LIMO: Less is More for Reasoning

Yixin Ye[*] Zhen Huang[*] Yang Xiao Ethan Chern Shijie Xia Pengfei Liu[†]

SJTU, SII, GAIR

Abstract

We present a fundamental *discovery* that challenges our understanding of how complex reasoning emerges in large language models. While conventional wisdom suggests that sophisticated reasoning tasks demand extensive training data (often $> 100,000$ examples), we demonstrate a striking phenomenon: complex mathematical reasoning abilities can be effectively elicited with *surprisingly* few examples. This finding challenges not only the assumption of massive data requirements but also the common belief that supervised fine-tuning primarily leads to memorization rather than generalization. Through comprehensive experiments, our proposed model LIMO demonstrates unprecedented performance and efficiency in mathematical reasoning. With merely **817** curated training samples, LIMO achieves **57.1%** accuracy on the highly challenging AIME benchmark and **94.8%** on MATH, improving the performance of previous strong SFT-based models from 6.5% to 57.1% on AIME and from 59.2% to 94.8% on MATH, while only using 1% of the training data required by previous approaches. Most remarkably, LIMO demonstrates exceptional out-of-distribution generalization, achieving **40.5%** absolute improvement across **10** diverse benchmarks, outperforming models trained on 100x more data, directly challenging the prevailing notion that SFT inherently leads to memorization rather than generalization. Synthesizing these pioneering results, we propose the **Less-Is-More Reasoning Hypothesis (LIMO Hypothesis)**: *In foundation models where domain knowledge has been comprehensively encoded during pre-training, sophisticated reasoning capabilities can emerge through minimal but precisely orchestrated demonstrations of cognitive processes*. This hypothesis posits that the *elicitation threshold* for complex reasoning is not inherently bounded by the complexity of the target reasoning task, but fundamentally determined by two key factors: (1) the completeness of the model's encoded knowledge foundation during pre-training, and (2) the effectiveness of post-training examples, which serve as "cognitive templates" that show the model how to effectively utilize its existing knowledge base to solve complex reasoning tasks. To facilitate reproducibility and future research in data-efficient reasoning, we release LIMO as a comprehensive open-source suite at https://github.com/GAIR-NLP/LIMO.
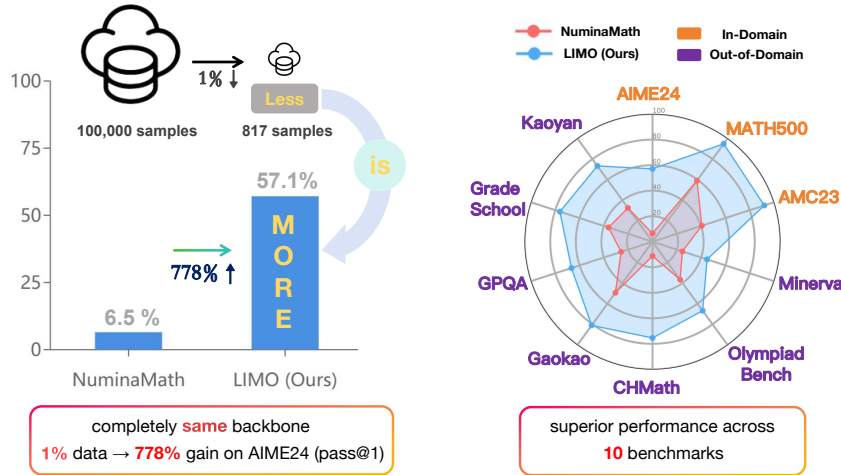
Figure 1: LIMO achieves substantial improvement over NuminaMath with fewer samples while excelling across diverse mathematical and multi-discipline benchmarks.

---

[*] Co-first authors
[†] Corresponding author

1 Introduction

Complex reasoning has long been considered one of the most challenging capabilities to instill in large language models (LLMs). While recent work has shown that LLMs can be effectively aligned with user preferences through relatively small amounts of instruction data (Zhou et al., 2024a), teaching models to reason—particularly in mathematics and programming—is widely believed to require vastly more training examples (Paster et al., 2023; Yue et al., 2024). This conventional wisdom stems from the inherent complexity of reasoning tasks, which demand multi-step logical deduction, domain knowledge application, and structured solution paths. The resulting paradigm typically involves training on tens or hundreds of thousands of examples (Yu et al., 2024; Li et al., 2024b), based on two fundamental assumptions: first, that mastering such complex cognitive processes requires extensive supervised demonstrations, and second, that supervised fine-tuning leads primarily to memorization rather than true generalization (Zhang et al., 2024; Xu et al., 2024; Chu et al., 2025).

While this approach has shown success, it imposes substantial computational costs and data collection burdens. More importantly, we argue this data-intensive paradigm may no longer be necessary. Recent advances have fundamentally transformed how LLMs acquire, organize, and utilize reasoning knowledge, suggesting the possibility of a more efficient approach. Two key developments in particular have created the conditions for a fundamental reimagining of how we approach reasoning in LLMs:

1. **Knowledge Foundation Revolution**: Modern foundation models now incorporate unprecedented amounts of mathematical content during pre-training (Qwen et al., 2025; Yang et al., 2024; Wang et al., 2024). For example: Llama 2's total training data across all domains was 1.8T tokens (Touvron et al., 2023), while Llama 3 used 3.7T tokens just for mathematical reasoning (Grattafiori et al., 2024). This suggests that contemporary LLMs may already possess rich mathematical knowledge in their parameter space, transforming the challenge from knowledge acquisition to knowledge elicitation.

2. **Inference-time Computation Scaling Revolution**: The emergence of techniques scaling longer reasoning chains has revealed that effective reasoning requires substantial computational space during inference. Recent works (OpenAI et al., 2024; Qin et al., 2024; Huang et al., 2024) have shown that allowing models to generate extended reasoning chains significantly improves their reasoning ability. In essence, inference-time computation provides the crucial *cognitive workspace* where models can systematically unpack and apply their pre-trained knowledge.

