ATLAS: A High-Difficulty, Multidisciplinary Benchmark for Frontier Scientific Reasoning

ATLAS Teams[1*]

[*]Shanghai AI Laboratory

The rapid advancement of Large Language Models (LLMs) has led to performance saturation on many established benchmarks, questioning their ability to distinguish frontier models. Concurrently, existing high-difficulty benchmarks often suffer from narrow disciplinary focus, oversimplified answer formats, and vulnerability to data contamination, creating a fidelity gap with real-world scientific inquiry. To address these challenges, we introduce **ATLAS** (**A**GI-Oriented **T**estbed for **L**ogical **A**pplication in **S**cience), a large-scale, high-difficulty, and cross-disciplinary evaluation suite composed of approximately 800 original problems. Developed by domain experts (PhD-level and above), ATLAS spans seven core scientific fields: mathematics, physics, chemistry, biology, computer science, earth science, and materials science. Its key features include: (1) **High Originality and Contamination Resistance**, with all questions newly created or substantially adapted to prevent test data leakage; (2) **Cross-Disciplinary Focus**, designed to assess models' ability to integrate knowledge and reason across scientific domains; (3) **High-Fidelity Answers**, prioritizing complex, open-ended answers involving multi-step reasoning and LaTeX-formatted expressions over simple multiple-choice questions; and (4) **Rigorous Quality Control**, employing a multi-stage process of expert peer review and adversarial testing to ensure question difficulty, scientific value, and correctness. We also propose a robust evaluation paradigm using a panel of LLM judges for automated, nuanced assessment of complex answers. Preliminary results on leading models demonstrate ATLAS's effectiveness in differentiating their advanced scientific reasoning capabilities. We plan to develop ATLAS into a long-term, open, community-driven platform to provide a reliable "ruler" for progress toward Artificial General Intelligence. The project is released at: https://github.com/open-compass/ATLAS

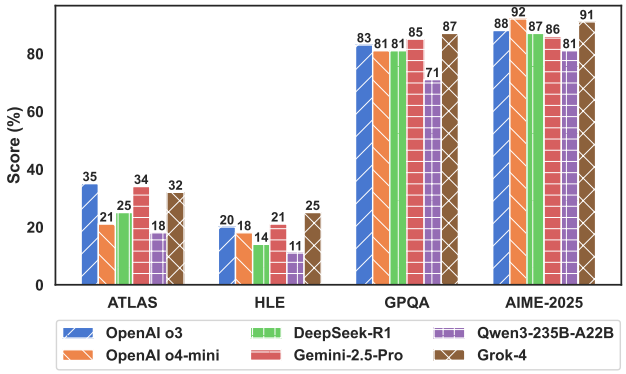

Figure 1: Reasoning LLMs performance comparison between ATLAS and other commonly used reasoning benchmarks.

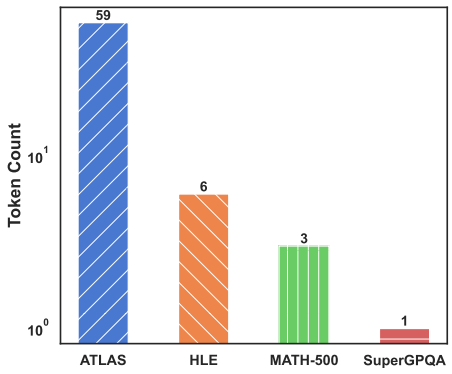

Figure 2: Average final answer token length for mainstream reasoning datasets.

[1]Detailed contributor list in Section G.

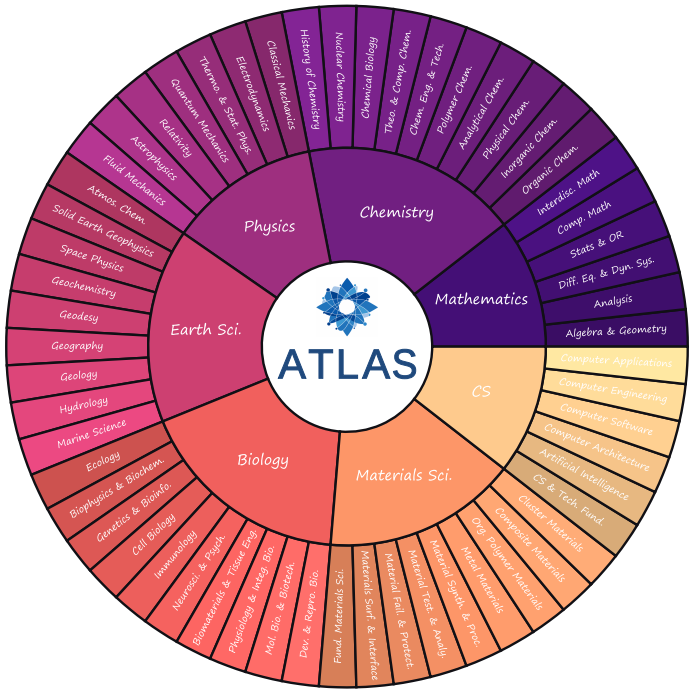

Figure 3: Overview of ATLAS, which contains 7 stem subjects and 57 corresponding sub-fields.

1. Introduction

1.1. Benchmark Saturation Phenomenon

In recent years, the advancement of Large Language Models (LLMs) has been remarkable, with their performance on various natural language processing tasks approaching or even surpassing human levels. However, this rapid progress has resulted in a significant issue: the "benchmark saturation" of standardized evaluation sets. Many benchmarks previously regarded as "gold standards", such as MMLU, are now easily surpassed by state-of-the-art models with accuracies exceeding 90% (Phan et al., 2025; Hendrycks et al., 2021a). This phenomenon reduces the effectiveness of these benchmarks in distinguishing the true capabilities of different models, particularly the subtle differences among cutting-edge models. A prominent example is the MATH dataset; upon its release in 2021, the leading model achieved a score of less than 10%. Within just three years, top models have achieved scores over 90% (Hendrycks et al., 2021b). This situation underscores the urgent need for a new generation of more challenging evaluation tools to accurately assess and propel the continuous development of AI capabilities.

1.2. Evaluation Needs for Frontier Scientific Reasoning

The next substantial breakthrough in artificial intelligence is anticipated to involve solving complex, high-value real-world problems, with scientific discovery as a central focus (Wang et al., 2023). AI for Science (AI4S) seeks to expedite the scientific research process through AI, necessitating models to have not only a robust knowledge base but also advanced, multi-step, and interdisciplinary reasoning skills (Reddy and Shojaee, 2025; Yu and Jin, 2025; Gottweis et al., 2025). To guide and assess the development of models in this strategic direction, it is essential to construct evaluation benchmarks that specifically test these capabilities. ATLAS is developed for this purpose, aiming to serve as a

"touchstone" for the AI4S domain, accurately reflecting the scientific reasoning abilities of models.

1.3. Limitations of Existing High-Difficulty Benchmarks

To tackle the challenge of benchmark saturation, the research community has developed several high-difficulty evaluation sets. While these initiatives have made significant contributions, they also exhibit limitations. Some benchmarks, despite their difficulty, are overly narrow in scope. For instance, MATH (Hendrycks et al., 2021b), MathBench (Liu et al., 2024a) and OlympiadBench (He et al., 2024) predominantly focus on mathematics or physics competition problems, hindering comprehensive evaluation of a model's integrated reasoning capabilities across diverse scientific domains. Conversely, benchmarks with broader coverage, such as Humanity's Last Exam (HLE) (Phan et al., 2025) and SuperGPQA (Team et al., 2025), while extremely challenging, are designed to assess general academic knowledge and are not specifically tailored for the deep, integrated scientific reasoning essential in the AI4S domain. Moreover, many existing benchmarks (e.g., AGIEval (Zhong et al., 2024), OlympiadBench (He et al., 2024)) derive problems from public exam or competition question banks, posing a persistent risk of data contamination. Models might score highly by having encountered similar or identical problems during training, reflecting memorization rather than authentic reasoning abilities. One of the primary objectives of ATLAS is to fundamentally resolve this issue through a rigorous original problem-setting approach.

Another limitation is that, for ease of verification, much of the existing work converts problems into multiple-choice questions and simple symbolic expressions (Hendrycks et al., 2021b; Rein et al., 2023; Team et al., 2025; Phan et al., 2025). This method has resulted in a disconnect between benchmarks and real-world questions, particularly in scientific domains (Miller and Tang, 2025). In an era of rapid LLM capability expansion, benchmarks should not be confined to easily verifiable problems. ATLAS is designed to preserve real-world problems and solutions, encompassing multiple sub-questions and complex natural and symbolic language expressions, to provide a more realistic and effective evaluation of a model's scientific capabilities. To address the evaluation bottleneck, we propose an effective, transferable and scalable LRM-as-Judge-based (Chiang et al., 2023; Zheng et al., 2023) evaluation workflow, wherein Large Reasoning Models serve as judge models. We anticipate that as model capabilities advance, the effectiveness of our workflow will be further improved. Concurrently, ATLAS is poised to significantly contribute to the development of LRM-as-Judge.

