Toshniwal et al., 2024a]	General Math	✓	14M	Oct-2024

4.1.1 Constructing High-Quality Demonstrations

In SFT, the quality and structural characteristics of training data fundamentally determined the behavioral patterns and computational capabilities of language models. As SFT directly modified model parameters through demonstration-based learning, model performance exhibited high sensitivity to the diversity, relevance, and correctness properties of fine-tuning datasets. The construction of high-quality instructional data that accurately reflected target applications and alignment objectives constituted a critical requirement for successful model adaptation. This section examines algorithmic approaches to data construction and curation for reasoning tasks. A comprehensive overview of relevant datasets is provided in Table 2.

Recent research by [Ho et al., 2023, Magister et al., 2023, Guo et al., 2025] demonstrated that constructing training data using strong LLMs significantly improved the mathematical reasoning capabilities of smaller LLMs. OpenMathInstruct-1 [Toshniwal et al., 2024b] augmented synthetic data generation through code-interpreter solutions produced by GPT-4. Yue et al. [2024a] utilized GPT-4 to generate Program-of-Thought (PoT) rationales, thereby enhancing tool usage capabilities. Beyond simple answer expansion, leveraging LLM knowledge for question bootstrapping enhanced problem coverage. [Liu et al., 2023, Li et al., 2024a] employed strong LLMs to generate semantically similar questions and corresponding answers. However, the generated questions exhibited limited diversity due to textual and conceptual similarity constraints. Li et al. [2024b] and Yang et al. [2024b] expanded question sets through diverse modification techniques, including numerical alterations, conceptual modifications, and complexity augmentation.

Yu et al. [2024] bootstrapped mathematical question generation through multiple techniques: Rephrasing, Self-Verification [Weng et al., 2023], FOBAR [Jiang et al., 2024], and answer augmentation. Huang et al. [2024] extracted key concepts from existing datasets and utilized these concepts alongside original problems as guidelines for generating novel questions. Bansal et al. [2024] demonstrated that data generated by smaller models exhibited greater distributional diversity. Ding et al. [2024] employed smaller models to generate questions de novo without seed data dependencies, utilizing complex augmentation constraints. Adarsh et al. [2024] ensured diversity by combining generative outputs from multiple smaller models, while Li et al. [2025b] actively preserved diversity in fine-tuned models through targeted methodologies.

Beyond high-quality solution generation, datasets incorporating erroneous reasoning enabled LLMs to develop error detection and correction capabilities essential for advanced mathematical reasoning. An et al. [2023] enhanced mathematical reasoning through the incorporation of error-correction data during the fine-tuning phase. This error-correction data, generated by GPT-4, included error identification, correction processes, and final answer generation. Liang et al. [2023] employed a teacher LLM to identify weaknesses in student LLMs and generated targeted problems for training dataset augmentation. He et al. [2023] provided comprehensive training signals including reasoning processes, foundational knowledge, and common error patterns during answer generation. Dai et al. [2024] generated dual chains of thought encompassing both correct and incorrect reasoning paths from teacher models, utilizing minimum edit distance to identify critical reasoning steps. This approach emphasized learning fundamental reasoning mechanisms rather than superficial fine-tuning.

4.1.2 Constructing Demonstrations in Long Chain-of-Thought Format

Recent studies demonstrated that enabling LLMs to generate extended CoT sequences during test-time inference significantly enhanced reasoning accuracy [Brown et al., 2024, Snell et al., 2024]. Fine-tuning methodologies increasingly adopted the long CoT paradigm, wherein allocating additional computational resources to CoT reasoning during both training and inference phases yielded consistent performance improvements. Through the construction and utilization of long CoT demonstrations, supervised fine-tuning enabled LLMs to acquire the capability of generating extended CoT outputs that exhibited diverse reasoning processes, including interactive exploration and self-reflection mechanisms.

Deepseek R1 [Guo et al., 2025], extending Deepseek R1 Zero, collected high-quality cold-start data and implemented a structured Markdown format. The system defined output format as $\langle \text{response}_{\text{process}} \rangle \langle \text{summary} \rangle$, wherein reasoning processes preceded summary generation of reasoning paths. This architectural design enhanced output readability and interpretability. Kimi K1.5 [Team et al., 2025] utilized a high-quality long CoT dataset, employing SFT as a warmup phase that improved the generation of logically coherent and detailed responses. LIMO [Ye et al., 2025] and s1 [Muennighoff et al., 2025] challenged the necessity of large sample sizes, demonstrating that minimal sample sets successfully activated reasoning capabilities in foundational LLMs. Satori [Shen et al., 2025] introduced a critic model for constructing multi-step demonstrations with reflection mechanisms, facilitating enhanced multi-step reasoning capabilities in trained models.

4.2 Reinforcement Learning

Reinforcement learning has been employed to enhance model generation capabilities through CoT reasoning. Prior to generating final answers y , models produced intermediate CoT reasoning steps $z \sim \pi_{\theta}(\cdot|x)$. Early reinforcement learning implementations in mathematical reasoning focused on optimizing standard CoT generation through outcome reward models and process reward models. Recent developments have adapted reinforcement learning techniques to optimize long CoT generation, primarily utilizing rule-based rewards to ensure coherent extended reasoning trajectories. The fundamental RL approach remains consistent across both applications, with the primary distinction lying in the reward design and the target output format—whether optimizing for standard CoT or extended long CoT reasoning processes.

4.2.1 Reward Modeling for Reasoning

In the application of reinforcement learning to improve LLMs, the reward model constitutes a critical component. A reward model assigns a numerical score $r_{\theta}(x, y) \in [0, 1]$ to estimate the probability that a solution y or intermediate step is correct for a given problem x . Reward models can be categorized into three primary types: outcome reward models (ORM), process reward models (PRM), and rule-based reward systems. This section examines each category and their respective contributions to reinforcement learning training stability.

Outcome Supervised Reward The standard approach for training ORMs in reasoning tasks involved fine-tuning an LLM as a classifier on datasets containing correct and incorrect solutions. These datasets were annotated either by human evaluators or generated from frozen LLMs during self-improvement processes, utilizing binary cross-entropy loss [Cobbe et al., 2021, Li et al., 2022]. Given a reward-modeling dataset $\mathcal{D}_{\text{RM}} = \mathcal{D}_{\text{incorrect}} \cup \mathcal{D}_{\text{correct}}$, discriminative reward models were trained according to:

$$\mathcal{L}(\theta, \mathcal{D}_{\text{RM}}) = -\mathbb{E}_{(x, y^+) \sim \mathcal{D}_{\text{correct}}} [\log r_{\theta}(x, y^+)] - \mathbb{E}_{(x, y^-) \sim \mathcal{D}_{\text{incorrect}}} [\log (1 - r_{\theta}(x, y^-))],$$

where $r_{\theta}(x, y) = \sigma(z_{\text{cls}})$ and $z_{\text{cls}} = \text{logit}_{\theta}(\text{cls} | y, x)$. Here, y^+ denoted correct solutions, y^- denoted incorrect solutions, and cls corresponded to a special vocabulary token. Zhang et al. [2024c] employed a balanced data mixture between correct ($\mathcal{D}_{\text{correct}}$) and incorrect ($\mathcal{D}_{\text{incorrect}}$) problem-solution pairs. He et al. [2024b] proposed a methodological shift from binary classification loss to preference-based loss for verifier training.

Multiple studies addressed the reduction of human annotation requirements through the LLM-as-a-verifier approach, wherein off-the-shelf LLMs evaluated solutions via prompting [Madaan et al., 2023, Welleck et al., 2023, Zhang et al., 2024a]. These LLMs assigned rewards or penalties to both outcomes and intermediate steps based on reasoning quality and alignment with predefined criteria. While these methods demonstrated effectiveness in language tasks, their performance in mathematical problem-solving remained limited. Zhang et al. [2024c] proposed GenRM, which outperformed discriminative verifiers by integrating CoT reasoning into the verification process.

In applications, ORMs evaluated responses at generation completion to guide policy optimization. Deepseek-Math [Shao et al., 2024] leveraged an ORM to optimize policies using the GRPO algorithm, while Qwen2.5-Math [Yang et al., 2024b] employed a hybrid reward mechanism combining ORM with rule-based rewards to enhance training stability and performance.

Process Supervised Reward Recent investigations by Lightman et al. [2023] indicated that PRMs outperformed ORMs in reasoning tasks. PRMs ($P \times S \rightarrow \mathbb{R}^+$) assigned scores to individual reasoning steps within solution s , typically trained using:

$$\mathcal{L}_{\text{PRM}} = \sum_{i=1}^K \left(y_{s_i} \log r_{s_i} + (1 - y_{s_i}) \log(1 - r_{s_i}) \right)$$

where y_{s_i} represented the ground truth label for step s_i (the i -th step of s), r_{s_i} denoted the sigmoid score assigned by the PRM to step s_i , and K indicated the total number of reasoning steps in s .

Given the labor-intensive nature of process reward annotation, Zhang et al. [2024b] developed methods for learning process rewards through final reward guidance. REFINER [Paul et al., 2024] provided structured feedback on reasoning errors through intermediate step evaluation. Setlur et al. [2024] trained process reward models to measure the likelihood of future correct response generation by incorporating additional policies. TSMC utilized intermediate target distributions for resampling during Monte Carlo processes [Feng et al., 2024b]. Luo et al. [2024] proposed a divide-and-conquer style MCTS algorithm for efficient collection of high-quality process data.

Compared to outcome rewards, process rewards provided more detailed feedback, demonstrating greater potential to enhance generators [Wu et al., 2023b]. Wang et al. [2024d] utilized automatically constructed PRMs to supervise LLMs through step-by-step PPO. Implicit PRM [Yuan et al., 2024] extended ORM training by implicitly learning process labels without requiring additional annotations. Lin et al. [2025] introduced a training framework combining process-level and outcome-level binary feedback to guide LLMs toward more reliable reasoning trajectories.

Rule-Based Reward In mathematical reasoning contexts, rule-based rewards were designed based on the verifiability of final answers and intermediate reasoning steps. The reward function evaluated whether the final answer exactly matched the ground truth solution, returning a binary reward: $r = R(y^*, y)$, where $r = 1$ if and only if the model's final answer exactly matched the ground truth y^* . Format rewards could additionally be incorporated to distinguish reasoning paths from final answers [Guo et al., 2025].

Recent studies [Guo et al., 2025, Shen et al., 2025, Luo et al., 2025, Yue et al., 2025] increasingly focused on rule-based rewards due to their provision of accurate and reliable reward signals that supported stable reinforcement learning training. Compared to learned reward models, rule-based rewards mitigated issues such as reward hacking by providing deterministic feedback grounded in task-specific logic. ReFT [Trung et al., 2024] explored rule-based RL with SFT warm-up, achieving significantly superior performance compared to SFT alone in mathematical domains. Deepseek R1 [Guo et al., 2025] employed a rule-based reward function and achieved continued improvement beyond 8,000 training steps, ultimately attaining performance comparable to or exceeding OpenAI's o1.

