

An Empirical Study on Reasoning and Generalization in Large Language Models

Lu Xiang, Yang Zhao, Yaping Zhang, Zixuan Ren, Xingquan Zhang, Zhiyuan Chen, Chengqing Zong*

State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, CAS, Beijing, China

lu.xiang@nlpr.ia.ac.cn, yang.zhao@nlpr.ia.ac.cn,
yaping.zhang@nlpr.ia.ac.cn, renzixuan2021@ia.ac.cn,
zhangxingquan2022@ia.ac.cn, chenzhiyuan2023@ia.ac.cn,
cqzong@nlpr.ia.ac.cn

Recent reasoning-focused large language models (LLMs) (e.g., GPT-o1 and DeepSeek-R1) combine chain-of-thought prompting with reinforcement learning to generate explicit, multi-step derivations. However, their practical advantages over standard instruction-tuned models remain underexplored. In this paper, we present a comprehensive empirical study to address these gaps. We compare reasoning-focused and non-reasoning LLMs across several critical dimensions: (1) their performance on mathematical and general reasoning benchmarks; (2) their generalization capabilities across 30 sub-disciplines spanning six scientific domains; (3) the qualitative and quantitative differences in their underlying solution processes; and (4) the impact of model compression via distillation on core reasoning capabilities. Crucially, to reliably conduct our in-depth process analysis, we address the critical issue of LLM-as-a-Judge reliability by proposing and validating a multi-level evaluation framework against human experts. Our experimental results demonstrate that reasoning-focused models significantly outperform non-reasoning counterparts on logic-intensive tasks such as mathematics and physics, while the latter remain competitive on factual or observational problems (e.g., biology, geography). Specifically, our validated process analysis reveals distinct styles: DeepSeek-R1 generates more elaborate reasoning processes, correlating with greater effectiveness on harder tasks, despite potential redundancy. In contrast, GPT-o1 exhibits greater conciseness and frequently presents solutions with superior clarity and efficiency, particularly on simpler problems, while struggling with complex scenarios. Additionally, reasoning capabilities can be partially preserved through distillation, with distilled reasoning models consistently outperforming their non-reasoning counterparts across all domains. These findings provide new insights into when and how explicit reasoning contributes to LLM performance, offering practical guidance for model development and deployment.

* Corresponding authors

1. Introduction

Large language models (LLMs) have showcased remarkable progress across a wide range of tasks, from natural language understanding to complex reasoning (Jahin et al. 2025; Chen et al. 2025; Xiang et al. 2025; Zhang et al. 2025). Prominent models, such as OpenAI’s GPT series (OpenAI 2023), Alibaba’s Qwen series (Qwen 2024), Google’s Gemini series (Cloud 2025), MetaAI’s LLaMA series (Grattafiori et al. 2024), and DeepSeek’s DeepSeek series (DeepSeek-AI 2024, 2025), have not only driven the evolution of natural language processing (NLP) technologies but also facilitated intelligent solutions for practical applications, significantly influencing the field and shaping the way we interact with language. Among these, DeepSeek-R1 (DeepSeek-AI 2025) and GPT-o1 (OpenAI 2024b) stand out as a new class of reasoning-focused LLMs, explicitly optimized for reasoning-intensive tasks.

Building upon these advances, recent work has categorized LLMs into two broad classes according to the **explicitness** of their reasoning:

(1) **Reasoning-focused LLMs**, or **Large Reasoning Models (LRMs)** (Xu et al. 2025; Chen et al. 2025; Li et al. 2025) are designed to perform explicit multi-step reasoning through long Chain-of-Thought (long CoT) (Wei et al. 2022), often trained via supervised fine-tuning (SFT) and reinforcement learning (RL) on curated reasoning data. Crucially, the specific RL strategies for reasoning typically follow two main paradigms: result-based supervision, which rewards only the final answer’s correctness (DeepSeek-AI 2025), and process supervision, which rewards the soundness of each intermediate step (Lightman et al. 2024). Notable LRMs include GPT-o1 (OpenAI 2024b) and DeepSeek-R1. To increase the model’s accessibility and practicality, DeepSeek has also released a series of distilled variants of R1, including both Qwen-based and LLaMA-based versions, which are optimized for deployment in resource-constrained environments while preserving the core reasoning capabilities (Guo et al. 2024; Shao et al. 2024).

(2) **Non-reasoning LLMs**, such as GPT-4o (OpenAI 2023), LLaMA (Grattafiori et al. 2024), DeepSeek-V3, Qwen2.5 series (Qwen 2024), rely on general instruction tuning and next-token prediction (Devlin et al. 2019) without explicit alignment to reasoning tasks. Their output often contains short or implicit reasoning patterns, lacking the consistency and interpretability required for complex problem solving. While non-reasoning models excel in naturalistic generation and factual recall, they tend to underperform in reasoning benchmarks (Min et al. 2024; Qin et al. 2024).

Despite the recent surge in reasoning-capable models, there remains limited empirical evidence comparing these models under controlled conditions. In particular, GPT-o1 and DeepSeek-R1, two prominent reasoning models that are thought to exemplify these different RL approaches, have rarely been thoroughly compared. Moreover, few studies have examined these models alongside their non-reasoning counterparts (e.g., DeepSeek-V3, GPT-4o). Consequently, it remains unclear **when** reasoning alignment can bring tangible benefits, **how** the underlying reasoning chains differ, **whether** such capabilities can be retained after compression, and **to what extent** explicit reasoning improves cross-domain generalization.

This challenge is further exacerbated by the lack of reliable methodologies for evaluating reasoning quality. While using LLMs as automated judges (LLM-as-a-Judge) is common practice, this method often suffers from significant reliability and validity issues, including susceptibility to various biases, inconsistency across runs, and a frequent failure to adhere to complex instructions (Zheng et al. 2023; Wang et al. 2024a). This limitation becomes especially evident when evaluating qualitative aspects of the reasoning process, such as the soundness of an inference, the coherence of an argument,

or the clarity of an explanation. Without rigorous, human-validated rubrics for such complex, qualitative judgments, there is little empirical support for the reliability of these LLM-based evaluators, thereby raising serious concerns about the validity of their assessments (Chehbouni et al. 2025).

To address these fundamental gaps, we design and conduct a series of multi-dimensional experiments aimed at answering the following research questions:

- **Task-level advantage:** On which specific tasks do reasoning-focused models truly outperform non-reasoning ones, and on which do they fall short?
- **Reasoning Process difference:** How do the reasoning chains of GPT-o1 and DeepSeek-R1 differ in depth, redundancy, and self-consistency?
- **Compression effect:** To what extent do distilled variants (e.g., R1-Distill-Qwen/LLaMA) retain the reasoning skill of the parent model?
- **Cross-domain generalization:** Does the capacity for explicit reasoning enhance a model’s ability to generalize across diverse domains?

To this end, this paper presents a novel empirical study of DeepSeek-R1, GPT-o1, the distilled variants of DeepSeek-R1, and other state-of-the-art reasoning and non-reasoning LLMs. Specifically, we contribute the following:

- We provide a **broad and systematic empirical comparison between reasoning-focused and non-reasoning LLMs across a diverse set of domains and tasks**, including mathematics, science, code generation, strategic question answering, and temporal reasoning. Crucially, this also includes a comprehensive analysis of **cross-domain generalization** across 30 scientific sub-domains to understand how explicit reasoning benefits broader adaptability. This comparison is uniquely strengthened by a controlled experiment leveraging the Qwen3 model series (Yang et al. 2025), which allows us to isolate the specific impact of the explicit reasoning process by comparing its thinking and non-thinking modes.
- We propose, define, and validate a **rigorous evaluation framework for assessing the quality of reasoning processes**. By creating a detailed, operationalized rubric and performing a human annotation study to calibrate the reliability of our automated metrics, we address a critical methodological flaw in current LLM-as-a-Judge research.
- We conduct a **deep multi-dimensional comparative analysis of DeepSeek-R1 and GPT-o1**. This analysis moves beyond simple accuracy metrics to examine their underlying characteristics, such as solution complexity, logical coherence, and token efficiency, thereby uncovering their respective strengths and limitations in problem-solving.
- We analyze the **reasoning retention and trade-offs in distilled variants of DeepSeek-R1**, comparing models like R1-Distill-Qwen and R1-Distill-LLaMA to their full-scale counterpart. This allows us to examine which aspects of reasoning degrade under compression and what remains transferable.

Through these detailed empirical comparisons, our work contributes new insights distinct from prior reviews or surveys, and clarifies how and when explicit reasoning benefits LLM performance. The rest of this paper is organized as follows: Section 2 introduces the key models involved in this study, including their architectures and reasoning capabilities. Section 3 details our experimental design, research questions, and the benchmarks employed. Section 4 presents our multi-level evaluation framework and, critically, the results of its rigorous human-in-the-loop validation. Section 5 presents benchmarking results from existing evaluations and our controlled experiments. Section

6 provides an in-depth empirical study of DeepSeek-R1 and GPT-o1, including reasoning granularity, complexity, and evaluation using multiple LLM-based judges. Section 7 examines model generalization across domains, comparing reasoning and non-reasoning models across 30 scientific sub-domains, and across key general-purpose reasoning tasks. Section 8 presents a broader discussion of our findings and future research directions. Finally, Section 9 concludes the paper.

2. LLM Paradigms and Model Descriptions

This section introduces the specific models selected for our experimental analysis. Instead of attempting to comprehensively catalog all existing LLMs, we focus on a representative subset of models that allows us to systematically investigate reasoning capabilities and generalization performance under controlled conditions.

2.1 Model Architecture and Training

2.1.1 Common Transformer Architectures. Both DeepSeek and OpenAI’s GPT models fundamentally build upon the transformer architecture introduced by [Vaswani et al. \(2017\)](#). This architecture relies on self-attention mechanisms to process input sequences and has proven remarkably effective across natural language processing tasks ([Zhao, Zhang, and Zong 2023](#)). Despite their common foundation, significant architectural variations exist between model families.

The typical transformer-based LLM architecture includes token embeddings, positional encodings, multiple layers of self-attention mechanisms, feed-forward neural networks, and normalization layers ([Touvron et al. 2023](#)). The decoder-only architecture has become the dominant approach for generative language models, though encoder-decoder architectures remain relevant for certain specialized applications.

2.1.2 Variations in Training Data, Optimization Objectives, and Scale. Although pre-training datasets for proprietary models remain undisclosed, open-weight architectures offer valuable insights into data composition. Foundation models like DeepSeek-V3 are pre-trained on massive corpora enriched with mathematical and code-oriented content to establish robust baseline capabilities ([DeepSeek-AI 2024, 2025](#)). Crucially, during post-training, reasoning-focused models such as DeepSeek-R1 utilize hundreds of thousands of curated reasoning trajectories generated via rejection sampling to cultivate explicit multi-step deduction ([DeepSeek-AI 2025](#)). In contrast, non-reasoning models are primarily aligned using diverse human-annotated conversations and general instruction-following examples to ensure broad applicability ([Ouyang et al. 2022](#)).

Beyond the underlying data, optimization objectives have evolved from simple next-token prediction to more sophisticated approaches. Models like DeepSeek-R1 incorporate specialized loss functions that explicitly reward reasoning pathways ([DeepSeek-AI 2025; Guo et al. 2024](#)), while others employ reinforcement learning from human feedback (RLHF) or reinforcement learning from AI feedback (RLAIF) to align outputs with human preferences or designated reasoning patterns ([Ouyang et al. 2022](#)).

Scale variations between models significantly impact performance characteristics. The parameter counts range from more compact 7B-8B parameter models designed for resource-constrained environments to larger 70B+ parameter versions optimized for maximum capability ([Hoffmann et al. 2022](#)). The relationship between scale and reasoning capability is not strictly linear, with architectural innovations and training methodology sometimes enabling smaller models to outperform larger ones on specific reasoning tasks ([Jiang et al. 2023](#)).

2.2 Reasoning-focused vs. Non-reasoning LLMs

LLMs can be broadly categorized into reasoning-focused and non-reasoning architectures, a distinction that has become increasingly important in the evolution of LLM capabilities.

Reasoning-focused LLMs are explicitly designed to demonstrate enhanced logical reasoning, multi-step problem-solving, and complex inference capabilities (Zhong et al. 2024; Yax, Anlló, and Palminteri 2024; Yu et al. 2024; Sun et al. 2023). LLMs in this category include the OpenAI o1 and o3 series, DeepSeek-R1, Gemini 2.0 Flash Thinking, and Grok 3. These models attempt to mimic human-like cognitive processes by incorporating specialized training procedures that promote systematic analysis, hypothesis testing, and structured problem-solving. Operationally, this reasoning manifests as the generation of a long Chain-of-Thought (CoT) trajectory. This CoT mechanism enables the models to manipulate abstract concepts, establish causal relationships, and articulate intermediate logical derivations before producing a final solution (Wei et al. 2022; Chen et al. 2025).

Non-reasoning LLMs, by contrast, excel at tasks requiring extensive knowledge retrieval, pattern recognition, and generation based on probabilistic continuations of text (Zhao et al. 2023). Example models in this category include GPT-4o, Claude 3.5 Sonnet, DeepSeek-V3, and the Qwen 2.5 series (Qwen 2024), among others. These models leverage statistical correlations within their training data to produce contextually appropriate responses without necessarily engaging in explicit reasoning processes. Their strength lies in efficient information retrieval, linguistic fluency, and the ability to generate coherent and contextually relevant content across diverse domains. While they may exhibit emergent reasoning abilities at sufficient scale, their architecture and training methodology are not specifically optimized for step-by-step reasoning processes.

This distinction is not binary but rather exists on a spectrum, with many modern models incorporating elements of both paradigms (Bubeck et al. 2023). The categorization is further complicated by the emergence of reasoning capabilities as an emergent property at scale, even in models not explicitly designed for reasoning tasks.

2.3 Overview of Models

To facilitate a systematic comparison, we select representative models from both reasoning-focused and non-reasoning LLMs. The chosen models differ in architecture, training objectives, and deployment scope, and include both proprietary and open-source systems. Specifically, we consider:

- **Reasoning-focused models**, including proprietary models such as GPT-o1, the open-source DeepSeek-R1, and its distilled variants.
- **Non-reasoning models**, including DeepSeek-V3 and GPT-4o, which are tuned for general-purpose tasks without explicit reasoning alignment.
- A controllable model series, Qwen3, which can operate in both *thinking* and *non-thinking* modes, providing a unique testbed for controlled experiments.

In the following subsections, we provide an overview of each model, highlighting its architecture, training strategy, and relevance to our comparative analysis.

2.3.1 DeepSeek-R1. DeepSeek-R1 represents a significant advancement in reasoning-focused language models. Its architecture incorporates specialized attention mechanisms that enhance multi-step logical inference capabilities (Liu et al. 2024). The model demonstrates particular strength in mathematical reasoning, algorithmic problem-solving, and tasks requiring structured logical deduction. This capability stems from both architectural

innovations and the composition of its training data, which includes a substantial proportion of scientific and technical content ([DeepSeek-AI 2025](#)).

The training methodology of DeepSeek-R1 is distinct in its adoption of a **result-based reward** paradigm, which primarily assesses the correctness of the final answer. This reward mechanism encourages the model to explore valid reasoning paths without being overly constrained by demonstration formats. DeepSeek-R1 achieves state-of-the-art or near state-of-the-art performance on a diverse range of reasoning benchmarks. This includes mathematical evaluations such as GSM8K, MATH, and AIME, as well as coding and complex scientific reasoning tasks like LiveCodeBench and GPQA Diamond. Its empirical strength lies in its ability to carry out multi-step symbolic manipulation, algebraic reasoning, and structured problem-solving, which reflects both architectural enhancements and multi-stage training strategies.

2.3.2 GPT-o1. OpenAI’s GPT-o1 represents a significant evolution in enhancing systematic reasoning. While its specific architectural details remain proprietary, OpenAI’s official report states that the model is trained using a *large-scale reinforcement learning algorithm* designed to teach it how to productively use its chain of thought. Through this RL process, GPT-o1 learns to refine its thinking process, try different problem-solving strategies, and recognize and correct its own mistakes ([OpenAI 2024b](#)).

Although the exact nature of GPT-o1’s reinforcement learning paradigm remains undisclosed, available evidence and community consensus suggest an alignment with **process-based reward** methodologies ([Lightman et al. 2024](#)). It is also often speculated that its ability to explore different reasoning paths involves mechanisms analogous to Monte Carlo Tree Search (MCTS) ([Christiano et al. 2017](#); [Zhou et al. 2023](#)). In our study, we categorize GPT-o1 as a model whose training appears heavily influenced by reinforcing a sound and traceable reasoning process. GPT-o1 balances high-quality reasoning with broad knowledge coverage, maintaining strong performance not only on benchmark reasoning tasks but also on general QA and language understanding benchmarks.

2.3.3 DeepSeek-v3. DeepSeek-v3 ([DeepSeek-AI 2024](#)) represents an alternative approach that prioritizes extensive knowledge retrieval and contextual memory over specialized reasoning mechanisms. The model is built on a **Mixture-of-Experts (MoE)** transformer architecture ([Dai et al. 2024](#)), where a subset of experts is activated per token, enabling the model to scale to over 100B parameters effectively while keeping inference efficient. Beyond this architecture, DeepSeek-V3 uses an auxiliary-loss-free strategy and a multi-token prediction training objective to enhance its capabilities.

Through pre-training on large diverse and high-quality tokens, DeepSeek-V3 excels at knowledge-intensive tasks. Importantly, DeepSeek-V3 also serves as the foundation model for DeepSeek-R1. Through supervised fine-tuning on reasoning-focused datasets and reinforcement learning, DeepSeek-R1 builds on DeepSeek-V3’s language and knowledge capabilities to gain strong reasoning capability.

2.3.4 GPT-4o. GPT-4o is OpenAI’s general-purpose LLM designed to handle text, image, and audio inputs while integrating vast knowledge with remarkable fluency and speed ([OpenAI 2024](#)). Although its exact training procedures remain undisclosed, it is widely understood to rely on supervised fine-tuning and RLHF to align with user preferences for helpfulness, informativeness, and fluency ([Shelby et al. 2023](#)).

While GPT-4o can generate step-by-step explanations when explicitly prompted, it is fundamentally categorized as a versatile foundation model rather than a reasoning-specialized one. Thus, its default responses tend to favor direct answers over explicit

Table 1: Overview of Key Models

Model	Paradigm	Key Features	RL strategy	Primary Strengths
DeepSeek-R1	Reasoning-focused	Reflective reasoning and self-verification steps for math and coding	Result-based (Final answer accuracy) + Format rewards	Advanced problem-solving, math competitions, subject-based QA
GPT-o1	Reasoning-focused	Proprietary large-scale Transformer; partial reliance on short CoT for multi-step tasks	Process-based reward + MCTS (speculated)	General-purpose NLP; smaller math problems; code explanations
DeepSeek-v3	Non-Reasoning	Mixture-of-Experts	SFT only	Factual QA; Long-context tasks
GPT-4o	Non-Reasoning (multimodal)	Proprietary architecture; Multimodal	User preference alignment	General-purpose QA; multimodal input
Distilled DeepSeek-R1 Models	Reasoning-focused	Distilled from DeepSeek-R1	Distillation+ DeepSeek-R1 pipeline	Efficient reasoning under low resources
Qwen3	Controllable (Reasoning / Non-Reasoning)	Unified framework with switchable operational modes	Hybrid RL (Rule-based, Model-based, & Preference)	flexible for complex reasoning and rapid QA

reasoning trajectories. Consequently, although excelling on knowledge-intensive benchmarks, it is generally less robust than reasoning-focused models on tasks requiring sustained symbolic manipulation, formal deduction, or complex mathematical reasoning.