We hypothesize that successful reasoning emerges from the synergy of these two factors: rich pre-trained knowledge and sufficient computational resources at inference time. These developments collectively suggest a striking possibility: if models possess rich reasoning knowledge and are given adequate computational space, then activating their reasoning capabilities may require only a small number of high-quality training samples that encourage extended deliberation, rather than massive fine-tuning datasets. Building on this insight, we propose the **Less-Is-More Reasoning (LIMO) Hypothesis**. This hypothesis identifies two critical factors that determine the *elicitation threshold* for complex reasoning: (1) the latent presence of prerequisite knowledge within the model's parameter space, and (2) the effectiveness of minimal exemplars in demonstrating systematic problem-solving processes that encourage extended deliberation. Critically, this suggests that the sample efficiency of eliciting advanced reasoning is not inherently bounded by the complexity of the target reasoning task, but rather by the completeness of the model's encoded knowledge foundation and its exposure to training samples that effectively utilize the inference-time computation space.

Through comprehensive experiments, we demonstrate that LIMO achieves 57.1% accuracy on the highly challenging AIME benchmark and 94.8% on MATH with merely 817 training samples, demolishing previous strong SFT-based models while using just 1% of their training data. Most remarkably, these benefits generalize across a diverse spectrum of previously unseen scenarios, with LIMO consistently outperforming models trained on 100x more data by 40.5% absolute improvement. This discovery has profound implications for artificial intelligence research: it suggests that even competition-level complex reasoning abilities can be effectively elicited through minimal but curated training samples. More fundamentally, it points to a promising technical pathway toward AGI - any sophisticated reasoning capability, no matter how complex, could potentially be activated with minimal samples given two key conditions: (1) sufficient domain knowledge embedded during pre-training, and (2) optimal cognitive reasoning chains for activation. This represents not merely an argument for data efficiency, but a fundamental insight into how complex reasoning capabilities emerge in large language models.

The main contributions of this work are: (1) We establish the LIMO hypothesis, demonstrating that complex reasoning capabilities can be elicited through surprisingly small datasets (hundreds of examples) by leveraging rich mathematical knowledge in pre-trained models and detailed reasoning chains. (2) We provide systematic empirical evidence challenging current assumptions about scaling laws in reasoning tasks, showing that benefits generalize robustly to out-of-distribution problems and suggesting the acquisition of genuine reasoning capabilities rather

Table 1: Comparative Analysis: Less-is-More Phenomena in Language Models

Aspect	General Alignment (LIMA)	Complex Reasoning (LIMO)
Core Capability	Response format and style adaptation for general-purpose interaction	Multi-step logical inference and complex cognitive reasoning
Knowledge Foundation	• General text corpus sufficient • Social interaction patterns • Basic world knowledge	• Diverse reasoning paradigms and problem-solving approaches • Rich context for exploring alternative solutions • Deep conceptual connections across domains
Computation Requirements	• Fixed-length generation sufficient • Single-pass processing adequate • Limited context window acceptable	• Scalable inference-time computation essential • Extended reasoning chain support required • Large cognitive workspace necessary
Historical Prerequisites	Emerged in 2023, requiring only: • Base models with general knowledge • Basic prompt engineering techniques	Emerged in 2025, requiring convergence of: • Advanced reasoning architectures • Inference-time scaling revolution

Training Data Quality	• **Question Design:**
	  – Common interaction scenarios
	  – Standard task diversity
	  – Basic instruction following
	• **Solution Quality:**
	  – Clear communication style
	  – Format consistency
	  – Appropriate tone
		• **Question Design:**
		  – High-difficulty problems fostering complex reasoning
		  – Problems deviating from training distribution
		  – Cross-domain knowledge integration challenges
		• **Solution Quality:**
		  – Optimal structure with adaptive step granularity
		  – Strategic cognitive scaffolding for reasoning
		  – Rigorous verification throughout solution

than superficial pattern matching. (3) We identify critical factors for effective reasoning elicitation, particularly the synergy between pre-trained knowledge foundations and test-time computation scaling, providing insights into how these advances can be combined to achieve superior reasoning performance with minimal fine-tuning data. (4) We release a comprehensive open-source suite including our fine-tuned models, evaluation pipelines, training code, and carefully curated datasets with varying quality levels. This release enables systematic investigation of data efficiency in complex reasoning and facilitates reproducibility of our findings, while providing valuable resources for future research in this direction.

2 Phenomena Rethinking: Less-is-More and RL Scaling

The emergence of LIMO represents a paradigm shift in how we conceptualize and activate complex reasoning capabilities in large language models. This section examines two key comparisons that illuminate the fundamental nature of this advance: first, contrasting LIMO with LIMA to understand how Less-is-More principles extend from *general alignment* to *complex reasoning*; and second, comparing LIMO with reinforcement learning (RL) scaling approaches to highlight distinct philosophical perspectives on developing reasoning capabilities. Through these analyses, we aim to establish a deeper understanding of how complex cognitive abilities emerge in language models and the conditions that enable their efficient activation.

2.1 LIMO vs LIMA

The emergence of *Less-is-More phenomena* in LLMs represents a fundamental shift in our understanding of how complex capabilities can be elicited with minimal data. While LIMA (Zhou et al., 2024a) first demonstrated this phenomenon in the context of general alignment, extending this principle to complex mathematical reasoning presents unique challenges and requirements. This section examines the critical developments that enable Less-is-More for reasoning, analyzing the essential differences between alignment and reasoning scenarios, and providing insights into the conditions necessary for efficient capability activation in large language models.

Knowledge Foundation Revolution The past two years have witnessed a transformation in how language models acquire and organize mathematical knowledge. While LIMA could rely on general text corpora for alignment,

Table 2: Comparative Analysis of LIMO and RL Scaling Approaches

Aspect	RL Scaling (e.g., o1, R1)	LIMO
First Principle	An implementation of the general principle: searching for optimal reasoning trajectories through RL	The fundamental principle: reasoning capabilities exist and need to be activated by high-quality reasoning trajectories
Solution Nature	Discovers reasoning trajectories through extensive RL-based exploration	Directly constructs high-quality reasoning trajectories based on cognitive understanding
Core Challenge	How to efficiently search for effective reasoning trajectories in a large solution space	How to identify and construct optimal reasoning trajectories that activate existing capabilities
Methodology	Implicit trajectory discovery through large-scale RL optimization	Explicit trajectory design through cognitive templates
Search Strategy	Broad exploration of solution space using computational resources	Targeted exploration guided by cognitive principles
Resource Efficiency	Resource-intensive search process	Resource-efficient direct construction
Generalization	Through extensive sampling of trajectory space	Through understanding of fundamental reasoning patterns

LIMO's success builds upon the rich mathematical content now embedded in modern foundation models through specialized pre-training (Wang et al., 2024). This specialized knowledge foundation serves as a prerequisite for efficient reasoning capability activation.