1.4. Our Contributions

This study aims to overcome the aforementioned challenges by constructing ATLAS. Its core contributions can be summarized in the following four points:

1. **ATLAS:** We release a new, highly challenging evaluation benchmark containing approximately 800 expert-created original problems. The benchmark focuses on multidisciplinary scientific reasoning, with a target difficulty set to a pass rate of less than 20% for current state-of-the-art models, to effectively measure the true capabilities of frontier models. ATLAS preserves real-world problems and solutions for realistic and effective evaluation of scientific capabilities.
2. **A Rigorous, Contamination-Resistant Construction Pipeline:** We detail an innovative, multi-stage data generation and validation process. This process (as shown in Figure 4) deeply integrates the wisdom of human experts with the adversarial testing of large models, ensuring the originality, high quality, and high difficulty of the problems from the source.
3. **A Transferable and Scalable Evaluation Workflow:** We present a streamlined and scalable evaluation workflow that utilizes LRM-as-Judge paradigm. This approach facilitates efficient and automated evaluations, allowing researchers and practitioners to assess the reasoning

capabilities of their models and conduct reinforcement learning on real-world benchmarks.

4. **A Sustainable Evaluation Platform:** The release of ATLAS is the first step in our long-term plan. Our ultimate goal is to build a community-driven collaborative platform that continuously generates and releases high-quality evaluation sets, thereby enabling the long-term, dynamic tracking of progress toward Artificial General Intelligence (AGI).

2. Related Work

The swift advancement of LLMs necessitates a concurrent evolution in evaluation benchmarks, driving them towards increased difficulty, breadth, and methodological rigor. We contextualize our research by examining three significant trends: the transition from broad-coverage to human-centric assessments, the emergence of frontier-difficulty reasoning benchmarks, and the creation of specialized STEM evaluations. Finally, we also discuss the related work of LLM-as-Judge, which is highly related to the evaluation work of ATLAS.

2.1. From Broad Coverage to Human-Centric Benchmarks

Initial comprehensive benchmarks like **MMLU** (Hendrycks et al., 2021a), with its multiple-choice questions across 47 subjects, have become "saturated" by state-of-the-art models, reducing their ability to differentiate frontier capabilities (Phan et al., 2025). In response, human-centric benchmarks emerged, drawing questions from high-stakes standardized tests to ensure quality and relevance to human cognition. For instance, **AGIEval** (Zhong et al., 2024) uses questions from exams like the SAT and Gaokao, and **C-Eval** (Huang et al., 2023) focuses on Chinese academic disciplines. While valuable, these benchmarks are constrained by the difficulty of their source material and face a significant, unavoidable risk of data contamination, as test questions are often public (Brown et al., 2020; Li et al., 2024b). Our work directly mitigates these issues through expert-authored, original problems.

2.2. The Rise of Frontier-Difficulty Reasoning Benchmarks

To address the limitations of existing tests, a new generation of benchmarks aims to create problems at the frontier of machine capabilities, emphasizing originality and resistance to search engine-based solutions. **GPQA** (Rein et al., 2023) exemplifies this with graduate-level questions whose creation process by multiple domain experts makes them demonstrably “Google-proof”: experts achieved 65% accuracy while skilled non-experts with web access only reached 34% (Bowman and Dahl, 2021). Similarly, **Humanity’s Last Exam (HLE)** (Phan et al., 2025) employs nearly 1,000 experts and uses state-of-the-art models as an adversarial filter to ensure its 2,500 questions are challenging even for top models like Gemini 2.5 Pro (21.64% accuracy). ATLAS adopts the rigorous methodological principles of GPQA and HLE—such as expert-driven creation and adversarial filtering (Bras et al., 2020; Kiela et al., 2021)—but narrows the focus from general knowledge to the specific, high-value domain of AI for Science.

A parallel research thrust has created deep, specialized benchmarks for core STEM disciplines. The **MATH** dataset (Hendrycks et al., 2021b) was a landmark, providing challenging competition math problems with step-by-step solutions that have been pivotal for research (Wang et al., 2024). To escalate the difficulty, **OlympiadBench** (He et al., 2024) incorporates problems from international Olympiads and the most difficult Gaokao questions. The extremely poor performance of models like GPT-4V (17.97%) on this benchmark highlights its immense challenge and reveals critical failure modes in SOTA models (Huang et al., 2024). ATLAS draws inspiration from the focus on deep, complex reasoning inherent in these specialized benchmarks.

2.3. Synthesis and Positioning of ATLAS

ATLAS synthesizes the strengths of these three distinct trends. We adopt the methodological rigor and originality-first principles of frontier-difficulty benchmarks like GPQA. We draw on the focus on deep, complex reasoning from specialized STEM benchmarks like OlympiadBench. Our core contribution is to apply these principles to a novel and strategically important domain: a broad but coherent suite of AI for Science subjects. In doing so, ATLAS fills a critical gap, providing a tool to measure and drive progress on the integrated reasoning skills vital for the next generation of scientific discovery (Luo et al., 2025; Zheng et al., 2025). The table below provides a comparative summary.

Table 1: High-Level Comparison of Benchmark Goals and Scope. We summarize prominent high-difficulty reasoning benchmarks, comparing their primary scientific goals and the scope of subjects they cover.

Benchmark	Primary Goal	Subject Scope
MMLU	To measure multitask knowledge acquired during pretraining via zero- and few-shot evaluation across a wide range of subjects.	Covers 57 diverse subjects across STEM, humanities, and social sciences, from elementary to professional level.
GPQA	To support scalable oversight research with "Google-proof" questions that are difficult for non-experts but verifiable by experts.	448 graduate-level multiple-choice questions in Biology, Physics, and Chemistry.
SuperGPQA	To scale GPQA-style evaluation to underrepresented and specialized "long-tail" academic disciplines.	Over 26,000 questions covering 285 graduate-level subjects, expanding well beyond traditional STEM.
OlympiadBench	To test complex, multi-step reasoning on problems requiring creative and rigorous solutions, mirroring top-tier human competition.	Olympiad-level problems in Mathematics and Physics sourced from international and national competitions (e.g., IMO, IPhO).
HLE	To measure knowledge at the frontier of human academic inquiry, addressing the performance saturation of earlier benchmarks like MMLU.	Highly challenging questions across Math, Science, Humanities, and CS, including multimodal and short-answer formats.
ATLAS (Ours)	**To measure frontier AI for Science (AI4S) reasoning, focusing on tasks central to the scientific discovery process with a rigorous "human in loop" pipeline.**	**Covers 7 core AI4S subjects and 47 subfields, with natural used questions for science research.**

2.4. Additional Discussion of LLM-as-Judge

Evaluating LLMs is now a central research focus, given their expanding deployment across applications ranging from natural-language processing to decision making. Conventional metrics often overlook the semantic and contextual subtleties of open-ended LLM outputs. Human assessment, while more reliable, is labor-intensive, costly, and hard to scale. The "LLM-as-a-Judge" paradigm has been proposed to overcome these limitations: an advanced LLM appraises the outputs of another model, yielding a scalable and economical proxy for human judgment. Foundational studies (Zheng et al., 2023; Xie et al., 2023) delineate both the promise and the limitations of this strategy; recent surveys (Chang et al., 2024; Gu et al., 2024; Li et al., 2024a) chart future directions. Continued progress will depend on mitigating bias, inconsistency, and prompt sensitivity to unlock the full potential of LLM-as-a-Judge systems.

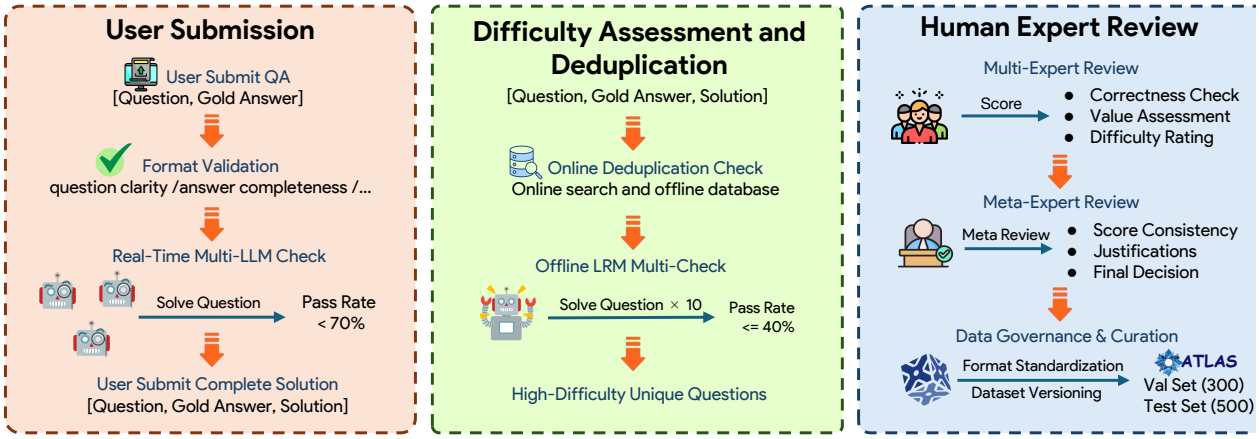

Figure 4: Overview of ATLAS construction pipeline.