2.3.5 Distilled DeepSeek-R1 Models. Distilled models with reduced parameters serve a crucial role in the LLM ecosystem by enabling deployment in resource-constrained environments. The distillation process typically involves the smaller student model learning to mimic the output distributions of the larger teacher model across a carefully selected set of training examples (Hinton, Vinyals, and Dean 2015; Xu et al. 2024b). DeepSeek Team released distilled 1.5B, 7B, 8B, 14B, 32B, 70B variants of the DeepSeek-R1 based on Qwen2.5 and LLaMA3 series (DeepSeek-AI 2025). These models undergo a two-stage distillation process: 1) Offline distillation from the teacher model’s chain-of-thought outputs; 2) Result-reward RL fine-tuning, mirroring the original R1 pipeline.

As reported in DeepSeek-AI (2025), these distilled models perform exceptionally well on benchmarks and significantly outperform previous open-source models and comparable to o1-mini. Besides, they offer significantly reduced inference cost and memory consumption compared to DeepSeek-R1. They are thus attractive for deployment scenarios where reasoning capability is needed but full-scale models are impractical.

2.3.6 Qwen3 (thinking vs. non-thinking). A critical challenge in comparing reasoning and non-reasoning models lies in isolating the specific effects of reasoning-focused tuning. To address this challenge, our study utilizes the Qwen3 series of models, which has integrated two distinct operational modes within a unified framework. This built-in, dual-mode capability allows us to guide the exact same model to operate in two distinct ways (Yang et al. 2025):

- **Thinking Mode:** When engaged for complex problem-solving, this mode activates the model’s multi-step reasoning capabilities. The model first generates an explicit chain-of-thought before producing the final answer.
- **Non-Thinking Mode:** In this mode, the model performs as a conventional instruction-tuned large language model, delivering rapid and direct responses.

By embedding these two modes within the exact same model, this framework enables a rigorously controlled comparison. It allows us to isolate and measure the performance gains specifically attributable to the explicit reasoning process, making Qwen3 a uniquely valuable resource for our study.

Table 1 presents a comparative overview of the six models discussed above. It summarizes their design paradigm, key features, reinforcement learning strategies, and primary strengths, highlighting the distinguishing characteristics among the models.

3. Experimental Design and Benchmarks

3.1 Research Questions

Our empirical investigation is driven by the following research questions:

RQ1: How do reasoning-focused models differ from non-reasoning models in performance across diverse tasks? This question aims to investigate the necessity and utility of explicit reasoning capabilities in LLMs. By comparing reasoning-focused models (e.g., DeepSeek-R1, GPT-o1) with non-reasoning models (e.g., DeepSeek-V3, GPT-4o), and a controlled experiment using the Qwen3 series, we examine whether explicit reasoning plays a critical role in achieving superior performance across different tasks.

RQ2: What are the qualitative and quantitative differences in the reasoning processes of DeepSeek-R1 and GPT-o1 on reasoning tasks? This question moves beyond final-answer accuracy to dissect the underlying problem-solving behaviors. We investigate how the solution paths of DeepSeek-R1 and GPT-o1 diverge in terms of step granularity, logical consistency, error types, and token efficiency. To ensure the reliability of this analysis, we conduct our evaluation using a rigorously defined and human-validated framework (detailed in Section 4).

RQ3: To what extent can the reasoning ability of DeepSeek-R1 be preserved through model distillation? Smaller models distilled from DeepSeek-R1 (e.g., 7B and 8B variants) offer a promising path for deploying reasoning capabilities in resource-constrained environments. This investigation focuses on two key aspects: (1) how much of DeepSeek-R1’s original reasoning performance is preserved through distillation, and (2) whether the reasoning knowledge distilled from DeepSeek-R1 can be used to enhance non-reasoning base models.

RQ4: To what extent do reasoning capabilities generalize across diverse domains? Recent reasoning-focused models, such as DeepSeek-R1 and GPT-o1, have been extensively optimized on mathematical reasoning datasets. This raises an important question: Does a model’s proficiency in mathematics serve as a direct predictor of its performance across adjacent scientific disciplines (e.g., physics, chemistry, biology) and general-purpose tasks like coding, temporal reasoning, and commonsense inference? Addressing this clarifies whether enhanced mathematical reasoning constitutes a core, transferable capability or whether success across these varied domains depends more fundamentally on specialized, domain-specific adaptations beyond general reasoning abilities.

3.2 Experimental Design and Benchmarks

To address our research questions, we designed a multi-dimensional experimental comparison examining model capabilities across three primary axes:

(1) Overall Performance Comparison: Contrasting reasoning versus non-reasoning models using both aggregated public benchmarks and a controlled experiment with the Qwen3 series to isolate the impact of explicit reasoning.

(2) In-depth Process Analysis: A detailed qualitative and quantitative comparison of the solution processes of DeepSeek-R1 and GPT-o1, conducted using our validated evaluation framework.

(3) Cross-Domain Generalization: Evaluating model performance across 30 scientific sub-domains and key general-purpose reasoning tasks.

To comprehensively evaluate performance, our study employs a suite of benchmarks organized into general-purpose, mathematical, and scientific reasoning categories.

3.2.1 General-Purpose Reasoning Benchmarks. We incorporate three benchmarks representing distinct facets of general-purpose reasoning:

- HumanEval (Chen et al. 2021): A widely-used benchmark for evaluating code generation. It consists of 164 hand-written programming problems, each with a function signature, docstring, and several unit tests. Models are tasked with completing the function body.
- StrategyQA (Geva et al. 2021): A challenging question-answering benchmark designed to test a model’s ability to perform multi-step, inferential reasoning over commonsense and world knowledge. The dataset contains yes/no questions where the reasoning process is not explicitly provided and must be inferred.
- TimeBench (Chu et al. 2024): A comprehensive benchmark for evaluating temporal reasoning. It aggregates multiple datasets into a suite of tasks that assess a model’s understanding of temporal expressions, event ordering, duration, and causality.

3.2.2 Mathematical Reasoning Benchmarks. We evaluate mathematical reasoning performance using four well-established datasets, each covering different levels of difficulty and mathematical domains:

- GSM8K (Cobbe et al. 2021): A dataset of 8,500 grade school math word problems requiring multi-step reasoning, which has become a standard benchmark for evaluating LLM mathematical reasoning capabilities.
- MATH (Hendrycks et al. 2021b): A collection of 12,500 challenging mathematics problems drawn from competitions, covering algebra, geometry, probability, and other advanced topics. It tests deeper mathematical reasoning beyond arithmetic.
- OlympiadBench (He et al. 2024): A collection of mathematical olympiad problems designed to test advanced problem-solving capabilities and creative mathematical reasoning. It represents some of the most challenging problems for current LLMs.
- AIME 2024 (AI-MO 2024): The most recent American Invitational Mathematics Examination problems, providing contemporary mathematical reasoning challenges not included in model training data, given our models’ knowledge cutoff dates.

From each benchmark, we carefully select a representative subset of problems to ensure a balanced evaluation. These subsets are chosen for their broad coverage of mathematical domains, and representation of different complexity levels, ensuring a comprehensive assessment of LLMs’ symbolic manipulation, logical reasoning, and problem-solving abilities (Pradipta Gema et al. 2024; Setlur et al. 2024; Zhang et al. 2024).

3.2.3 Scientific Benchmarks. To assess reasoning and generalization in specialized scientific domains, we employ several domain-specific benchmarks that go beyond general pretraining knowledge:

- GPQA Diamond (Rein et al. 2024): A subject-structured benchmark designed to evaluate factual and reasoning performance across science-heavy academic

disciplines. It contains high-quality, curriculum-aligned multiple-choice questions organized into hierarchical subject and subfield groupings, allowing fine-grained assessment of model capabilities in domain-specific contexts.

- **SuperGPQA (M-A-P Team 2025):** A newly developed benchmark containing graduate-level questions across multiple disciplines, designed specifically to test the upper bounds of LLM capabilities in specialized knowledge domains.

These datasets are particularly relevant to **RQ4**, as they enable us to examine whether models trained with mathematical or general reasoning supervision generalize effectively to structured, discipline-specific tasks. To facilitate this analysis, we constructed a subject-organized test set by sampling from both GPQA and SuperGPQA across six core scientific domains (mathematics, physics, chemistry, astronomy, geography, and biology), with five subfields under each. This enables detailed performance analysis at both the domain and subdomain levels.

3.3 Data Contamination and Considerations

A potential challenge in evaluating LLMs is data contamination, where benchmarks used for evaluation may have been inadvertently included in the models' vast pre-training corpora (Xu et al. 2024a; Deng et al. 2024). While complete elimination is difficult for web-scale models, we employed several mitigation strategies:

- **Use of Recent Benchmarks:** We included benchmarks created after the likely knowledge cutoff dates of the evaluated models, specifically AIME 2024 and the SuperGPQA dataset (M-A-P Team 2025). Problems from these sources are highly unlikely to have been part of the models' pre-training data, offering a cleaner signal of genuine reasoning capabilities.
- **Focus on Process-Level Analysis:** A significant portion of our analysis evaluates process-level metrics (e.g., step consistency, complexity) rather than just final answer accuracy. While final answers might be memorized, consistently reproducing a specific, multi-step reasoning process is substantially harder, making this analysis less susceptible to contamination.
- **Inclusion of Diverse and Specialized Benchmarks:** Our evaluation suite includes 30 fine-grained scientific sub-domains from GPQA and SuperGPQA. The niche and specific nature of many of these problems makes them less likely to have been prevalent in standard pre-training mixtures compared to canonical benchmarks like GSM8K, further reducing the risk of direct memorization.

Despite these mitigation efforts, we acknowledge that potential conceptual overlap between our benchmarks and the models' training data cannot be entirely ruled out. This remains an inherent limitation of current LLM evaluations.

4. Definition and Validation of the Evaluation Framework

To reliably answer our research questions, particularly those requiring an in-depth analysis of the quality of reasoning processes (RQ2), we must first establish a trustworthy evaluation framework. As LLMs are increasingly adopted as evaluators in reasoning analysis research (Wei et al. 2022; Lightman et al. 2024), valid concerns over evaluation soundness and reliability have grown (Wei et al. 2025). In our work, LLMs perform complex analytical tasks, including decomposing solution narratives and scoring their quality at both a step-by-step and holistic level. Since these LLM-based evaluations support

Xiang et al. An Empirical Study on Reasoning and Generalization in Large Language Models

Question

The 9 members of a baseball team went to an ice cream parlor after their game. Each player had a singlescoop cone of chocolate, vanilla, or strawberry ice cream. At least one player chose each flavor, and the number of players who chose chocolate was greater than the number of players who chose vanilla, which was greater than the number of players who chose strawberry. Let N be the number of different assignments of flavors to players that meet these conditions. Find the remainder when N is divided by 1000.

Reasoning Process

Okay, so there's this problem about 9 members of a baseball team going to an ice cream parlor. Each player gets a single scoop of either chocolate, vanilla, or strawberry ... We need to find the number of different assignments of flavors to players that meet these conditions, then take that number modulo 1000.

First, let me parse the problem. We have three flavors: ...

But here, we have constraints. Let's think about how to model this ...

Alright, let's start with step 1: finding all possible triples (c, v, s) ...

Wait, $c + v = 6$ here, since total is 9 ...

Wait, maybe I missed some triples. Let me verify ...

Wait, wait, something's wrong here ...

Wait, no. Hold on. Wait, total sum is 9 ...

Wait, that seems straightforward. But let me double-check ...

But wait, is that all? Are there only three triples? Wait, perhaps I missed some triples. Let me verify again ...

...

Solution Process

The problem involves determining the number of different assignments of flavors (chocolate, vanilla, strawberry) to 9 baseball team members, given that each flavor is chosen by at least one player, and the number of players choosing chocolate is greater than those choosing vanilla, which in turn is greater than those choosing strawberry. We need to find this number modulo 1000.

First, we identify the valid triples (c, v, s) where $c + v + s = 9$, $c > v > s \geq 1$:

1. For $s = 1$:

- $c + v = 8$ with $c > v \geq 2$

- Valid pairs: $(6, 2, 1)$ and $(5, 3, 1)$

2. For $s = 2$:

- $c + v = 7$ with $c > v \geq 3$

- Valid pair: $(4, 3, 2)$

No valid triples exist for $s \geq 3$ as they would exceed the total number of players.

Next, we calculate the multinomial coefficients for each valid triple:

1. For $(6, 2, 1)$:

$$\frac{9!}{6! \cdot 2! \cdot 1!} = \frac{362880}{720 \cdot 2 \cdot 1} = 252$$

2. For $(5, 3, 1)$:

$$\frac{9!}{5! \cdot 3! \cdot 1!} = \frac{362880}{120 \cdot 6 \cdot 1} = 504$$

3. For $(4, 3, 2)$:

$$\frac{9!}{4! \cdot 3! \cdot 2!} = \frac{362880}{24 \cdot 6 \cdot 2} = 1260$$

Summing these values gives the total number of assignments:

$$252 + 504 + 1260 = 2016$$

Finally, we find the remainder when 2016 is divided by 1000:

$$2016 \bmod 1000 = 16$$

Thus, the remainder when N is divided by 1000 is $\boxed{16}$.

Figure 1: The illustration of Reasoning Process and Solution Process

several key findings in this paper, we conducted a rigorous validation study to assess their reliability. This framework serves to validate all LLM-based evaluation components, ensure reproducibility, and clarify the reliability of our automated judgments.

4.1 Defining the Evaluation Protocol

The foundation of our framework is a comprehensive protocol that specifies the objects of analysis, the evaluation dimensions, and the automated tools used for measurement.

4.1.1 Core Concepts: Reasoning vs. Solution Process. A core aspect of our study is the fine-grained analysis of a model's problem-solving narrative. We distinguish between two key components of this narrative, as illustrated in Figure 1.

- **Reasoning Process:** This refers to the model's intermediate, exploratory thinking before it commits to a structured solution. It includes self-reflection, hypothesis testing, and error correction that may or may not directly appear in the final output.

- **Solution Process:** This refers to the explicit steps the model presents to derive the final answer. This begins when the model transitions from reasoning to constructing a concrete solution, which is the primary object of our step-level quality evaluation.

4.1.2 Evaluation Dimensions and Rubric. The evaluation of multi-step reasoning is an active area of research, with existing approaches often focusing on a single level of granularity. Some methods evaluate the **final outcome**, such as the correctness of the final answer (Cobbe et al. 2021). Another significant line of work analyzes the **reasoning steps** (Xia et al. 2025), using techniques like parsing-based methods (Prasad et al. 2023), embedding-based metrics (Golovneva et al. 2023), or process reward models (Lightman et al. 2024). A third approach makes **holistic judgments** about the entire reasoning process, often through prompting-based methods (Zeng et al. 2025; Wang et al. 2025).

However, these approaches tend to operate in isolation, lacking a unified framework that systematically connects fine-grained, step-level properties to the overall solution quality. To address this gap, this paper proposes and empirically validates a multi-level evaluation framework. We create a comprehensive evaluation framework by defining two distinct but interconnected levels of analysis:

- **The Micro-level:** Designed to assess the atomic properties of individual reasoning steps, such as their logical consistency and correctness.
- **The Macro-level:** Designed to evaluate the holistic qualities of the entire solution, such as its overall efficiency and conceptual soundness.

This dual-level structure allows us to not only judge the final output but also diagnose the specific strengths and weaknesses within the reasoning process itself. The complete rubric for both levels, presented in Table 2, serves as the annotation guide for our human annotators and the core instruction set for the LLM judge.

4.1.3 LLM-based Analysis Pipeline. Our evaluation protocol utilizes a two-stage LLM-based analysis pipeline to systematically assess the Solution Process based on the dimensions defined above.

Stage 1: Solution Decomposition. To enable fine-grained analysis of the Solution Process, the pipeline first employs LLM-based decomposers (QwQ-plus and DeepSeek-R1-Distill-Qwen-14B) to divide the narrative solution into a sequence of discrete steps. This is achieved by prompting the decomposers with the instruction presented in Appendix A. The prompt directs the LLM to identify atomic logical units, including calculations, deductions, and meta-reasoning actions, and output them in a structured JSON format. An example of this decomposition is shown in Figure 2.

Stage 2: Quality Evaluation. Once the solution is decomposed into structured steps, the pipeline employs an LLM judge (QwQ-plus) to assess quality based on our multi-level evaluation framework (Table 2). This stage involves two types of evaluation:

- **Step-level Quality Evaluation:** Assigning scores to individual reasoning steps for micro-level properties.
- **Holistic Quality Evaluation:** Assigning scores to the complete solution across macro-level dimensions and providing an Overall Score.

Both evaluations are guided by detailed, rubric-based prompts provided to the LLM judge. The complete prompts are presented in full in Appendix B (Micro-level) and Appendix C (Macro-level).