4.2.2 Reinforcement Learning

Inspired by reinforcement learning from human feedback (RLHF), reinforcement learning was introduced for mathematical reasoning enhancement in LLMs [Trung et al., 2024, Kazemnejad et al., 2024, Gehring et al., 2024, Setlur et al., 2024, Li et al., 2024e]. While supervised fine-tuning relied on offline datasets and exhibited susceptibility to compounding errors due to distributional shifts, RL mitigated these limitations through online, reward-driven optimization. In response evaluation frameworks, outcome rewards assessed entire responses by evaluating final outcome confidence levels. Process rewards provided scores at each reasoning step conclusion, delivering more comprehensive and informative supervision signals.

LLM-Specific RL Algorithms Proximal Policy Optimization (PPO) [Schulman et al., 2017] emerged as the predominant RL method for RLHF following its adoption in ChatGPT. However, as a general-purpose RL algorithm, PPO required an additional critic model, substantially increasing computational costs and GPU memory consumption. To address these limitations, ReMax [Li et al., 2024f] pioneered the elimination of the critic model by recognizing that reinforcement learning for LLMs exhibited simpler properties than

general reinforcement tasks. ReMax leveraged the REINFORCE algorithm and introduced greedy sampling responses to calculate reward baselines, maintaining training stability without the computational overhead of a critic model.

Building upon this approach, RLOO [Ahmadian et al., 2024] similarly eliminated the critic model but employed Monte Carlo sampling for improved baseline estimation. GRPO [Shao et al., 2024] adopted the PPO objective while calculating advantages from group-normalized rewards, achieving comparable performance without the critic model overhead. Reinforce++ [Hu, 2025] employed similar techniques with additional optimizations for training efficiency. Furthermore, addressing the challenge of rapidly decreasing policy entropy during training, which limited exploration capabilities, DAPO [Yu et al., 2025] introduced the Clip-Higher strategy, building upon GRPO to maintain exploration throughout the training process.

Reinforcement Learning for Long CoT Long CoT reinforcement learning represented a paradigm shift in optimizing LLMs by leveraging RL with rule-based rewards to enhance extended reasoning capabilities. While OpenAI’s o1 model demonstrated the effectiveness of allocating increased computational resources during both training and inference, implementation details remained proprietary. DeepSeek R1 achieved comparable or superior performance through a purely RL-based approach, catalyzing renewed research interest in long CoT RL methodologies. The fundamental breakthrough involved emulating System 2-style reasoning through three critical factors:

- **Golden Reward:** Previous vanilla RL methods frequently encountered reward model saturation, wherein reward model inaccuracies induced reward hacking behaviors. DeepSeek R1 addressed this limitation by employing rule-based rewards determined through comparison with ground truth answers, enabling more reliable correctness evaluation and stable exploration dynamics.
- **Scaling CoT Length:** Extending CoT from short to long sequences enhanced the model’s exploration capabilities and strengthened intrinsic reasoning abilities. Increased generation length enabled exploration of diverse response spaces and facilitated the emergence of advanced capabilities including verification and reflection mechanisms, promoting deeper analytical thinking.
- **Pure RL Training:** DeepSeek R1-zero implemented pure RL training by bypassing the SFT phase entirely, directly applying reinforcement learning to the base model. This approach enabled exploration of broader distributional spaces beyond the constraints imposed by SFT datasets, thereby maximizing exploration potential [Zeng et al., 2025].

Regarding reward mechanism design, R1 experimented with PRM integration but encountered intermediate step labeling challenges. Alternative approaches emerged: Kimi K1.5 [Team et al., 2025] utilized ORM and GenRM [Ankner et al., 2024, Gao et al., 2024] for reward generation, continuing training after SFT with long CoT data to achieve performance comparable to o1. Satori [Shen et al., 2025] explicitly integrated verifier and search capabilities within a unified model architecture, internalizing reasoning mechanisms to enhance problem-solving efficiency. Self-Reward methodologies [Xiong et al., 2025b, Pang et al., 2023] incorporated self-critique and verification capabilities through SFT, subsequently enhancing these abilities via RL to enable robust reasoning and self-improvement during inference. RFTT [Zhang et al., 2025] integrated multiple cognitive capabilities through specialized tokens such as `(analyze)`, `(verify)`, and `(refine)`, enabling structured long CoT construction.

4.2.3 Direct Preference Optimization

To address the complexity of online RL optimization, Direct Preference Optimization (DPO) [Rafailov et al., 2023] was proposed as an alternative approach that directly utilized offline pair-wise preference data for model optimization. This methodology significantly streamlined the training pipeline and enhanced training stability. Unlike traditional methods such as RLHF, which required training a reward model followed by policy model optimization, DPO eliminated the necessity for separate reward model training, substantially simplifying the training process. DPO demonstrated particular effectiveness in optimizing language models through alignment with human preferences, facilitating more efficient model fine-tuning [Dubey et al., 2024].

Pang et al. [2024] modified the DPO loss function by introducing an additional negative log-likelihood term. Through iterative optimization of reasoning paths via comparative evaluation of winning and losing pairs, this approach achieved improvements in mathematical reasoning tasks. SimPO [Meng et al., 2024] incorporated average log probability calculations, effectively aligning with model generation while eliminating dependency on reference models, thereby improving computational and memory efficiency.

In the context of process-level optimization, Step-DPO [Lai et al., 2024] extended DPO by specifically targeting long-chain reasoning, producing notable improvements in mathematical word problem solving. Additionally, Zeng et al. [2024] modified DPO to optimize policies at the token level, further refining LLM alignment with human preferences.

Limitations of DPO. Despite its widespread adoption as an alternative to RLHF, DPO fundamentally operates within a supervised learning paradigm³ rather than a true reinforcement learning framework, resulting in several critical limitations. First, DPO lacks the exploration capabilities inherent to RL methods, as it cannot utilize out-of-preference data or policy-generated samples to discover novel high-quality responses beyond the training distribution [Li et al., 2024e, Xu et al., 2024]. Second, theoretical analyses revealed that DPO exhibits asymmetric optimization dynamics, decreasing the probability of dispreferred responses faster than increasing preferred ones, which explains its sensitivity to SFT quality and tendency to hinder learning capacity [Feng et al., 2024a]. Third, empirical studies demonstrated that DPO’s implicit reward model generalizes poorly under distribution shifts [Li et al., 2024e, Lin et al., 2024a], with accuracy drops reaching 7% in out-of-distribution settings compared to explicit reward models [Lin et al., 2024a]. These collective findings indicate that while DPO offers computational efficiency advantages, it sacrifices the exploratory benefits, generalization capabilities, and adaptive flexibility that characterize true reinforcement learning approaches.

4.3 Test Time Inference with Structural Search

Test-time scaling through increased generation burden has been shown to enhance model mathematical reasoning capabilities [Wu et al., 2024b, Snell et al., 2024]. While majority voting [Wang et al., 2023c] enables answer derivation through extensive data collection, complex mathematical problems with high token budgets necessitate tree search methods for exhaustive solution space exploration. The distinction between these approaches is illustrated in Figure 5.

In the decision-making process of LLMs, structured tree search integration enhances planning through systematic exploration. This approach enables LLMs to concurrently maintain and investigate multiple potential solutions during generation, dynamically assessing current states while strategically planning ahead or backtracking to refine reasoning paths. Yao et al. [2023] proposed the Tree-of-Thoughts (ToT) framework, integrating depth-first search (DFS) and breadth-first search (BFS) to enable systematic exploration and backtracking over intermediate reasoning steps. Besta et al. [2024b] extended ToT by introducing the Graph of Thoughts (GoT) structure, supporting cyclic and interconnected reasoning steps that capture complex dependencies and enable iterative refinement processes. Forest-of-Thought (FoT) [Bi et al., 2024] further advanced this paradigm by integrating multiple reasoning trees, facilitating collective decision-making for solving complex mathematical problems with enhanced reliability and efficiency.

However, these methods typically relied on heuristic or simplistic search strategies, potentially limiting efficiency in large search spaces. To address this limitation, Monte Carlo Tree Search (MCTS) emerged as a powerful decision-making algorithm for enhanced search efficiency. MCTS constructs structural trees for optimal path exploration, where each node represents a reasoning state [Browne et al., 2012, Chaslot et al., 2008]. During tree expansion, future states undergo simulation to identify valuable nodes, followed by value function updates through the backup process to refine node evaluations [Coulom, 2006].

Recent research demonstrated successful MCTS integration with LLMs. Reasoning-via-Planning (RAP) [Hao et al., 2023] integrated MCTS to enhance reasoning by repurposing LLMs as both world models for state prediction and reasoning agents for action generation. AlphaMath [Chen et al., 2024a] employed MCTS during inference-time reasoning, integrating value models with LLMs to autonomously generate process supervision and step-level evaluation signals within MCTS rollouts. LLaMA-Berry [Zhang et al., 2024a] combined MCTS with iterative Self-Refine [Madaan et al., 2023], dynamically optimizing reasoning paths through exploration and self-critique, guided by pairwise preference reward models (PPRM) that globally evaluated solution quality via Enhanced Borda Count—a method combining basic Borda Count algorithms with transitive closure of preferences computed using the Floyd-Warshall algorithm [Warshall, 1962].

Beyond inference-time applications, MCTS integration during training improved exploration efficiency. Wan et al. [2024] proposed TS-LLM, leveraging learned value functions and AlphaZero-like algorithms to guide

³Here we categorize DPO as a supervised learning paradigm rather than reinforcement learning or offline reinforcement learning, as it fundamentally operates through likelihood maximization (and minimization for negative samples) over fixed preference datasets.

LLMs during training. rStar-Math [Guan et al., 2025] utilized MCTS for generating candidate reasoning steps, guided by process preference models evaluating step-wise quality through multi-turn training.

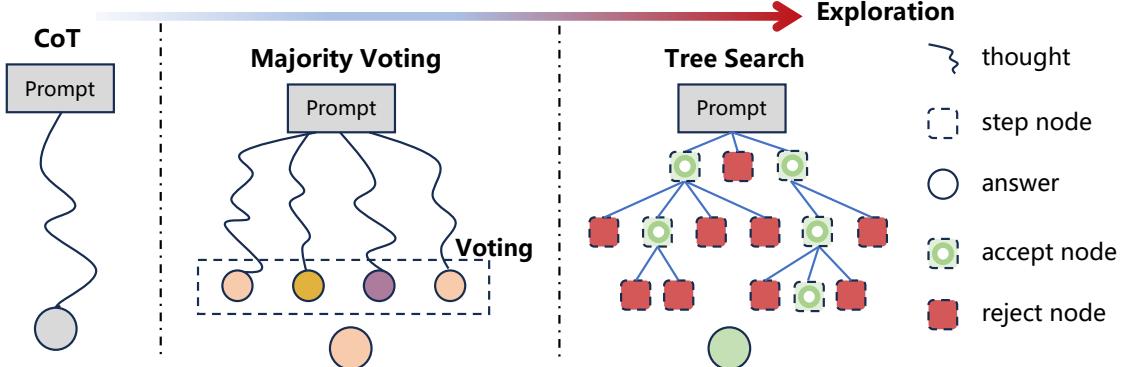


Figure 5: Schematic illustrating of different method.