3. ATLAS Construction Pipeline

In the current AI era, we recognize that the value of an evaluation benchmark depends not only on the questions it contains but, more importantly, on the methodological rigor in their creation. The methodology for constructing evaluation benchmarks has evolved from the straightforward data collection seen in MMLU to the complex, multi-stage, human-machine collaborative processes used by projects like GPQA and HLE, which are critical to ensuring their effectiveness and credibility (Rein et al., 2023; Team et al., 2025). In line with these advancements, we have designed and implemented a rigorous multi-stage process to systematically ensure the high quality, difficulty, and originality of the problems.

3.1. Question Design for Real-World Fidelity

Our question design prioritizes realism over evaluation convenience and adheres to the following principles:

- **Question Types.** We focus on short-answer and fill-in-the-blank formats. Over 50% of the questions are **compound questions** with multiple sub-parts. This structure tests a model's ability to manage complex instructions, maintain long-range context, and perform multi-step reasoning, exposing weaknesses that single questions cannot.
- **Answer Complexity.** Answers are designed to be complex entities, such as a full LaTeX equation ($\int_0^\infty e^{-x^2}dx = \frac{\sqrt{\pi}}{2}$), a list of chemical products, or a short, high-difficulty but simplified proof.
- **Bilingualism.** All questions in ATLAS are available in both English and Chinese to support the global research community, a practice shared by other frontier benchmarks like OlympiadBench.

3.2. Core Design Principles

Our construction process is based on the following four core principles:

- **Frontier Difficulty and Originality.** To combat benchmark saturation and data contamination (Phan et al., 2025; Rein et al., 2023), all problems are newly-authored or substantially re-engineered by domain experts. Questions are targeted at a graduate-level or higher difficulty, explicitly testing complex, multi-step reasoning rather than information retrieval.

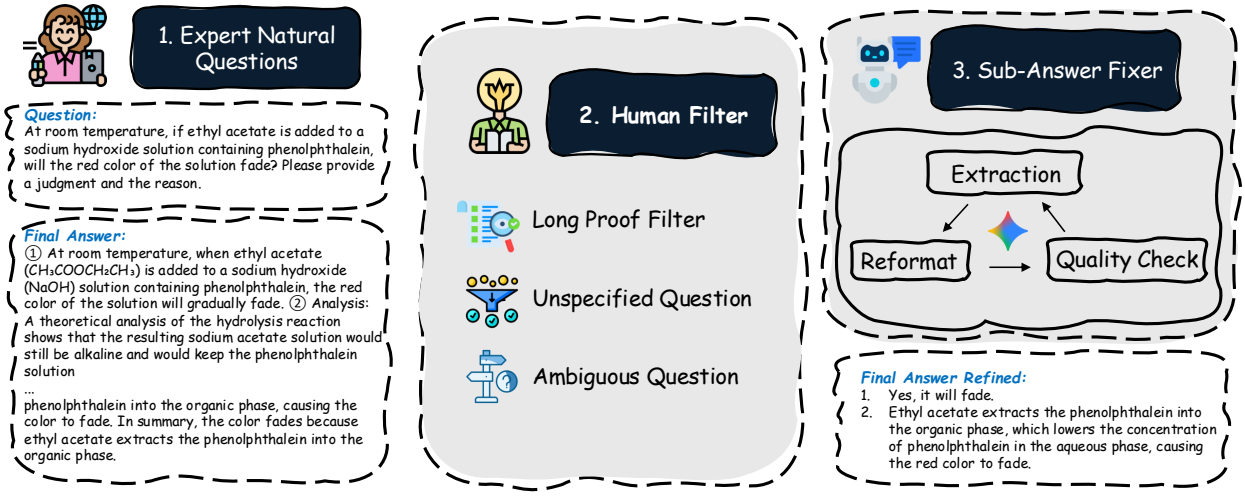

Figure 5: Overview of how ATLAS refine the final answer for natural questions.

- **Hybrid Human-AI Quality and Difficulty Calibration.** We employ a rigorous, multi-stage validation pipeline. This includes a two-tiered, anonymous expert review process for scientific accuracy and clarity, inspired by methodologies from GPQA (Rein et al., 2023). Crucially, this is augmented by an adversarial filtering stage where only problems that state-of-the-art models (e.g., DeepSeek-R1) fail to solve with high frequency (e.g., $\leq$40% accuracy) are retained, ensuring the benchmark remains at the frontier of AI capabilities.
- **Objective and Complex Answer Formulation.** To mirror real-world scientific outputs and prevent guessing, we eschew simple multiple-choice or short-string answers. Every problem features a single, objectively verifiable answer, often expressed in complex formats such as LaTeX equations, chemical formulas, or structured multi-part responses, demanding generative reasoning rather than simple recognition.

3.3. Data Generation and Quality Assurance Workflow

Figure 4 details our data construction and quality assurance workflow. This process combines the deep domain knowledge of human experts with the computational power of large models, forming a powerful "dual-filter" system to ensure that every problem admitted to the final database possesses both high quality and high difficulty. The workflow includes the following key stages:

- **Stage 1: Expert-Sourced Problem Generation and Pre-screening.** Our process begins with Ph.D.-level experts from more than 25 different institutions crafting original problems that demand multi-step, and often cross-disciplinary, reasoning. Each submission includes a canonical solution and detailed steps. These problems then undergo an automated pre-screening pipeline, which normalizes formatting and performs a similarity check against a vast offline database of existing problems to filter out duplicates and ensure novelty.
- **Stage 2: Adversarial Filtering and Iterative Refinement.** To ensure the questions in our benchmark are both novel and sufficiently challenging, we implement a rigorous pipeline for adversarial filtering and iterative refinement. This process consists of two main phases: originality verification and difficulty calibration. First, to mitigate data contamination, each submission undergoes an originality assessment. We employ a Retrieval-Augmented Generation (RAG) system to screen submissions against a comprehensive corpus of web content, academic papers, and existing benchmarks. This system first retrieves the top-K most semantically similar entries. Subsequently,

an LLM is used to evaluate these retrieved items and assign both redundancy and originality scores to the submission. Only submissions that meet a high originality threshold advance to the next stage. Following the originality check, problems are evaluated for difficulty by an ensemble of state-of-the-art LRMs. We apply a stringent adversarial criterion: a problem is accepted only if these LRMs achieve a solution accuracy of 40% or less over ten attempts. This strict standard ensures that the final problems robustly challenge current AI capabilities. Problems that do not meet this difficulty threshold are returned to the human experts. They can then choose to discard the problem or iteratively refine it to increase its complexity before resubmission, creating a closed-loop quality enhancement process.

- **Stage 3: Multi-Layered Human Validation and Final Ingestion.** Problems that pass the adversarial filtering undergo a rigorous, multi-stage manual quality inspection. Each problem is sent to three anonymous peer reviewers in the same domain for a double-blind evaluation of its correctness, clarity, and difficulty. Discrepancies in reviews are resolved by a senior meta reviewer who makes a final determination. Finally, before being admitted to the benchmark, a last check is performed against online search engines to confirm the problem has not been publicly disclosed. Only problems that clear every stage of this comprehensive validation process are accepted into the final database. Detailed information about human review stage can refer to Section C.
- **Stage 4: Final Answer Refinement and Verification.** Following the rigorous validation of the problems, as shown in Figure 5, a final stage is dedicated to refining the expert-provided answers to ensure maximum clarity, correctness, and pedagogical value. This process, illustrated in the provided figure, transforms the initial expert solutions into a canonical format suitable for the benchmark. The refinement pipeline consists of three key steps: First, a LLM agent performs **Extraction**, decomposing the initial, often verbose, answer into its fundamental components, such as the direct judgment (e.g., "the color will fade") and the scientific reasoning. Next, the extracted components undergo a semi-automated **Quality Check** and **Reformatting** process. During this step, the agent verifies the factual and scientific accuracy of the underlying reasoning—for instance, correcting an initial hypothesis of hydrolysis to the correct mechanism of solvent extraction as shown in the example. Concurrently, the answer is restructured into a clear, step-by-step format, eliminating ambiguities and extraneous details. This ensures that the final refined answer is not only correct but also presented in a structured and easily digestible manner, thereby enhancing its utility for precise model evaluation.