Xiang et al. An Empirical Study on Reasoning and Generalization in Large Language Models

Table 2: Rubric for Human Annotation and LLM Evaluation

Evaluation Level	Dimension	Scoring Rubric
Micro-Level (single step)	Logical Consistency	2 (Highly Consistent) : The step’s deduction strictly and explicitly follows from the preceding steps or given premises. 1 (Basically Consistent): The step is logically sound but requires the reader to infer a minor, obvious connection. 0 (Inconsistent): The step represents a logical leap, a contradiction, or is based on unstated assumptions.
	Correctness	1 (Correct): The mathematical operation, formula application, or logical deduction is entirely correct. 0 (Incorrect): The step contains any mathematical or logical error.
	Step Clarity	2 (Very Clear): The language and notation are unambiguous, professional, and easy to understand. 1 (Basically Clear): The step is understandable, but its phrasing or conciseness could be improved. 0 (Unclear): The step is confusing, vague, or uses ambiguous terminology.
	Redundancy & Efficiency	1 (Necessary & Efficient): The step is essential and does not take an unnecessarily convoluted path. 0 (Redundant or Inefficient): The step repeats previous information or takes an obviously inefficient detour
Macro-Level (complete solution)	Answer Accuracy	1 (Correct): The final answer is completely correct and meets format requirements. 0 (Incorrect): The final answer is incorrect or in the wrong format.
	Reasoning Step Accuracy	2 (Good): The core method and logical chain are correct, with at most minor, non-critical flaws. 1 (Acceptable): The core method is generally correct, but the process has noticeable logical gaps or several minor errors that undermine its rigor. 0 (Poor): The solution contains at least one major logical or conceptual error that derails its validity, or is fundamentally wrong.
	Solution Efficiency	2 (Good): The solution path is direct and does not contain obvious inefficient behaviors. The method used is standard and appropriate for the problem, even if it might not be the most clever possible solution. 1 (Acceptable): The solution path can be followed, but it includes obvious inefficient behaviors (e.g., conducting large-scale brute force enumeration when available formulas can be used; taking unnecessary long detours). 0 (Poor): The solution path is extremely convoluted and inefficient. The method is absurdly complex for the problem, demonstrating a clear lack of strategic insight.
	Step Clarity	2 (Good): The presentation is well-structured and easy to follow. The reader can understand the logic effortlessly. 1 (Acceptable): The solution is understandable but requires significant effort from the reader to connect the steps or decipher the notation. 0 (Poor): The presentation is messy, disorganized, or confusing, making the logic nearly impossible to reconstruct.
	Mathematical Concept Application	2 (Good): Correctly identifies and flawlessly applies the core mathematical concept(s). A minor error in application is acceptable if the core concept is right. 1 (Acceptable): Uses a concept from the correct general field, but applies the wrong specific theorem/method (conceptual confusion). 0 (Poor): Applies a concept from a completely inappropriate field of mathematics, or fails to apply any recognizable concept.

4.2 Human Annotation for Validation

To validate the reliability of the LLM-based analysis pipeline described above, we conducted a rigorous human annotation study employing trained annotators with graduate-level backgrounds in computer science and mathematics. The protocol involved two distinct validation tasks focused on different components of the pipeline.

4.2.1 Task 1: Decomposition Validation. The primary goal of this task was to validate the reliability of the two LLM-based step decomposers (QwQ-plus and DeepSeek-R1-Distill-Qwen-14B). We created a validation set comprising 40 problems, stratified by difficulty (20 from GSM8K and 20 from AIME). For each problem, a solution generated by either DeepSeek-R1 or GPT-o1 was selected. To evaluate both decomposers efficiently,

Question

The 9 members of a baseball team went to an ice cream parlor after their game. Each player had a singlescoop cone of chocolate, vanilla, or strawberry ice cream. At least one player chose each flavor, and the number of players who chose chocolate was greater than the number of players who chose vanilla, which was greater than the number of players who chose strawberry. Let N be the number of different assignments of flavors to players that meet these conditions. Find the remainder when N is divided by 1000.

Solution Process

The problem involves determining the number of different assignments of flavors (chocolate, vanilla, strawberry) to 9 baseball team members, given that each flavor is chosen by at least one player, and the number of players choosing chocolate is greater than those choosing vanilla, which in turn is greater than those choosing strawberry. We need to find this number modulo 1000.

First, we identify the valid triples (c, v, s) where $c+v+s=9$, $c > v > s \geq 1$:

1. For $s = 1$:
 - $c + v = 8$ with $c > v \geq 2$
 - Valid pairs: $(6, 2, 1)$ and $(5, 3, 1)$

2. For $s = 2$:
 - $c + v = 7$ with $c > v \geq 3$
 - Valid pair: $(4, 3, 2)$

No valid triples exist for $s \geq 3$ as they would exceed the total number of players.

Next, we calculate the multinomial coefficients for each valid triple:

$$1. \text{ For } (6, 2, 1): \frac{9!}{6! \cdot 2! \cdot 1!} = \frac{362880}{720 \cdot 2 \cdot 1} = 252$$

$$2. \text{ For } (5, 3, 1): \frac{9!}{5! \cdot 3! \cdot 1!} = \frac{362880}{120 \cdot 6 \cdot 1} = 504$$

$$3. \text{ For } (4, 3, 2): \frac{9!}{4! \cdot 3! \cdot 2!} = \frac{362880}{24 \cdot 6 \cdot 2} = 1260$$

Summing these values gives the total number of assignments:

$$252 + 504 + 1260 = 2016$$

Finally, we find the remainder when 2016 is divided by 1000:

$$2016 \bmod 1000 = 16$$

Thus, the remainder when N is divided by 1000 is $\boxed{16}$.

Format Solution Steps

```
[
  {
    "step_number": 1,
    "action": "Determine valid triples for s=1: c + v = 8 with c > v ≥ 2, yielding (6,2,1) and (5,3,1)"
  },
  {
    "step_number": 2,
    "action": "Determine valid triples for s=2: c + v = 7 with c > v ≥ 3, yielding (4,3,2)"
  },
  {
    "step_number": 3,
    "action": "Confirm no valid triples exist for s ≥ 3 as they exceed total players"
  },
  {
    "step_number": 4,
    "action": "Calculate multinomial coefficient for (6,2,1): 9!/(6!2!1!)=252"
  },
  {
    "step_number": 5,
    "action": "Calculate multinomial coefficient for (5,3,1): 9!/(5!3!1!)=504"
  },
  {
    "step_number": 6,
    "action": "Calculate multinomial coefficient for (4,3,2): 9!/(4!3!2!)=1260"
  },
  {
    "step_number": 7,
    "action": "Sum coefficients: 252 + 504 + 1260 = 2016"
  },
  {
    "step_number": 8,
    "action": "Compute 2016 mod 1000, resulting in 16"
  }
]
```

Solution
Breakdown

Figure 2: Illustration of the LLM-based solution decomposition process. The left presents the full solution generated by *DeepSeek-R1*. The right demonstrates the automated breakdown of this solution into individual steps by an LLM (*QwQ-plus*). Each step is structured as a JSON object, where the `step_number` and `action` fields encapsulate the core reasoning of that specific step.

we applied QwQ-plus to decompose the solutions for 20 problems (10 GSM8K, 10 AIME), and DeepSeek-R1-Distill-Qwen-14B to decompose the solutions for the other 20 problems (10 GSM8K, 10 AIME). This resulted in 40 unique machine-generated decompositions (one per problem).

Two trained annotators then independently rated the *reasonableness* of each of these 40 decompositions using a 3-point scale (2=Highly Reasonable, 1=Basically Reasonable, 0=Unreasonable). This judgment was guided by the following three core criteria:

- **Completeness:** Whether the decomposition captured all key logical steps from the original solution without omission. An unreasonable omission would be leaving out a critical action like a derivation or a substitution.
- **Faithfulness:** Whether the decomposition accurately represented the logic and intent of the original solution without distortion or misinterpretation. An unfaithful step would misrepresent the original solution's logic, for example, describing a proof by contradiction as a direct proof.
- **Atomicity:** Whether each step represented a single, coherent logical action rather than improperly merging multiple distinct steps.

Inter-annotator agreement was calculated to assess the consistency of these judgments. The results of this validation are presented in Section 4.3.1.

4.2.2 Task 2: Quality Validation. The goal of this task was to create a human gold standard for solution quality and use it to calibrate the reliability of our LLM judge (QwQ-plus) for both step-level and holistic quality evaluation. We selected a subset of 20 problems from the Task 1 set (10 GSM8K, 10 AIME). Crucially, for each of these 20 problems, we included the solutions generated by both DeepSeek-R1 and GPT-o1, resulting in a total of 40 solution pairs for quality assessment. The QwQ-plus decomposer (validated in Task 1) was used to generate the step-by-step breakdown for all 40 solutions, yielding a total of 295 individual steps for micro-level analysis. The annotation process involved two distinct stages and two separate groups of trained annotators each:

Stage 1: Human Gold-Standard Creation. One group of three annotators focused exclusively on micro-level scoring. Using the detailed rubric (Table 2), they scored each of the 295 steps across the four micro-level dimensions (Logical Consistency, Correctness, Step Clarity, Redundancy & Efficiency). The second group of two annotators focused exclusively on macro-level quality scoring. They scored each of the 40 complete solutions across the five macro-level dimensions and provided an Overall Score. Disagreements within each group were resolved through discussion to produce the final human gold-standard dataset for both micro and macro levels.

Stage 2: LLM-as-a-Judge Evaluation. We then prompted the QwQ-plus LLM judge to score the exact same set of 295 steps (micro-level) and 40 solutions (macro-level) using the identical rubric embedded in its instructions (Appendix B & C). The reliability evaluation of the LLM judge employed distinct methodologies for each level, as detailed in Section 4.3. For the micro-level analysis, agreement with the human gold standard was calculated. For the macro-level analysis, reliability was assessed by measuring the internal consistency of a group comprising the LLM judge and human annotators.

4.3 Validation Results

This section presents the results of our validation study and defines the policy for using automated metrics based on their demonstrated reliability.

4.3.1 Validation of LLM-based Decomposition. The validation of our step decomposition models was conducted to ensure their reliability for the complexity analysis in Section 6.3. Two trained annotators independently rated the reasonableness of 40 machine-generated decompositions (20 by each model on a stratified sample) using the 3-point scale. The results of this validation are summarized in Table 3. The metrics in the table were calculated as follows:

- **Average Reasonableness Score:** For each decomposition, a final gold score was obtained by averaging the scores from the two annotators. The value reported is the average of these gold scores over the 20 samples evaluated for each model.
- **% of Decompositions Rated as Highly Reasonable:** This represents the percentage of the gold scores for each model that received a top score of 2.
- **Inter-Annotator Agreement (IAA):** To measure the consistency between the two annotators, we calculated Cohen’s Kappa, correcting for chance agreement.

As shown in Table 3, the human validation confirms that both LLM-based decomposers are reliable for segmenting solution processes. Both QwQ-plus and DeepSeek-R1-Distill-Qwen-14B achieved high average reasonableness scores (1.875 and 1.750 out of

Table 3: Validation Results for Step Decomposition Models. *Average Reasonableness Score* is the weighted average score across all samples. *% of Decompositions Rated as Highly Reasonable* denotes the percentage of samples that received a top score of 2.

Metric	QwQ-plus	DeepSeek-R1-Distill-Qwen-14B
Average Reasonableness Score (0-2 scale)	1.875	1.750
% of Decompositions Rated as Highly Reasonable	85.00%	70.00%
Inter-Annotator Agreement (Cohen’s Kappa)	0.773	0.733

Table 4: Human Inter-Annotator Agreement (Micro-Level). Agreement among three annotators on 295 steps.

Dimension (Micro-Level)	Scale	Fleiss’ Kappa
Correctness	0-1	0.869
Logical Consistency	0-2	0.556
Step Clarity	0-2	0.580
Redundancy & Efficiency	0-1	0.638

2, respectively), suggesting that their decompositions were generally judged by human annotators to be reasonable. Furthermore, the substantial Inter-Annotator Agreement for both models (Cohen’s Kappa > 0.73) demonstrates that the annotation criteria were applied consistently, reinforcing the reliability of these ratings. QwQ-plus showed slightly stronger performance, with a higher average score, a greater percentage of decompositions rated as Highly Reasonable, and slightly higher annotator agreement. Nevertheless, the strong performance of both models provides robust evidence supporting their validity for use in our subsequent step complexity analysis (Section 6.3).

4.3.2 Calibrating LLM Judges for Micro-level Quality Evaluation. This validation addresses whether the LLM judges used for our micro-level quality analysis (Sections 6.3.2) are reliable. The validation proceeded in two stages: first, we validated our rubric by measuring human annotator agreement, and second, we calibrated the LLM judge by comparing its scores to the resulting human gold standard.

Inter-Annotator Agreement (IAA). Before validating the LLM, we first had to prove that our evaluation rubric (Table 2) is itself reliable. As described in Task 2, a group of three annotators independently scored all 295 steps. We calculated the Inter-Annotator Agreement (IAA) among these three annotators using **Fleiss’ Kappa** κ .

The results in Table 4 show strong agreement across the board. *Correctness* achieved Almost Perfect agreement (Fleiss’ $\kappa=0.869$), *Redundancy & Efficiency* achieved Substantial agreement (Fleiss’ $\kappa=0.638$), and the more subjective dimensions of *Logical Consistency* and *Step Clarity* both achieved Moderate agreement (Fleiss’ $\kappa=0.556$ and Fleiss’ $\kappa=0.580$, respectively). These Moderate scores are reasonable and expected, as these 3-point (0-2) scales require annotators to make finer subjective judgments (e.g., distinguishing Highly Consistent from Basically Consistent) compared to the more objective, binary *Correctness* dimension. This confirms our rubric is well-defined and can be applied consistently by human experts, establishing a reliable ground truth.

Table 5: LLM-as-a-Judge (QwQ-plus) vs. Human Gold Standard (Micro-Level, N=295 steps). **MAE** (Mean Absolute Error) measures the average magnitude of the errors between the LLM’s score and the gold standard. **RMSE** (Root Mean Square Error) also measures error, but penalizes larger errors more heavily.

Dimension	Accuracy	Cohen’s Kappa	MAE	RMSE
Logical Consistency	0.8712	0.4068	0.2068	0.6032
Correctness	0.7966	0.2416	0.2034	0.4510
Step Clarity	0.9390	0.2554	0.0746	0.3189
Redundancy & Efficiency	0.8339	0.3638	0.1661	0.4076

LLM Judge Calibration for Micro-level Quality Evaluation. Next, we compared the QwQ-plus LLM judge’s scores against the human gold standard for the 295 steps. The results are summarized in Table 5.

Nuanced Overall Performance. The LLM judge’s performance (Table 5) is nuanced and reveals several key characteristics. First, the **high Accuracy** scores (79-94%) across all dimensions are a **positive indicator**. We are evaluating solutions from high-performing LLMs, so the majority of steps are indeed expected to be correct and clear. This high accuracy demonstrates that the LLM-as-a-Judge correctly identifies the proper score for this vast majority of straightforward cases, establishing a strong baseline performance.

Using Cohen’s Kappa to Understand Strengths and Limitations. Since high accuracy can be misleading in unbalanced datasets, we employ **Cohen’s Kappa** to measure agreement. The resulting Cohen’s Kappa scores (0.242 to 0.407), which indicate slight to moderate agreement, allow for a more precise analysis of the LLM’s relative strengths and limitations.

- **Relative Strengths:** The LLM performs best in assessing *Logical Consistency* (Cohen’s $\kappa=0.4068$), achieving Moderate agreement. This is a **promising result**, approaching the human Moderate baseline (Fleiss’ $\kappa=0.556$) and suggesting a genuine, if imperfect, capability in assessing logical flow. Its Fair agreement on *Redundancy & Efficiency* (Cohen’s $\kappa=0.3638$) also shows a useful, non-random signal.
- **Identified Limitations:** The primary challenges for the LLM lie in the *Correctness* (Cohen’s $\kappa=0.2416$) and *Step Clarity* (Cohen’s $\kappa=0.2554$) dimensions. The Slight agreement here, especially for *Correctness*, indicates that while the LLM is right most of the time (per Accuracy), it struggles with the critical task of reliably identifying the minority of steps that are incorrect. This is its key limitation.
- **Error Profile:** Error metrics (*MAE* and *RMSE*) corroborate these findings. For *Step Clarity*, near-perfect Accuracy (93.9%) and low MAE (0.075) occur primarily because the LLM defaults to the majority class. However, the low Cohen’s Kappa (0.255) confirms its inability to meaningfully distinguish between actual clarity levels.

4.3.3 Calibrating LLM Judges: Macro-Level Solution Quality. We then validated the **macro-level quality evaluation** task, where the LLM judge evaluated 40 complete solutions. This calibration is essential as it supports the macro-level solution quality analysis presented in Section 6.4.

Stage 1: Human Inter-Annotator Agreement (Macro-Level). As described in Task 2, two annotators scored the 40 solutions on the five macro-level dimensions. We measured agreement using two metrics: **Exact Agreement** (the percentage of times annotators gave the exact same score) and **Cohen’s Kappa**. As shown in Table 6, the human annotators

Table 6: Human Inter-Annotator Agreement (Macro-Level). Agreement between two annotators on 40 solutions.

Dimension	Scale	Exact Agreement	Cohen's Kappa
Answer Accuracy	0-1	1	1
Reasoning Step Accuracy	0-2	0.9500	0.8718
Solution Efficiency	0-2	0.8250	0.6447
Step Clarity	0-2	0.850	0.6708
Mathematical Concept Application	0-2	0.9000	0.7203
Overall		0.9050	0.7815

Table 7: Macro-Level Agreement: Human-Human vs. Human-Human-LLM Group. (40 solutions)

Dimension	Human-Human Agreement Cohen's Kappa	Human-Human-LLM Group Fleiss' Kappa
Answer Accuracy	1.0000	1.0000
Reasoning Step Accuracy	0.8718	0.7807
Solution Efficiency	0.6447	0.6014
Step Clarity	0.6708	0.6666
Mathematical Concept Application	0.7203	0.7672

achieved exceptionally high agreement on all dimensions. The *Exact Agreement* was 82.5% or higher across the board. The more robust *Cohen's Kappa* scores confirm this strong reliability: *Answer Accuracy* reached perfect agreement (Cohen's $\kappa=1.0$), *Reasoning Step Accuracy* achieved Almost Perfect agreement (Cohen's $\kappa=0.872$), and all other subjective dimensions achieved Substantial agreement (Cohen's $\kappa \geq 0.645$). This confirms the high reliability and clarity of our macro-level rubric.

Stage 2: LLM-as-a-Judge Calibration (Macro-Level). To calibrate the LLM judge's performance at solution level, we treated it as a **third annotator** alongside the two human annotators. We then calculated the **Fleiss' Kappa** for this entire three-member human-machine group. The results are presented in Table 7. The Fleiss' Kappa scores for the human-machine group remained at Substantial or Almost Perfect levels (0.601 to 1.0) across all dimensions. For example, on *Solution Efficiency*, the human-only Cohen's κ was 0.645, and the human-machine Fleiss' κ was extremely close at 0.601. On *Step Clarity*, the scores were nearly identical (0.671 vs. 0.667). Notably, on *Mathematical Concept Application*, the group including the LLM still achieved a Substantial Fleiss' κ of 0.767. This minimal change in agreement scores strongly indicates that the LLM's scoring patterns are **highly consistent and aligned with those of the trained human experts**. It effectively acted as a reliable third annotator, validating its use for the macro-level analysis in Section 6.4.1.

4.4 Reliability Conclusion and Policy

Based on the validation results from both the micro-level (Section 4.3.2) and macro-level (Section 4.3.3), we draw the following conclusions: **1) The Evaluation Framework and Rubric are Reliable.** The high Inter-Annotator Agreement (IAA) scores among human experts at both the micro (Table 4) and macro (Table 6) levels demonstrate that our multi-level rubric is well-defined and quantifiable. **2) The LLM-as-a-Judge Has Specific Strengths and Limitations.** Our calibration reveals a critical, level-dependent difference:

- **Strong Macro-Level Reliability:** At the **macro-level**, the LLM judge proves to be highly reliable. Its judgments are shown to be strongly consistent with human experts, maintaining Substantial group agreement (Fleiss' $\kappa > 0.60$) across all subjective dimensions (Table 7).
- **Specific Micro-Level Limitation:** In contrast, at the **micro-level**, the LLM's primary weakness is exposed. It struggles to reliably identify individual incorrect steps, as shown by the Slight agreement on *Correctness* (Cohen's $\kappa=0.242$).