4.4 Self-Improvement

The construction of extensive rationale datasets for inducing language model rationale generation in downstream tasks presents significant resource constraints in terms of both time and cost. Self-improvement methodologies have emerged as a viable alternative, leveraging the inherent capabilities of LLMs to refine CoT reasoning processes. A fundamental challenge in this approach concerns the generation and selection of high-quality responses to ensure methodological reliability and effectiveness.

During the inference phase, several approaches have been developed to enhance response quality. Self-Refine [Madaan et al., 2023] implements an iterative refinement mechanism utilizing feedback loops. Self-Verification [Weng et al., 2023] incorporates answer validation procedures to mitigate errors in generated responses. RISE [Qu et al., 2024] facilitates inference-time self-improvement through fine-tuning on distilled recursive introspection data.

Contemporary research has systematically investigated the enhancement of LLM capabilities through fine-tuning on self-synthesized responses, which can be categorized according to two distinct methodological approaches. The first approach emphasizes the generation of superior solution quality. STaR [Zelikman et al., 2022] employs few-shot prompting with ground truth answers to generate enhanced rationales. Quiet-STaR [Zelikman et al., 2024] implements token-level thought rationale generation to improve prediction accuracy. REST^{EM} [Singh et al., 2024] applies the Expectation-Maximization algorithm [Anthony et al., 2017] for iterative answer quality refinement. Both rStar-Math [Guan et al., 2025] and ReST-MCTS* [Zhang et al., 2024b] incorporate Monte Carlo Tree Search to augment exploration capabilities. SIRLC [Pang et al., 2023] utilizes the LLM as an intrinsic reward model for reinforcement learning-based training. ReGenesis [Peng et al., 2024] constructs abstract-to-concrete response pathways through self-synthesized guidelines, demonstrating robust generalization to out-of-distribution tasks. Auto-CEI [Zhao et al., 2024] employs expert iteration [Anthony et al., 2017] for near-policy reasoning path exploration, implementing error correction mechanisms to minimize cumulative errors.

The second approach addresses the selection of high-quality responses from generated candidates. Huang et al. [2023] applies self-consistency methods [Wang et al., 2023c] for high-confidence response identification. V-STaR [Hosseini et al., 2024] implements a DPO-based verifier for solution correctness assessment. Both rStar-Math and ReST-MCTS* incorporate process reward guidance to ensure step-level error mitigation and reasoning trajectory reliability.

4.5 Utilizing External Knowledge

Large language models, despite demonstrating substantial reasoning capabilities, exhibit systematic limitations in tasks requiring precise computational accuracy or access to real-time information, frequently manifesting as hallucinations or factual inaccuracies. Recent methodological developments have addressed these constraints through two principal approaches: external tool integration and retrieval-augmented generation (RAG).

External tools comprise specialized functional modules accessed through standardized APIs or web interfaces, including web search engines, symbolic computation systems, and code execution environments. These tools extend model capabilities beyond static parameter encoding by enabling real-time information retrieval, precise mathematical computation, and dynamic program execution. In contexts involving complex mathematical expressions or temporally-sensitive queries, LLMs demonstrate increased susceptibility to generating plausible yet factually incorrect responses, a limitation that external tool integration effectively addresses. Das et al. [2024] developed a template-based decision framework for tool invocation timing, though this approach lacks support for sequential or interdependent tool utilization. Tang et al. [2023] implemented a self-instruct paradigm for diverse tool-use scenario generation, similarly constrained by limited multi-tool coordination capabilities. Das et al. [2024] established a comprehensive framework integrating knowledge retrieval (Bing Web Search), program generation and execution (Python), and symbolic computation (WolframAlpha API), demonstrating measurable performance improvements on mathematical benchmarks. Nevertheless, mathematical problem-solving remains a significant challenge, as current state-of-the-art LLMs lack universally applicable strategies for reliable external tool integration, necessitating development of more adaptive methodologies.

Retrieval-augmented generation (RAG) represents an alternative approach for incorporating external knowledge, retrieving relevant documents to enhance accuracy, reduce hallucinations in knowledge-intensive tasks, and improve result verifiability [Gao et al., 2023b, Asai et al., 2023]. Complex mathematical problems frequently require specialized or temporally current information not encoded in model parameters, establishing RAG as an essential complementary technique. Basic RAG implementations retrieve documents based on query embedding similarity, directly appending retrieved content to input prompts. Advanced RAG methodologies incorporate pre-retrieval operations (e.g., query reformulation) and post-retrieval processes (e.g., relevance re-ranking, content filtering) to optimize retrieved information quality and applicability [Gao et al., 2023b]. Through external knowledge integration, RAG facilitates access to relevant theorems, definitions, and established problem-solving strategies beyond model parametric knowledge.

Empirical evidence demonstrates RAG effectiveness in mathematical reasoning contexts. Yang et al. [2025] established that knowledge base retrieval effectively guides accurate proof generation. Levonian et al. [2023] documented improvements in faithfulness and grounding, particularly for theorem-dependent problems. Dixit and Oates [2024] demonstrated that step-wise retrieved references enhance mathematical word problem performance through concrete fact anchoring of reasoning processes. Recent methodological advances [Jin et al., 2025, Chen et al., 2025a] have explored reinforcement learning approaches for training LLMs in dynamic information retrieval and integration. Active exploration and verification mechanisms in RAG-driven reasoning constitute promising research directions, as enhanced model capabilities for external knowledge identification, assessment, and integration hold significant potential for advancing mathematical problem-solving performance.

5 Discussions

The rapid advancement of mathematical reasoning in LLMs has revealed fundamental insights about the nature of reasoning capabilities, training methodologies, and inherent limitations. This section synthesizes key findings and implications across multiple dimensions of this evolving field.

Fundamental Nature of Chain-of-Thought Reasoning CoT represents more than a prompting technique. It fundamentally enables models to learn and replicate structured procedures or algorithms rather than merely predicting outcomes. Evidence from diverse domains supports this interpretation: in navigation tasks, Yang et al. [2022] and Lehnert et al. [2024] demonstrated that CoT helps LLMs internalize planning and search algorithms through step-by-step decomposition. Similarly, the SYNAPSE framework [Zheng et al., 2024b] applies CoT to web agent tasks through trajectory-as-exemplar prompting, enabling models to infer multi-step procedures from contextual demonstrations. This procedural learning capability positions CoT as a blueprint for reliable algorithmic problem-solving.

The expressive power of CoT, however, faces theoretical boundaries. Research by Merrill and Sabharwal [2024] establishes that the number of intermediate steps critically determines a transformer's computational capability: logarithmic steps provide modest enhancements, linear steps enable recognition of all regular languages when combined with projected pre-norm, while polynomial-time problems require computationally expensive polynomial steps. Alternative architectures, such as looped transformers [Yang et al., 2024c], offer potential solutions by incorporating iterative characteristics while maintaining efficiency with significantly reduced parameter counts.

Training Paradigm Trade-offs and Challenges The evolution from supervised fine-tuning to reinforcement learning reveals critical trade-offs in model development. While SFT significantly enhances instruction-following capabilities through alignment with human-annotated data, it simultaneously induces a diversity-alignment dilemma. Extensive SFT reduces generation diversity [Li et al., 2025b, Wang et al., 2024a], potentially causing mode collapse where models produce consistently similar outputs. This diversity reduction proves particularly detrimental for downstream reinforcement learning applications requiring broad exploration spaces [Zeng et al., 2025]. Recent evidence suggests that bypassing SFT in favor of direct RL optimization preserves exploratory capabilities, yielding superior performance in reasoning tasks [Guo et al., 2025, Zeng et al., 2025].

Reward modeling presents another fundamental challenge. While Process Reward Models demonstrate superiority over Outcome Reward Models in reasoning tasks, obtaining high-quality reward signals remains prohibitively expensive. Traditional approaches rely on costly human-annotated datasets like PRM800k [Lightman et al., 2023]. Recent automated annotation methods using Monte Carlo sampling and MCTS [Wang et al., 2024d] or process preference reward models [Guan et al., 2025] reduce annotation costs but remain vulnerable to reward hacking and struggle with generating precise reward scores [Guo et al., 2025].

Boundaries of Reasoning Improvement A notable insight emerging from recent studies concerns the fundamental ceiling of reasoning improvements through RL. The insight lies in how different RL training paradigms affect model performance, as measured by the Pass@k metric. For example, vanilla RL tends to improve Pass@1 performance while leaving Pass@k largely unchanged [Shao et al., 2024]. In contrast, long CoT RL leads to gains in Pass@k metrics, reflecting enhanced quality and diversity across the top responses [Zeng et al., 2025]. This improvement is attributed to longer generated sequences, which allow for the exploration of a broader set of reasoning pathways.

Meanwhile, multiple investigations [Gandhi et al., 2025, Chang et al., 2025, Zeng et al., 2025, Zhao et al., 2025, Yue et al., 2025] consistently report that such gains are primarily a result of activating and leveraging reasoning capabilities already present in the base models, rather than instilling fundamentally new abilities. For instance, Gandhi et al. [2025] and Chang et al. [2025] show that models such as Qwen inherently possess strong verification and backtracking skills, which RL simply helps to reveal. Similarly, Yue et al. [2025] empirically demonstrates that the upper bounds of sample performance during RL optimization are dictated by the capabilities of the base model, as reflected in Pass@k scores. These findings suggest that RL serves more as a mechanism for unlocking latent knowledge than for expanding a model’s reasoning capacity.

Nonetheless, it is questionable whether Pass@k (and the choice of k in particular) is an appropriate and comprehensive measure of both base and fine-tuned model capabilities. If the metric does not fully capture the nuances of reasoning ability, then conclusions drawn solely from Pass@k may be limited in their validity.

Integration with Structural Search The integration of structural search methods with LLMs represents a paradigm shift from single-path generation to systematic exploration with strategic backtracking capabilities. Structural search provides a solid approach for reasoning. This hybrid approach, combining classical algorithms with modern language models, demonstrates significant potential for tackling complex mathematical challenges [Yao et al., 2023, Guan et al., 2025]. However, the fundamental constraint imposed by base model capabilities suggests that future advances may require innovations in pre-training methodologies or architectural designs rather than solely relying on fine-tuning optimization techniques.

These findings collectively indicate that while current methods effectively unlock and organize existing model capabilities, transcending these boundaries will likely require fundamental advances in how models acquire and represent mathematical knowledge during initial training phases. The field stands at a critical juncture where understanding these limitations can guide more targeted research efforts toward genuinely expanding reasoning capabilities rather than merely optimizing their expression.

6 Potential Directions in Mathematical Reasoning for LLMs

Mathematical reasoning constitutes a fundamental component in the progression of Large Language Models toward Artificial General Intelligence. Despite substantial advances, significant challenges persist in computational approaches to mathematical problem-solving. This section examines three critical research directions: (1) extending LLM performance boundaries (Sections 6 and 6), (2) enhancing reasoning efficiency (6), and (3) enabling cross-domain reasoning generalization (Section 6). Each direction presents distinct challenges requiring systematic investigation of current limitations and potential solutions.