Computation Capability Revolution   A crucial distinction between LIMA and LIMO lies in their computational requirements. While LIMA's alignment tasks could be accomplished with fixed-length generation and single-pass processing, LIMO's reasoning tasks demand extensive computation space for multi-step deliberation. The emergence of inference-time scaling techniques (OpenAI et al., 2024; Qin et al., 2024) provided the necessary "cognitive workspace" where models can systematically unpack and apply their pre-trained knowledge.

Synergistic Convergence   The timing of LIMO's discovery reflects the necessary convergence of these two revolutions. The two-year gap between LIMA and LIMO represents not just the time needed for better pre-trained models, but the essential wait for inference-time computation breakthroughs. This convergence enables a phenomenon we call the *Reasoning Elicitation Threshold*: when models possess both rich domain knowledge and sufficient computation space, complex reasoning capabilities can be activated with minimal but precise demonstrations.

Implications for Future Research   This comparative analysis reveals Less-is-More not merely as an advocacy for using fewer data, but as a fundamental principle governing the efficient elicitation of model capabilities. The success of LIMO demonstrates that when essential prerequisites (knowledge foundation and computation framework) are met, complex capabilities can be elicited with remarkable data efficiency. This insight suggests a new research direction: systematically identifying the prerequisites and optimal activation conditions for different capabilities. Future work should explore whether other advanced capabilities (e.g., planning, creative problem-solving) can achieve similar efficiency once their corresponding knowledge and computation foundations are established. The Less-is-More principle thus serves as both a theoretical framework for understanding capability emergence and a practical guide for pursuing data-efficient capability development across various domains.

2.2   LIMO vs RL Scaling

The emergence of two distinct approaches to developing reasoning capabilities in large language models - RL Scaling and LIMO - represents a fundamental divergence in how we understand and enhance model intelligence. RL Scaling, exemplified by works like o1 (OpenAI, 2024), DeepSeek-R1 (Guo et al., 2025), approaches the challenge from an engineering optimization perspective. It assumes that reasoning capabilities need to be extensively trained into models through large-scale reinforcement learning. While effective, this approach essentially treats RL as a broad search mechanism to discover effective reasoning patterns through massive computational resources. In contrast, LIMO introduces a more foundational perspective: reasoning capabilities are already latent within

pre-trained models, embedded during the pre-training phase. The key challenge shifts from "training" to "elicitation" - finding precise cognitive templates that can elicit these innate abilities.

From this perspective, RL Scaling approaches like DeepSeek-R1 can be viewed as specific implementations of this principle, using reinforcement learning as a mechanism to search for such trajectories. While both approaches ultimately seek high-quality reasoning solutions, LIMO offers a more principled and direct path through explicit trajectory design, while RL Scaling discovers these trajectories through extensive computational exploration. This reframing suggests that various methods, including RL, expert design, or hybrid approaches, could all be understood and evaluated within LIMO's framework as different strategies for discovering optimal reasoning trajectories.

3 LIMO Dataset

3.1 The LIMO Hypothesis

We formalize the Less-Is-More Reasoning (LIMO) Hypothesis as follows: In foundation models where domain knowledge has been comprehensively encoded during pre-training, sophisticated reasoning capabilities can emerge through minimal but precisely orchestrated demonstrations of cognitive processes. This hypothesis rests on two fundamental premises: (I) The latent presence of prerequisite knowledge within the model's parameter space (II) The quality of reasoning chains that precisely decompose complex problems into detailed, logical steps, making the cognitive process explicit and traceable. To validate this hypothesis, we propose a systematic approach to construct a high-quality, minimal dataset that can effectively elicit the model's inherent reasoning capabilities.

3.2 Problem Definition

In this paper, we focus on the reasoning tasks with verifiable answer. Given a question $q \in \mathcal{Q}$, where $\mathcal{Q}$ represents the space of reasoning problems, the goal is to generate an answer $a \in \mathcal{A}$ and a reasoning chain $r \in \mathcal{R}$. We define a reasoning chain r as a sequence of intermediate steps $\{s_1, s_2, ..., s_n\}$, where each step s_i represents a logical deduction that bridges the gap between the question and the final answer.

Formally, we can represent the reasoning process as a function f:

$$f : \mathcal{Q} \to \mathcal{R} \times \mathcal{A} \tag{1}$$

Therefore, the quality of the resulting dataset $\mathcal{D}$ is determined by two fundamental yet multifaceted components: (1) the quality of questions $q \in \mathcal{Q}$, which encompasses factors such as diversity in problem-solving approaches, appropriate difficulty levels to challenge model capabilities, and the breadth of knowledge domains covered, and (2) the quality of solutions $(r, a) \in \mathcal{R} \times \mathcal{A}$, which encompasses aspects including pedagogical value, logical coherence, and methodological rigor. Questions should be designed to encourage sophisticated reasoning patterns and knowledge integration, while solutions should demonstrate clear logical progression and serve as effective learning examples. These interrelated quality dimensions, among others, guide our systematic data curation process detailed in the following sections.

3.3 High-Quality Data Curation

Our data curation process focuses on constructing a high-quality dataset $\mathcal{D} = \{(q_i, r_i, a_i)\}_{i=1}^N$ where N is intentionally kept small to validate our LIMO hypothesis.

3.3.1 Question Selection

We hypothesize that high-quality questions $q \in \mathcal{Q}$ should naturally elicit extended reasoning processes. Our selection criteria include the following:

- **Level of difficulty** We prioritize challenging problems that foster complex reasoning chains, diverse thought processes, and knowledge integration, enabling LLMs to effectively leverage pre-trained knowledge for high-quality inference.
- **Generality** Problems that deviate more from the model's training distribution can better challenge its fixed thinking patterns, encourage exploration of new reasoning approaches, thus expanding its inference search space.
- **Knowledge Diversity** The selected problems should cover various mathematical domains and concepts, requiring the model to integrate and connect distant knowledge during problem-solving.

To implement these criteria effectively, we first assembled a comprehensive pool of candidate problems from various established datasets: NuminaMath-CoT, featuring meticulously annotated problems from high school to advanced competition levels; AIME historical examination problems, known for its extremely challenging and integrative problems spanning multiple mathematical domains; MATH (Hendrycks et al., 2021), encompassing various competitive mathematics problems from prestigious contests; and several other sources of mathematical problems.