Conclusion and Usage Policy: Based on these findings, our policy differentiates the reliability of the LLM-as-a-Judge according to the granularity of assessment. Given the demonstrated high alignment with human experts, the judge is considered a **reliable tool** for the macro-level quality scoring. Conversely, the LLM-as-a-Judge cannot be used as an independent, reliable evaluation tool for micro-level analysis. **However, considering its high accuracy (79-94%) and its demonstrated, non-random agreement** on other key dimensions (like *Logical Consistency*), the micro-level analysis still holds significant value as an **exploratory** assessment for preliminary trend discovery.

5. Reasoning vs. Non-reasoning LLMs

To explore how reasoning-focused models differ from non-reasoning models across various tasks (**RQ1**), this section presents a two-stage performance analysis. We first aggregate publicly available benchmark results to establish a high-level overview of the current performance landscape. Following this, we conduct a novel, controlled experiment using the Qwen3 model series to more rigorously isolate and quantify the specific impact of the explicit reasoning process.

5.1 Benchmark Aggregation

To provide a broad context for our study, we begin by analyzing a wide range of benchmark results as reported in official model releases and technical reports. Instead of conducting new evaluations for this initial overview, we aggregate and juxtapose these publicly available results to identify general trends and motivate our deeper, controlled experiments. This aggregation aims to create a unified view of the current performance landscape across knowledge question answering, mathematics, code generation, and alignment tasks. Table 8 presents the performance for various LLMs on different tasks. We have adopted these benchmark results from existing technical reports (DeepSeek-AI 2025; Wu et al. 2025; Yang et al. 2025) to facilitate our analysis. An analysis of the reported data reveals several consistent trends.

On Reasoning-Intensive Tasks. Reasoning-focused models, such as DeepSeek-R1 and GPT-o1, demonstrate consistent superiority over their non-reasoning counterparts on benchmarks that require structured reasoning, such as mathematics, science, and coding. For example, DeepSeek-R1 scores 79.8% on AIME 2024 and 97.3% on MATH-500, outperforming the strongest non-reasoning baseline (DeepSeek-V3) by more than 40 and 7 percentage points, respectively. This suggests that reasoning alignment, achieved through chain-of-thought supervision and reinforcement learning, brings clear benefits in tasks where multi-step reasoning is critical to success.

On Knowledge-Centric and Open-Ended Tasks. Conversely, when benchmarks focus more on factual recall or open-ended tasks, the performance gap narrows and occasionally reverses. Non-reasoning models like DeepSeek-V3 and GPT-4o often remain highly competitive. For example, GPT-o1 leads on MMLU (91.8%), while DeepSeek-V3

Table 8: Performance comparison of Reasoning and Non-Reasoning LLMs. These results are sourced from existing technical reports. The highest scores are presented in bold, whereas the second-best scores are underlined.

Benchmark	Reasoning Models				Non-Reasoning Models	
	DeepSeek -R1	GPT-o1	R1-Distill -Qwen -Base-32B	R1-Distill -LLaMA -70B-Instruct	DeepSeek -V3	GPT-4o
<i>General Tasks</i>						
MMLU (Hendrycks et al. 2021a)	90.8	91.8	-	-	88.5	87.2
MMLU-Pro (Wang et al. 2024b)	84.0	-	-	-	<u>75.9</u>	72.6
MMLU-redux (Gema et al. 2025)	92.9	92.8	88.2	89.3	89.1	88.0
SimpleQA (OpenAI 2024a)	30.1	47.0	-	-	24.9	<u>38.2</u>
C-Eval (Huang et al. 2023)	91.8	85.5	82.2	71.8	86.5	76.0
C-SimpleQA (He et al. 2024)	63.7	-	-	-	68.0	58.7
<i>Mathematics & Science Tasks</i>						
AIME 2024 (AI-MO 2024)	79.8	<u>79.2</u>	72.6	70.0	39.2	9.3
MATH-500	97.3	96.4	94.3	94.5	90.2	74.6
GPQA (Rein et al. 2024)	71.5	75.7	62.1	65.2	59.1	49.9
<i>Coding Tasks</i>						
LiveCodeBench (Jain et al. 2024)	65.9	<u>63.4</u>	57.2	57.5	36.2	32.9
Codeforces(Percentile)	96.3	96.6	93.4	91.4	58.7	23.6
Codeforces(rating)	2029	2061	1691	1633	1134	759
SWE Verified	49.2	48.9	-	-	42.0	38.8
Aider-Polyglot	<u>53.3</u>	61.7	-	-	49.6	16.0
<i>Alignment Tasks</i>						
IF-Eval (Zhou et al. 2023)	83.3	92.6	72.5	79.3	86.1	84.3
Arena-Hard (Li et al. 2024)	92.3	<u>92.1</u>	60.8	60.6	85.5	80.4
AlpacaEval2.0 (Dubois et al. 2024)	87.6	-	-	-	70.0	51.1
WritingBench (Wu et al. 2025)	8.55	8.15	6.13	6.08	<u>7.95</u>	<u>8.16</u>

outperforms reasoning models on C-SimpleQA and ranks a close second on C-Eval. In alignment-oriented evaluations like IF-Eval and Arena-Hard, DeepSeek-V3 remains highly competitive. These results indicate that the extensive breadth of pre-training corpora and instruction-tuning data can compensate for, or even surpass, the advantages of explicit reasoning alignment in tasks that do not require lengthy deductive chains.

Impact of Distillation on DeepSeek-R1. The analysis of DeepSeek-R1’s distilled variants shows that core reasoning capabilities can be effectively preserved even with significantly reduced parameter sizes. For instance, R1-Distill-Qwen-Base-32B achieves 94.3 on MATH-500 compared to 97.3 for DeepSeek-R1, and 93.4 versus 96.3 on Codeforces (percentile), yet it continues to outperform all non-reasoning models. The R1-Distill-LLaMa-70B-Instruct also shows a performance drop but maintains the qualitative characteristics of the full model, such as excelling in logic-centric tasks while being less competitive in general or open-ended tasks. This finding confirms that a significant portion of reasoning ability transfers effectively via knowledge distillation.

5.2 Controlled Experiment with the Qwen3 Model Series

To rigorously test the hypothesis that the explicit reasoning process is a causal factor for the performance gap on structured tasks, we conducted a controlled experiment. We leveraged the Qwen3 series, which features a unified framework with distinct *thinking* (reasoning) and *non-thinking* (direct answering) modes (see Section 2.3.6).

Xiang et al. An Empirical Study on Reasoning and Generalization in Large Language Models

Table 9: Performance of the Qwen3 Model Series in *Thinking* and *Non-Thinking* Modes. These results are sourced from the official Qwen3 technical report (Yang et al. 2025).

Benchmark	Thinking				Non-Thinking			
	Qwen3-8B	Qwen3-14B	Qwen3-32B	Qwen3-235B-A22B	Qwen3-8B	Qwen3-14B	Qwen3-32B	Qwen3-235B-A22B
<i>General Tasks</i>								
MMLU-redux	87.5	88.6	90.9	92.7	79.5	82.0	85.7	89.2
C-Eval	83.4	86.2	87.3	89.6	77.9	81.0	83.3	86.1
<i>Mathematics & Science Tasks</i>								
AIME 2024	76.0	79.3	81.4	85.7	29.1	31.7	31.0	40.1
AIME 2025	67.3	70.4	72.9	81.5	20.9	23.3	20.2	24.7
MATH-500	97.4	96.8	97.2	98.0	87.4	90.0	88.6	91.2
GPQA	62.0	64.0	68.4	71.1	39.3	54.8	54.6	62.9
<i>Coding Tasks</i>								
LiveCodeBench	57.5	63.5	65.7	70.7	22.8	29.0	31.3	35.3
Codeforces(Percentile)	95.6	95.3	97.7	98.2	52.4	64.1	71.0	75.7
Codeforces(rating)	1785	1766	1977	2056	1110	1267	1353	1387
<i>Alignment Tasks</i>								
IF-Eval	85.0	85.4	85.0	83.4	83.0	84.8	83.2	83.2
Arena-Hard	85.8	91.7	93.8	95.6	79.6	86.3	92.8	96.1
AlignBench v1.1	8.46	8.56	8.72	8.94	8.38	8.52	8.58	8.91
WritingBench	7.59	7.80	7.90	8.03	7.15	7.24	7.54	7.70

5.2.1 Performance from the Qwen3 Technical Report. To contextualize this empirical setup, we first analyze the performance of the Qwen3 series as documented in its official technical report (Yang et al. 2025). The performance, summarized in Table 9, facilitates a preliminary assessment of the capabilities and typical performance margin between the model’s *thinking* and *non-thinking* modes across standard benchmarks.

Performance on Reasoning-Intensive Tasks. Table 9 reveals a pronounced trend in tasks requiring complex reasoning. On both Mathematics & Science and Coding benchmarks, the *thinking* mode consistently yields substantial performance gains over the *non-thinking* mode. For example, on AIME 2024 and LiveCodeBench, scores achieved by the thinking mode frequently double or even triple those of its non-thinking counterpart. This consistent pattern provides strong preliminary evidence that an explicit chain-of-thought process is crucial for success in these logic-intensive domains.

Performance on General and Alignment Tasks. In contrast, the performance margin narrows significantly on General and Alignment Tasks. While the thinking mode still holds an edge on benchmarks like MMLU-redux and Arena-Hard, performance on others, such as IF-Eval, is nearly identical. This suggests that for tasks relying heavily on retrieved knowledge, instruction following, or stylistic alignment, the benefits of an explicit reasoning process are less pronounced.

Overall, the publicly reported results strongly support the hypothesis that the utility of an explicit reasoning mode is highly task-dependent. This motivates our subsequent controlled experiment, which further investigates this trend on a targeted subset of reasoning tasks central to our study.

5.2.2 Our Controlled Experiment on Specific Reasoning Tasks. While publicly reported results provide a valuable overview, they cover a broad range of benchmarks that may not directly map to the specific reasoning types central to our research questions. To enable a

Table 10: Performance of Qwen3 Models in Our Controlled Experiment (Mean $\pm$ Std Dev across 3 runs). The *Gain* column represents the absolute performance change, calculated by subtracting the *Non-Thinking* score from the *Thinking* score.

Benchmark Category	Qwen3-4B			Qwen3-30B		
	Non-Thinking	Thinking	Gain	Non-Thinking	Thinking	Gain
Mathematics	54.83 $\pm$ 0.85	77.83 $\pm$ 0.62	23.00 $\uparrow$	61.83 $\pm$ 0.85	80.83 $\pm$ 1.55	19.00 $\uparrow$
Code (base)	88.00 $\pm$ 1.63	96.00 $\pm$ 1.63	8.00 $\uparrow$	96.67 $\pm$ 0.94	95.33 $\pm$ 0.94	1.34 $\downarrow$
Code (extended)	82.67 $\pm$ 1.89	90.67 $\pm$ 2.49	8.00 $\uparrow$	86.00 $\pm$ 0.00	87.33 $\pm$ 2.49	1.33 $\uparrow$
Temporal Reasoning	40.12 $\pm$ 0.41	51.12 $\pm$ 0.31	11.00 $\uparrow$	43.17 $\pm$ 0.16	55.71 $\pm$ 0.85	12.54 $\uparrow$
Knowledge Reasoning	67.33 $\pm$ 0.94	66.67 $\pm$ 3.40	0.66 $\downarrow$	66.00 $\pm$ 1.63	72.00 $\pm$ 1.63	6.00 $\uparrow$

fine-grained analysis and rigorously test the hypothesis that an explicit reasoning process is a causal factor for performance differences, we conducted a controlled experiment.

Experimental Setup. We selected two models from the Qwen3 series of different sizes, **Qwen3-4B**¹ and **Qwen3-30B**², to assess whether the impact of the reasoning mode is consistent across scales. Our evaluation suite was constructed by sampling from four key reasoning categories: (1) a challenging Mathematics set comprising 200 problems (50 from each of GSM8K, MATH, Olympiad, and AIME); (2) 50 Code Generation problems from HumanEval, where solutions were evaluated using the EvalPlus framework (Liu et al. 2023) to yield both base (HumanEval) and extended (HumanEval⁺) test cases; (3) a comprehensive Temporal Reasoning set covering 15 subtasks of TimeBench (50 samples each)³; and (4) 50 Knowledge Reasoning problems from StrategyQA. For each problem, both models were prompted in their *thinking* and *non-thinking* modes. To account for generation variability, each prompt was executed three times, and we report the average performance and standard deviation across these three runs.

Experimental Results. The results of our controlled experiment, presented in Table 10, provide a detailed view of the impact of explicit reasoning. The *thinking* mode demonstrates a decisive advantage on highly structured tasks like Mathematics and Temporal Reasoning. In contrast, its effect on Code Generation and Knowledge Reasoning is more complex and reveals a strong dependency on model scale.

Across both model sizes, the most significant and consistent performance gains were observed on the Mathematics benchmark, with an increase of 23.00% for the 4B model and 19.00% for the 30B model. A similarly strong improvement was seen on Temporal Reasoning, with gains of 11.00% and 12.54% for the 4B and 30B models, respectively. This provides strong causal evidence that **for tasks requiring complex, multi-step logical deduction, an explicit chain-of-thought process is a critical factor for success.**

The impact on Code Generation reveals a more intricate, scale-dependent pattern. The smaller 4B model showed a consistent improvement of 8% on both HumanEval (Code-base) and HumanEval⁺ (Code-extended). However, the larger 30B model exhibited mixed results: a slight performance degradation (-1.34%) on HumanEval, and only a marginal gain (+1.33%) on HumanEval⁺. This suggests that **for highly capable models on certain tasks, a forced, explicit reasoning chain might interfere with already well-tuned direct generation capabilities.**

1 <https://huggingface.co/Qwen/Qwen3-4B>

2 <https://huggingface.co/Qwen/Qwen3-30B-A3B>

3 We excluded the *commonsense/situatedgen* subtasks from TimeBench as they are generative in nature, which does not align with the discriminative evaluation format of the other selected tasks.

The results on Knowledge Reasoning also reveal a distinct scale-dependent effect. The *thinking* mode slightly hindered the performance of the 4B model (-0.66%), but provided a moderate +6.00% boost to the larger 30B model. This suggests that leveraging a structured thinking process on less structured, knowledge-intensive tasks might **require a certain level of model capacity**.

5.3 Key Insights

Aggregating the findings from both the high-level benchmark comparison and our controlled experiment, several key insights emerge regarding the role and impact of explicit reasoning in LLMs:

Explicit Reasoning Provides a Quantifiable Advantage on Structured Tasks. The general trend in our aggregation of public data (Section 5.1), which shows reasoning-focused models excelling on tasks like mathematics and coding, is strongly confirmed by our controlled experiment (Section 5.2). The substantial performance gains when activating Qwen3's *thinking* mode on Mathematics and Temporal Reasoning provide clear causal evidence. This demonstrates that the performance gap is not merely an architectural artifact, but a direct result of the explicit reasoning methodology itself.

The Benefits of Explicit Reasoning are Task-Specific and Scale-Dependent. Both analyses clearly define the boundaries of the reasoning advantage. Our aggregated data shows that the performance gap narrows significantly on knowledge-centric and alignment tasks. This is further substantiated by our controlled experiment, where the *thinking* mode yielded modest gains on StrategyQA. Moreover, our experiment reveals a more complex, scale-dependent interaction: the *thinking* mode resulted in a slight performance decrease for a highly capable model on a task it already excels at (e.g., the 30B model on HumanEval), while being crucial for enabling a larger model to succeed on knowledge-based problems. This indicates that explicit reasoning is not a universal solution and its effectiveness is modulated by task nature and intrinsic model capacity.

Reasoning Capabilities can be Effectively Preserved via Distillation. As established in our analysis of public benchmarks (Section 5.1), the reasoning skills of a powerful parent model can be successfully transferred to smaller, distilled variants. These compressed models retain a significant portion of their parent's capabilities, consistently outperforming non-reasoning models of comparable or even larger scale on reasoning-intensive tasks. This illustrates that knowledge distillation is an effective method for creating efficient yet powerful reasoning models, partially answering RQ3.

Summary Highlights

- Reasoning-focused models demonstrate a clear and quantifiable advantage on structured tasks like mathematics and temporal reasoning.
- The benefit of explicit reasoning is highly task-specific and diminishes significantly on knowledge-intensive, less structured tasks.
- The impact of a reasoning mode is complex and scale-dependent, with potential performance degradation in highly capable models on tasks where they already demonstrate strong proficiency.
- Reasoning capabilities can be effectively preserved in smaller, distilled models, offering a practical path to deploying efficient yet powerful reasoners.

Table 11: Performance comparison between DeepSeek-R1 and GPT-o1 across four datasets (Mean $\pm$ Std Dev across 3 runs). The accuracy rates represent the percentage of correct answers for 50 randomly selected problems from each dataset, totaling 200 questions.

	GSM8K	MATH	Olympiad	AIME	Overall
DeepSeek-R1	99.33 $\pm$ 0.94	82.67 $\pm$ 0.94	56.67 $\pm$ 0.94	80.00 $\pm$ 2.83	79.67 $\pm$ 0.85
GPT-o1	100.00 $\pm$ 0.00	82.00 $\pm$ 2.83	47.33 $\pm$ 0.94	78.00 $\pm$ 3.27	76.83 $\pm$ 1.03

6. Comparative Analysis: DeepSeek-R1 vs. GPT-o1

Building upon the performance overview in Section 5, this section shifts the analytical focus from *what* models achieve to *how* they achieve it. We conduct a multi-dimensional comparative analysis of two leading reasoning models, DeepSeek-R1 and GPT-o1, to investigate their underlying reasoning processes (RQ2). To ensure a rigorous and transparent analysis, this examination is anchored in our suite of mathematical benchmarks, where the reasoning steps are explicit and verifiable.

6.1 Overall Performance Comparison

We evaluated DeepSeek-R1 and GPT-o1 on a total of 200 mathematical problems. These problems were sampled by selecting 50 problems from each of four well-established benchmarks representing increasing levels of difficulty: GSM8K (elementary arithmetic) (Cobbe et al. 2021), MATH (high-school to early-undergraduate algebra, geometry, combinatorics) (Hendrycks et al. 2021b), OlympiadBench (Olympiad-level problems) (He et al. 2024), and AIME 2024 (top-tier U.S. high-school competition problems) (AI-MO 2024). To ensure statistical robustness, all experiments were conducted with three runs, and we report the average performance and standard deviation across these three runs.