This unique dual-filter system, which combines adversarial LLM filtering with multi-stage manual review, provides a robust operational definition for "high-quality difficulty". The LLM filtering ensures that problems are challenging for machines, building on the successful experiences of projects like HLE (Phan et al., 2025). The multi-stage manual review ensures that this difficulty stems from the problem's scientific depth and complexity, rather than from flaws, ambiguities, or reliance on obscure knowledge, aligning with GPQA's focus on human expert performance (Rein et al., 2023). This hybrid methodology is essential for ensuring the long-term validity and credibility of ATLAS.

4. Dataset Analysis: The ATLAS Corpus

The first phase of the ATLAS project has resulted in the ATLAS corpus, which contains approximately 800 high-quality scientific reasoning problems selected through our rigorous process. This section provides a detailed analysis of this dataset from both quantitative and qualitative perspectives.

4.1. Quantitative Overview

ATLAS covers seven core disciplines of AI for Science. To provide a comprehensive evaluation, we have established several sub-fields under each discipline and ensured a balanced distribution of text-only and multimodal problems. The table below shows the detailed statistical distribution of the dataset.

Table 2: Statistical Details of the Benchmarks

Category	Sub-category	Count / Percentage
ATLAS (Breakdown by Subject)		**798**
	Physics	175
	Materials Sci.	140
	Chemistry	117
	Earth Sci.	109
	Biology	102
	Mathematics	94
	Computer Science	61
ATLAS (Breakdown by Question Type)		**100%**
	Calculation & Derivation	71.4%
	Selection & Judgment	12.2%
	Explanatory & Descriptive	10.2%
	Structured & Composite	6.1%

4.2. Qualitative Examples

To give readers a more intuitive feel for the characteristics of the problems in ATLAS, we present a few representative examples from different subjects shown in Question 4.1 and Question 4.2.

Question 4.1: Mathematics Example

- **Sub-field:** Algebra and Geometry
- **Problem:** Let p be an odd prime, and let $m \geq 0$ and $N \geq 1$ be integers. Let Λ be a free $\mathbb{Z}/p^N\mathbb{Z}$-module of rank $2m+1$, and let

$$(,) : \Lambda \times \Lambda \to \mathbb{Z}/p^N\mathbb{Z}$$

be a perfect symmetric $\mathbb{Z}/p^N\mathbb{Z}$-bilinear form. Here, 'perfect' means that the induced map

$$\Lambda \to \mathrm{Hom}_{\mathbb{Z}/p^N\mathbb{Z}}(\Lambda, \mathbb{Z}/p^N\mathbb{Z}), \quad x \mapsto (x, \cdot)$$

is an isomorphism. Find the number of elements in the set

$$\{x \in \Lambda \mid (x,x) = 0\}$$

as a function of p, m, N.
- **Solution:** For each integer $0 \leq n \leq N$, let $\Lambda(n) := \{x \in \Lambda \mid (x,x) \in p^n\mathbb{Z}/p^N\mathbb{Z}\}$. Let $C(n) := |\Lambda(n)|$. We want to compute $C(N)$. It is trivial that $C(0) = |\Lambda| = p^{(2m+1)N}$ ... We can establish two claims: 1. For $n \geq 2$, the multiplication-by-p map $\Lambda(n-2)/p^{n-1}\Lambda \to \Lambda(n)''/p^n\Lambda$ is a bijection. 2. For $n \geq 2$, the map $\Lambda(n)'/p^n\Lambda \to \Lambda(n-1)'/p^{n-1}\Lambda$ is p^{2m}-to-1. These claims lead to the recurrence relation: $C(n) = C(n)' + C(n)'' = p^{-(2m+1)}C(n-2) + p^{(2m+1)(N-1)-(n-1)}(p^{2m}-1)$. Solving this recurrence yields the final result for $C(N)$.

- **Refined Final Answer:** $p^{(2m+1)r+2m(N-2r)} + \frac{p^{(2m+1)r}-1}{p^{(2m+1)}-1} p^{(2m+1)r-1+2m(N-2r)}(p^{2m}-1)$, where $r := \lfloor N/2 \rfloor$.
- **Source Organization:** Fudan University

Question 4.2: Biology Example

- **Sub-field:** Immunology
- **Problem:** Background: In the innate immune system, RIG-I-like receptor (RLR) family proteins recognize viral RNA in the cytoplasm, triggering the downstream mitochondrial antiviral signaling protein (MAVS). MAVS acts as a signaling adapter, recruiting multiple proteins to form the MAVS signalosome, which activates transcription factors IRF3 and NF-κB, inducing the expression of type I and type III interferons (IFNs) and other antiviral genes.
  1. What is the core RNA-binding region of MAVS?
  2. In the interaction mechanism between the key adapter protein MAVS (mitochondrial antiviral signaling protein) and cellular RNA in innate immunity, what part of the cellular mRNA does MAVS directly bind to via its central disordered domain to regulate downstream antiviral signal transduction of RIG-I-like receptors (RLRs)?
  3. Treatment with RNase disrupts the stability of what complex, and reduces what property of transcription factors like IRF3 and NF-κB p65, indicating that cellular RNA is crucial for the activation and formation of the MAVS signalosome?
- **Refined Final Answer:**
  1. Central disordered region
  2. 3'UTR (3' Untranslated Region)
  3. MAVS signalosome complex; phosphorylation level
- **Source Organization:** Shanghai Jiao Tong University School of Medicine

ATLAS distinguishes itself by focusing on problems that demand a synthesis of expert-level domain knowledge and complex, multi-step reasoning chains. The generative, short-answer format fundamentally prevents "guessing" and forces models to construct answers from first principles. For instance, the mathematics problem shown requires not just recalling definitions from abstract algebra but performing a multi-step, non-trivial derivation involving recurrence relations over a finite ring $\mathbb{Z}/p^N\mathbb{Z}$. This probes a model's ability to manipulate abstract symbolic structures.

Furthermore, the problems often necessitate causal and mechanistic reasoning, as seen in the biology example. To answer correctly, a model must navigate the complex cascade of the MAVS signaling pathway, identifying specific molecular components (central disordered region), their binding targets (3'UTR), and the functional consequences of their interactions (phosphorylation). This requires integrating disparate facts into a coherent causal model, a hallmark of true scientific understanding. Many problems in the corpus are not self-contained; they implicitly assume a knowledge base equivalent to that of an advanced undergraduate or graduate student in the field. This knowledge-intensive nature, combined with the demand for rigorous, generative reasoning, establishes ATLAS as a challenging and realistic benchmark for evaluating the capabilities of next-generation AI models in the scientific domain.

4.3. Language and Structural Characteristics

The problems in ATLAS are also challenging in terms of language and structure. The average length of a problem statement is about 65 words, but some problems describing complex scenarios can exceed 200 words. The answer format is short answer or fill-in-the-blank, requiring the model to generate precise text, numerical values, or mathematical expressions in LaTeX format. The extensive use of LaTeX (especially in physics and mathematics problems) places higher demands on the model's ability to generate and understand symbols. Compared to multiple-choice questions, this generative evaluation method can more effectively prevent models from scoring by guessing, thus more accurately reflecting their reasoning and expression abilities.

5. Evaluation and Performance

In this section, we conduct an extensive evaluation of ATLAS. We firstly establish a standardized evaluation framework to assess the performance of LLMs. Subsequently, we evaluate several leading LLMs and provide a comprehensive analysis.

5.1. Setup

LLMs. We encompass a representative series of frontier large reasoning models for evaluation, incorporating both closed-source proprietary models and prominent open-source models. The examined closed LRMs include: OpenAI GPT-5 (OpenAI, 2025), OpenAI o3 (OpenAI, 2024), OpenAI o4-mini, Gemini-2.5-Pro (Comanici et al., 2025), Grok-4 (xAI, 2025), and Doubao-Seed-1.6-thinking (ByteDance Seed, 2025), as well as open-source LRMs such as DeepSeek-V3.1 (DeepSeek-AI et al., 2024), GPT-OSS-120B (Agarwal et al., 2025), DeepSeek-R1-0528 (DeepSeek-AI et al., 2025), Qwen3-235B-A22B (Yang et al., 2025), Qwen3-235B-A22B-2507 (Yang et al., 2025), and GLM-4.5 (Zeng et al., 2025).

Judge as Reasoning. As previously noted, the answers in ATLAS comprise multiple responses, along with complex natural language and symbolic descriptions, which complicate the assessment of the alignment between model predictions and true answers using rule-based heuristic methods (Li et al., 2024a; Zheng et al., 2023; Liu et al., 2025). To address this challenge, we regard the evaluation of ATLAS as a complex reasoning task, employing prominent large reasoning models to evaluate the model prediction results. In this paper, we utilize two models, OpenAI o4-mini (OpenAI, 2024) and GPT-OSS-120B (Agarwal et al., 2025), as Judge models.