From this rich initial collection, we employed a systematic multi-stage filtration process. Beginning with an initial pool of tens of millions of problems, we first applied a baseline difficulty filter using Qwen2.5-Math-7B-Instruct (Yang et al., 2024), eliminating problems that this model could solve correctly in a few attempts. This process helped establish a preliminary difficulty threshold. Subsequently, we subjected the remaining problems to a more rigorous evaluation using state-of-the-art reasoning models including R1, DeepSeek-R1-Distill-Qwen-32B (Guo et al., 2025), and models from Huang et al. (2024), retaining only problems where even these most capable models achieved success rates below certain threshold through multiple sampling iterations. Finally, to maintain corpus diversity, we employed strategic sampling techniques that balanced representation across mathematical domains and complexity levels while avoiding conceptual redundancy. This meticulous selection process ultimately yielded **817** carefully curated problems from an initial pool of tens of millions of candidates, with the selected problems collectively satisfying our stringent quality criteria while spanning a rich spectrum of mathematical reasoning challenges.

3.3.2 Reasoning Chain Construction

Beyond high-quality questions, the quality of solutions plays a pivotal role in the training phase of large language models. To curate high-quality solutions, we adopted a comprehensive selection strategy. We began by gathering official solutions for problems where available, complemented by solutions authored by both human experts and AI specialists. Additionally, we leveraged state-of-the-art reasoning models, including DeepSeek R1, DeepSeek-R1-Distill-Qwen-32B (Guo et al., 2025), and Qwen2.5-32b-Instruct, to generate diverse solution approaches. Furthermore, following the methodology proposed in O1-Journey-Part2 (Huang et al., 2024), we utilized self-distillation techniques based on Qwen2.5-32b-Instruct to create additional model variants, which were then used to generate supplementary problem responses. These responses were then filtered according to the correctness of the answers to establish a baseline collection of valid solutions. Subsequently, all the authors conducted a comprehensive analysis of these filtered solutions through collaborative examination. Through careful observation and systematic review, we identified several key characteristics that distinguish high-quality reasoning chains:

- **Optimal Structural Organization**: The solution exhibits clear and well-organized structural formatting, with adaptive granularity in step decomposition. Particularly, it allocates more tokens and detailed elaboration at crucial reasoning junctures while maintaining concise expressions for straightforward steps. This self-adaptive approach to step granularity ensures that complex transitions receive appropriate attention while avoiding unnecessary verbosity in simpler deductions.
- **Effective Cognitive Scaffolding**: High-quality solutions provide strategic educational support by gradually building understanding through carefully structured explanations. This includes progressive concept introduction, clear articulation of key insights at critical points, and thoughtful bridging of conceptual gaps, making complex reasoning processes more accessible and learnable.
- **Rigorous Verification**: High-quality solutions incorporate extremely frequent verification steps throughout the reasoning process. This includes validating intermediate results, cross-checking assumptions, and confirming the logical consistency of each deduction, thereby ensuring the reliability of the final answer.

Based on these identified characteristics, we developed a hybrid approach combining rule-based filtering and LLM-assisted curation to select high-quality solutions for each question identified in the previous section. This systematic process ensures that each selected solution adheres to our established quality criteria while maintaining consistency across the dataset. By focusing on a minimal yet meticulously curated set of reasoning chains, we embody the core principle of **Less-Is-More**: high-quality demonstrations, rather than sheer data volume, are key to unlocking complex reasoning capabilities. The resulting dataset $\mathcal{D}$ consists of carefully curated triples (q, r, a), where each reasoning chain r satisfies our quality criteria. By maintaining these stringent standards while limiting the dataset size $|\mathcal{D}|$, we aim to demonstrate that high-quality demonstrations, rather than large quantities of training data, are crucial for unlocking complex reasoning capabilities.

4 Methodology

Based on the **Less-Is-More** principle, a model with substantial reasoning knowledge from pre-training and the ability to perform long-chain reasoning at test time can develop robust reasoning abilities. After training on only a few hundred instances of SFT data, the model learns to integrate meta-reasoning tasks into a cohesive reasoning chain.

4.1 Training Protocol

We fine-tune Qwen2.5-32B-Instruct using supervised fine-tuning on our LIMO dataset. The training process employs full-parameter fine-tuning with DeepSpeed ZeRO-3 optimization (Rajbhandari et al., 2020) and FlashAttention-2 (Dao, 2023), with a sequence length limit of 16,384 tokens.

4.2   Evaluation Framework

In-domain Evaluation   To comprehensively assess the models' performance across various reasoning capabilities, we have established a diverse evaluation framework encompassing both traditional and novel benchmarks. Our primary evaluation suite includes several well-established mathematical competitions and benchmarks: the American Invitational Mathematics Examination **(AIME24)**, **MATH500** (Hendrycks et al., 2021), and the American Mathematics Competitions **(AMC23)**.

Out-of-distribution Evaluation   To rigorously evaluate the models' performance on out-of-distribution (OOD) tasks, we carefully selected benchmarks that differ from our training data in various aspects. These benchmarks can be categorized into three distinct groups:

- **Diverse Mathematical Competitions**: We furthur selected **OlympiadBench** (He et al., 2024), which represents a distinct distribution of mathematical challenges to test models' OOD performance.
- **Novel Multilingual Benchmarks**: To minimize data contamination, we constructed several benchmarks using the most recent examination problems: **CHMath** from the 2024 Chinese High School Mathematics League Competition, **Gaokao** from China's 2024 National College Entrance Examination, **Kaoyan** from Chinese Graduate School Entrance Examinations, and **GradeSchool**, our newly developed benchmark for elementary mathematical reasoning. Notably, all problems in these benchmarks are written in **Chinese, while our training data contains no Chinese problems**. This introduces an additional OOD dimension, assessing not only the model's ability to generalize across problem distributions but also its cross-lingual reasoning capabilities when confronted with unseen languages.
- **Multi-disciplinary Benchmarks**: To assess broader generalization capabilities beyond mathematics (our training domain), we incorporated **Miverva** (Lewkowycz et al., 2022) (which includes undergraduate-level STEM problems) and **GPQA** (Rein et al., 2023). These benchmarks evaluate reasoning abilities across multiple disciplines and cognitive levels, providing insights into the model's capacity to transfer mathematical reasoning skills to broader contexts.