6.1.1 Overall Accuracy Across Datasets. The overall results, as shown in Table 11, reveal a clear and nuanced performance landscape. On foundational benchmarks such as GSM8K and MATH, both models achieve high and comparable accuracy, demonstrating robust capabilities in solving standard school-level mathematical problems. However, the performance diverges significantly on the more challenging competition-level problems. **DeepSeek-R1 demonstrates a consistent and superior capability across both high-difficulty benchmarks.** It establishes a substantial advantage on the Olympiad-style problems, outperforming GPT-o1 by a significant margin of over 9 percentage points. It also maintains a measurable lead on the highly structured AIME problems.

The models' diverging performance on these two benchmarks is noteworthy, as they test distinct, complementary reasoning skills. **AIME problems are considered highly structured**, requiring the derivation of a single integer answer through complex but often procedural applications of known theorems. In contrast, **Olympiad-style problems are more diverse and open-ended**, as they cover a wider range of abstract topics and typically demand the construction of a complete proof (Prove that...) or the exploration of all possible solutions (Find all...), a process that relies more on creative and abstract reasoning pathways.

DeepSeek-R1's superior performance on both of these distinct high-difficulty domains indicates a strong and versatile capability in advanced mathematical reasoning.

Xiang et al. An Empirical Study on Reasoning and Generalization in Large Language Models

Table 12: Performance Breakdown on OlympiadBench by Problem Type (Mean $\pm$ Std Dev across 3 runs). The #Num row specifies the number of problems sampled from each problem type within the selected 50 problems.

	Algebra	Combinatorics	Geometry	Number Theory
#Num	19	12	7	12
DeepSeek-R1	64.91 $\pm$ 2.48	50.00 $\pm$ 0.00	57.14 $\pm$ 0.00	50.00 $\pm$ 0.00
GPT-o1	54.39 $\pm$ 2.48	41.67 $\pm$ 0.00	33.33 $\pm$ 6.73	50.00 $\pm$ 0.00

This notable difference in performance, particularly on OlympiadBench, prompts a more granular analysis of the specific problem types within this benchmark.

6.1.2 Problem-Type Analysis on OlympiadBench. To understand the source of DeepSeek-R1’s significant advantage on OlympiadBench, we further analyze model accuracy by problem type: Algebra, Combinatorics, Geometry, and Number Theory. This analysis helps to pinpoint the specific areas where the models’ reasoning capabilities diverge.

As shown in Table 12, the breakdown reveals that DeepSeek-R1’s lead is comprehensive and particularly pronounced in **Algebra** and **Geometry**. In Algebra, which often requires rigorous, step-by-step symbolic manipulation, DeepSeek-R1 outperforms GPT-o1 by more than 10 percentage points. Even more strikingly, it establishes a lead of nearly 24 percentage points in Geometry, a domain that often requires spatial and abstract reasoning. While both models show comparable performance on **Number Theory**, DeepSeek-R1 also maintains a clear advantage in **Combinatorics**.

These performance differences across specific mathematical subjects further highlight DeepSeek-R1’s particular strengths in procedural and symbolic tasks. This finding provides a compelling reason to further investigate the hypothesis introduced earlier that the two models may indeed possess fundamentally different reasoning styles. To move beyond performance scores and explore the underlying qualitative differences, we proceed with a detailed analysis of their solution processes in the following sections.

6.2 Evaluation of Reasoning Redundancy

Beyond overall accuracy, we perform a fine-grained analysis of the multi-step reasoning capabilities of DeepSeek-R1 and GPT-o1. This involves evaluating their token usage, solution complexity, and step-level reasoning quality. As illustrated in Figure 1, we distinguish between two critical phases in problem-solving: the **reasoning process** and the **solving process**. To quantify the degree of reasoning redundancy in model outputs, we compare the token usage patterns of DeepSeek-R1 and GPT-o1 across different datasets. We introduce three metrics to capture this:

Reasoning Token Length: This metric refers to the number of tokens generated by the model during the reasoning process. It helps measure how much thought or intermediate reasoning is included in the model’s response⁴.

Solution Token Length: This measures the length of tokens in the solving process.

⁴ Because the GPT-o1 API does not expose the raw text of its reasoning process, we approximate its reasoning token length by subtracting the number of visible solution tokens from the total generated tokens.

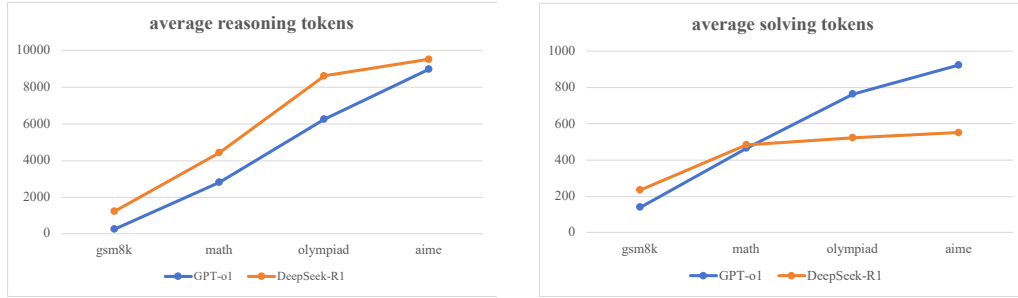


Figure 3: Average tokens used by DeepSeek-R1 and GPT-o1 (Y-axis) across four mathematical datasets (X-axis). The Left plots the Reasoning Tokens; the Right plots the Solving Tokens. (Mean across 3 runs).

Token Compression Ratio: It measures the efficiency of token usage in the transition from reasoning to the final solution. A higher compression ratio indicates that a greater proportion of the reasoning process is compressed, potentially indicating redundancy or inefficiency in the reasoning phase, as less of the reasoning is directly translated into the solution. The token compression ratio is calculated as:

$$\text{Compression Ratio} = 1 - \left(\frac{\text{Solving Tokens}}{\text{Reasoning Tokens}} \right) \quad (1)$$

6.2.1 Token Usage Analysis. As illustrated in Figure 3, DeepSeek-R1 consistently generates significantly longer **reasoning chains** than GPT-o1 across all datasets. The token gap is most pronounced on the elementary GSM8K problems, where DeepSeek-R1 uses nearly five times as many tokens for its exploratory thinking. This confirms that DeepSeek-R1 engages in a substantially more elaborate and verbose reasoning phase before committing to a final answer.

However, the token usage in the final **solution process** reveals a more complex and insightful pattern. On simpler benchmarks (GSM8K and MATH), DeepSeek-R1’s final solutions are also longer. In a notable reversal, on the most difficult competition-level problems (Olympiad and AIME), DeepSeek-R1’s solution process becomes significantly more concise than GPT-o1’s, using substantially fewer tokens to present its final derivation. This indicates that *despite DeepSeek-R1 engaging in significantly more extensive reasoning, it achieves comparable efficiency in transforming this reasoning into the final solution.*

6.2.2 Token Compression Ratio Analysis. To quantify the relationship between the exploratory thinking and the final solution, we introduce the **Token Compression Ratio**. This metric provides insight into how efficiently models distill their reasoning process into final solutions. A higher ratio indicates a more significant compression from thought to solution. As shown in Table 13, **DeepSeek-R1 exhibits a consistently higher compression ratio across all datasets.**

This high compression ratio of DeepSeek-R1, however, carries a dual interpretation that appears to be highly dependent on task difficulty. On simpler benchmarks such as GSM8K, where both models achieve near-perfect accuracy, DeepSeek-R1’s exceptionally

Table 13: Token compression ratio of GPT-o1 and DeepSeek-R1 (%)

	GPT-o1	DeepSeek-R1
GSM8K	47.76	81.07
MATH	83.49	89.10
Olympiad	87.81	93.94
AIME	89.72	94.21

high compression ratio (81.07%) indicates that its elaborate reasoning process may involve **considerable redundant exploration**. In such cases, the extended reasoning trajectory constitutes an unnecessary computational overhead for problems that could be resolved more efficiently.

In contrast, on more challenging, competition-level problems such as those from the Olympiad and AIME datasets, this same high compression ratio emerges as a signature of an effective and productive reasoning strategy. Here, **DeepSeek-R1's more elaborate exploratory phase is not mere redundancy**; instead, it appears to be a productive process that allows the model to explore a wider solution space and ultimately converge on a more efficient and direct final solution path. This higher-level efficiency in the final solution process may be a key factor contributing to its superior performance on the difficult mathematical problems.

Summary Highlights

- DeepSeek-R1 consistently employs a more verbose, exploratory reasoning phase than GPT-o1, resulting in significantly higher token compression ratios.
- This elaborate reasoning demonstrates a dual characteristic: it appears as inefficient redundancy on simpler problems, yet evolves into a productive and effective strategy on complex, competition-level tasks.
- On challenging tasks, this productive exploration allows DeepSeek-R1 to converge upon more concise final solutions, which may be a key factor in its superior performance.

6.3 Step-Level Analysis of the Solution Process

Having analyzed the high-level token usage, we now conduct a more fine-grained, step-level analysis of the Solution Process. This examination is divided into two parts: first, a quantitative analysis of the structural complexity based on the number of steps, and second, a qualitative analysis of each step's content. Both analyses are performed on a focused sample of 100 problems, comprising 50 from GSM8K and 50 from AIME.

6.3.1 Solution Steps Comparison. To assess the **complexity** of the solution process for math problems, we decompose the logical steps involved in the solving process. Here, we define complexity as the number of logical steps, operations, or inferences required to reach the final answer. The decomposition of solutions into steps was performed using our two LLM-based decomposers (QwQ-plus and DeepSeek-R1-Distill-Qwen-14B). As established in our validation study (Section 4.3.1), both of these tools were shown to be highly reliable, providing a valid basis for this analysis.

Analysis of Step Count. Table 14 summarizes the average number of solution steps generated by GPT-o1 and DeepSeek-R1, as decomposed by the two models. The results

Table 14: Average number of solution steps generated by GPT-o1 and DeepSeek-R1 on GSM8K and AIME datasets, as analyzed by two step-decomposition LLMs (QwQ-plus and DeepSeek-R1-Distill-Qwen-14B).

Step-Decomposition LLM	Dataset	Avg. Steps (GPT-o1)	Avg. Steps (DeepSeek-R1)
QwQ-plus	GSM8K	4.66	4.94
	AIME	6.86	7.26
DeepSeek-R1-Distill-Qwen-14B	GSM8K	4.76	4.78
	AIME	8.26	7.8

reveal that the perceived complexity, in terms of step count, can vary slightly depending on the decomposition tool used, but general trends emerge. On the elementary GSM8K problems, both models produce solutions with a very similar number of steps, regardless of which decomposer is used. This suggests that for simpler problems, the solution structures are relatively consistent.

However, on the more challenging AIME problems, the picture is more nuanced. The QwQ-plus decomposer identifies slightly more steps in DeepSeek-R1’s solutions compared to GPT-o1’s. Conversely, the DeepSeek-R1-Distill-Qwen-14B decomposer finds slightly fewer steps in DeepSeek-R1’s solutions. **Critically, neither decomposer shows a dramatic difference in step count between the two models on AIME.**

This finding is particularly insightful when contrasted with the performance results (Table 11), where DeepSeek-R1 demonstrated superior accuracy on AIME. It strongly suggests that **the sheer number of steps is not a reliable predictor of success on complex problems.** DeepSeek-R1 achieves higher accuracy despite having a roughly comparable number of solution steps to GPT-o1, with only minor variations observed depending on the specific decomposition method used. This reinforces the idea that DeepSeek-R1’s more elaborate exploratory phase (Section 6.2.2) might lead to a more *effective* or *robust* final solution path, even if that path isn’t always numerically shorter.

6.3.2 Step Quality Analysis. Building upon the previous analysis of the solution process, we conduct a fine-grained evaluation of individual steps for the same 100 sampled problems. This micro-level assessment employed the QwQ-plus LLM. The evaluation was guided by a detailed prompt designed around our validated rubric (Table 2), with the complete prompt provided in Appendix B. Given the judge’s established limitations (Section 4.3.2), and in accordance with our usage policy (Section 4.4), this analysis is explicitly framed as **exploratory**, capturing LLM-perceived quality rather than definitive ground truth. We evaluate each step generated by DeepSeek-R1 and GPT-o1 across four micro-level dimensions: Logical Consistency, Correctness, Step Clarity, and Redundancy & Efficiency. The average scores for each dimension are presented in Table 15. The results suggest subtle but consistent differences in the quality characteristics of the steps.

Logical Soundness. The LLM judge consistently assigned very high Logical Consistency and Correctness scores to steps from both models across both benchmarks. According to the judge’s scores, DeepSeek-R1 appeared to maintain a marginal but consistent advantage on the more challenging AIME problems in both dimensions. This suggests that the LLM perceived DeepSeek-R1’s process as slightly more robust when tackling complex derivations, although both were rated highly.

Presentation and Efficiency. The LLM’s scores indicate interesting trade-offs in clarity and efficiency. On the simpler GSM8K benchmark, both models received identical

Table 15: Performance comparison of **GPT-o1** and **DeepSeek-R1** on the GSM8K and AIME datasets across four evaluation dimensions.

Dimensions	GSM8k		AIME	
	GPT-o1	DeepSeek-R1	GPT-o1	DeepSeek-R1
Logical Consistency	1.83	1.87	1.91	1.92
Correctness	0.98	0.98	0.92	0.93
Step Clarity	1.96	1.96	1.89	1.94
Redundancy & Efficiency	0.99	0.93	0.97	0.98

high scores for Step Clarity, but GPT-o1 was rated higher by the judge on Redundancy & Efficiency, indicating a perceived more direct path. However, this pattern appeared to reverse on the AIME benchmark. Here, DeepSeek-R1 received noticeably higher Step Clarity scores and also slightly surpassed GPT-o1 in the Redundancy & Efficiency ratings.

These findings suggest that, in the LLM judge’s assessment, DeepSeek-R1’s more elaborate exploratory phase (Section 6.2.2) not only leads to a final solution path that is sometimes structurally simpler (Section 6.3.1) but is also composed of steps ultimately judged as clearer and marginally more efficient on complex problems. This perceived high quality at the step level may contribute significantly to DeepSeek-R1’s superior overall accuracy on challenging benchmarks like AIME.

Summary Highlights

- Step counts are similar between models on simpler problems, but diverge slightly on complex AIME tasks, with DeepSeek-R1 maintaining a comparable number of steps while achieving higher accuracy.
- Based on the LLM judge’s exploratory assessment, DeepSeek-R1’s individual solution steps show higher logical consistency and correctness, particularly on challenging AIME problems.
- DeepSeek-R1’s steps also achieve higher clarity and efficiency scores on AIME problems, reversing the pattern observed on simpler tasks.

6.4 Macro-Level Evaluation of Solution Quality

Having analyzed the complexity and quality at the step level, we now broaden our scope to evaluate the holistic quality of the complete solution process. This section provides a macro-level assessment, first by scoring each model’s solutions individually based on our validated framework, and then through a more exploratory pairwise comparison to highlight their relative strengths and weaknesses.

6.4.1 Macro-Level Quality Assessment. We begin by assessing the overall quality of each solution generated by DeepSeek-R1 and GPT-o1 independently. This analysis leverages the QwQ-plus LLM judge, whose reliability in applying the macro-level dimensions of our rubric (Table 2) was established through calibration against human ratings (Section 4.3.3). Importantly, to compare the quality of successful reasoning paths rather than just the ability to find the correct answer, this analysis focuses exclusively on solutions that achieved the correct final answer (i.e., Answer Accuracy = 1). The judge scored these correct solutions across the five macro-level dimensions: **Answer Accuracy**, **Reasoning Step Accuracy**, **Solution Efficiency**, **Step Clarity** (Holistic), and **Mathematical Concept**

Table 16: Average Macro-Level Quality Scores for Correct Solutions Only (Assessed by QwQ-plus). Scores reflect the macro-level assessment and include the five rubric dimensions plus a separate, holistic Overall Score (0-2 scale) provided by the LLM judge.

Dimensions	GSM8k		AIME	
	GPT-o1	DeepSeek-R1	GPT-o1	DeepSeek-R1
answer accuracy	1.00	1.00	1.00	1.00
reasoning step accuracy	2.00	2.00	1.98	2.00
solution efficiency	2.00	1.94	2.00	1.97
step clarity	1.98	1.88	1.78	1.71
mathematical concept application	1.80	2.00	1.76	2.00
overall_score	1.80	1.77	1.76	1.76

Application. In addition to these rubric-based scores, the judge was asked to provide a separate, holistic Overall Score (0-2 scale) reflecting its comprehensive assessment of the solution’s quality. The prompt used for this evaluation is detailed in Appendix C.

Analysis of Holistic Scores for Correct Solutions. The average scores for these successful solutions, presented in Table 16, provide insights into the qualitative differences between the models’ effective problem-solving strategies.

DeepSeek-R1 Excels in Reasoning Rigor. Consistent with its superior step-level correctness (Section 6.3.2), DeepSeek-R1 achieves higher average scores on **Reasoning Step Accuracy**, particularly on the challenging AIME benchmark, where it attains a perfect score of 2.00. It also demonstrates a clear advantage in **Mathematical Concept Application** across both datasets, again achieving a perfect score on AIME. These results reinforce that DeepSeek-R1’s successful reasoning pathways are generally more logically sound and conceptually grounded, especially when addressing complex problems.

GPT-o1 Favored for Presentation and Efficiency. In contrast, GPT-o1 consistently receives higher ratings for **Solution Efficiency** and **Step Clarity (Holistic)** on both GSM8K and AIME. This indicates that, even when DeepSeek-R1 exhibits greater logical rigor, GPT-o1’s solutions are frequently perceived by the LLM judge as more direct and easier to follow in terms of presentation. This observation aligns with the model’s previously noted tendency toward more concise reasoning paths.

Notably, despite these differences across evaluation dimensions, the final **Overall Score** assigned by the LLM judge is highly similar between the two models. GPT-o1 holds a slight advantage on GSM8K (1.80 vs. 1.77), while both models achieve identical scores (1.76) on AIME. This convergence suggests that, when successful, both models produce high-quality solutions that balance different strengths, making a definitive judgment of overall superiority complex even for an LLM evaluator.

6.4.2 Comparative Pairwise Evaluation. To gain further qualitative insights into the models’ relative strengths and weaknesses when both successfully solve a problem, we conducted an additional, exploratory analysis using a pairwise comparison setup. This analysis focused exclusively on the subset of problems where both DeepSeek-R1 and GPT-o1 produced the correct final answer.

In this task, the QwQ-plus LLM judge was presented with both models’ correct solutions for the same problem side-by-side. It was prompted to output a structured JSON containing: (1) an analysis of the strengths and weaknesses for each solution, explicitly linked to the five macro-level criteria, and (2) a final decision on which solution

Table 17: Pairwise Comparison Win Rates for Correct Solutions (LLM Judge Preference).

Benchmark	DeepSeek-R1 win	GPT-o1 win	Tie
GSM8k	12.00%	86.00%	2.00%
AIME	25.71%	74.28%	0.00%

was qualitatively better overall (winner: Model1, Model2, or Tie), along with a detailed justification. The full prompt for this pairwise task is provided in Appendix D.