Metrics. Referencing typical reasoning tasks such as code (Chen et al., 2021) and mathematics (Hendrycks et al., 2021b; OpenAI, 2024; DeepSeek-AI et al., 2025), we report the average accuracy across multiple inferences and G-Pass@k (Liu et al., 2024b) to assess the stability of LLMs' performance.

Implementation Details. For the closed-source LLMs (i.e., LRMs), we utilize the official API to obtain predictions, while for the open-source LLMs, we deploy them using serving frameworks like SGLang (Zheng et al., 2024) and vLLM (Kwon et al., 2023) for inference. We set the maximum number of generation tokens for each LLM to 32,768, and the sampling temperature is established at 0.6. For each question, we generate 4 predictions. The complete experiment roughly consumed all the API quota worth $3,000, as well as hundreds of GPU Hours.

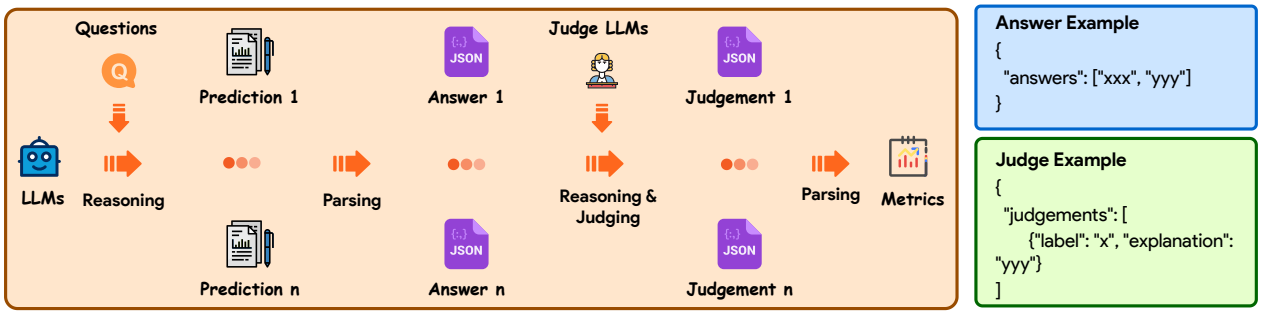

Figure 6: Overview of the evaluation workflow. During the evaluation process, the LLM is prompted to provide formatted predictions, from which the answers are extracted and input into the Judge LLMs for the computation of evaluation metrics.

5.2. Evaluation Workflow

Considering the complexity of evaluation of ATLAS, it is difficult to evaluate the model's performance through simple and conventional evaluation processes. We propose a comprehensive and user-friendly evaluation framework for assessing ATLAS based on the OpenCompass (Contributors, 2023) repository. As illustrated in Figure 6, the evaluation workflow consists of the following steps: 1) Prediction Generation; 2) Answer Parsing; 3) Judgment Generation; 4) Judgment Parsing.

Step 1: Prediction Generation. We provide each LLM (i.e., LRM) a detailed instruction to generate predictions as shown in Prompt E.1. The LLM is prompted to solve the given question step by step and must put their final answers in the JSON format. The advantages of this approach are as follows, which facilitates the extraction of answers, particularly for questions with multiple sub-questions.

Step 2: Answer Parsing. After obtaining the predictions from the LLM, we parse the JSON-formatted answers to extract the final answers.

Step 3: Judgment Generation. During this step, we input the original question, the parsed answer, and the ground truth into the Judge LLM. The Judge LLM generates assessments based on the instructions provided in Prompt E.2. Our instructions direct the LLM to evaluate the correctness of each sub-answer while considering reasonable error, providing its judgments for each sub-answer in JSON format.

Step 4: Judgment Parsing. Similar to the parsing of the answers, we parse the JSON-formatted judgments and calculate the evaluation metrics.

Our evaluation workflow offers insights into the assessment of complex problems. It not only provides judgment results for the overall outcome but also delivers fine-grained judgment results, which are beneficial for applications in Reinforcement Learning with Verifiable Rewards.

5.3. Quantitative Results

In this section, we present the quantitative results and analysis of the evaluation conducted on the validation set of ATLAS. For more detailed results on the test set of ATLAS, please refer to Section F.

Table 3: The performance of various LLMs on the validation set of ATLAS, as judged by GPT-OSS-120B, is sorted by average accuracy. Each LLM is prompted to generate four predictions, and we report the average accuracy as well as the mG-Pass@$\{2, 4\}$ scores. A high mG-Pass score indicates a high level of stability across multiple predictions.

Model	#Tokens	Accuracy (%) ↑	mG-Pass@2 (%) ↑	mG-Pass@4 (%) ↑
OpenAI GPT-5-High	32k	42.9	34.7	32.1
Gemini-2.5-Pro	32k	35.3	25.3	23.4
Grok-4	32k	34.1	25.8	24.1
OpenAI o3-High	32k	33.8	24.0	22.3
DeepSeek-R1-0528	32k	26.4	16.1	14.1
Qwen3-235B-A22B-2507	32k	26.1	18.4	16.9
Doubao-Seed-1.6-thinking	32k	26.1	18.1	16.8
DeepSeek-V3.1	32k	25.3	16.3	15.0
OpenAI o4-mini	32k	22.4	15.1	13.5
GPT-OSS-120B-High	32k	21.7	14.1	12.8

Overall Performance. The evaluation performance show in Table 3 on the validation set of ATLAS reveals that: 1) OpenAI GPT-5-High stands out as the top-performing model, achieving the highest accuracy (42.9%) and exhibiting strong prediction stability, with mG-Pass@2 at 34.7% and mG-Pass@4 at 32.1%; 2) The mG-Pass scores corroborate the accuracy results, indicating that models with higher accuracy typically exhibit greater stability across multiple predictions; 3) A notable performance disparity exists between the top-tier models (OpenAI GPT-5, OpenAI o3, Gemini-2.5-Pro, Grok-4). Specifically, OpenAI o3-High ranks second with an accuracy of 35.3%, maintaining stability with mG-Pass@2 at 25.3% and mG-Pass@4 at 23.4%. Gemini-2.5-Pro closely follows, recording an accuracy of 34.1% and mG-Pass scores of 25.8% (@2) and 24.1% (@4), indicating competitive stability. The remaining models, such as DeepSeek-R1-0528 (26.4%), DeepSeek-V3.1 (25.3%), Qwen3-235B-A22B-2507 (26.1%), Doubao-Seed-1.6-thinking (26.1%), OpenAI o4-mini (22.4%), and GPT-OSS-120B-High (21.7%) exhibit progressively lower accuracy and stability. Notably, some open-source LLMs like DeepSeek-R1-0528 and Qwen3-235B-A22B-2507 still deliver competitive results compared to other proprietary systems in the lower tier.

Subject Performance. Figure 7 demonstrates the performance of different LLMs across different subjects of ATLAS's validation set. Across all subjects and metrics, OpenAI GPT-5 consistently leads, achieving the highest accuracy and mG-Pass scores by a clear margin. Gemini-2.5-Pro also produces competitive results, particularly in Physics, Chemistry, and Biology. Grok-4 shows notable strength in CS, achieving the highest performance in this domain. In contrast, Qwen3-235B-A22B-2507 and Qwen3-235B-A22B generally exhibit the lowest scores across most subjects and metrics, indicating weaker performance in this ATLAS evaluation. OpenAI o3 and DeepSeek-V3.1 display moderate performance, while Doubao-Seed-1.6-thinking yields mixed results, performing relatively well in some areas but lagging in others. For Specific Subjects:

- **Chemistry**: OpenAI GPT-5 clearly dominates, followed by Gemini-2.5-Pro, while Grok-4 and Doubao-Seed-1.6-thinking perform moderately well.
- **CS**: Grok-4 significantly outperforms all other models, achieving the highest accuracy and mG-Pass scores, with GPT-5 and o3 trailing behind.
- **Earth Sci.**: OpenAI GPT-5 leads with the highest accuracy and stability, while o3 and Gemini-2.5-Pro achieve moderate performance.