Performance metrics   We evaluate performance using the pass@1 metric across our suite of benchmarks. All evaluations are conducted in a Zero-shot Chain-of-Thought (CoT) setting to better assess the model's reasoning capabilities. For benchmarks including MATH500, OlympiadBench, Gaokao, Kaoyan, GradeSchool, MinervaMath, and GPQA, we employ a straightforward approach using greedy decoding with a single sample to assess correctness. However, for the smaller benchmarks containing fewer than 50 problems each (specifically AIME24, AMC23, and CHMATH), we implement a more thorough evaluation protocol, generating 16 samples with a temperature setting of 0.7 and calculating the unbiased pass@1 metric as introduced in Chen et al. (2021). For problems where answers are well-structured numerical values, we directly apply rule-based evaluations to check for mathematical equivalence. For more complex answer formats—such as expressions, equations, or structured solutions—we leverage an LLM-based evaluator, which we have validated for high reliability. Throughout all evaluations, we maintain a maximum output length of 32,768 tokens to minimize the potential for output truncation, ensuring our assessment captures complete problem-solving attempts. Additionally, when evaluating LIMO, we observed that inference-time scaling occasionally results in repetitive patterns at the end of lengthy outputs. In such cases, we extract the most likely final answer from the model's response for evaluation to ensure accurate assessment of its problem-solving capabilities.

5   Experiment

5.1   Baselines

We compare LIMO against a comprehensive set of baselines with the following prominent models:

- **OpenAI-o1-preview (OpenAI, 2024)**, a large language model that has demonstrated advanced mathematical reasoning abilities across various complex tasks.
- **QwQ-32B-Preview (Team, 2024b)**, a model specifically designed for mathematical problem-solving with strong reasoning capabilities.
- **Qwen2.5-32B-Instruct**, which serves as our base model for comparative analysis.

For evaluation, we use the OpenAI API to access OpenAI-o1-preview, while using VLLM (Kwon et al., 2023) to deploy other open-weight models (e.g. QwQ-32B-Preview). To ensure fair comparison, all models follow the same evaluation protocol with identical inference hyper-parameters.

To investigate the impact of training data efficiency, we conduct comparative experiments using mainstream open-source reasoning datasets for supervised fine-tuning on our base model. For a fair comparison, all experiments use the same LLM backbone as LIMO, ensuring that performance differences are solely attributable to the training data characteristics.

- **OpenThoughts-114k**:[1] A synthetic reasoning dataset containing 114k examples covering mathematics, science, coding, and puzzles. The solutions follow a structured reasoning format generated by DeepSeek-R1.
- **NuminaMath-100k**: A randomly selected 100k subset of NuminaMath-CoT, featuring mathematical problems ranging from Chinese high school exercises to international mathematics olympiad competitions. Each solution follows a Chain of Thought (CoT) format (Wei et al., 2022).

These datasets contain substantially more samples than LIMO's training set (817 examples), allowing us to examine the relationship between data quantity and model performance.

5.2   Main Results

Our experimental results demonstrate LIMO's superior performance across both in-domain and out-of-domain tasks, as shown in Table 3.

In-domain Performance   On in-domain tasks, LIMO achieves the best results across all benchmarks. For AIME24, LIMO achieves 57.1% accuracy, outperforming QwQ-32B-Preview (50.0%) and OpenAI-o1-preview (44.6%) by significant margins. Most notably, on MATH500, LIMO achieves 94.8% accuracy, surpassing QwQ-32B-Preview (89.8%) and OpenAI-o1-preview (85.5%). The performance gap is even more pronounced on AMC23, where LIMO reaches 92.0% accuracy compared to QwQ-32B-Preview's 83.6%.

Out-of-domain Generalization   LIMO demonstrates strong generalization capabilities across diverse out-of-domain tasks. On OlympiadBench, LIMO achieves 66.8% accuracy, significantly outperforming QwQ-32B-Preview (58.5%) and the base model (45.3%). Similar improvements are observed on other challenging benchmarks such as CHMath (75.4% vs 68.5%) and GradeSchool (76.2% vs 63.8%). Notably, LIMO maintains competitive performance even on GPQA, where it achieves 66.7% accuracy, close to OpenAI-o1-preview's leading score of 73.3%.

Comparison with Larger Datasets   Our experiments reveal that despite larger scale, both baseline datasets underperform compared to LIMO. NuminaMath-100k shows significant degradation (32.3% vs. base model's 49.9%) due to uncurated reasoning chains, while OpenThoughts-114k achieves suboptimal results (58.3%) probably due to unfocused problem selection. In contrast, LIMO's carefully curated 817 problems yield superior performance (72.8%), demonstrating that targeted selection and high-quality annotations are more crucial than data quantity for developing robust reasoning capabilities.

Overall Performance   LIMO achieves the highest average performance of 72.1% across all benchmarks, substantially outperforming OpenAI-o1-preview (67.8%), QwQ-32B-Preview (66.4%), and other baselines. This comprehensive evaluation demonstrates that LIMO's carefully curated training approach with just 817 examples can outperform models trained on datasets that are orders of magnitude larger.

5.3   Analysis

5.3.1   RQ1: Impact of Reasoning Chain Quality

To gain a deeper understanding of why Less-Is-More achieves such remarkable results, we investigate the quality of reasoning chains (CoT). A fundamental question naturally arises: what characteristics define a high-quality reasoning chain that leads to superior model performance? To address this, we conducted a controlled comparative study examining how solutions of varying quality for the same problem statements affect the performance of models trained on them.

Setup   To conduct this analysis, we selected 500 problems from the LIMO dataset. The selection was based on the intersection of problems for which the models used in rejection sampling exhibited performance differences and those with corresponding human-annotated solutions, ensuring consistency across comparisons. For these 500 problems, we collected and categorized solutions into five distinct quality levels based on our comprehensive evaluation framework. These solutions were sourced from various origins, including human experts, AI specialists, and model-generated responses, then classified strictly based on their reasoning quality rather than their source.