Quantitative Win/Tie Rate Analysis. We first analyze the quantitative outcome of the pairwise comparison by examining the distribution of the *winner* decisions made by the LLM judge. The results, summarized in Table 17, provide a high-level view of the judge’s preference between the two models specifically when both arrived at the correct answer.

The win rates reveal a strong and perhaps surprising preference for GPT-o1’s solutions in this head-to-head comparison, especially given the findings from the individual scoring (Section 6.4.1). While the individual holistic scores suggested DeepSeek-R1 held slight advantages in reasoning rigor and concept application (or was at least comparable overall on AIME), the LLM judge overwhelmingly favored GPT-o1 when forced to choose a single winner between two correct solutions (86% win rate on GSM8k, 74% on AIME).

This discrepancy highlights a potentially crucial difference between the two evaluation modes. The individual scoring task likely encourages strict adherence to the defined rubric, where DeepSeek-R1’s strengths in logical soundness and concept application are rewarded. In contrast, the pairwise comparison, especially when comparing two functionally correct answers, may implicitly shift the judge’s focus towards more holistic and perhaps presentation-oriented qualities. Factors such as Step Clarity and perceived Solution Efficiency (conciseness), where GPT-o1 consistently showed an edge even in the individual scores, might become more decisive in a comparative judgment. This preference aligns with the models’ known stylistic differences: DeepSeek-R1’s emphasis on thorough, potentially verbose exploration versus GPT-o1’s tendency towards clearer, more direct exposition.

Given the complexity and inherent subjectivity of pairwise judgments, especially without direct human validation for this specific task, this finding is best interpreted as an exploratory insight. It suggests that while DeepSeek-R1’s successful solutions might be more rigorous according to a strict rubric, GPT-o1’s successful solutions are often perceived as qualitatively superior in terms of presentation and conciseness by the LLM judge when a direct choice must be made.

Qualitative Analysis of Judge Feedback. Beyond the win rates, the structured textual feedback provided by the LLM judge (strengths, weaknesses, and justification) offers a richer source of qualitative insight into the models’ perceived stylistic differences when both produce correct solutions. To systematically analyze this textual data, we identified frequently occurring N-grams (phrases) within the judge’s feedback, focusing particularly on those linked to our macro-level evaluation criteria.

Characterizing Model Strengths and Weaknesses. As shown in Table 18, the N-gram analysis reveals distinct profiles for each model, highlighting the trade-offs perceived by the LLM judge even among correct solutions:

Table 18: Top N-grams from LLM Judge Feedback (Pairwise Comparison of Correct Solutions). Shows the most frequent phrases cited by the LLM judge when describing strengths and weaknesses for each model. **Counts in parentheses indicate frequency. Square brackets [...] show optional words to complete similar phrases. Sums like (4+3) aggregate counts of closely related N-grams.**

Model	Top Phrases Cited as Strengths (Count)	Top Phrases Cited as Weaknesses (Count)
DeepSeek-R1	- without/avoiding unnecessary complexity (11+4) - all logically sound (8) - concise without unnecessary [complexity] (6+6) - logically sound mathematically (5+5) - correctly applies addition (5) - making easy follow (5) - all logically follow/lead (5+5) - logically free errors (4+4+4)	- reduce readers unfamiliar (6+4) - may confuse readers (5+3+3) - making less transparent (4) - brief lack explicit [explanations] (4+3) - making harder follow (3) - includes redundant recalculate (2)
GPT-o1	- all logically sound [mathematically] (11+5+5) - correctly applies [addition/subtraction/all] (5+4+3+3) - logically free errors (4+4) - making [process] easy follow (3+3+3) - clearly demonstrates understanding (3) - concise without redundancy (3)	- slightly less concise [than] (10+3+3) - slightly less efficient [extra] (8+3+3) - does affect correctness (5) - explicit intermediate calculations (3) - making less transparent (2) - lacks explicit verification (2)

DeepSeek-R1 was frequently praised for its **efficiency and conciseness**, dominated by phrases like *without/avoiding unnecessary complexity* (15 mentions) and *concise without unnecessary [complexity]* (12 mentions). Its **logical soundness** emerged as another core strength, evidenced by *logically sound mathematically* (10 mentions) and *logically free errors* (12 mentions). Conversely, its primary weaknesses centered on potential reader confusion, with phrases like *reduce readers unfamiliar* (10 mentions) and *may confuse readers* (11 mentions) occurring often, alongside concerns about a lack of explicit detail or transparency (e.g., *making less transparent*, 4 mentions; *brief lack explicit [explanations]*, 7 mentions). This suggests that while DeepSeek-R1 generates direct and logically sound solutions, they risk being difficult to follow due to omitted intermediate steps.

GPT-o1, conversely, received its highest praise for **logical soundness and correctness**, driven by phrases such as *all logically sound [mathematically]* (21 mentions) and *correctly applies [...]* (15 mentions). Its strength in **clarity** was also prominent, reflected in *making [process] easy follow* (9 mentions). However, its primary weaknesses involved being **less concise or efficient**, frequently cited via *slightly less concise [than]* (16 mentions) and *slightly less efficient [extra]* (14 mentions). Secondary critiques noted unnecessary *explicit intermediate calculations* (3 mentions) and minor verification issues (e.g., *lacks explicit verification*, 2 mentions). Interestingly, the phrase *does affect correctness* (5 mentions) also appeared, likely as a qualifier indicating that these inefficiencies were minor. This profile depicts GPT-o1's correct solutions as rigorous and well-explained, but potentially more verbose or less direct than DeepSeek-R1's.

Table 19: Summary of K-Means Clustering Results (K=4) on Pairwise Justifications. *Size (N)* represents the number of justifications belonging to the cluster; *Core Keywords/Phrases* lists terms characteristic of the theme; and *Representative Examples* provide two justification examples.

Cluster	Size (N)	Core Keywords/Phrases	Representative Examples
0	21	correct model, model solution, qualitatively better, solutions correct, correct model solution	1: While both solutions are correct and efficient in reaching the answer, Model 1 (Solution A) is qualitatively better overall due to superior step clarity and thoroughness ... 2: While both solutions are accurate and correct, Model 2 (Solution B) is qualitatively better due to superior step clarity and holistic understandability...
1	24	solution efficiency, solutions accurate, superior solution efficiency, accurate model, superior solution	1: While both solutions are accurate, Model 2 (Solution B) is qualitatively superior. It excels in Solution Efficiency by avoiding ... 2: Both solutions are accurate, but Model 2 is superior in Solution Efficiency and Step Clarity due to its concise, streamlined approach (no redundant ...
2	18	correct answer, step accuracy, reasoning step accuracy, reasoning step, correct answer model	1: While both solutions reach the correct answer, Model 2 is superior overall. It excels in Reasoning Step Accuracy by explicitly addressing ... 2: Both solutions (Model 1 and Model 2) are identical in their reasoning, accuracy, and efficiency. They both correctly calculate ...
3	22	efficient model, model solution, step clarity holistic, clarity holistic, efficient model solution	1: While both solutions are accurate and efficient, Model 2 (Solution B) is qualitatively better due to superior Step Clarity (Holistic). It explicitly ... 2: While both solutions are accurate and efficient, Model 2 (Solution B) outperforms in Step Clarity (Holistic) by explicitly deriving overtime hours ...

This qualitative analysis provides crucial context for the win rates observed in Section 6.4.2. Although DeepSeek-R1 achieved slightly higher individual scores in rigor-focused dimensions (Section 6.4.1) and was recognized for its concise paths, the LLM judge **overwhelmingly preferred GPT-o1** in head-to-head comparisons. The N-gram data indicates that this preference stems from the judge heavily prioritizing GPT-o1’s **clarity, explicit explanations, and ease of following**, even when acknowledging its potential lack of conciseness. Conversely, DeepSeek-R1 was likely penalized for its lack of transparency and potential to confuse readers, despite its perceived efficiency and logical soundness.

These findings highlight a critical behavior of the LLM judge in pairwise evaluations: **presentation quality and comprehensibility are disproportionately weighted, potentially eclipsing subtle advantages in logical rigor or path efficiency**. This exploratory insight underscores the complexity of automated evaluation and points towards potential biases favoring well-explained, albeit potentially less optimal, solutions.

Analyzing Justifications via Clustering. To gain deeper insight into the LLM judge’s decision-making process when choosing a winner, we further analyzed the free-text justification provided for each pairwise comparison. We employed TF-IDF vectorization (using bigrams and trigrams) followed by K-Means clustering (K=4) to identify the dominant themes or categories of reasons cited by the judge when explaining its preference. The resulting clusters, along with their core differentiating phrases and representative examples, are presented in Table 19.

Based on the core phrases and examples, we interpret the four distinct themes driving the judge's decisions as follows:

Cluster 0: General Qualitative Preference. This cluster (21 entries) contains more general statements of overall superiority ("qualitatively better"). While less specific, the examples often link this preference back to clarity (as in the representative example) or efficiency, suggesting overlap with Clusters 3 and 1.

Cluster 1: Emphasis on Efficiency. This large cluster (24 entries) centers justifications around Solution Efficiency, highlighting aspects like fewer steps or a streamlined approach, frequently coupled with accuracy.

Cluster 2: Focus on Reasoning Steps/Accuracy. This cluster (18 entries) uniquely focuses on the quality within the reasoning process, such as Reasoning Step Accuracy, even when both final answers are correct.

Cluster 3: Emphasis on Clarity. With 22 entries, this cluster clearly prioritizes Step Clarity (Holistic) as the deciding factor, often explicitly mentioned alongside accuracy and efficiency.

Mapping these justification themes to the winning model (Table 17, where GPT-o1 had a high win rate) reveals a clear pattern: the judge's decisions were primarily driven by Superior Clarity (Cluster 3) and Superior Efficiency (Cluster 1). These align directly with the perceived strengths of GPT-o1 identified in the N-gram analysis (clarity, ease of following) and its favorable comparison regarding efficiency when correctness was maintained. While Cluster 2 (Reasoning Steps) reflects DeepSeek-R1's strengths in rigor, it represents a smaller portion of the justifications, suggesting this factor was less decisive in the pairwise choice between two correct solutions. Cluster 0 serves as a general bucket confirming overall preference often tied back to clarity or efficiency.

While acknowledging the exploratory nature of this analysis, the combination of N-gram frequencies and the K=4 justification clusters paints a consistent and clearer picture. Despite DeepSeek-R1 scoring slightly higher on some individual rigor-focused metrics (Section 6.4.1) and being cited for path efficiency in N-grams, the LLM judge, when forced to choose between two correct solutions, **overwhelmingly preferred GPT-o1**. Both the N-gram analysis and the justification clusters (particularly the prominence of **Clarity** in Cluster 3 and **Efficiency** in Cluster 1) strongly suggest this preference stems from the judge heavily prioritizing **presentation quality, perceived efficiency, and comprehensibility**. DeepSeek-R1, despite its logical soundness and potential path efficiency, was likely penalized in the comparative setting for weaknesses related to reader confusion and lack of transparency. This highlights the complexity of automated pairwise evaluation and suggests a potential bias towards solutions that are not only correct but also presented in a way deemed most easily digestible, where GPT-o1 appears to excel in the judge's view.

Summary Highlights

- When analyzing correct solutions individually, DeepSeek-R1 scores higher on reasoning rigor and concept application, while GPT-o1 scores higher on presentation clarity and efficiency.
- In direct pairwise comparisons between correct solutions, the LLM judge overwhelmingly prefers GPT-o1, highlighting a potential evaluation bias.
- Qualitative analysis suggests this pairwise preference is driven by the judge heavily weighting clarity, explicit explanations, and presentation quality, where GPT-o1 excels.

7. Analysis of Cross-Domain Generalization

Having established the performance differences between reasoning and non-reasoning models (Section 5) and analyzed the underlying process variations (Section 6), this section investigates the crucial question of generalization (RQ4). We address this by examining model performance across two distinct categories: specialized scientific domains and general-purpose reasoning tasks.

7.1 Complex Science Evaluation

In this section, we conduct a thorough evaluation of both reasoning-focused and non-reasoning LLMs across a variety of scientific disciplines, including mathematics, physics, chemistry, astronomy, geography, and biology. This analysis addresses the *scientific domain aspect* of RQ4, investigating whether proficiency in mathematics serves as a foundational capability for broader scientific problem-solving.

7.1.1 Experimental Setup. We construct a scientific domain benchmark covering 30 subfields across six disciplines: mathematics, physics, chemistry, astronomy, geography, and biology. From each sub-discipline, we sample 50 mid-to-high difficulty multiple-choice questions, totaling 1,500 questions. All samples are drawn from GPQA (Rein et al. 2024) and SuperGPQA (M-A-P Team 2025). The disciplines and their corresponding sub-disciplines are as follows:

- **Mathematics:** Mathematical Analysis, Geometry and Topology, Number Theory, Probability and Statistics, Combinatorics.
- **Physics:** Electrodynamics, Thermodynamics and Statistical Physics, Acoustics, Particle Physics and Nuclear Physics, Mechanics.
- **Chemistry:** Physical Chemistry, Analytical Chemistry, Electrochemistry, Organic Chemistry, Chemical Engineering.
- **Astronomy:** Cosmology, Solar System Sciences, Astrophysics, Stellar and Interstellar Evolution, Astronomy Observations and Techniques.
- **Geography:** Mineralogy and Petrology, Structural Geology, Atmospheric Physics and Environment, Meteorology, Physical Oceanography.
- **Biology:** Genetics, Ecology, Botany, Physiology, Microbiology.

This structured sampling allows us to investigate whether explicit reasoning capabilities contribute to improved model performance in specialized scientific contexts.

7.1.2 Performance on Scientific Benchmark. To understand how reasoning capabilities impact model generalization in specialized scientific contexts, we compare the performance of reasoning and non-reasoning models across 30 sub-disciplines spanning six major scientific domains. A summary of the results is provided in Table 20.

Overall Trends. From the results, **DeepSeek-R1** demonstrates the most consistent and superior performance across a wide range of domains, particularly in mathematically demanding fields like **mathematics**, **physics**, and **chemistry**. **GPT-o1** and **QwQ-32B** also perform competitively but lag slightly behind in technical sub-domains. Among the non-reasoning models, **DeepSeek-V3** performs relatively well across several fields and outperforms reasoning models in a few applied sub-disciplines, such as **acoustics**, **physical chemistry**, and **meteorology**.

Table 20: Accuracy comparison between reasoning and non-reasoning models across 30 scientific sub-disciplines. Each sub-discipline includes 50 questions.

Sub-discipline	Reasoning Models					Non-Reasoning Models			
	DeepSeek-R1	GPT-o1	QwQ-32B	R1-Distill -Qwen -Base-32B	R1-Distill -LLaMA -70B-Instruct	DeepSeek-v3	GPT-4o	Qwen2.5 -72B -Instruct	LLaMA-3.3 -70B -Instruct
Mathematics									
Mathematical Analysis	0.875	0.700	0.750	0.450	0.750	0.725	0.250	0.400	0.225
Geometry and Topology	0.850	0.650	0.775	0.450	0.775	0.600	0.325	0.350	0.250
Number Theory	0.900	0.650	0.750	0.450	0.700	0.600	0.325	0.450	0.375
Probability and Statistics	0.825	0.550	0.675	0.525	0.500	0.650	0.475	0.525	0.475
Combinatorics	0.750	0.625	0.600	0.425	0.525	0.600	0.175	0.375	0.300
Physics									
Electrodynamics	0.900	0.850	0.775	0.525	0.675	0.850	0.625	0.400	0.475
Thermodynamics and Statistical Physics	0.775	0.750	0.650	0.375	0.550	0.725	0.525	0.500	0.475
Acoustics	0.550	0.500	0.325	0.200	0.375	0.600	0.325	0.250	0.225
Particle Physics and Nuclear Physics	0.650	0.575	0.525	0.250	0.500	0.650	0.450	0.300	0.400
Mechanics	0.550	0.475	0.450	0.350	0.475	0.575	0.325	0.325	0.325
Chemistry									
Physical Chemistry	0.800	0.700	0.650	0.300	0.525	0.825	0.500	0.425	0.375
Analytical Chemistry	0.725	0.650	0.675	0.500	0.575	0.800	0.450	0.350	0.500
Electrochemistry	0.550	0.575	0.375	0.300	0.400	0.575	0.275	0.275	0.250
Organic Chemistry	0.425	0.450	0.250	0.325	0.225	0.400	0.300	0.250	0.275
Chemical Engineering	0.800	0.625	0.375	0.275	0.450	0.700	0.350	0.375	0.225
Astronomy									
Cosmology	0.725	0.550	0.525	0.450	0.575	0.650	0.450	0.475	0.375
Solar System Sciences	0.725	0.600	0.525	0.400	0.500	0.625	0.300	0.475	0.325
Astrophysics	0.525	0.625	0.450	0.400	0.425	0.625	0.450	0.275	0.400
Stellar and Interstellar Evolution	0.600	0.650	0.450	0.325	0.325	0.600	0.375	0.450	0.250
Astronomy Observations and Techniques	0.725	0.650	0.550	0.425	0.500	0.700	0.575	0.500	0.425
Geography									
Mineralogy and Petrology	0.550	0.600	0.475	0.350	0.325	0.675	0.325	0.500	0.400
Structural Geology	0.575	0.550	0.475	0.350	0.450	0.725	0.400	0.525	0.400
Atmospheric Physics and Environment	0.500	0.550	0.450	0.375	0.450	0.550	0.475	0.375	0.300
Meteorology	0.800	0.625	0.550	0.575	0.525	0.825	0.450	0.375	0.275
Physical Oceanography	0.625	0.600	0.425	0.325	0.375	0.700	0.375	0.450	0.375
Biology									
Genetics	0.675	0.550	0.450	0.300	0.525	0.650	0.350	0.425	0.325
Ecology	0.475	0.375	0.425	0.350	0.400	0.550	0.325	0.250	0.400
Botany	0.675	0.675	0.450	0.400	0.500	0.650	0.375	0.425	0.500
Physiology	0.575	0.675	0.575	0.400	0.425	0.525	0.475	0.375	0.325
Microbiology	0.475	0.525	0.475	0.450	0.375	0.625	0.475	0.475	0.350

DeepSeek-R1 Performance Across Disciplines. Among all evaluated models, DeepSeek-R1 consistently ranks highest across most disciplines, particularly those that demand symbolic reasoning and multistep deduction. It achieves the highest average accuracy

in mathematics (0.84), followed by physics (0.685), and chemistry and astronomy (both at 0.66). Notably, subfields such as Number Theory and Electrodynamics see DeepSeek-R1 reach an accuracy of 90%, closely followed by Mathematical Analysis (87.5%) and Geometry and Topology (85%). This strong performance highlights the model's proficiency in structured reasoning tasks that involve formal deduction, equation manipulation, and logical multi-step reasoning. By contrast, in subfields such as Organic Chemistry, Ecology, and Microbiology, its accuracy falls below 50%, indicating that DeepSeek-R1's reasoning-centric architecture may be less effective in domains where success depends more on factual knowledge, domain-specific heuristics, or empirical data memorization.