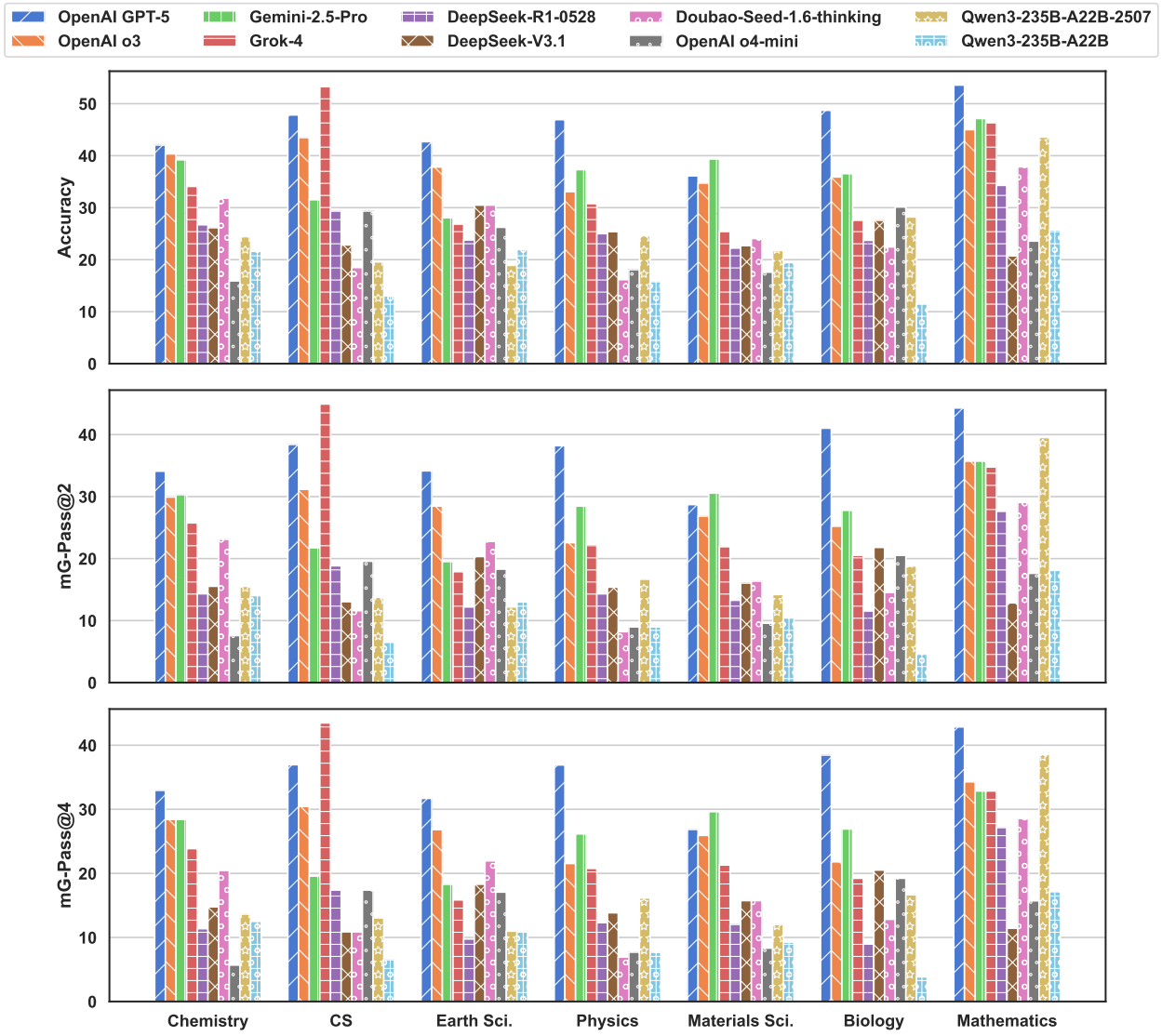

Figure 7: The performance of different LLMs across different subjects of ATLAS's validation set.

- **Physics**: GPT-5 achieves the strongest results, with Gemini-2.5-Pro and o3 also showing competitive performance.
- **Materials Sci.**: GPT-5 dominates this domain, while Gemini-2.5-Pro and o3 form the second tier of performance.
- **Biology**: GPT-5 again leads by a wide margin, with Gemini-2.5-Pro and Grok-4 performing moderately well.
- **Mathematics**: GPT-5 achieves the highest performance, followed by Qwen3-235B-A22B-2507, which shows competitive results, while Gemini-2.5-Pro also performs strongly.

Finally, we observe that the mG-Pass@$\{2, 4\}$ scores generally align with the trend of accuracy, indicating that models with higher accuracy also demonstrate greater inference stability. GPT-5's consistently high mG-Pass@$\{2, 4\}$ scores across all domains underscore its leading performance, while Grok-4's dominance in CS highlights its particular strength in this field. Conversely, the lower mG-Pass@$\{2, 4\}$ scores for models such as Qwen3-235B-A22B indicate not only reduced accuracy but also less stable or more variable predictions.

Table 4: Answer extraction error rate of different LLMs on ATLAS.

Model	Truncation Rate(%)	JSON Parse Error Rate (%) ↓
OpenAI o4-mini	0.00	0.00
OpenAI o3	1.58	0.00
DeepSeek-R1-0528	2.16	0.00
Gemini-2.5-Pro	3.49	0.00
Doubao-Seed-1.6-thinking	8.22	0.08
Grok-4	10.38	0.00

Table 5: The performance of selected LLMs on the validation set of ATLAS under 64k and 32k coutput budget, as judged by GPT-OSS-120B, is sorted by average accuracy.

Model	#Tokens	Accuracy (%) ↑	mG-Pass@2 (%) ↑	mG-Pass@4 (%) ↑
OpenAI GPT-5-High	64k	43.7	35.2	33.6
	32k	42.9	34.7	32.1
OpenAI o3-High	64k	36.6	26.6	25.3
	32k	33.8	24.0	22.3
Gemini-2.5-Pro	64k	36.6	27.7	26.1
	32k	35.3	25.3	23.4
Qwen3-235B-A22B-2507	64k	27.6	20.5	19.3
	32k	26.1	18.4	16.9
DeepSeek-V3.1	64k	26.4	17.3	15.5
	32k	25.3	16.3	15.0

5.4. Further Analysis

Output Budget and Answer Extraction. During the evaluation process, the inability to extract answers can detrimentally affect the model's performance. In our evaluation workflow, the main extraction errors originate from prediction truncation and JSON parsing errors. Consequently, we have documented the rates of answer extraction errors across various LLMs, as illustrated in Table 4. A notable observation from the table is that almost models achieved a 0.00% JSON Parse Error rate. This result is excellent, indicating that once an answer is generated, the JSON output structure remains consistently valid across all evaluated LLMs. This result implies robust parsing capabilities or careful adherence to JSON formatting of current salient LLMs, which is crucial for automated judgment and verification in data synthesis and reinforcement learning. Conversely, the Truncation Rate reveals significant variations in performance. OpenAI o4-mini is exceptional, exhibiting a 0.00% Truncation Rate, indicating it never truncates its answers, thereby ensuring complete responses—an essential characteristic for applications requiring comprehensive information. OpenAI o3 also demonstrates very low truncation rates at 1.58%, indicating that it rarely truncates its answers. While not perfect, these rates are commendable, suggesting that the majority of its answers are complete. DeepSeek-R1-0528 and Gemini-2.5-Pro display moderate truncation rates of 2.16% and 3.49%, respectively. The models with the highest truncation rates are Doubao-Seed-1.6-thinking at 8.22% and Grok-4 at 10.38%. These statistics are concerning, as they imply that over 8% and 10% of their generated answers are truncated, respectively. This suggests that the models may produce excessively lengthy chains of thought and necessitate improvements in their efficiency. To illustrate the impact of output budget, we conducted experiments with a token budget of 64k compared to 32k, as detailed in

Table 6: Summary of the primary error categories of ATLAS. We randomly sample 200 judge explanations of erroneous predictions to identify and summarize the most frequent error modes.

Error Category	Description	Proportion (%)
Numerical discrepancies	Numerical values differ from the standard beyond acceptable tolerance (e.g., ± 0.1).	27.0
Mathematical errors	Incorrect formulas, equations, or mathematical expressions (e.g., wrong coefficients, terms).	16.5
Missing components	Omission of required elements (e.g., terms in equations, methods, or parts of multi-part answers).	13.0
Structural mismatches	Answers differ in format, structure, or functional form from the standard answer.	11.0
Incorrect methods	Use of wrong approaches, techniques, or assumptions to solve the problem.	8.5
Contradictory conclusions	Answers directly oppose the standard answer's conclusion or assertion (e.g., "Yes" vs. "No").	7.0
Unit errors	Missing, incorrect, or inconsistent units in the answer.	5.0
Incorrect reasoning	Flawed logic or explanations that do not align with the standard answer's reasoning.	4.5
Misinterpretation of concepts	Misunderstanding of key principles or concepts required for the answer.	4.0
Incomplete answers	Partially correct answers lacking essential details or components.	3.5

Table 3. As presented in Table 5, while most LLMs exhibit improved performance with increased limits from 32k to 64k output tokens, this extension in output length incurs significant inference overhead, particularly given the parameter size of contemporary LLMs. This underscores the importance of enhancing the inference efficiency of LLMs.