Quality Measure   Following the principles outlined in Section 3.3.2, we took a holistic approach to categorize the reasoning chains into five quality levels (L1-L5, with L5 being the highest). Our assessment focused on several key aspects: how well the steps were organized and connected, whether important logical transitions were properly explained, and if the solution included self-verification steps to check the work. Using these general guidelines, we classified L5 solutions as those showing excellent organization with clear, well-explained steps and thorough self-verification. L4 solutions were also well-structured but perhaps with slightly less rigorous checking. L3 solutions showed decent organization but sometimes skipped over explaining crucial logical leaps. L2 solutions

---

[1] https://github.com/open-thoughts/open-thoughts

Table 3: **Comparison of model performance (pass@1) across various mathematical reasoning benchmarks** Models include state-of-the-art LLMs (OpenAI-o1-preview, QwQ-32B-Preview), our base model (Qwen2.5-32B-Instruct), and models fine-tuned on different datasets. Training data sizes are shown in parentheses. Best results for each benchmark are shown in bold. Our proposed LIMO model (highlighted in blue) achieves superior performance despite using significantly fewer training examples (817) compared to other fine-tuned models (more than 100k).

Datasets	OpenAI-o1-preview	Qwen2.5-32B-Instruct	QwQ-32B-preview	OpenThoughts (114k)	NuminaMath (100k)	LIMO ours(817)
In Domain						
AIME24	44.6	16.5	50.0	50.2	6.5	**57.1**
MATH500	85.5	79.4	89.8	80.6	59.2	**94.8**
AMC23	81.8	64.0	83.6	80.5	40.6	**92.0**
Out of Domain						
OlympiadBench	52.1	45.3	58.5	56.3	36.7	**66.8**
CHMath	50.0	27.3	68.5	74.1	11.2	**75.4**
Gaokao	62.1	72.1	80.1	63.2	49.4	**81.0**
Kaoyan	51.5	48.2	70.3	54.7	32.7	**73.4**
GradeSchool	62.8	56.7	63.8	39.0	36.2	**76.2**
Minerva	**47.1**	41.2	39.0	41.1	24.6	44.9
GPQA	**73.3**	48.0	65.1	42.9	25.8	66.7
AVG.	61.1	49.9	66.9	58.3	32.3	**72.8**

often provided abbreviated reasoning without much explanation, while L1 solutions typically just listed basic steps with minimal elaboration and rarely included any verification.

Results   Our training results (Figure 2) strongly correlate with the reasoning chain quality levels. Models trained on L5 quality reasoning chains achieved the highest performance on both AIME24 and MATH500, demonstrating the effectiveness of well-structured, detailed, and self-verified reasoning. Performance consistently decreased with each quality level, with L4 and L3 showing moderate success, while L2 and L1 resulted in notably lower performance. These results empirically validate our quality assessment framework and highlight the crucial role of high-quality reasoning chains in model performance. Specifically, we observed that the performance gap between L5 and L1 solutions was substantial - approximately 15 percentage points on AIME24 and 12 percentage points on MATH500. This significant difference suggests that the quality of reasoning chains plays a far more important role in model performance than previously assumed, reinforcing the importance of carefully curating training data to include well-structured, thorough solutions.
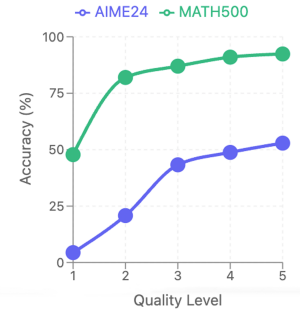
Figure 2: Comparison of models trained on reasoning chains of different quality levels.

5.3.2   RQ2: Impact of Question Quality

We hypothesize that more challenging problems foster complex reasoning chains, diverse thought processes, and enhanced knowledge integration, enabling LLMs to better leverage pre-trained knowledge for high-quality inference. To validate this hypothesis, we investigate how question quality affects the reasoning capabilities of models fine-tuned on these questions and their corresponding solutions.

Setup   We selected three sets of problems of similar size but increasing difficulty, constructing solutions in a consistent manner to form three training datasets. Our findings indicate that models trained on more challenging datasets exhibit superior reasoning performance. Specifically, we sampled three sets of problems, each containing 500 samples, from MATH and AIME:

- **Simple-500**: 500 simple problems randomly selected problems from MATH levels 1 and 2.
- **Complex-500**: 500 complex problems randomly selected problems from MATH levels 3, 4, and 5.
- **Advanced-500**: 500 advanced problems randomly selected problems from past AIME tests.

9

To rigorously establish the increasing difficulty of these sets, we evaluated various LLMs on them, observing a decline in accuracy and an increase in the average length of correctly generated reasoning chains.

We then used DeepSeek-R1 to generate solutions, which represent the highest-quality solutions available, for each problem set, forming the training data for fine-tuning Qwen2.5-32B-Instruct.

Results   We evaluate all three fine-tuned models on the AIME2024 and MATH500 benchmarks to assess their reasoning performance. The results (Figure 3) indicate that modifying the selection of problems alone leads to a 16% improvement in accuracy on the challenging AIME2024 benchmark, reaching 51.5%. Furthermore, despite the absence of in-domain training data, the model fine-tuned on Advanced-500 outperforms the other two models, achieving an accuracy of 91.2% on the MATH500 benchmark. This result suggests that the improvement in reasoning ability due to increased problem difficulty generalizes across datasets.
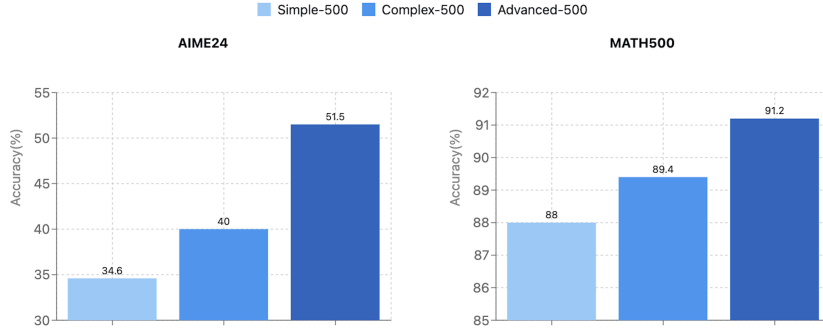
Figure 3: Performance comparison on MATH and AIME benchmarks between models trained on different question quality: Simple-500, Complex-500, and Advanced-500.

5.3.3   RQ3: LLM Backbone

Building on our LIMO hypothesis, which emphasizes the importance of latent prerequisite knowledge within the model's parameter space, we examine how different pre-training data affect a model's capacity to leverage minimal exemplars for mathematical reasoning. This investigation allows us to assess the first key factor of our hypothesis: the role of pre-trained knowledge in enabling complex reasoning capabilities.