Enhancing Reasoning through RL Exploration The emergence of DeepSeek-R1 demonstrates the transformative potential of long CoT reinforcement learning in advancing mathematical reasoning capabilities, particularly for multi-step problems including competition-level mathematics and complex logical derivations. While RL optimization aims to maximize base model capabilities, current implementations remain substantially below theoretical performance ceilings. Empirical evidence indicates that policies exhibit tendency toward reinforcing common reasoning patterns with limited exploration diversity [Xiong et al., 2025a], resulting in convergence to local optima. The fundamental challenge involves efficient exploration of diverse reasoning pathways within constrained sampling budgets.

A promising research direction involves **leveraging compact representation spaces for exploration**. This approach constructs latent actions within low-dimensional spaces to guide token generation more efficiently. For instance, BWArea [Jia et al., 2024] and CoLA [Jia et al., 2025] utilize future information to infer latent actions, enabling operation within compact action spaces that facilitate more comprehensive information gathering compared to conventional token-level optimization. This design paradigm offers potential for achieving more efficient exploration while maintaining computational tractability.

Knowledge-Augmented Reasoning Complex mathematical problems extending beyond model parametric knowledge—including open problems and mathematical conjectures—necessitate integration of external knowledge sources. The incorporation of external tools and retrieval-augmented generation represents a critical direction for expanding problem-solving capabilities [Shen, 2024, Gao et al., 2023b]. External tools encompassing calculators, code interpreters, and geometric visualizers provide essential augmentation layers. The key capability requirement involves determining optimal timing and methodology for external resource invocation, requiring models to accurately assess their limitations and strategically leverage external assistance.

A primary research direction involves **training LLMs for effective external module interaction**. Recent investigations have explored optimization of interaction strategies through long CoT RL to minimize erroneous or unproductive external calls during reasoning processes [Jin et al., 2025, Chen et al., 2025a, Li et al., 2025a]. However, scaling challenges emerge as the number of available modules increases, each presenting unique capabilities and constraints. The **learnware paradigm** [Zhou and Tan, 2024] offers a structured solution through formalized registration, retrieval, and reuse mechanisms for specialized modules. By encapsulating external interactions into modular capabilities, this approach provides scalable pathways for enhancing reasoning performance across diverse task domains.

Optimality of Reasoning Paths Long CoT generation frequently produces sub-optimal reasoning trajectories due to optimization processes prioritizing final answer correctness over path efficiency [Sui et al., 2025]. Outcome-based reward functions evaluate terminal outputs while neglecting intermediate reasoning quality, treating responses as correct regardless of logical errors within reasoning chains. This approach fails to capture the importance of intermediate steps for achieving computational efficiency and maintaining logical consistency. Consequently, LLMs often generate unnecessarily lengthy or convoluted inference paths despite reaching correct conclusions [Wang et al., 2025].

Addressing this limitation requires **verification-aware optimization strategies**. Formal language integration with built-in proof systems (e.g., Lean) enables models to perform self-verification of intermediate steps, ensuring logical consistency while improving reasoning efficiency. Additionally, external tool integration for intermediate step validation provides mechanisms for solution correctness verification. However, effective tool utilization within long CoT frameworks for accuracy enhancement remains an active area of investigation requiring further theoretical and empirical development.

Reasoning Generalization to Open Domains While LLMs demonstrate cross-domain generalization within mathematics such as solving AIME24 problems following MATH dataset training-robust generalization to open-ended domains presents fundamental challenges. Empirical investigations reveal that LLMs internalize abstract, compositional reasoning structures termed meta-skills, enabling flexible reuse of high-level cognitive patterns across contexts [Guo et al., 2024, Didolkar et al., 2024]. However, effective triggering and utilization of these meta-skills in complex, open-ended scenarios remains unresolved.

CoT prompting facilitates meta-level problem-solving pattern learning, providing limited task adaptation flexibility. This approach proves insufficient for broader generalization due to models' inability to internalize foundational reasoning principles including systematic abstraction and ideological thinking. Consequently, models rely on surface-level statistical correlations rather than adaptive reasoning for dynamic tasks such as scientific hypothesis generation. Domain-specific training becomes essential for addressing open-domain ap-

plications, with the primary challenge residing in generalizable reward function design. Unlike mathematics with clear binary feedback, complex real-world scenarios demand nuanced reward mechanisms for effective higher-order reasoning skill development.

Two promising research directions emerge: (1) **Reward model scaling**, inspired by pretraining scaling laws, involves expanding reward model training scales through learning from diverse pretraining datasets; (2) **Generative reward models** leverage generative architectures to construct reward functions, applying long CoT RL for progressive cultivation of complex, generalizable critic capabilities. These approaches offer potential pathways for bridging the gap between mathematical reasoning proficiency and general-purpose problem-solving capabilities.

7 Conclusion

This survey provides a comprehensive examination of mathematical reasoning in Large Language Models, organizing recent advances within a unified framework that distinguishes between comprehension (problem understanding) and generation (solution synthesis) capabilities. We analyze the progression from training-free prompting techniques to sophisticated fine-tuning and inference-time scaling methods, revealing key insights about the current state and future directions of the field.

We identify three critical challenges defining the current frontier: (1) efficient exploration within constrained sampling budgets during RL optimization, (2) sub-optimal reasoning trajectories despite correct answers, and (3) limited generalization from mathematical to open-domain problem-solving. Promising research directions include compact representation spaces for exploration, verification-aware optimization, and scalable reward modeling.

As mathematical reasoning remains a critical benchmark for evaluating genuine understanding in the pursuit of AGI, this survey aims to serve as both a comprehensive reference for researchers and an accessible entry point for newcomers. We hope this work accelerates progress by providing a clear understanding of achievements, limitations, and the path forward in this vital area of AI research.

References

- American invitational mathematics examination (aime). <https://www.maa.org/math-competitions/aime>, 2024. Mathematical Association of America.
- S. Adarsh, K. Shridhar, C. Gulcehre, N. Monath, and M. Sachan. SIKeD: Self-guided iterative knowledge distillation for mathematical reasoning. *CoRR*, abs/2410.18574, 2024.
- A. Ahmadian, C. Cremer, M. Gallé, M. Fadaee, J. Kreutzer, O. Pietquin, A. Üstün, and S. Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.
- J. Ahn, R. Verma, R. Lou, D. Liu, R. Zhang, and W. Yin. Large language models for mathematical reasoning: Progresses and challenges. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, Student Research Workshop*, 2024.
- Z. Allen-Zhu and Y. Li. Physics of language models: Part 3.1, knowledge storage and extraction. In *Proceedings of the 41th International Conference on Machine Learning*, 2024.
- A. Amini, S. Gabriel, P. Lin, R. Koncel-Kedziorski, Y. Choi, and H. Hajishirzi. MathQA: Towards interpretable math word problem solving with operation-based formalisms. *CoRR*, abs/1905.13319, 2019.
- S. An, Z. Ma, Z. Lin, N. Zheng, J.-G. Lou, and W. Chen. Learning from mistakes makes LLM better reasoner. *CoRR*, abs/2310.20689, 2023.
- Z. Ankner, M. Paul, B. Cui, J. D. Chang, and P. Ammanabrolu. Critique-out-loud reward models. *CoRR*, abs/2408.11791, 2024.
- T. Anthony, Z. Tian, and D. Barber. Thinking fast and slow with deep learning and tree search. In *Advances in Neural Information Processing Systems 31*, 2017.
- I. Arcuschin, J. Janiak, R. Krzyzanowski, S. Rajamanoharan, N. Nanda, and A. Conmy. Chain-of-Thought reasoning in the wild is not always faithful. *CoRR*, abs/2503.08679, 2025.
- S. Arora and B. Barak. *Computational Complexity - A Modern Approach*. Cambridge University Press, 2009.

- S. Arora, Y. Li, Y. Liang, T. Ma, and A. Risteski. A latent variable model approach to PMI-based word embeddings. *CoRR*, abs/1502.03520, 2019.
- A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *Proceedings of the 12th International Conference on Learning Representations*, 2023.
- A. Askell, Y. Bai, A. Chen, D. Drain, D. Ganguli, T. Henighan, A. Jones, N. Joseph, B. Mann, N. DasSarma, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, J. Kernion, K. Ndousse, C. Olsson, D. Amodei, T. B. Brown, J. Clark, S. McCandlish, C. Olah, and J. Kaplan. A general language assistant as a laboratory for alignment. *CoRR*, abs/2112.00861, 2021.
- J. Austin, A. Odena, M. Nye, M. Bosma, H. Michalewski, D. Dohan, E. Jiang, C. Cai, M. Terry, Q. Le, et al. Program synthesis with large language models. *CoRR*, abs/2108.07732, 2021.
- Z. Azerbayev, H. Schoelkopf, K. Paster, M. D. Santos, S. M. McAleer, A. Q. Jiang, J. Deng, S. Biderman, and S. Welleck. Llemma: An open language model for mathematics. In *Proceedings of the 12th International Conference on Learning Representations*, 2024.
- Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862, 2022.
- H. Bansal, A. Hosseini, R. Agarwal, V. Q. Tran, and M. Kazemi. Smaller, weaker, yet better: Training LLM reasoners via compute-optimal sampling. *CoRR*, abs/2408.16737, 2024.
- M. Besta, N. Blach, A. Kubicek, R. Gerstenberger, M. Podstawski, L. Gianinazzi, J. Gajda, T. Lehmann, H. Niewiadomski, P. Nyczyk, and T. Hoefler. Graph of thoughts: Solving elaborate problems with large language models. In M. J. Wooldridge, J. G. Dy, and S. Natarajan, editors, *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, 2024a.
- 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 38th AAAI Conference on Artificial Intelligence*, 2024b.
- Z. Bi, K. Han, C. Liu, Y. Tang, and Y. Wang. Forest-of-Thought: Scaling test-time compute for enhancing LLM reasoning. *CoRR*, abs/2412.09078, 2024.
- D. Bobrow et al. Natural language input for a computer problem solving system. *Ph. D. Thesis, Department of Mathematics*, 1964.
- B. Brown, J. Juravsky, R. Ehrlich, R. Clark, Q. V. Le, C. Ré, and A. Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *CoRR*, abs/2407.21787, 2024.
- C. B. Browne, E. Powley, D. Whitehouse, S. M. Lucas, P. I. Cowling, P. Rohlfschagen, 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*, 4(1):1–43, 2012.
- E. Y. Chang, Y. Tong, M. Niu, G. Neubig, and X. Yue. Demystifying long chain-of-thought reasoning in LLMs. *CoRR*, abs/2502.03373, 2025.
- G. Chaslot, S. Bakkes, I. Szita, and P. Spronck. Monte-Carlo tree search: A new framework for game AI. In *Proceedings of the 22nd AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 2008.
- G. Chen, M. Liao, C. Li, and K. Fan. AlphaMath almost zero: Process supervision without process. In *Advances in Neural Information Processing Systems 38*, 2024a.
- M. Chen, T. Li, H. Sun, Y. Zhou, C. Zhu, F. Yang, Z. Zhou, W. Chen, H. Wang, J. Z. Pan, et al. Learning to reason with search for LLMs via reinforcement learning. *CoRR*, abs/2503.19470, 2025a.
- Q. Chen, L. Qin, J. Liu, D. Peng, J. Guan, P. Wang, M. Hu, Y. Zhou, T. Gao, and W. Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *CoRR*, abs/2503.09567, 2025b.
- S. Chen, B. Li, and D. Niu. Boosting of thoughts: Trial-and-error problem solving with large language models. In *Proceedings of the 12th International Conference on Learning Representations*, 2024b.
- W. Chen, M. Yin, M. Ku, P. Lu, Y. Wan, X. Ma, J. Xu, X. Wang, and T. Xia. TheoremQA: A theorem-driven question answering dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.
- Y. Chen, J. Benton, A. Radhakrishnan, J. Uesato, C. Denison, J. Schulman, A. Somani, P. Hase, M. Wagner, F. Roger, et al. Reasoning models don't always say what they think. *CoRR*, abs/2505.05410, 2025c.