Error Category. To provide valuable insights into areas where improvement efforts would have the most impact, we analyze the errors in the prediction results, as summarized in Table 6. The results reveal that numerical discrepancies are the most prevalent error category, accounting for 27.0% of all errors. This finding suggests that precision in numerical outputs or calculations poses a significant challenge. Following numerical discrepancies, mathematical errors represent the second largest category at 16.5%, indicating difficulties with the application of correct formulas, equations, or expressions. Missing components (13.0%) and structural mismatches (11.0%) are also significant error types, underscoring issues related to completeness and adherence to expected output formats. Additionally, incorrect methods and reasoning account for a notable proportion, suggesting substantial room for improvement in the current LLMs' professional knowledge within the scientific domain.

Judge Model. Given that our evaluation is highly related to the judge model used, we analyze the performance of various advanced reasoning models used as judge models. As demonstrated in Table 7, we compare the performance evaluations conducted by Qwen3-235B-A22B and GPT-OSS-

Table 7: The performance comparison between judged by Qwen3-235B-A22B and GPT-OSS-120B is as follows: a "+" subscript indicates that the score judge by Qwen3-235B-A22B outperforms that by GPT-OSS-120B, while a "-" signifies the opposite.

Model	Accuracy (%) ↑	mG-Pass@2 (%) ↑	mG-Pass@4 (%) ↑
OpenAI o3	$30.3_{-5.3}$	$19.3_{-6.4}$	$17.3_{-6.9}$
Gemini-2.5-Pro	$31.3_{-3.6}$	$21.0_{-3.4}$	$19.3_{-3.2}$
Grok-4	$30.7_{-2.2}$	$21.8_{-3.2}$	$19.9_{-3.6}$
DeepSeek-R1-0528	$22.0_{-3.8}$	$12.9_{-2.6}$	$11.3_{-2.1}$

120B, respectively. Across nearly all models and metrics, Qwen3-235B-A22B typically allocates lower scores than GPT-OSS-120B in the domains of Accuracy, mG-Pass@2, and mG-Pass@4. To investigate these differences, we conducted a case analysis comparing the judgments of GPT-OSS-120B and Qwen3-235B-A22B. As demonstrated in Case 5.1, which involves a computer science question in the validation set of ATLAS, OpenAI o3 produced the prediction $t_n = 2n \ln n(1 + o(1))$. In the context of algorithm complexity in computer science, $\log$ and $\ln$ are equivalent, verifying the correctness of OpenAI o3's prediction. However, Qwen3-235B-A22B fails to acknowledge this equivalence, leading to an incorrect judgment. Additionally, in Case 5.2, which shows a Materials Sci. question, OpenAI o3 accurately predicts the outcomes for two sub-questions, whereas Qwen3-235B-A22B erroneously assumed that the prediction "yes,no" pertained to a single question, resulting in an incorrect judgment. Lastly, as illustrated in Case 5.3, involving a physics question, OpenAI o3 provided an answer of 1.6×10^2N, exhibiting a relative error of 0.376% compared to the standard answer, thus meeting the permissible error conditions specified in our judge prompt (Prompt E.2). Consequently, it should be deemed correct. However, Qwen3-235B-A22B erroneously identified the absolute error as the relative error, resulting in an incorrect judgment. The case analysis results indicate that reasoning models with advanced capabilities also exhibit superior judgment accuracy. Compared to GPT-OSS-120B, Qwen3-235B-A22B is more susceptible to errors in knowledge and semantic understanding. Nonetheless, assessing the effectiveness of judgment models requires further investigation, which falls outside the scope of this paper. We hope the introduction of ATLAS will encourage the community to advance research on judgment models for questions that are difficult to verify.

Case 5.1: Case in CS subject

Question:

Given an array $A[1 : n]$ of n distinct numbers, recall the *quicksort* algorithm to sort them into an ascending order.

- First we let $p \leftarrow A[1]$ be the pivot.
- Then we use p to rearrange $A[1 : n]$ in the following way:
  - For some $k \in [n]$, we have $A[k] = p$;
  - For all $i \in \{1, 2, \ldots, k-1\}$, we have $A[i] < p$;
  - For all $i \in \{k+1, \ldots, n\}$, we have $A[i] > p$.
  - Then recursively sort $A[1 : k-1]$ and $A[k+1; n]$.

Assume our algorithm used in step 2 to compute $A[1 : k-1]$ and $A[k+1 : n]$ is *stable*. That is, if $A[i]$ and $A[j]$ are both smaller or larger than p and $i < j$, then after the rearrangement, $A[i]$ still appear before $A[j]$.

Clearly the performance of the algorithm depends on the input array A. In this problem, we analyze its *average complexity*. Suppose that the input array A is chosen uniformly at

random from the $n!$ permutations of the n distinct numbers. Let t_n be the average times of comparisons during the execution of quicksort (the average is taken over all inputs).
Find the general expression for t_n. (In the problem, we do care about the constant appearing in the main term. As a result, your expression should be of the form $t_n = f(n)(1 + o(1))$ for some $f(n)$).

Ground Truth:
$t_n = 2n \log n(1 + o(1))$

Model Prediction (OpenAI o3):
$t_n = 2n \ln n(1 + o(1))$

Judgement by Qwen3-235B-A22B:
label: **B** (✘)

explanation: The candidate's answer uses natural logarithm (ln) while the standard answer uses log (typically base 2 in algorithm analysis). The expressions differ by a constant factor ($\ln n$ vs. $\log n$), which violates the requirement to match the constant in the main term. The standard answer requires the form $2n \log n(1 + o(1))$, but the candidate's answer $2n \ln n(1 + o(1))$ is not semantically equivalent due to the logarithm base discrepancy.

Judgement by OpenAI o4-mini:
label: **A** (✔)

explanation: The candidate's expression $t_n = 2(n+1)H_n - 4n = 2n \ln n + O(n) = 2n \ln n(1 + o(1))$ is asymptotically equivalent to the standard $2n \log n(1 + o(1))$.

Case 5.2: Case in Materials Sci. subject

Question:
Does the recovery of the material include a decrease in dislocation density and a reduction in grain boundary area?
Ground Truth:
The reduction in dislocation density belongs to recovery, while the decrease in grain boundary area does not belong to recovery.
Model Prediction (OpenAI o3):
yes, no
Judgement by Qwen3-235B-A22B:
label: **C** (✘)
explanation: The candidate's answer contains multiple conflicting selections ('yes' and 'no'), making it invalid under the 'REPETITIVE' criterion as it includes contradictory outputs in a single response.
Judgement by OpenAI o4-mini:
label: **A** (✔)
explanation: The candidate correctly answered 'yes' for dislocation density reduction and 'no' for grain boundary area reduction, matching the standard answer.

Case 5.3: Case in Physics subject

Question:
A bowl is symmetrically inverted and immersed in a dense fluid with a specific gravity $SG = 15.6$. The immersion depth is $200\,\text{mm}$ (measured along the centerline from the edge of the bowl). The height of the bowl is $80\,\text{mm}$, and the fluid rises $20\,\text{mm}$ inside the bowl. The bowl is made of a clay formulation with a specific gravity $SG = 6.1$, an inner diameter of $D = 100\,\text{mm}$, and the volume of the bowl itself is approximately $0.9\ L$. Calculate the force F required to hold the bowl in its current position.
Ground Truth:
$F = 159.4\text{N}$
Model Prediction (OpenAI o3):
$1.6 \times 10^2\text{N}$ (downward)
Judgement by Qwen3-235B-A22B:
label: **B** (✘)
explanation: The candidate's answer $1.6 \times 10^2\text{N} = 160\text{N}$ differs from the standard answer 159.4N by 0.6N, exceeding the ± 0.1 tolerance for numerical equivalence. While the direction (downward) is physically reasonable, the standard answer does not include directional information, and the magnitude discrepancy invalidates strict equivalence.
Judgement by OpenAI o4-mini:
label: **A** (✔)
explanation: The candidate's value of $1.6 \times 10^2\text{N}(160\text{N})$ is within the acceptable tolerance of the standard 159.4N. The added direction does not conflict with the standard.

6. Discussion and Future Work

6.1. Implications of the Results

The benchmark results from ATLAS clearly indicate that while large models exhibit astonishing capabilities in many areas, there remains a huge gap between them and human experts in scientific domains that require deep, rigorous, and comprehensive reasoning. This finding is significant for our understanding and planning of the path toward Artificial General Intelligence (AGI). It suggests that true general intelligence lies not only in linguistic fluency and breadth of knowledge but, more importantly, in mastering the structured, verifiable, and highly complex reasoning paradigm of science. The evaluation results of ATLAS provide us with a sober and quantitative measure of how far we are from achieving AI capable of reliable scientific discovery. This aligns with the vision of projects like HLE, which aim to provide a "roadmap" for future research (Phan et al., 2025).