Setup   To isolate the impact of pre-training while controlling for model architecture and fine-tuning procedures, we conduct experiments using two 32B-parameter variants of the Qwen model family: Qwen1.5-32B-Chat (Team, 2024a) and Qwen2.5-32B-Instruct (the base model of LIMO). Both models share the same architecture and parameter count, while Qwen2.5 demonstrates significant improvements in pre-training data quality, particularly in mathematical and code-related data, compared to its predecessor. We SFT both models using identical LIMO datasets and evaluation protocols, assessing their performance on the AIME2024 and MATH500 benchmarks.

Results   Our experiments reveal that the choice of pre-trained model dramatically impacts reasoning performance, as demonstrated in 4. LIMO, built on Qwen2.5-32B-Instruct, significantly outperforms its predecessor across both benchmarks. On the challenging AIME2024 test, LIMO achieves 57.1% accuracy, a remarkable 47.1 percentage point improvement over Qwen1.5-32B-Instruct's 10.0%. Similarly, on MATH500, LIMO demonstrates exceptional performance with 94.8% accuracy, surpassing Qwen1.5-32B-Instruct by 34.4 percentage points. These substantial improvements suggest that the enhanced pre-training in Qwen2.5 creates a stronger foundation for mathematical reasoning. The results align with our LIMO hypothesis, indicating that richer pre-trained knowledge within the model's parameter space enables more effective utilization of minimal exemplars during fine-tuning.

5.3.4   Case study

Qualitative Analysis   Fig. 5 compares responses generated by Qwen2.5-32B-Instruct, DeepSeek-R1, and LIMO. **LIMO achieves capabilities and behaviors comparable to DeepSeek-R1, despite using minimal data and compute resources (only 817 training samples).** Notably, LIMO demonstrates strong self-reflection and long chain-of-thought generation capabilities. LIMO verifies its own statements ("Wait, 24 minutes is 0.4 hours? Wait, no. Wait, 60 minutes is 1 hour, so 24 minutes is 24/60, which is 0.4 hours") and validates its calculations ("But let me check again. Maybe I made a mistake in calculations."). Furthermore, it learns to allocate additional tokens (compute) for detailed complex equation-solving ("Now let's compute the left side, · · · , multiply both sides by 2) to prevent errors. In contrast, the base model Qwen2.5-32B-Instruct exhibits limitations in its reasoning process, being unable to correct inaccurate statements and failing to cross verify equation 2 in its solution. **These results**

Figure 4: Impact of Pre-trained Model Choice on Mathematical Reasoning Performance

Figure 5: Comparison between the responses generated by Qwen2.5-32B-Instruct, DeepSeek-R1, and LIMO

strongly support the LIMO hypothesis: With minimal but high-quality post-training examples, models can be empowered into strong reasoners.

Quantitative Analysis   Table 4 demonstrates the differences between models trained with varying sample quality. We observe a general trend where increasing post-training example quality leads to models that generate longer responses with more lines. Additionally, these higher-quality models employ more self-reflecting transitions (e.g., wait, perhaps, maybe, therefore) to allocate additional inference tokens (compute) for deeper thinking.

6   Background and Related Work

6.1   Evolution of Mathematical Reasoning in LLMs

Large-scale training data has been the driving force behind the development of reasoning abilities in LLMs. In the pretraining phase, the reasoning ability of LLMs can be enhanced by relevant corpora (Wang et al., 2024; Azerbayev et al., 2024; Paster et al., 2023; Shao et al., 2024). These curated corpora can be composed of multiple sources, such as textbooks, scientific papers, and mathematical code, which capture diverse human cognitive patterns used to solve problems. In the post-training phase, a line of research focuses on curating large-scale instruction data to teach LLMs to reason (Yue et al., 2023, 2024; Li et al., 2024a). This includes scaling the number of questions and their corresponding solutions. The scaling approach is promising and has achieved significant performance gains. However, the reasoning ability gained through this method has been criticized for relying on the memorization of fixed patterns rather than achieving true generalization (Mirzadeh et al., 2024; Zhang et al., 2024). For example, Mirzadeh et al. (2024) finds that LLMs exhibit noticeable variance when responding to different instantiations of the same question, and their performance declines when only the numerical values in the question are altered. This raises doubts about the generalization capability of SFT methods (Chu et al., 2025) and whether LLMs can be true reasoners rather than mere knowledge retrievers (Kambhampati, 2024).

6.2   Test-time Scaling and Long Chain Reasoning

Instead of focusing on scaling model parameters and training data (Kaplan et al., 2020), recent work has shifted to exploring test-time scaling (OpenAI, 2024; Snell et al., 2024), i.e., increasing the number of tokens to improve

Data Quality Level	Avg. Tokens per response	Avg. Lines per response	Top 10 Frequently Occurring Keywords (in order)
Level 1	230	9.21	since, however, number, let, thus, which, get, two, triangle, theta
Level 2	444.88	50.68	number, need, times, which, find, list, thus, since, triangle, sum
Level 3	4956.11	375.60	**perhaps**, **alternatively**, **consider**, number, **wait**, which, sides, need, equal, seems
Level 4	4726.97	354.87	**wait**, which, number, **perhaps**, therefore, let, since, **maybe**, sides, two
Level 5	5290.26	239.29	**wait**, therefore, which, number, since, lets, two, sides, let, **maybe**

Table 4: Statistical analysis of models trained with examples of varying data quality. This table presents three key metrics: average token count per response, average line count per response, and frequently occurring keywords in model-generated responses. Keywords associated with reasoning transitions and uncertainty are highlighted in **bold**, with common stop words (e.g., "a", "the") excluded to focus on substantive language patterns. Notable differences in response length and keyword usage patterns suggest varying levels of reasoning complexity.

performance. This can be achieved by augmenting LLMs with methods such as parallel sampling (Brown et al., 2024; Wang et al., 2022; Li et al., 2022) or symbolic tree search (Hao et al., 2023; Chen et al., 2024; Yao et al., 2023) to enhance reasoning ability. Furthermore, OpenAI (2024); Guo et al. (2025) explore training LLMs using reinforcement learning to generate long CoT, which often include self-reflection, verification, and backtracking—processes commonly employed by humans when solving complex problems. This approach not only innovates the training paradigm for LLMs but also provides a new form of training data to augment their reasoning ability. Our work demonstrates that this long CoT exhibits high-quality characteristics in eliciting the inherent reasoning abilities of LLMs.