- E. Chern, H. Zou, X. Li, J. Hu, K. Feng, J. Li, and P. Liu. Generative AI for math: Abel. <https://github.com/GAIR-NLP/abel>, 2023.
- P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, pages 4299–4307, 2017.
- K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021.
- A. Collins and E. Koechlin. Reasoning, learning, and creativity: frontal lobe function and human decision-making. *PLoS Biology*, 10(3):e1001293, 2012.
- R. Coulom. Efficient selectivity and backup operators in Monte-Carlo tree search. In *Proceedings of the 5th International Conference on Computers and Games*, 2006.
- C. Dai, K. Li, W. Zhou, and S. Hu. Beyond imitation: Learning key reasoning steps from dual chain-of-thoughts in reasoning distillation. *CoRR*, abs/2405.19737, 2024.
- D. Das, D. Banerjee, S. Aditya, and A. Kulkarni. MATHSENSEI: A tool-augmented large language model for mathematical reasoning. *CoRR*, abs/2402.17231, 2024.
- A. Didolkar, A. Goyal, N. R. Ke, S. Guo, M. Valko, T. Lillicrap, D. Rezende, Y. Bengio, M. Mozer, and S. Arora. Metacognitive capabilities of LLMs: An exploration in mathematical problem solving. *CoRR*, abs/2405.12205, 2024.
- Y. Ding, X. Shi, X. Liang, J. Li, Q. Zhu, and M. Zhang. Unleashing reasoning capability of LLMs via scalable question synthesis from scratch. *CoRR*, abs/2410.18693, 2024.
- P. Dixit and T. Oates. SBI-RAG: Enhancing math word problem solving for students through schema-based instruction and retrieval-augmented generation. *CoRR*, 2410.13293, 2024.
- A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Rozière, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. M. Kloumann, I. Misra, I. Evtimov, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. van der Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnstun, J. Saxe, J. Jia, K. V. Alwala, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, and et al. The Llama 3 herd of models. *CoRR*, abs/2407.21783, 2024.
- E. A. Feigenbaum, J. Feldman, et al. *Computers and thought*, volume 37. McGraw-Hill, 1963.
- D. Feng, B. Qin, C. Huang, Z. Zhang, and W. Lei. Towards analyzing and understanding the limitations of DPO: A theoretical perspective. *CoRR*, abs/2404.04626, 2024a.
- G. Feng, B. Zhang, Y. Gu, H. Ye, D. He, and L. Wang. Towards revealing the mystery behind Chain of Thought: A theoretical perspective. In *Advances in Neural Information Processing Systems* 37, 2023.
- S. Feng, X. Kong, S. Ma, A. Zhang, D. Yin, C. Wang, R. Pang, and Y. Yang. Step-by-step reasoning for math problems via twisted sequential monte carlo. *CoRR*, abs/2410.01920, 2024b.
- C. R. Fletcher. Understanding and solving arithmetic word problems: A computer simulation. *Behavior Research Methods, Instruments, & Computers*, 17(5):565–571, 1985.
- A. D. Friederici. The brain basis of language processing: from structure to function. *Physiological Reviews*, 91(4):1357–1392, 2011.
- K. Gandhi, A. Chakravarthy, A. Singh, N. Lile, and N. D. Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *CoRR*, abs/2503.01307, 2025.
- B. Gao, Z. Cai, R. Xu, P. Wang, C. Zheng, R. Lin, K. Lu, J. Lin, C. Zhou, W. Xiao, J. Hu, T. Liu, and B. Chang. LLM critics help catch bugs in mathematics: Towards a better mathematical verifier with natural language feedback. *CoRR*, abs/2406.14024, 2024.
- L. Gao, A. Madaan, S. Zhou, U. Alon, P. Liu, Y. Yang, J. Callan, and G. Neubig. PAL: Program-aided language models. In *Proceedings of the 40th International Conference on Machine Learning*, 2023a.
- Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, and H. Wang. Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997, 2023b.

- J. Gehring, K. Zheng, J. Copet, V. Mella, T. Cohen, and G. Synnaeve. RLEF: Grounding code LLMs in execution feedback with reinforcement learning. *CoRR*, abs/2410.02089, 2024.
- S. Ghosh, C. K. R. Evuru, S. Kumar, U. Tyagi, O. Nieto, Z. Jin, and D. Manocha. Visual description grounding reduces hallucinations and boosts reasoning in LVLMs. *CoRR*, abs/2405.15683, 2024.
- D. Goldwasser and D. Roth. Learning from natural instructions. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, 2014.
- 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. *CoRR*, abs/2501.04519, 2025.
- 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. *CoRR*, abs/2501.12948, 2025.
- S. Guo, A. Didolkar, N. R. Ke, A. Goyal, F. Huszár, and B. Schölkopf. Learning beyond pattern matching? assaying mathematical understanding in LLMs. *CoRR*, abs/2405.15485, 2024.
- S. Hao, Y. Gu, H. Ma, J. J. Hong, Z. Wang, D. Z. 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.
- 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. *CoRR*, abs/2402.14008, 2024a.
- M. He, Y. Shen, W. Zhang, Z. Tan, and W. Lu. Advancing process verification for large language models via tree-based preference learning. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2024b.
- N. He, H. Lai, C. Zhao, Z. Cheng, J. Pan, R. Qin, R. Lu, R. Lu, Y. Zhang, G. Zhao, et al. TeacherLM: Teaching to fish rather than giving the fish, language modeling likewise. *CoRR*, abs/2310.19019, 2023.
- D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. *CoRR*, abs/2009.03300, 2020.
- D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *The 35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.
- G. Hinton. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.
- N. Ho, L. Schmid, and S. Yun. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.
- A. Hosseini, X. Yuan, N. Malkin, A. Courville, A. Sordoni, and R. Agarwal. V-STaR: Training verifiers for self-taught reasoners. *CoRR*, abs/2402.06457, 2024.
- J. Hu. REINFORCE++: A simple and efficient approach for aligning large language models. *CoRR*, abs/2501.03262, 2025.
- J. Huang, S. S. Gu, L. Hou, Y. Wu, X. Wang, H. Yu, and J. Han. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.
- Y. Huang, X. Liu, Y. Gong, Z. Gou, Y. Shen, N. Duan, and W. Chen. Key-point-driven data synthesis with its enhancement on mathematical reasoning. *CoRR*, abs/2403.02333, 2024.
- S. Imani, L. Du, and H. Shrivastava. MathPrompter: Mathematical reasoning using large language models. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics*, pages 37–42, 2023.
- C. Jia, P. Wang, Z. Li, Y.-C. Li, Z. Zhang, N. Tang, and Y. Yu. BWArea model: Learning world model, inverse dynamics, and policy for controllable language generation. *CoRR*, abs/2405.17039, 2024.
- C. Jia, Z. Li, P. Wang, Y.-C. Li, Z. Hou, Y. Dong, and Y. Yu. Controlling large language model with latent actions. *CoRR*, abs/2503.21383, 2025.
- W. Jiang, H. Shi, L. Yu, Z. Liu, Y. Zhang, Z. Li, and J. Kwok. Forward-backward reasoning in large language models for mathematical verification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.
- Z. Jie and W. Lu. Leveraging training data in few-shot prompting for numerical reasoning. *CoRR*, abs/2305.18170, 2023.

- 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. *CoRR*, abs/2503.09516, 2025.
- 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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.
- A. Joulin. Fasttext.zip: Compressing text classification models. *CoRR*, abs/1612.03651, 2016.
- A. Kazemnejad, M. Aghajohari, E. Portelance, A. Sordoni, S. Reddy, A. Courville, and N. L. Roux. VinePPO: Unlocking RL potential for LLM reasoning through refined credit assignment. *CoRR*, abs/2410.01679, 2024.
- T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 36*, 2022.
- R. Koncel-Kedziorski, S. Roy, A. Amini, N. Kushman, and H. Hajishirzi. MAWPS: A math word problem repository. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*, 2016.
- T. Kwiatkowski, E. Choi, Y. Artzi, and L. Zettlemoyer. Scaling semantic parsers with on-the-fly ontology matching. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013.
- 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. *CoRR*, abs/2406.18629, 2024.
- L. Lehnert, S. Sukhbaatar, D. Su, Q. Zheng, P. Mcvay, M. Rabbat, and Y. Tian. Beyond A*: Better planning with Transformers via search dynamics bootstrapping. *CoRR*, abs/2402.14083, 2024.
- Z. Levonian, C. Li, W. Zhu, A. Gade, O. Henkel, M.-E. Postle, and W. Xing. Retrieval-augmented generation to improve math question-answering: Trade-offs between groundedness and human preference. *CoRR*, abs/2310.03184, 2023.
- A. Lewkowycz, A. Andreassen, D. Dohan, E. Dyer, H. Michalewski, V. V. Ramasesh, A. Slone, C. Anil, I. Schlag, T. Gutman-Solo, Y. Wu, B. Neyshabur, G. Gur-Ari, and V. Misra. Solving quantitative reasoning problems with language models. *CoRR*, abs/2206.14858, 2022.
- C. Li, W. Wang, J. Hu, Y. Wei, N. Zheng, H. Hu, Z. Zhang, and H. Peng. Common 7b language models already possess strong math capabilities. *CoRR*, abs/2403.04706, 2024a.
- C. Li, Z. Yuan, H. Yuan, G. Dong, K. Lu, J. Wu, C. Tan, X. Wang, and C. Zhou. MuggleMath: Assessing the impact of query and response augmentation on math reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024b.
- J. Li, E. Beeching, L. Tunstall, B. Lipkin, R. Soletskyi, S. Huang, K. Rasul, L. Yu, A. Q. Jiang, Z. Shen, et al. NuminaMath: The largest public dataset in AI4Maths with 860k pairs of competition math problems and solutions. *Hugging Face repository*, page 9, 2024c.
- Q. Li, L. Cui, X. Zhao, L. Kong, and W. Bi. GSM-Plus: A comprehensive benchmark for evaluating the robustness of LLMs as mathematical problem solvers. *CoRR*, abs/2402.19255, 2024d.
- X. Li. A survey on LLM test-time compute via search: Tasks, LLM profiling, search algorithms, and relevant frameworks. *CoRR*, abs/2501.10069, 2025.
- X. Li, H. Zou, and P. Liu. ToRL: Scaling tool-integrated RL. *CoRR*, abs/2503.23383, 2025a.
- 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. *CoRR*, abs/2206.02336, 2022.
- Y. Li, K. Sreenivasan, A. Giannou, D. Papailiopoulos, and S. Oymak. Dissecting Chain-of-Thought: Compositionality through in-context filtering and learning. In *Advances in Neural Information Processing Systems 37*, 2023.
- Z. Li, T. Xu, and Y. Yu. When is RL better than DPO in RLHF? a representation and optimization perspective. In *The Second Tiny Papers Track at International Conference on Learning Representations*, 2024e.
- Z. Li, T. Xu, Y. Zhang, Z. Lin, Y. Yu, R. Sun, and Z.-Q. Luo. ReMax: A simple, effective, and efficient reinforcement learning method for aligning large language models. In *The 41st International Conference on Machine Learning*, 2024f.
- Z. Li, C. Chen, T. Xu, Z. Qin, J. Xiao, Z.-Q. Luo, and R. Sun. Preserving diversity in supervised fine-tuning of large language models. In *Proceedings of the 13th International Conference on Learning Representations*, 2025b.