DeepSeek-R1 vs. GPT-o1. In direct comparison, DeepSeek-R1 outperforms GPT-o1 in most formal science categories, particularly in mathematics and physics. This indicates a significant advantage in symbolic reasoning and structured problem-solving tasks. In domains such as chemistry, astronomy, geography, and biology, the two models perform comparably, with slight advantages shifting across sub-disciplines depending on task nature. For instance, GPT-o1 shows relatively stronger performance in organic chemistry and physiology, while R1 dominates in chemical engineering and genetics.

Performance of Non-Reasoning Models. Among non-reasoning models, DeepSeek-V3 stands out with surprisingly strong performance, significantly outperforming GPT-4o in nearly all disciplines. Notably, it outperforms reasoning models in a number of applied or observational science fields, such as acoustics, physical chemistry, analytical chemistry, and meteorology, often even surpassing some reasoning models in those domains. These results suggest that MoE-based large-scale architectures like DeepSeek-V3, can develop effective domain generalization abilities without explicit reasoning training, likely due to their extensive pretraining on diverse factual corpora.

The Impact of Distillation. A particularly noteworthy finding is that reasoning-oriented distillation not only enhances model performance but also enables smaller models to match or even surpass significantly larger non-reasoning baselines. The distilled model R1-Distill-LLaMA-70B-Instruct consistently outperforms the original LLaMA-3.3-70B-Instruct across all six scientific disciplines, despite sharing similar architecture and parameter scale. This indicates that reasoning supervision can yield substantial improvements, even when model capacity remains constant. Similarly, the distilled model R1-Distill-Qwen-Base-32B exhibits competitive performance against the larger Qwen2.5-72B-Instruct. These comparisons demonstrate that distilled reasoning models can achieve superior generalization in specialized scientific domains compared to their non-reasoning counterparts, emphasizing the value and efficiency of reasoning-focused fine-tuning pipelines (partially addressing RQ3).

7.2 Generalization to General-Purpose Reasoning Domains

To further assess the breadth of generalization, we extend our analysis to three key general-purpose reasoning domains: code generation, temporal reasoning, and knowledge-based commonsense reasoning. This addresses the *general-purpose aspect* of RQ4. We evaluate the performance of our primary reasoning models (DeepSeek-R1, GPT-o1) against strong non-reasoning counterparts (DeepSeek-V3, GPT-4o) using the same sampled datasets as in the Qwen3 controlled experiment: HumanEval (Chen et al. 2021)(using EvalPlus (Liu et al. 2023) for base and extended tests), (Chu et al. 2024) (15

Table 21: Performance on General-Purpose Reasoning Benchmarks (Mean $\pm$ Std Dev across 3 runs)

Benchmark	Reasoning Models		Non-Reasoning Models	
	DeepSeek-R1	GPT-o1	DeepSeek-V3	GPT-4o
Code (base)	97.33 $\pm$ 0.94	98.00 $\pm$ 0.00	99.33 $\pm$ 0.94	96.67 $\pm$ 0.94
Code (extended)	94.67 $\pm$ 0.94	92.00 $\pm$ 0.00	94.00 $\pm$ 0.00	90.67 $\pm$ 0.94
Temporal Reasoning	57.04 $\pm$ 0.12	47.54 $\pm$ 0.68	60.00 $\pm$ 0.54	53.42 $\pm$ 0.42
Knowledge Reasoning	84.67 $\pm$ 1.89	83.33 $\pm$ 0.94	67.33 $\pm$ 2.49	85.33 $\pm$ 2.49

subtasks, excluding generative ones), and StrategyQA (Geva et al. 2021). All results reported are the average across three runs.

The findings across these diverse domains, shown in Table 21, reveal a more complex and competitive landscape regarding the benefits of reasoning alignment when compared against strong non-reasoning models. The distinct advantages observed for explicit reasoning in the controlled Qwen3 experiment (Section 5.2) do not consistently manifest with the same magnitude in this broader comparison.

Code Generation (HumanEval): All top models perform extremely well. Surprisingly, the non-reasoning **DeepSeek-V3 achieves the highest score** on the base HumanEval tests (Code (base)), surpassing both reasoning models. On the more challenging EvalPlus tests (Code (extended)), DeepSeek-R1 takes a slight lead, closely followed by DeepSeek-V3, while GPT-o1 and GPT-4o lag slightly behind. While reasoning models remain strong, particularly DeepSeek-R1 on harder cases, the overall advantage is minimal, and DeepSeek-V3 proves highly competitive. This suggests that large-scale training, potentially combined with MoE architectures, is highly effective for code generation, yielding benefits comparable or superior to those from explicit reasoning alignment in this domain.

Temporal Reasoning (TimeBench): DeepSeek-R1 shows a clear and significant advantage on this task, outperforming both GPT-o1 and GPT-4o by a notable margin. This finding aligns with the Qwen3 experiment and supports the hypothesis that explicit, structured reasoning is highly beneficial for tasks involving temporal logic and event sequences. Interestingly, the non-reasoning DeepSeek-V3 significantly outperforms all other models and the non-reasoning GPT-4o performs considerably better than the reasoning-focused GPT-o1, indicating that factors beyond explicit reasoning alignment (perhaps scale or specific training data) also play a crucial role in this domain. GPT-o1’s relatively weaker performance here is a noteworthy observation.

Knowledge Reasoning (StrategyQA): Consistent with findings in Section 5, DeepSeek-V3 performs poorly on this knowledge-intensive benchmark. Here, the non-reasoning GPT-4o achieves the highest score, slightly surpassing both DeepSeek-R1 and GPT-o1. This strongly reinforces the conclusion that success on tasks heavily reliant on accessing and synthesizing broad factual knowledge depends primarily on the model’s underlying knowledge base and alignment training, where explicit reasoning provides little to no advantage, and where models like DeepSeek-V3 may be lacking.

These results provide a crucial qualification to the findings from the Qwen3 experiment. While activating explicit reasoning within the same model (Qwen3) yields significant gains on structured tasks, the advantage of *specifically reasoning-aligned models* (like DeepSeek-R1) over *strong generalist models* is **much more nuanced and domain-specific**, especially when considering highly capable models like DeepSeek-V3 and

GPT-4o. DeepSeek-R1 maintains a strong edge in complex mathematical reasoning and outperforms GPT-o1 and GPT-4o in temporal reasoning, but the advantage is minimal in base code generation and reversed in knowledge reasoning. Furthermore, DeepSeek-V3's exceptional performance, even leading in temporal reasoning over DeepSeek-R1, demonstrates that architectural choices (like MoE) and large-scale training can sometimes match or exceed the benefits of explicit reasoning alignment even in structured domains. This suggests that while reasoning alignment is a valuable technique, particularly evident in DeepSeek-R1's specialized strengths, factors such as massive scale, comprehensive training data, and potentially specific architectural designs are equally critical in determining top-tier performance across a broad range of general-purpose reasoning tasks.

Summary Highlights

- Enhanced mathematical reasoning capabilities demonstrate strong generalization to other domains requiring structured logic, including formalized sciences (e.g., physics, chemistry) and general tasks (e.g., code generation, temporal reasoning).
- Conversely, this generalization advantage diminishes markedly in knowledge-intensive domains (e.g., biology, commonsense QA), where large-scale, non-reasoning models remain highly competitive.
- Broad domain generalization depends on the intricate interplay among reasoning alignment, model scale, architectural design, and knowledge integration, rather than being solely determined by reasoning alignment alone.

8. Discussion and Future Direction

This paper presented a comprehensive empirical study comparing reasoning-focused and non-reasoning LLMs. Our investigation utilized benchmark aggregation, controlled experiments, and in-depth process analysis via a validated evaluation framework. Based on our findings, we discuss key insights and propose future research directions.

8.1 Summary of Key Findings

Reasoning Advantage is Context-Dependent. Explicit reasoning capabilities provide substantial but context-dependent benefits (RQ1). Our controlled experiment (Section 5.2) confirmed significant performance improvements on structured tasks (mathematics, coding, temporal reasoning) when activating Qwen3's thinking mode. However, this advantage diminished or disappeared on less structured, knowledge-intensive tasks (StrategyQA), where models like GPT-4o exhibited leading performance (Section 7.2). Scale-dependency and potential interference effects further indicate that explicit reasoning is not universally optimal, suggesting a need for adaptive reasoning strategies.

Divergent Reasoning Styles and Trade-offs. DeepSeek-R1 and GPT-o1 exhibit distinct reasoning styles (RQ2, Section 6). DeepSeek-R1 employs a verbose exploratory phase. Our exploratory micro-level analysis (Section 6.3.2) perceived this process as leading to concise, logically robust final solutions on complex problems, correlating with its superior performance on challenging benchmarks (AIME, Olympiad). GPT-o1 often produces solutions perceived as clearer or more efficiently presented. This divergence suggests a

potential trade-off associated with training methodology: DeepSeek-R1’s result-based reward may foster robustness, while GPT-o1’s inferred process-based reward might prioritize interpretability.

Reasoning Evaluation: Accuracy Limits and Judge Biases. Evaluating reasoning requires moving beyond final accuracy. Our validated multi-level rubric (Section 4) enables reliable holistic (macro-level) assessment and exploratory fine-grained (micro-level) analysis. However, our exploratory pairwise evaluation (Section 6.4.2) revealed potential LLM judge biases. When comparing correct solutions, the judge strongly preferred GPT-o1, likely weighting presentation clarity and perceived efficiency (qualities in which GPT-o1 often excelled) over subtle advantages in logical rigor. This critically highlights that automated evaluations, especially comparative ones, may harbor implicit biases needing validation against human judgment.

Generalization Limits: Interplay of Reasoning, Scale, and Knowledge. Reasoning skills honed on mathematics generalize effectively to other structured domains (formal sciences, coding, temporal reasoning), but less so to knowledge-intensive areas (biology, commonsense QA) (RQ4, Section 7). Notably, the strong performance of the non-reasoning DeepSeek-V3 in several structured domains (base code, temporal reasoning) challenges the notion that reasoning alignment is the sole path to high performance. Generalization appears to depend on a complex interplay between reasoning skills, model scale, architecture (e.g., MoE), and knowledge grounding.

Distillation Enhances Base Model Capabilities. Reasoning capabilities can be effectively transferred via distillation (RQ3, Section 5.1 and Section 7.1). Distilled variants of DeepSeek-R1 significantly outperformed their non-reasoning base models and remained competitive against larger generalist models on reasoning tasks. This validates distillation as an effective method for enhancing smaller models with advanced reasoning abilities, facilitating efficient deployment.

Limitations. While comprehensive, our study is subject to certain limitations. Data contamination remains a potential, though mitigated, confounder (Section 3.3). The benchmark suite, though diverse, is not exhaustive. The in-depth process analysis focused primarily on mathematics. Critically, only the rubric-based LLM judge underwent validation; the pairwise judge requires further calibration against human preferences. The set of models analyzed, while representative, is finite.

8.2 Future Directions

Our findings suggest several promising and specific avenues for future research, directly motivated by the complexities observed in this study:

- **Developing Adaptive Reasoning Mechanisms:** The observation that explicit reasoning is highly beneficial for some tasks (complex math) but potentially detrimental for others (e.g., highly optimized code generation in large models) strongly motivates research into adaptive reasoning. Future work should explore methods enabling models to dynamically determine when to employ a detailed chain-of-thought versus a more direct response pathway, perhaps based on perceived task complexity or confidence levels.

- **Advancing Reasoning Evaluation Beyond Simple Metrics:** Our work highlights critical challenges in evaluation. While our validated rubric improves reliability, the discrepancy between rubric scores and pairwise preferences underscores the need for methods that better capture nuanced trade-offs (e.g., rigor vs. clarity). Future research should focus on: (a) developing more sophisticated, validated process-level metrics that perhaps incorporate measures of exploration efficiency or path optimality, and (b) rigorously calibrating automated pairwise judges against systematic human preference studies to understand and mitigate potential biases towards factors like presentation style.
- **Disentangling Reasoning Alignment from Scale, Architecture, and Data:** The surprising strength of the non-reasoning DeepSeek-V3, particularly its leading performance in domains like temporal reasoning and base code generation, demands further investigation. Future work is needed to carefully disentangle the relative contributions of explicit reasoning alignment versus other factors like massive model scale, specific architectural choices (e.g., Mixture-of-Experts), and potentially curated pre-training data mixtures in achieving high performance on complex, structured tasks. Understanding this interplay is crucial for guiding the development of next-generation models.
- **Balancing Robustness and Interpretability in Reasoning Generation:** Our comparison between DeepSeek-R1 (more robust process) and GPT-o1 (clearer presentation) suggests different training paradigms optimize for different qualities. Future research could explore hybrid training methodologies that explicitly aim to cultivate models capable of generating reasoning pathways that are simultaneously logically sound, efficient, and easily interpretable by humans. This might involve combining result-based and process-based rewards, or developing novel fine-tuning techniques focused on both correctness and clarity.
- **Investigating Process-Level Effects of Distillation:** While our study confirmed performance retention in distilled models, future work could conduct a detailed process-level analysis comparing the chain length, consistency, and error patterns of distilled variants against the parent model to better understand the trade-offs during compression.
- **Investigating Distillation Effects on Reasoning Processes:** Our current work confirmed that distilled models preserve reasoning capabilities effectively in terms of final accuracy. However, our in-depth process analysis was focused on the parent models. Future work should bridge this gap by applying a process-level analysis to compare the distilled variants directly against the parent model. This would provide crucial insights into the qualitative trade-offs of compression and clarify which specific aspects of the reasoning process (e.g., robustness vs. conciseness) are preserved or lost during distillation.

9. Conclusion

This paper presents an empirical study on the reasoning and generalization capabilities of LLMs, centered on a comparison between reasoning-focused models (like DeepSeek-R1, GPT-o1) and their non-reasoning counterparts (like DeepSeek-V3, GPT-4o). Our key findings are as follows:

(1) **The advantage of explicit reasoning is highly context-dependent.** Our findings across multiple analyses, including benchmark synthesis, controlled experiments with Qwen3, and broad generalization tests, consistently show that reasoning-focused models demonstrate a decisive advantage in structured, logic-intensive domains (e.g., mathemat-

ics, physics, temporal reasoning). The Qwen3 experiment specifically provided strong causal evidence for this link. However, this advantage diminishes in knowledge-intensive or observational domains (e.g., biology, geography), where large-scale non-reasoning models remain highly competitive. This underscores that top-tier performance arises from a complex interplay of factors, not solely from reasoning alignment.

(2) Methodologically, our validation established the nuanced reliability of the LLM-as-a-Judge. We demonstrated a clear distinction based on evaluation granularity: while the LLM judge exhibits lower agreement on micro-level quality evaluation, it proves highly reliable for assessing macro-level solution quality. This finding underpins our *macro-reliable, micro-exploratory* analysis policy.

(3) Leading reasoning models exhibit distinct reasoning styles. Objectively, DeepSeek-R1 employs a significantly more verbose exploratory phase. However, it often achieves comparable structural complexity (step count) and sometimes greater conciseness (final solution token count) on difficult problems compared to GPT-o1. Our LLM judge assessments, interpreted through our validation policy, add further nuance. GPT-o1's solutions were reliably perceived as clearer at the macro-level. Conversely, our exploratory micro-level analysis suggested DeepSeek-R1's steps were perceived as more logically coherent. These process differences align with DeepSeek-R1's superior accuracy on challenging benchmarks.

(4) Reasoning capabilities can be effectively preserved via distillation. Distilled versions of DeepSeek-R1 successfully preserve much of the reasoning competency of their teacher model. While some degradation occurs, these distilled reasoning models consistently outperform their non-reasoning counterparts of similar or even larger scale. This confirms the feasibility of reasoning-preserving compression for efficiency-constrained applications.

These findings provide novel empirical evidence on the precise role and limitations of explicit reasoning in LLMs. Critically, they also offer methodological guidance for how to reliably validate and deploy automated evaluation frameworks in future research.

Appendix A: Solution Decomposition

Prompt for dividing solution process into separate steps

Please analyze the following solution process for a math problem and break it down into individual steps. Each step should be numbered and only include the core actions or reasoning, without diverging into unrelated details or additional explanations. Format each step as a JSON object with the keys "step_number" and "action", where "step_number" is the step number (e.g., 1, 2, 3, etc.) and "action" is the core action or reasoning in that step.

Output format:

Return the result as a JSON array, where each element is a JSON object representing a single step. Each JSON object should have two keys:

- "step_number": an integer indicating the step number (e.g., 1, 2, 3, etc.)
- "action": a string that describes the core action or reasoning of that step.

Solution process:

{solution}

Appendix B: Step-by-Step Evaluation

We provided QwQ-plus with the following evaluation prompt for the decomposed solution steps of each problem:

Prompt for step-by-step evaluation

You are an expert in analyzing mathematical reasoning, acting as a meticulous step-by-step evaluator. Your task is to evaluate each step of a given solution process based on a strict set of criteria.

For each step provided, you must evaluate the following four dimensions:

1. Logical Consistency (Score: 2, 1, or 0)

Definition: This dimension assesses if the step's deduction strictly and logically follows from the information presented in preceding steps or the initial problem statement.

Scoring Criteria:

- Score 2 (Highly Consistent): The step's deduction strictly and explicitly follows from the preceding steps or given premises
- Score 1 (Basically Consistent): The step is logically sound but requires inferring a small, obvious connection.
- Score 0 (Inconsistent): The step represents a logical leap, a contradiction, or is based on unstated assumptions.

2. Correctness (Score: 1 or 0)

Definition: This dimension assesses if the mathematical operations, formula applications, or logical deductions within this step are entirely correct.

Scoring Criteria:

- Score 1 (Correct): The mathematical operation, formula application, or logical deduction is entirely correct.
- Score 0 (Incorrect): The step contains any mathematical or logical error.

3. Step Clarity (Score: 2, 1, or 0)

Definition: This dimension assesses if the language and notation used in the step are unambiguous, well-structured, and easy for a qualified reader to understand.

Scoring Criteria:

- Score 2 (Highly Clear): The language and notation are unambiguous, professional, and easy to understand.
- Score 1 (Basically Clear): The step is understandable, but its phrasing or conciseness could be improved.
- Score 0 (Unclear): The step is confusing, vague, or uses ambiguous terminology.

4. Redundancy & Efficiency (Score: 1 or 0)

Definition: This dimension assesses if the step is both necessary for the final solution and is presented in a direct manner.

Scoring Criteria:

- Score 1 (Necessary & Efficient): The step is essential and does not take an unnecessarily convoluted path.
- Score 0 (Redundant or Inefficient): The step repeats previous information or takes an obviously inefficient detour.

Input: A JSON array of solution steps.

Output Format: Return a JSON array of objects, each containing the "step_number" and the scores for all four dimensions (e.g., "step_number": 1, "logical_consistency": 2, "correctness": 1, ...).