6.3   Data Efficiency in Language Models

Zhou et al. (2024a) demonstrates that with just 1,000 carefully curated prompts and responses, models can learn to follow specific formats and generalize well to unseen tasks. The findings emphasize the importance of quality over quantity in the alignment process. However, whether this lesson can be applied to reasoning tasks remains uncertain, given the potential high computational complexity of such tasks (Merrill and Sabharwal, 2024; Xiang et al., 2025). While some work on reasoning highlights the importance of quality during the curation of training data (Zhou et al., 2024b), the quantity of such data is still much larger compared to that in LIMA. Our work extends the ideology of LIMA to reasoning tasks by investigating what constitutes high-quality questions and solutions, and demonstrates that the reasoning ability of LLMs can be enhanced in a highly data-efficient manner.

7   Future Work

While LIMO demonstrates remarkable success in mathematical reasoning with minimal data, several promising directions remain for future exploration.

Domain Generalization: First, extending the LIMO hypothesis to broader reasoning domains represents a critical next step. While our work focuses on mathematical reasoning, the principles of high-quality reasoning chains could potentially generalize to scientific reasoning, logical deduction, and causal inference. Understanding how these principles transfer across domains could reveal universal patterns in effective reasoning. This exploration would require adapting our quality metrics and developing domain-specific evaluation frameworks, ultimately contributing to a more comprehensive theory of machine reasoning.

Theoretical Foundations: A deeper theoretical understanding of LIMO's success is also essential. Future research should focus on formalizing the relationship between pre-training knowledge, inference-time computation, and reasoning capabilities. This includes investigating the minimum threshold of pre-trained knowledge required for effective reasoning and developing mathematical models to predict the optimal balance between reasoning chain

quality and quantity. Such theoretical foundations could guide the development of more efficient training strategies and provide insights into the fundamental nature of machine reasoning.

Automated Assessment: The development of automated quality assessment tools represents another crucial direction. Current manual evaluation of reasoning chain quality, while effective, is time-consuming and difficult to scale. Future work should focus on creating automated systems that can evaluate and improve reasoning chain quality based on our proposed metrics. This could include developing algorithms that automatically enhance existing reasoning chains and generate high-quality ones with minimal human intervention, making the LIMO approach more accessible and scalable.

Multi-modal Integration: Cross-modal reasoning presents an exciting frontier for extending LIMO's principles. As real-world reasoning often involves multiple modalities, investigating how visual information and structured data can enhance mathematical reasoning capabilities is crucial. This research direction would require developing new quality metrics for multi-modal reasoning chains and understanding how different types of information can be effectively integrated into the reasoning process.

Real-world Impact: The application of LIMO principles to real-world scenarios deserves significant attention. Future work should focus on adapting these approaches to practical problems in education, scientific research, and industrial applications. This includes developing specialized versions of LIMO for specific domains and creating tools that help human experts generate high-quality reasoning chains for complex real-world problems. Such applications could significantly impact how we approach problem-solving in various fields.

Cognitive Science Bridge: Finally, integrating insights from cognitive science could provide valuable directions for improvement. Understanding the parallels between LIMO's reasoning patterns and human cognitive processes could inform the development of more effective reasoning strategies. This includes studying how different reasoning approaches affect model performance and generalization, and incorporating cognitive science principles into the design of reasoning chains. Such research could not only improve AI systems but also provide insights into human reasoning processes.

These future directions collectively aim to deepen our understanding of efficient reasoning in large language models while expanding their practical applications. By pursuing these paths, we can work toward more sophisticated, efficient, and widely applicable reasoning systems that better serve human needs across various domains.

8 Acknowledgement

References

[1] Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2024. Llemma: An open language model for mathematics.

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

[3] Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. 2024. Alphamath almost zero: process supervision without process. *ArXiv preprint*, abs/2405.03553.

[4] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

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

[6] Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

[7] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony

Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant

References

Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

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

[9] Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model.

[10] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*.

[11] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

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

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

[14] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

[15] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

[16] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857.

[17] Chen Li, Weiqi Wang, Jingcheng Hu, Yixuan Wei, Nanning Zheng, Han Hu, Zheng Zhang, and Houwen Peng. 2024a. Common 7b language models already possess strong math capabilities.

[18] Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif Rasul, Longhui Yu, Albert Q Jiang, Ziju Shen, et al. 2024b. Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. *Hugging Face repository*.

[19] Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. 2022. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097.

[20] William Merrill and Ashish Sabharwal. 2024. The expressive power of transformers with chain of thought.

[21] Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2024. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models.

[22] OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. 2024. Openai o1 system card.

[23] OpenAI. 2024. Learning to reason with llms, september 2024.

[24] Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, and Jimmy Ba. 2023. Openwebmath: An open dataset of high-quality mathematical web text.

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

[26] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

[27] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

[28] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

[29] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, YK Li, Yu Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

References

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

[31] Qwen Team. 2024a. Introducing qwen1.5.

[32] Qwen Team. 2024b. Qwq: Reflect deeply on the boundaries of the unknown.

[33] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

[34] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

[35] Zengzhi Wang, Xuefeng Li, Rui Xia, and Pengfei Liu. 2024. Mathpile: A billion-token-scale pretraining corpus for math.

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

[37] Violet Xiang, Charlie Snell, Kanishk Gandhi, Alon Albalak, Anikait Singh, Chase Blagden, Duy Phung, Rafael Rafailov, Nathan Lile, Dakota Mahan, Louis Castricato, Jan-Philipp Franken, Nick Haber, and Chelsea Finn. 2025. Towards system 2 reasoning in llms: Learning how to think with meta chain-of-thought.

[38] Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. 2024. Benchmarking benchmark leakage in large language models.

[39] An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. 2024. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

[40] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models.

[41] Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. Metamath: Bootstrap your own mathematical questions for large language models.

[42] Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mammoth: Building math generalist models through hybrid instruction tuning.

[43] Xiang Yue, Tuney Zheng, Ge Zhang, and Wenhu Chen. 2024. Mammoth2: Scaling instructions from the web.

[44] Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Charlotte Zhuang, Dylan Slack, Qin Lyu, Sean Hendryx, Russell Kaplan, Michele Lunati, and Summer Yue. 2024. A careful examination of large language model performance on grade school arithmetic.

[45] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024a. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

[46] Fan Zhou, Zengzhi Wang, Qian Liu, Junlong Li, and Pengfei Liu. 2024b. Programming every example: Lifting pre-training data quality like experts at scale.