- Z.-Z. Li, D. Zhang, M.-L. Zhang, J. Zhang, Z. Liu, Y. Yao, H. Xu, J. Zheng, P.-J. Wang, X. Chen, et al. From system 1 to system 2: A survey on reasoning large language models. *CoRR*, abs/2502.17419, 2025c.
- P. P. Liang, A. Zadeh, and L.-P. Morency. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10):1–42, 2024.
- Z. Liang, W. Yu, T. Rajpurohit, P. Clark, X. Zhang, and A. Kalyan. Let GPT be a math tutor: Teaching math word problem solvers with customized exercise generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.
- 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. *CoRR*, abs/2305.20050, 2023.
- Y. Lin, S. Seto, M. ter Hoeve, K. Metcalf, B.-J. Theobald, X. Wang, Y. Zhang, C. Huang, and T. Zhang. On the limited generalization capability of the implicit reward model induced by direct preference optimization. *CoRR*, abs/2409.03650, 2024a.
- Y.-T. Lin, D. Jin, T. Xu, T. Wu, S. Sukhbaatar, C. Zhu, Y. He, Y.-N. Chen, J. Weston, Y. Tian, A. Rahnama, S. Wang, H. Ma, and H. Fang. Step-KTO: Optimizing mathematical reasoning through stepwise binary feedback. *CoRR*, abs/2501.10799, 2025.
- Z. Lin, Z. Gou, Y. Gong, X. Liu, Y. Shen, R. Xu, C. Lin, Y. Yang, J. Jiao, N. Duan, et al. Rho-1: Not all tokens are what you need. *CoRR*, abs/2404.07965, 2024b.
- J. Lindsey, W. Gurnee, E. Ameisen, B. Chen, A. Pearce, N. L. Turner, C. Citro, D. Abrahams, S. Carter, B. Hosmer, J. Marcus, M. Sklar, A. Templeton, T. Bricken, C. McDougall, H. Cunningham, T. Henighan, A. Jermyn, A. Jones, A. Persic, Z. Qi, T. B. Thompson, S. Zimmerman, K. Rivoire, T. Conerly, C. Olah, and J. Batson. On the biology of a large language model. *Transformer Circuits Thread*, 2025.
- W. Ling, D. Yogatama, C. Dyer, and P. Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017.
- B. Liu, S. Bubeck, R. Eldan, J. Kulkarni, Y. Li, A. Nguyen, R. Ward, and Y. Zhang. TinyGSM: achieving > 80% on GSM8k with small language models. *CoRR*, abs/2312.09241, 2023.
- H. Liu, Y. Zhang, Y. Luo, and A. C. Yao. Augmenting math word problems via iterative question composing. *CoRR*, abs/2401.09003, 2024a.
- Z. Liu, H. Liu, D. Zhou, and T. Ma. Chain of thought empowers transformers to solve inherently serial problems. In *Proceedings of the 12th International Conference on Learning Representations*, 2024b.
- P. Lu, L. Qiu, W. Yu, S. Welleck, and K.-W. Chang. A survey of deep learning for mathematical reasoning. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics*, 2023.
- 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. *CoRR*, abs/2406.06592, 2024.
- M. Luo, S. Tan, J. Wong, X. Shi, W. Y. Tang, M. Roongta, C. Cai, J. Luo, T. Zhang, L. E. Li, R. A. Popa, and I. Stoica. DeepScaleR: Surpassing O1-Preview with a 1.5b model by scaling RL. <https://pretty-radio-b75.notion.site/DeepScaleR-Surpassing-O1-Preview-with-a-1-5B-Model-by-Scaling-RL-19681902c1468005bed8ca303013a4e2>, 2025. Notion Blog.
- 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*, 2023.
- L. C. Magister, J. Mallinson, J. D. Adamek, E. Malmi, and A. Severyn. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.
- J. Meadows and A. Freitas. A survey in mathematical language processing. *CoRR*, abs/2205.15231, 2022.
- Y. Meng, M. Xia, and D. Chen. SimPO: Simple preference optimization with a reference-free reward. *CoRR*, abs/2405.14734, 2024.
- J. Menick, M. Trebacz, M. Mikulak-Krupiński, E. Elsen, A. Valbusa, A. Raichuk, R. Kadlec, G. Irving, J. Kopecký, M. Ring, A. Glaese, N. Elhage, T. Hennigan, S. Saci, T. Cai, M. Fritz, A. Jones, D. Pfau, T. Pohlen, and O. Vinyals. Teaching language models to support answers with verified quotes. *CoRR*, abs/2203.11147, 2022.
- W. Merrill and A. Sabharwal. The expressive power of Transformers with Chain of Thought. In *Proceedings of the 12th International Conference on Learning Representations*, 2024.

- S.-Y. Miao, C.-C. Liang, and K.-Y. Su. A diverse corpus for evaluating and developing english math word problem solvers. *CoRR*, abs/2106.15772, 2021.
- S. Mirzadeh, K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio, and M. Farajtabar. GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. *CoRR*, abs/2410.05229, 2024.
- S. Mishra, M. Finlayson, P. Lu, L. Tang, S. Welleck, C. Baral, T. Rajpurohit, O. Tafjord, A. Sabharwal, P. Clark, and A. Kalyan. LILA: A unified benchmark for mathematical reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.
- 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. *CoRR*, abs/2501.19393, 2025.
- OpenAI. GPT-4 system card. <https://cdn.openai.com/papers/gpt-4-system-card.pdf>, March 2023.
- OpenAI. Learning to reason with LLMs. <https://openai.com/index/learning-to-reason-with-lm/>, 2024.
- L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 36*, 2022.
- J.-C. Pang, P. Wang, K. Li, X.-H. Chen, J. Xu, Z. Zhang, and Y. Yang. Language model self-improvement by reinforcement learning contemplation. In *International Conference on Learning Representations*, 2023.
- R. Y. Pang, W. Yuan, H. He, K. Cho, S. Sukhbaatar, and J. Weston. Iterative reasoning preference optimization. In *Advances in Neural Information Processing Systems 38*, 2024.
- K. Paster, M. Dos Santos, Z. Azerbayev, and J. Ba. OpenWebMath: An open dataset of high-quality mathematical web text. In *Proceedings of the 12th International Conference on Learning Representations*, 2024.
- A. Patel, S. Bhattacharya, and N. Goyal. Are NLP models really able to solve simple math word problems? *CoRR*, abs/2103.07191, 2021.
- D. Paul, M. Ismayilzada, M. Peyrard, B. Borges, A. Bosselut, R. West, and B. Faltings. REFINER: reasoning feedback on intermediate representations. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, 2024.
- X. Peng, C. Xia, X. Yang, C. Xiong, C.-S. Wu, and C. Xing. ReGenesis: LLMs can grow into reasoning generalists via self-improvement. *CoRR*, abs/2410.02108, 2024.
- B. Prystawski, M. Li, and N. D. Goodman. Why think step by step? Reasoning emerges from the locality of experience. In *Advances in Neural Information Processing Systems 36*, 2023.
- X. Qu, Y. Li, Z. Su, W. Sun, J. Yan, D. Liu, G. Cui, D. Liu, S. Liang, J. He, P. Li, W. Wei, J. Shao, C. Lu, Y. Zhang, X. Hua, B. Zhou, and Y. Cheng. A survey of efficient reasoning for large reasoning models: Language, multimodality, and beyond. *CoRR*, abs/2503.21614, 2025.
- Y. Qu, T. Zhang, N. Garg, and A. Kumar. Recursive introspection: Teaching language model agents how to self-improve. In *Advances in Neural Information Processing Systems*, 2024.
- A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- 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. In *Advances in Neural Information Processing Systems 37*, 2023.
- J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017.
- A. Setlur, C. Nagpal, A. Fisch, X. Geng, J. Eisenstein, R. Agarwal, A. Agarwal, J. Berant, and A. Kumar. Rewarding progress: Scaling automated process verifiers for LLM reasoning. *CoRR*, abs/2410.08146, 2024.
- 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. *CoRR*, abs/2402.03300, 2024.
- M. Shen, G. Zeng, Z. Qi, Z. Hong, Z. Chen, W. Lu, G. W. Wornell, S. Das, D. Cox, and C. Gan. Satori: Reinforcement learning with chain-of-action-thought enhances LLM reasoning via autoregressive search. *CoRR*, abs/2502.02508, 2025.
- Z. Shen. LLM with tools: A survey. *CoRR*, abs/2409.18807, 2024.