Appendix C: Macro-Level Quality Evaluation**Prompt for macro-level quality evaluation**

You are an expert in analyzing mathematical reasoning, acting as a meticulous solution evaluator. You are given the complete solution process for a math problem. The solution process consists of a series of steps, but you should evaluate the entire process as a whole. For the entire solution process, evaluate the following five dimensions:

1. Answer Accuracy (0-1)

Is the final answer derived from the solution process correct?

- 1 (Correct) = The final answer is completely correct and the format meets the requirements.
- 0 (Incorrect) = The final answer is wrong or the format is incorrect.

2. Reasoning Step Accuracy (0-2)

Evaluates the reliability and rigor of the entire reasoning chain.

- 2 (Good): The core method and logical chain are correct, with at most minor, non-critical flaws.
- 1 (Acceptable): The core method is generally correct, but the process has noticeable logical gaps or several minor errors that undermine its rigor.
- 0 (Poor): The solution contains at least one major logical or conceptual error that derails its validity, or is fundamentally wrong.

3. Solution Efficiency (0-2)

Judge whether the solution includes obvious and unnecessary clumsy methods, rather than whether it is the "optimal solution".

- 2 (Good): The solution path is direct and does not contain obvious inefficient behaviors. The method used is standard and appropriate for the problem, even if it might not be the most clever possible solution.
- 1 (Acceptable): The solution path can be followed, but it includes obvious inefficient behaviors (e.g., conducting large-scale brute force enumeration when available formulas can be used; taking unnecessary long detours).
- 0 (Poor): The solution path is extremely convoluted and inefficient. The method is absurdly complex for the problem, demonstrating a clear lack of strategic insight.

4. Step Clarity (0-2)

Evaluates the overall presentation, structure, and readability of the entire solution.

- 2 (Good): The presentation is well-structured and easy to follow. The reader can understand the logic effortlessly.
- 1 (Acceptable): The solution is understandable but requires significant effort from the reader to connect the steps or decipher the notation.
- 0 (Poor): The presentation is messy, disorganized, or confusing, making the logic nearly impossible to reconstruct.

5. Mathematical Concept Application (0-2)

Evaluates if the model chose the correct strategic approach (the right "map" or "set of tools") for the problem.

- 2 (Good): Correctly identifies and flawlessly applies the core mathematical concept(s). A minor error in application is acceptable if the core concept is right.
- 1 (Acceptable): Uses a concept from the correct general field, but applies the wrong specific theorem/method (conceptual confusion).
- 0 (Poor): Applies a concept from a completely inappropriate field of mathematics, or fails to apply any recognizable concept.

After evaluating each dimension, provide the reasons behind each score and then provide an overall score (average or weighted average of the dimensions) for the entire solution process.

Input: You will receive the problem description, the standard answer, the solution, and answer derived from solution.

Output Format: Output the results in JSON format with the following structure:

```
{
  "answer_accuracy": {
    "score": 1 | 0,
    "explanation": "Explanation of whether the answer is correct or not"
  },
  "reasoning_step_accuracy": {
    "score": 0 | 1 | 2,
    "explanation": "Explanation ..."
  },
  "solution_efficiency": {
    "score": 0 | 1 | 2,
    "explanation": "Explanation ..."
  },
  "step_clarity": {
    "score": 0 | 1 | 2,
    "explanation": "Explanation ..."
  },
  "mathematical_concept_application": {
    "score": 0 | 1 | 2,
    "explanation": "Explanation ..."
  },
  "overall_score": Overall_score
}
```

Appendix D: Comparative Pairwise Evaluation

The pairwise evaluation process was conducted by QwQ-plus for the mathematical problems. For each problem, both the solutions generated by DeepSeek-R1 and GPT-o1 were provided as input to the evaluator LLM. The evaluator LLM was instructed to perform a comparative analysis to identify the key strengths and weaknesses of each solution and determining which was qualitatively better overall. The prompt for pairwise evaluation is shown below:

Prompt for pairwise evaluation

You are an expert LLM judge specializing in the qualitative evaluation of scientific and mathematical problem-solving. Your goal is to provide a detailed pairwise comparison of two solutions.

You MUST base your entire evaluation on the following five macro-level criteria:

1. **Answer Accuracy:** Is the final answer derived from the solution process correct?
2. **Reasoning Step Accuracy:** Are the entire reasoning chain correct?
3. **Solution Efficiency:** Is the solution concise and direct, or is it overly complex and circuitous?

4. **Step Clarity (Holistic):** Is the solution well-structured, easy to follow, and understandable as a whole?

5. **Mathematical Concept Application:** Does the solution correctly identify and apply the relevant mathematical concepts?

Task:

You are given one problem and two competing solutions: Solution A (Model1) and Solution B (Model2). Perform the following steps:

1. Analyze Solution A, identifying its key strengths and weaknesses based on the criteria.
2. Analyze Solution B, identifying its key strengths and weaknesses based on the criteria.
3. Compare both solutions and determine which is qualitatively better overall.
4. Provide your final output in the specified JSON format.

Problem:

{problem}

Solution A (Model1):

{model1_solution}

Solution B (Model2):

{model2_solution}

Golden Answer:

{standard_answer}

Model1 Answer:

{model1_answer}

Model2 Answer:

{model2_answer}

Output Format (JSON Only):

```
{
  "solution_A_analysis": {
    "model_name": "Model1",
    "strengths": "[List key strengths, linking them to the 5 criteria. Be specific.]",
    "weaknesses": "[List key weaknesses, linking them to the 5 criteria. Be specific.]"
  },
  "solution_B_analysis": {
    "model_name": "Model2",
    "strengths": "[List key strengths, linking them to the 5 criteria. Be specific.]",
    "weaknesses": "[List key weaknesses, linking them to the 5 criteria. Be specific.]"
  },
  "decision": {
    "winner": "[Model1 | Model2 | Tie]",
    "justification": "[Provide a detailed justification for your choice, explicitly comparing the solutions against the 5 macro-level criteria.]"
  }
}
```

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant No. 62506359) and the General Office of National Language Commission Research Planning Committee (Grant No. ZDA145-18). We also appreciate the contributions of the following students for their valuable suggestions during this work: XinYing Deng, Huiqi Gan, Shubei Tang, Junhong Wu, Bingyao Xu, Donglei Yu, QiXuan Zhang and Wenwen Zhuang. We thank the anonymous reviewers for their insightful comments.

References

AI-MO. 2024. Aime 2024.

<https://huggingface.co/datasets/AI-MO/aimo-validation-aime>. Accessed: 26

Xiang et al. An Empirical Study on Reasoning and Generalization in Large Language Models

March 2025.

- Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv e-prints*, page arXiv:2303.12712.
- Chehbouni, Khaoula, Mohammed Haddou, Jackie Chi Kit Cheung, and Golnoosh Farnadi. 2025. Neither Valid nor Reliable? Investigating the Use of LLMs as Judges. *arXiv e-prints*, page arXiv:2508.18076.
- Chen, Mark, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, et al. 2021. Evaluating Large Language Models Trained on Code. *arXiv e-prints*, page arXiv:2107.03374.
- Chen, Qiguang, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025. Towards Reasoning Era: A Survey of Long Chain-of-Thought for Reasoning Large Language Models. *arXiv e-prints*, page arXiv:2503.09567.
- Christiano, Paul F, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.
- Chu, Zheng, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. 2024. TimeBench: A comprehensive evaluation of temporal reasoning abilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1204–1228, Association for Computational Linguistics, Bangkok, Thailand.
- Cloud, Google. 2025. Gemini 2.5 brings enhanced reasoning to enterprise use cases. <https://cloud.google.com/blog/products/ai-machine-learning/gemini-2-5-pro-flash-on-vertex-ai>. Accessed: 20 April 2025.
- Cobbe, Karl, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. *arXiv e-prints*, page arXiv:2110.14168.
- Dai, Damai, Chengqi Deng, Chenggang Zhao, R.x. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y.k. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. 2024. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1280–1297, Association for Computational Linguistics, Bangkok, Thailand.
- DeepSeek-AI. 2024. DeepSeek-V3 Technical Report. *arXiv e-prints*, page arXiv:2412.19437.
- DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv e-prints*, page arXiv:2501.12948.
- Deng, Chunyuan, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2024. Investigating data contamination in modern benchmarks for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8706–8719, Association for Computational Linguistics, Mexico City, Mexico.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Association for Computational Linguistics, Minneapolis, Minnesota.
- Dubois, Yann, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2024. Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators. *arXiv e-prints*, page arXiv:2404.04475.
- Gema, Aryo Pradipta, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile Van Krieken, and Pasquale Minervini. 2025. Are we done with MMLU? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5069–5096, Association for Computational Linguistics, Albuquerque, New Mexico.

- Geva, Mor, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.
- Golovneva, Olga, Moya Peng Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. ROSCOE: A suite of metrics for scoring step-by-step reasoning. In *The Eleventh International Conference on Learning Representations*.
- Grattafiori, Aaron, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Mathur, et al. 2024. The Llama 3 Herd of Models. *arXiv e-prints*, page arXiv:2407.21783.
- Guo, Daya, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y. Wu, Y. K. Li, Fuli Luo, Yingfei Xiong, and Wenfeng Liang. 2024. DeepSeek-Coder: When the Large Language Model Meets Programming – The Rise of Code Intelligence. *arXiv e-prints*, page arXiv:2401.14196.
- He, Chaoqun, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850, Association for Computational Linguistics, Bangkok, Thailand.
- He, Yancheng, Shilong Li, Jiaheng Liu, Yingshui Tan, Weixun Wang, Hui Huang, Xingyuan Bu, Hangyu Guo, Chengwei Hu, Boren Zheng, Zhuoran Lin, Xuepeng Liu, Dekai Sun, Shirong Lin, Zhicheng Zheng, Xiaoyong Zhu, Wenbo Su, and Bo Zheng. 2024. Chinese SimpleQA: A Chinese Factuality Evaluation for Large Language Models. *arXiv e-prints*, page arXiv:2411.07140.
- Hendrycks, Dan, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.
- Hendrycks, Dan, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *NeurIPS*.
- Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. *arXiv e-prints*, page arXiv:1503.02531.
- Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.
- Huang, Yuzhen, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: a multi-level multi-discipline chinese evaluation suite for foundation models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, Curran Associates Inc., Red Hook, NY, USA.
- Jahin, Afrar, Arif Hassan Zidan, Yu Bao, Shizhe Liang, Tianming Liu, and Wei Zhang. 2025. Unveiling the Mathematical Reasoning in DeepSeek Models: A Comparative Study of Large Language Models. *arXiv e-prints*, page arXiv:2503.10573.
- Jain, Naman, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code. *arXiv e-prints*, page arXiv:2403.07974.
- Jiang, Albert Q., Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv e-prints*, page arXiv:2310.06825.
- Li, Tianle, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2024. From Crowdsourced Data to High-Quality Benchmarks: Arena-Hard and BenchBuilder Pipeline. *arXiv e-prints*, page arXiv:2406.11939.
- Li, Zhong-Zhi, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhijiang Guo, Le Song, and Cheng-Lin Liu. 2025. From System 1 to System 2: A Survey of Reasoning Large Language Models. *arXiv e-prints*, page arXiv:2502.17419.
- Lightman, Hunter, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let’s verify step by step. In *The*

Xiang et al. An Empirical Study on Reasoning and Generalization in Large Language Models

- Twelfth International Conference on Learning Representations.
- Liu, Jiawei, Chunqiu Steven Xia, Yuyao Wang, and LINGMING ZHANG. 2023. Is your code generated by chatGPT really correct? rigorous evaluation of large language models for code generation. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Liu, Nelson F, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.
- M-A-P Team. 2025. SuperGPQA: Scaling LLM Evaluation across 285 Graduate Disciplines. *arXiv e-prints*, page arXiv:2502.14739.
- Min, Yingqian, Zhipeng Chen, Jinhao Jiang, Jie Chen, Jia Deng, Yiwen Hu, Yiru Tang, Jiapeng Wang, Xiaoxue Cheng, Huatong Song, Wayne Xin Zhao, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. 2024. Imitate, Explore, and Self-Improve: A Reproduction Report on Slow-thinking Reasoning Systems. *arXiv e-prints*, page arXiv:2412.09413.
- OpenAI. 2023. GPT-4 Technical Report. *arXiv e-prints*, page arXiv:2303.08774.
- OpenAI. 2024. GPT-4o System Card. *arXiv e-prints*, page arXiv:2410.21276.
- OpenAI. 2024a. Introducing simpleqa. <https://openai.com/index/introducing-simpleqa/>. Accessed: 26 March 2025.
- OpenAI. 2024b. Learning to reason with llms. <https://openai.com/index/learning-to-reason-with-llms/>. Accessed: 26 March 2025.
- Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Pradipta Gema, Aryo, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile van Krieken, and Pasquale Minervini. 2024. Are We Done with MMLU? *arXiv e-prints*, page arXiv:2406.04127.
- Prasad, Archiki, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. 2023. ReCEval: Evaluating reasoning chains via correctness and informativeness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10066–10086, Association for Computational Linguistics, Singapore.
- Qin, Yiwei, Xuefeng Li, Haoyang Zou, Yixiu Liu, Shijie Xia, Zhen Huang, Yixin Ye, Weizhe Yuan, Hector Liu, Yuanzhi Li, and Pengfei Liu. 2024. O1 Replication Journey: A Strategic Progress Report – Part 1. *arXiv e-prints*, page arXiv:2410.18982.
- Qwen. 2024. Qwen2.5 Technical Report. *arXiv e-prints*, page arXiv:2412.15115.
- Rein, David, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.
- Setlur, Amrith, Saurabh Garg, Xinyang Geng, Naman Garg, Virginia Smith, and Aviral Kumar. 2024. RL on incorrect synthetic data scales the efficiency of LLM math reasoning by eight-fold. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Shao, Zhihong, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv e-prints*, page arXiv:2402.03300.
- Shelby, Renee, Shalaleh Rismeni, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N’Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’23, page 723–741, Association for Computing Machinery, New York, NY, USA.
- Sun, Jiankai, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, Yue Wu, Wenhai Wang, Junsong Chen, Zhangyue Yin, Xiaozhe Ren, Jie Fu, Junxian He, Wu Yuan, Qi Liu, Xihui Liu, Yu Li, Hao Dong, Yu Cheng, Ming Zhang, Pheng Ann Heng, Jifeng Dai, Ping Luo, Jingdong Wang, Ji-Rong Wen, Xipeng Qiu, Yike Guo, Hui Xiong, Qun Liu, and Zhenguo Li. 2023. A Survey of Reasoning with Foundation Models. *arXiv e-prints*, page arXiv:2312.11562.
- Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Wang, Peiyi, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024a. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Association for Computational Linguistics, Bangkok, Thailand.
- Wang, Ruonan, Runxi Wang, Yunwen Shen, Chengfeng Wu, Qinglin Zhou, and Rohitash Chandra. 2025. Evaluation of LLMs for mathematical problem solving. *arXiv e-prints*, page arXiv:2506.00309.
- Wang, Yubo, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhui Chen. 2024b. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Wei, Hui, Shenghua He, Tian Xia, Fei Liu, Andy Wong, Jingyang Lin, and Mei Han. 2025. Systematic evaluation of LLM-as-a-judge in LLM alignment tasks: Explainable metrics and diverse prompt templates. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.
- Wei, Jason, 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.
- Wu, Yuning, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, and Fei Huang. 2025. WritingBench: A Comprehensive Benchmark for Generative Writing. *arXiv e-prints*, page arXiv:2503.05244.
- Xia, Shijie, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. 2025. Evaluating mathematical reasoning beyond accuracy. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI'25/IAAI'25/EAAI'25*, AAAI Press.
- Xiang, Lu, Yang Zhao, Yaping Zhang, and Chengqing Zong. 2025. A survey of large language models in discipline-specific research: Challenges, methods and opportunities. *Studies in Informatics and Control*, 34:5–24.
- Xu, Cheng, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. 2024a. Benchmark Data Contamination of Large Language Models: A Survey. *arXiv e-prints*, page arXiv:2406.04244.
- Xu, Fengli, Qianyu Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Chenyang Shao, Yuwei Yan, Qinglong Yang, Yiwen Song, Sijian Ren, Xinyuan Hu, Yu Li, Jie Feng, Chen Gao, and Yong Li. 2025. Towards Large Reasoning Models: A Survey of Reinforced Reasoning with Large Language Models. *arXiv e-prints*, page arXiv:2501.09686.
- Xu, Xiaohan, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024b. A Survey on Knowledge Distillation of Large Language Models. *arXiv e-prints*, page arXiv:2402.13116.
- Yang, An, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Lv, et al. 2025. Qwen3 Technical Report. *arXiv e-prints*, page arXiv:2505.09388.
- Yax, Nicolas, Hernán Anlló, and Stefano Palminteri. 2024. Studying and improving reasoning in humans and machines. *Communications Psychology*, 2(1):51.
- Yu, Fei, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2024. Natural language reasoning, a survey. *ACM Computing Surveys*, 56(12):1–39.
- Zeng, Zhongshen, Pengguang Chen, Shu Liu, Haiyun Jiang, and Jiaya Jia. 2025. MR-GSM8k: A meta-reasoning benchmark for large language model evaluation. In *The Thirteenth International Conference on Learning Representations*.
- Zhang, Hugh, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, William Song, Tiffany Zhao, Pranav Vishnu Raja, Charlotte Zhuang, Dylan Z Slack, Qin Lyu, Sean M. Hendryx, Russell Kaplan, Michele Lunati, and Summer Yue. 2024. A careful examination of large language model performance on grade school arithmetic. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Xiang et al. An Empirical Study on Reasoning and Generalization in Large Language Models

- Zhang, Yaping, Qixuan Zhang, Xingquan Zhang, Zhiyuan Chen, Wenwen Zhuang, Yupu Liang, Lu Xiang, Yang Zhao, Jiajun Zhang, Yu Zhou, et al. 2025. Hiscibench: A hierarchical multi-disciplinary benchmark for scientific intelligence from reading to discovery. [arXiv preprint arXiv:2512.22899](#).
- Zhao, Wayne Xin, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models. [arXiv e-prints](#), page arXiv:2303.18223.
- Zhao, Yang, Jiajun Zhang, and Chengqing Zong. 2023. Transformer: A general framework from machine translation to others. *Machine Intelligence Research*, 20(4):514–538.
- Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.
- Zhong, Tianyang, Zhengliang Liu, Yi Pan, Yutong Zhang, Yifan Zhou, Shizhe Liang, Zihao Wu, Yanjun Lyu, Peng Shu, Xiaowei Yu, Chao Cao, Jiang, et al. 2024. Evaluation of OpenAI o1: Opportunities and Challenges of AGI. [arXiv e-prints](#), page arXiv:2409.18486.
- Zhou, Denny, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.
- Zhou, Jeffrey, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-Following Evaluation for Large Language Models. [arXiv e-prints](#), page arXiv:2311.07911.