- F. Shi, M. Suzgun, M. Freitag, X. Wang, S. Srivats, S. Vosoughi, H. W. Chung, Y. Tay, S. Ruder, D. Zhou, D. Das, and J. Wei. Language models are multilingual chain-of-thought reasoners. *CoRR*, abs/2210.03057, 2022.
- F. Shi, X. Chen, K. Misra, N. Scales, D. Dohan, E. H. Chi, N. Schärli, and D. Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, 2023.
- A. Singh, J. D. Co-Reyes, R. Agarwal, A. Anand, P. Patil, X. Garcia, P. J. Liu, J. Harrison, J. Lee, K. Xu, A. T. Parisi, A. Kumar, A. A. Alemi, A. Rizkowsky, A. Nova, B. Adlam, B. Bohnet, G. F. 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. *Transactions on Machine Learning Research*, 2024.
- J. R. Slagle. Experiments with a deductive question-answering program. *Communications of the Association for Computing Machinery*, 8(12):792–798, 1965.
- C. Snell, J. Lee, K. Xu, and A. Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *CoRR*, abs/2408.03314, 2024.
- Z. Sprague, F. Yin, J. D. Rodriguez, D. Jiang, M. Wadhwa, P. Singhal, X. Zhao, X. Ye, K. Mahowald, and G. Durrett. To CoT or not to CoT? chain-of-thought helps mainly on math and symbolic reasoning. *CoRR*, abs/2409.12183, 2024a.
- Z. Sprague, F. Yin, J. D. Rodriguez, D. Jiang, M. Wadhwa, P. Singhal, X. Zhao, X. Ye, K. Mahowald, and G. Durrett. To CoT or not to CoT? chain-of-thought helps mainly on math and symbolic reasoning. *CoRR*, abs/2409.12183, 2024b.
- P. Srivastava, M. Malik, V. Gupta, T. Ganu, and D. Roth. Evaluating LLMs' mathematical reasoning in financial document question answering. In *Findings of the Association for Computational Linguistics*, 2024.
- N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. Learning to summarize from human feedback. *CoRR*, abs/2009.01325, 2020.
- Y. Sui, Y.-N. Chuang, G. Wang, J. Zhang, T. Zhang, J. Yuan, H. Liu, A. Wen, H. Chen, X. Hu, et al. Stop overthinking: A survey on efficient reasoning for large language models. *CoRR*, abs/2503.16419, 2025.
- Q. Tang, Z. Deng, H. Lin, X. Han, Q. Liang, B. Cao, and L. Sun. ToolAlpaca: Generalized tool learning for language models with 3000 simulated cases. *CoRR*, abs/2306.05301, 2023.
- Z. Tang, X. Zhang, B. Wang, and F. Wei. MathScale: Scaling instruction tuning for mathematical reasoning. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.
- G. Team and Google. Gemini: A family of highly capable multimodal models. https://storage.googleapis.com/deepmind-media/gemini_1_report.pdf, December 2023.
- 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. *CoRR*, abs/2501.12599, 2025.
- A. Templeton, T. Conerly, J. Marcus, J. Lindsey, T. Bricken, B. Chen, A. Pearce, C. Citro, E. Ameisen, A. Jones, H. Cunningham, N. L. Turner, C. McDougall, M. MacDiarmid, C. D. Freeman, T. R. Sumers, E. Rees, J. Batson, A. Jermyn, S. Carter, C. Olah, and T. Henighan. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024.
- S. Toshniwal, W. Du, I. Moshkov, B. Kisacanin, A. Ayrapetyan, and I. Gitman. OpenMathInstruct-2: Accelerating AI for math with massive open-source instruction data. *CoRR*, abs/2410.01560, 2024a.
- S. Toshniwal, I. Moshkov, S. Narendhiran, D. Gitman, F. Jia, and I. Gitman. OpenMathInstruct-1: A 1.8 million math instruction tuning dataset. In *Advances in Neural Information Processing Systems 38*, 2024b.
- 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. *CoRR*, abs/2302.13971, 2023.
- L. Q. 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*, 2024.
- A. Turing. Computing machinery and intelligence. 1950.
- R. Tutunov, A. Grosnit, J. Ziomek, J. Wang, and H. Bou-Ammar. Why can large language models generate correct chain-of-thoughts? *CoRR*, abs/2310.13571, 2023.

- 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 *Proceedings of the 41th International Conference on Machine Learning*, 2024.
- 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*, 2023a.
- E. Z. Wang, F. Cassano, C. Wu, Y. Bai, W. Song, V. Nath, Z. Han, S. M. Hendryx, S. Yue, and H. Zhang. Planning in natural language improves LLM search for code generation. *CoRR*, abs/2409.03733, 2024a.
- J. Wang, Q. Sun, X. Li, and M. Gao. Boosting language models reasoning with chain-of-knowledge prompting. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024b.
- K. Wang, J. Pan, W. Shi, Z. Lu, H. Ren, A. Zhou, M. Zhan, and H. Li. Measuring multimodal mathematical reasoning with MATH-Vision dataset. In *Advances in Neural Information Processing Systems 38*, 2024c.
- L. Wang, W. Xu, Y. Lan, Z. Hu, Y. Lan, R. K. Lee, and E. Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics*, 2023b.
- 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*, 2024d.
- T. Wang and W. Lu. Learning multi-step reasoning by solving arithmetic tasks. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics*, 2023.
- X. Wang and D. Zhou. Chain-of-thought reasoning without prompting. *CoRR*, abs/2402.10200, 2024.
- 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 *Proceedings of the 11th International Conference on Learning Representations*, 2023c.
- Y. Wang, X. Liu, and S. Shi. Deep neural solver for math word problems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017.
- Y. Wang, Q. Liu, J. Xu, T. Liang, X. Chen, Z. He, L. Song, D. Yu, J. Li, Z. Zhang, et al. Thoughts are all over the place: On the underthinking of o1-like LLMs. *CoRR*, abs/2501.18585, 2025.
- Z. Wang, X. Li, R. Xia, and P. Liu. MathPile: A billion-token-scale pretraining corpus for math. In *Advances in Neural Information Processing Systems 37*, 2024e.
- S. Warshall. A theorem on boolean matrices. *Journal of the Association for Computing Machinery*, 1962.
- J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022a.
- J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 36*, 2022b.
- T. Wei, J. Luan, W. Liu, S. Dong, and B. Wang. CMATH: can your language model pass chinese elementary school math test? *CoRR*, abs/2306.16636, 2023.
- S. Welleck, J. Liu, R. Le Bras, H. Hajishirzi, Y. Choi, and K. Cho. NaturalProofs: Mathematical theorem proving in natural language. In *Proceedings of the 1th Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.
- S. Welleck, X. Lu, P. West, F. Brahman, T. Shen, D. Khashabi, and Y. Choi. Generating sequences by learning to self-correct. In *Proceedings of the 11th International Conference on Learning Representations*, 2023.
- 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*, 2023.
- C. Wu, Z. Lin, W. Fang, and Y. Huang. A medical diagnostic assistant based on LLM. In *Health Information Processing Evaluation Track Papers*, 2024a.
- S. Wu, E. M. Shen, C. Badrinath, J. Ma, and H. Lakkaraju. Analyzing chain-of-thought prompting in large language models via gradient-based feature attributions. *CoRR*, abs/2307.13339, 2023a.
- Y. Wu, Z. Sun, S. Li, S. Welleck, and Y. Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models. *CoRR*, abs/2408.00724, 2024b.

- 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. In *Advances in Neural Information Processing Systems 37*, 2023b.
- W. Xiong, J. Yao, Y. Xu, B. Pang, L. Wang, D. Sahoo, J. Li, N. Jiang, T. Zhang, C. Xiong, et al. A minimalist approach to LLM reasoning: from Rejection Sampling to REINFORCE. *CoRR*, abs/2504.11343, 2025a.
- W. Xiong, H. Zhang, C. Ye, L. Chen, N. Jiang, and T. Zhang. Self-rewarding correction for mathematical reasoning. *CoRR*, abs/2502.19613, 2025b.
- 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. *CoRR*, abs/2501.09686, 2025.
- S. Xu, W. Fu, J. Gao, W. Ye, W. Liu, Z. Mei, G. Wang, C. Yu, and Y. Wu. Is DPO superior to PPO for LLM alignment? a comprehensive study. *CoRR*, abs/2404.10719, 2024.
- Y. Yan, J. Su, J. He, F. Fu, X. Zheng, Y. Lyu, K. Wang, S. Wang, Q. Wen, and X. Hu. A survey of mathematical reasoning in the era of multimodal large language model: Benchmark, method & challenges. *CoRR*, abs/2412.11936, 2024.
- A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, G. Dong, H. Wei, H. Lin, J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang, J. Ma, J. Yang, J. Xu, J. Zhou, J. Bai, J. He, J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng, R. Men, R. Gao, R. Lin, S. Wang, S. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou, X. Ren, X. Zhang, X. Wei, X. Ren, X. Liu, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Z. Zhang, Z. Guo, and Z. Fan. Qwen2 technical report. *CoRR*, 2024a.
- A. Yang, B. Zhang, B. Hui, B. Gao, B. Yu, C. Li, D. Liu, J. Tu, J. Zhou, J. Lin, et al. Qwen2.5-Math technical report: Toward mathematical expert model via self-improvement. *CoRR*, abs/2409.12122, 2024b.
- K. Yang, A. M. Swope, A. Gu, R. Chalamala, P. Song, S. Yu, S. Godil, R. J. Prenger, and A. Anandkumar. LeanDojo: Theorem proving with retrieval-augmented language models. In *Advances in Neural Information Processing Systems 36*, 2023.
- L. Yang, K. Lee, R. Nowak, and D. Papailiopoulos. Looped Transformers are better at learning learning algorithms. *CoRR*, abs/2311.12424, 2024c.
- S. Yang, D. Schuurmans, P. Abbeel, and O. Nachum. Chain of thought imitation with procedure cloning. In *Advances in Neural Information Processing Systems*, 2022.
- T. Yang, M. Yan, H. Zhao, and T. Yang. LemmaHead: RAG assisted proof generation using large language models. *CoRR*, abs/2501.15797, 2025.
- S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems 37*, 2023.
- T. Ye, Z. Xu, Y. Li, and Z. Allen-Zhu. Physics of language models: Part 2.2, how to learn from mistakes on grade-school math problems. *CoRR*, abs/2408.16293, 2024.
- Y. Ye, Z. Huang, Y. Xiao, E. Chern, S. Xia, and P. Liu. LIMO: Less is more for reasoning. *CoRR*, abs/2502.03387, 2025.
- 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*, 2024.
- H. Ying, Z. Wu, Y. Geng, J. Wang, D. Lin, and K. Chen. Lean workbook: A large-scale Lean problem set formalized from natural language math problems. *CoRR*, abs/2406.03847, 2024a.
- H. Ying, S. Zhang, L. Li, Z. Zhou, Y. Shao, Z. Fei, Y. Ma, J. Hong, K. Liu, Z. Wang, et al. InternLM-Math: Open math large language models toward verifiable reasoning. *CoRR*, abs/2402.06332, 2024b.
- L. Yu, W. Jiang, H. Shi, Y. Jincheng, Z. Liu, Y. Zhang, J. Kwok, Z. Li, A. Weller, and W. Liu. MetaMath: Bootstrap your own mathematical questions for large language models. In *Proceedings of the 12th International Conference on Learning Representations*, 2024.
- 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. *CoRR*, abs/2503.14476, 2025.
- 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. *CoRR*, abs/2412.01981, 2024.

- X. Yue, X. Qu, G. Zhang, Y. Fu, W. Huang, H. Sun, Y. Su, and W. Chen. MAmmoTH: Building math generalist models through hybrid instruction tuning. In *Proceedings of the 12th International Conference on Learning Representations*, 2024a.
- X. Yue, T. Zheng, G. Zhang, and W. Chen. Mammoth2: Scaling instructions from the web. *CoRR*, abs/2405.03548, 2024b.
- Y. Yue, Z. Chen, R. Lu, A. Zhao, Z. Wang, S. Song, and G. Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? *CoRR*, abs/2504.13837, 2025.
- E. Zelikman, Y. Wu, J. Mu, and N. Goodman. STaR: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems* 36, 2022.
- 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. *CoRR*, abs/2403.09629, 2024.
- W. Zeng, Y. Huang, Q. Liu, W. Liu, K. He, Z. Ma, and J. He. SimpleRL-Zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *CoRR*, abs/2503.18892, 2025.
- Y. Zeng, G. Liu, W. Ma, N. Yang, H. Zhang, and J. Wang. Token-level direct preference optimization. In *Proceedings of the 41th International Conference on Machine Learning*, 2024.
- B. Zhang, K. Zhou, X. Wei, X. Zhao, J. Sha, S. Wang, and J. Wen. Evaluating and improving tool-augmented computation-intensive math reasoning. In *Advances in Neural Information Processing Systems* 37, 2023a.
- D. Zhang, L. Wang, L. Zhang, B. T. Dai, and H. T. Shen. The gap of semantic parsing: A survey on automatic math word problem solvers. *CoRR*, abs/1808.07290, 2019.
- 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. *CoRR*, abs/2410.02884, 2024a.
- D. Zhang, S. Zhoubian, Z. Hu, Y. Yue, Y. Dong, and J. Tang. REST-MCTS*: LLM self-training via process reward guided tree search. *CoRR*, abs/2406.03816, 2024b.
- K. Zhang, Q. Yao, B. Lai, J. Huang, W. Fang, D. Tao, M. Song, and S. Liu. Reasoning with reinforced functional token tuning. *CoRR*, abs/2502.13389, 2025.
- L. Zhang, A. Hosseini, H. Bansal, M. Kazemi, A. Kumar, and R. Agarwal. Generative verifiers: Reward modeling as next-token prediction. *CoRR*, abs/2408.15240, 2024c.
- Y. Zhang, Y. Luo, Y. Yuan, and A. C. Yao. Autonomous data selection with language models for mathematical texts. *CoRR*, abs/2402.07625, 2024d.
- Z. Zhang, A. Zhang, M. Li, and A. Smola. Automatic chain of thought prompting in large language models. In *Proceedings of the 11th International Conference on Learning Representations*, 2023b.
- B. P. Zhangir Azerbayev, Edward Ayers. Proof-pile. <https://huggingface.co/datasets/hoskinson-center/proof-pile>, 2023.
- R. Zhao, A. Meterez, S. Kakade, C. Pehlevan, S. Jelassi, and E. Malach. Echo chamber: RL post-training amplifies behaviors learned in pretraining. *CoRR*, abs/2504.07912, 2025.
- 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. *CoRR*, abs/2303.18223, 2023.
- Z. Zhao, H. Dong, A. Saha, C. Xiong, and D. Sahoo. Automatic curriculum expert iteration for reliable LLM reasoning. *CoRR*, abs/2410.07627, 2024.
- 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. *CoRR*, abs/2412.06559, 2024a.
- G. Zheng, B. Yang, J. Tang, H. Zhou, and S. Yang. DDCoT: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In *Advances in Neural Information Processing Systems* 36, 2023.
- K. Zheng, J. M. Han, and S. Polu. MiniF2F: a cross-system benchmark for formal olympiad-level mathematics. *CoRR*, abs/2109.00110, 2021.
- L. Zheng, R. Wang, X. Wang, and B. An. Synapse: Trajectory-as-exemplar prompting with memory for computer control. In *Proceedings of the 12th International Conference on Learning Representations*, 2024b.
- Q. Zhong, K. Wang, Z. Xu, J. Liu, L. Ding, B. Du, and D. Tao. Achieving >97% on GSM8K: deeply understanding the problems makes LLMs better reasoners. *CoRR*, abs/2404.14963, 2024.
- G. Zhou, P. Qiu, C. Chen, J. Wang, Z. Yang, J. Xu, and M. Qiu. Reinforced MLLM: A survey on RL-based reasoning in multimodal large language models. *CoRR*, abs/2504.21277, 2025.

- H. Zhou, A. Nova, H. Larochelle, A. C. Courville, B. Neyshabur, and H. Sedghi. Teaching algorithmic reasoning via in-context learning. *CoRR*, abs/2211.09066, 2022.
- K. Zhou, B. Zhang, J. Wang, Z. Chen, W. X. Zhao, J. Sha, Z. Sheng, S. Wang, and J.-R. Wen. JiuZhang3.0: Efficiently improving mathematical reasoning by training small data synthesis models. *CoRR*, abs/2405.14365, 2024.
- Z.-H. Zhou and Z.-H. Tan. Learnware: Small models do big. *Science China Information Sciences*, 67(1):112102, 2024.
- D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving. Fine-tuning language models from human preferences. *CoRR*, abs/1909.08593, 2019.

A Mathematical Reasoning Evaluation Benchmarks

In this section, we introduce benchmarks that can be used to evaluate a model’s capabilities relevant to mathematical reasoning. An essential goal in evaluating mathematical reasoning models is to assess whether they exhibit capabilities comparable to, or exceeding, those of humans. To enable targeted evaluation, we propose categorizing datasets based on corresponding human levels, which allows us to assess different aspects of mathematical reasoning ability. These benchmarks are categorized by their difficulty levels, including elementary level, middle/high school level, university level and competition level.

A.1 Elementary Level

In the elementary level, MAWPS [Koncel-Kedziorski et al., 2016] is an online repository of 3,320 math word problems designed to evaluate models for solving such problems, covering topics such as basic arithmetic, algebra, and proportional reasoning. ASDiv-A [Miao et al., 2021] consists mainly of elementary-level English math word problems, which encompass a range of topics such as arithmetic, algebra, and basic geometry concepts. Based on ASDiv-A, SVAMP [Patel et al., 2021] is also composed of elementary-level math word problems, typically taught in grade four or lower. Math23K [Wang et al., 2017] is a collection of 23,161 Chinese elementary math word problems designed to evaluate the mathematical reasoning and problem-solving capabilities of models, particularly their ability to translate natural language into mathematical expressions and equations. MMLU [Hendrycks et al., 2020] is a dataset comprising 231,400 entries, with STEM tasks covering subjects such as mathematics, physics, computer science, and more. Chinese Elementary School Math Word Problems (CMATH) dataset [Wei et al., 2023], comprising 1.7k elementary school-level math word problems with detailed annotations, source from actual Chinese workbooks and exams.

A.2 Middle/High School Level

In the middle/high school level, GSM8K [Cobbe et al., 2021] is a collection of 8.5K high-quality linguistically diverse grade school math word problems, designed to evaluate the mathematical reasoning capabilities of models, particularly their ability to solve multi-step problems that require understanding and manipulation of numbers, operations, and basic mathematical concepts. This benchmark assesses models’ proficiency in areas such as basic arithmetic and algebra. Based on GSM8K dataset, GSM-HARD [Gao et al., 2023a], GSM-IC [Shi et al., 2023], GSM8K_robust [Chern et al., 2023], GSM-PLUS [Li et al., 2024d] add synthetic perturbation for further evaluation on models’ capabilities of robustness, especially on arithmetic variation and critical thinking. By translating 250 math problems from GSM8K into different languages, researchers obtained Multilingual Grade School Math(MGSM)[Hendrycks et al., 2021], with MGSM-zh commonly used as a Chinese dataset.

A.3 University Level

In university level benchmark, MathQA [Amini et al., 2019] is gathered by using a new representation language to annotate over the AQuA-RAT [Ling et al., 2017] dataset with fully-specified operational programs. MathQA-Python [Austin et al., 2021] is a Python adaptation of the MathQA benchmark, designed to evaluate LLMs’ ability to synthesize code that solves mathematically complex problems described in natural language. OCW [Lewkowycz et al., 2022] is a dataset of over 200 undergraduate-level questions in science and mathematics from MIT’s OpenCourseWare to measure the model’s quantitative reasoning abilities in a CoT context. The SAT [Azerbayev et al., 2024]dataset contains 32 math questions that do not contain figures from the May 2023 College Board SAT examination.

Additionally, there are benchmarks focusing on theorem proving and formal mathematics. For example, TheoremQA [Chen et al., 2023] is a theorem-driven question-answering benchmark designed to assess LLMs’ abilities to apply domain-specific theorems in solving complex problems. The LeanDojo [Yang et al., 2023] benchmark evaluates the theorem-proving capabilities of models by assessing their ability to generalize to novel theorems, which are not reliant on previously encountered premises. The benchmark includes 98,734 theorems/proofs extracted from mathlib, covering topics primarily at the university level. The miniF2F [Zheng et al., 2021] benchmark consists of 488 formal mathematics problems of Olympiad-level difficulty, offering a challenging and comprehensive evaluation of neural theorem-proving capabilities.

Table 3: Mathematical reasoning benchmarks. **Synth.** indicates synthetic data. See tags below.

(M) = MWP, (A) = Arithmetic, (G) = Geometry, (T) = Theorem Proving

Level	Dataset	Domain	Synth.	Language	Size	Release Time
Elementary School	MAWPS[Koncel-Kedziorski et al., 2016]	(M)	X	En	3,320	Jun-2016
	ASDiv-A[Miao et al., 2021]	(M)	X	En	2,305	Jul-2020
	SVAMP[Patel et al., 2021]	(M)	✓	En	1,000	Apr-2021
	Math23K[Wang et al., 2017]	(M)	X	Zh	23,161	Sept-2017
	CMATH[Wei et al., 2023]	(M)	X	Zh	1,689	Jun-2023
Middle/High School	GSM8K[Cobbe et al., 2021]	(M)	X	En	8.5k	Oct-2021
	GSM8K-PLUS[Li et al., 2024d]	(M)	✓	En	10,552	Feb-2024
	LILA[Mishra et al., 2022]	(M) / (A) / (G) / (T)	X	En	133,815	Oct-2022
	MGSM[Shi et al., 2022]	(M)	X	Multi	250	Oct-2022
University	AQuA[Ling et al., 2017]	(M)	X	En	100,000	May-2017
	MathQA[Amini et al., 2019]	(M)	X	En	37,297	May-2019
	SAT[Azerbayev et al., 2024]	(M) / (A)	X	En	32	Oct-2023
	OCW[Lewkowycz et al., 2022]	(M)	✓	En	200	Jun-2022
	TheoremQA[Chen et al., 2023]	(G) / (T)	✓	En	800	May-2023
	LeanDojo[Yang et al., 2023]	(T)	X	En	98,734	Jun-2023
	miniF2F[Zheng et al., 2021]	(T)	✓	En	488	Aug-2021
Competition	MATH[Hendrycks et al., 2021]	(M) / (A) / (G) / (T)	X	En	12,500	Mar-2021
	AIME24[aim, 2024]	(M) / (A) / (G) / (T)	X	En	30	Feb-2024
	Olympicbench[He et al., 2024a]	(M) / (A) / (G) / (T)	X	En	8.48k	Feb-2024
	MATH-Vision[Wang et al., 2024c]	(M) / (A) / (G)	X	En	3,040	Feb-2024
Mixed	MMLU-STEM[Hendrycks et al., 2020]	(M) / (A)	✓	En	3,153	Sep-2020

A.4 Competition Level

MATH [Hendrycks et al., 2021] consists of 12,500 competition-level mathematics problems spanning a range of topics, including algebra, geometry, calculus, and number theory. miniF2F [Zheng et al., 2021] is also an Olympiad-level benchmark. AIME24 (American Invitational Mathematics Examination 2024) [aim, 2024] consists of 30 problems from the 2024 American Mathematics Invitational, designed to evaluate models' performance on complex mathematical problems. OlympiadBench [He et al., 2024a] is a bilingual, multimodal benchmark designed for Olympiad-level science competitions, featuring 8,952 problems from mathematics and physics Olympiads, including the Chinese college entrance exam. MATH-Vision [Wang et al., 2024c] provides a comprehensive and diverse set of challenges for evaluating Large Multimodal Models' (LMM) mathematical reasoning abilities in mathematical reasoning within visual contexts.