6.2. Limitations of ATLAS

We also recognize the limitations of our current work. First, the scale of the initial dataset, with about 800 problems, while involving a huge investment in the quality of each problem, is smaller in total number than some larger-scale benchmarks. Second, the current version of ATLAS is predominantly composed of Chinese and seven core scientific subjects. These limitations are the starting point for our future work and also reflect our commitment to academic rigor.

6.3. The ATLAS Platform: A Future Roadmap

The ATLAS project is not a one-time data release but the beginning of a long-term, continuous construction plan. Our vision is to build an open, collaborative ATLAS platform to continuously promote the development of AI scientific reasoning capabilities. The future roadmap includes:

- **Continuous Content Updates:** We plan to regularly release new problem packs to keep pace with the rapid development of models and prevent "overfitting" of the evaluation benchmark. This will ensure that ATLAS can serve as an effective tool for measuring the capabilities of frontier models over the long term.
- **Expanding the Scope of Evaluation:** Future versions will gradually expand their coverage to include more scientific fields (such as neuroscience, pharmacy, environmental science, etc.), more languages (especially English), and possibly new task formats (such as hypothesis generation, experimental design, literature review, etc.), to more comprehensively evaluate the scientific capabilities of models.
- **Building a Community Collaboration Ecosystem:** We will establish a collaborative platform to invite domain experts from around the world to participate in the problem creation and review process. By drawing on the community collaboration models of projects like HLE (Phan et al., 2025), we can gather a broader range of wisdom to ensure the continuous high-quality development of ATLAS and make it a public resource truly owned and maintained by both the scientific and AI communities.

7. Conclusion

In response to the challenges of benchmark saturation and data contamination faced by current large model evaluations, this paper proposes and constructs a new high-difficulty, multidisciplinary scientific reasoning benchmark—ATLAS. Through a rigorous process combining expert-original problem creation, adversarial model filtering, and multi-stage blind review, we have ensured the high standards of originality, quality, and difficulty of ATLAS. The initial dataset focuses on the core areas of AI for Science, presented in Chinese, filling an important gap in the existing evaluation ecosystem.

Systematic evaluation of the most advanced current models shows that all models perform poorly on ATLAS, confirming the benchmark's effectiveness as a "touchstone" for frontier capabilities and revealing significant deficiencies in the deep scientific reasoning of the current AI. Through in-depth analysis of model performance and error patterns, we provide specific diagnostic information and research directions for future AI model improvements.

We believe that ATLAS and its future continuous development will provide a valuable and reliable tool for measuring and guiding the progress of AI capabilities in the key area of scientific discovery, thereby promoting the arrival of Artificial General Intelligence more robustly and clearly.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. gpt-oss-120b & gpt-oss-20b model card. *CoRR*, abs/2508.10925, 2025. 5.1

Samuel R. Bowman and George E. Dahl. What will it take to fix benchmarking in natural language understanding? In *NAACL-HLT*, pages 4843–4855. Association for Computational Linguistics, 2021. 2.2

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 1078–1088. PMLR, 2020. 2.2

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020. 2.1

ByteDance Seed. Introduction to techniques used in seed1.6. https://seed.bytedance.com/en/seed1_6, 2025. 5.1

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3): 39:1–39:45, 2024. 2.4

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang,

Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *CoRR*, abs/2107.03374, 2021. 5.1

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/. 1.3

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit S. Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric Chu, Toby Boyd, Brad Hekman, Aaron Parisi, Chaoyi Zhang, Kornraphop Kawintiranon, Tania Bedrax-Weiss, Oliver Wang, Ya Xu, Ollie Purkiss, Uri Mendlovic, Ilaï Deutel, Nam Nguyen, Adam Langley, Flip Korn, Lucia Rossazza, Alexandre Ramé, Sagar Waghmare, Helen Miller, Nathan Byrd, Ashrith Sheshan, Raia Hadsell Sangnie Bhardwaj, Pawel Janus, Tero Rissa, Dan Horgan, Sharon Silver, Ayzaan Wahid, Sergey Brin, Yves Raimond, Klemen Kloboves, Cindy Wang, Nitesh Bharadwaj Gundavarapu, Ilia Shumailov, Bo Wang, Mantas Pajarskas, Joe Heyward, Martin Nikoltchev, Maciej Kula, Hao Zhou, Zachary Garrett, Sushant Kafle, Sercan Arik, Ankita Goel, Mingyao Yang, Jiho Park, Koji Kojima, Parsa Mahmoudieh, Koray Kavukcuoglu, Grace Chen, Doug Fritz, Anton Bulyenov, Sudeshna Roy, Dimitris Paparas, Hadar Shemtov, Bo-Juen Chen, Robin Strudel, David Reitter, Aurko Roy, Andrey Vlasov, Changwan Ryu, Chas Leichner, Haichuan Yang, Zelda Mariet, Denis Vnukov, Tim Sohn, Amy Stuart, Wei Liang, Minmin Chen, Praynaa Rawlani, Christy Koh, JD Co-Reyes, Guangda Lai, Praseem Banzal, Dimitrios Vytiniotis, Jieru Mei, and Mu Cai. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261, 2025. 5.1

OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass, 2023. 5.2

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, and Wangding Zeng. Deepseek-v3 technical report. *CoRR*, abs/2412.19437, 2024. 5.1

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng,

Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025. 5.1

Juraj Gottweis, Wei-Hung Weng, Alexander N. Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, Yossi Matias, Andrew Carroll, Kavita Kulkarni, Nenad Tomasev, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R. D. Costa, José R. Penadés, Gary Peltz, Yunhan Xu, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Towards an AI co-scientist. *CoRR*, abs/2502.18864, 2025. 1.2

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. A survey on llm-as-a-judge. *CoRR*, abs/2411.15594, 2024. 2.4

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. Olympiadbench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *ACL*, pages 3828–3850. Association for Computational Linguistics, 2024. 1.3, 2.2

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *ICLR*. OpenReview.net, 2021a. 1.1, 2.1

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS Datasets and Benchmarks*, 2021b. 1.1, 1.3, 2.2, 5.1

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *NeurIPS*, 2023. 2.1

Zhen Huang, Zengzhi Wang, Shijie Xia, Xuefeng Li, Haoyang Zou, Ruijie Xu, Run-Ze Fan, Lyumanshan Ye, Ethan Chern, Yixin Ye, Yikai Zhang, Yuqing Yang, Ting Wu, Binjie Wang, Shichao Sun, Yang Xiao, Yiyuan Li, Fan Zhou, Steffi Chern, Yiwei Qin, Yan Ma, Jiadi Su, Yixiu Liu, Yuxiang Zheng, Shaoting Zhang, Dahua Lin, Yu Qiao, and Pengfei Liu. Olympicarena: Benchmarking multi-discipline cognitive reasoning for superintelligent AI. In *NeurIPS*, 2024. 2.2

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In *NAACL-HLT*, pages 4110–4124. Association for Computational Linguistics, 2021. 2.2

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *SOSP*, pages 611–626. ACM, 2023. 5.1

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *CoRR*, abs/2412.05579, 2024a. 2.4, 5.1

Yucheng Li, Yunhao Guo, Frank Guerin, and Chenghua Lin. An open-source data contamination report for large language models. In *EMNLP (Findings)*, pages 528–541. Association for Computational Linguistics, 2024b. 2.1

Hongwei Liu, Zilong Zheng, Yuxuan Qiao, Haodong Duan, Zhiwei Fei, Fengzhe Zhou, Wenwei Zhang, Songyang Zhang, Dahua Lin, and Kai Chen. Mathbench: Evaluating the theory and application proficiency of llms with a hierarchical mathematics benchmark. In *ACL (Findings)*, pages 6884–6915. Association for Computational Linguistics, 2024a. 1.3

Junnan Liu, Hongwei Liu, Linchen Xiao, Ziyi Wang, Kuikun Liu, Songyang Gao, Wenwei Zhang, Songyang Zhang, and Kai Chen. Are your llms capable of stable reasoning? *CoRR*, abs/2412.13147, 2024b. 5.1

Shudong Liu, Hongwei Liu, Junnan Liu, Linchen Xiao, Songyang Gao, Chengqi Lyu, Yuzhe Gu, Wenwei Zhang, Derek F Wong, Songyang Zhang, et al. Compassverifier: A unified and robust verifier for llms evaluation and outcome reward. *arXiv preprint arXiv:2508.03686*, 2025. 5.1

Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. LLM4SR: A survey on large language models for scientific research. *CoRR*, abs/2501.04306, 2025. 2.3

Justin K. Miller and Wenjia Tang. Evaluating LLM metrics through real-world capabilities. *CoRR*, abs/2505.08253, 2025. 1.3

OpenAI. Introducing openai o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/, 2024. Accessed: 2024-12. 5.1

OpenAI. Gpt-5 and the new era of work. https://openai.com/index/gpt-5-new-era-of-work/, 2025. Accessed: 2025-08. 5.1

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Daron Anderson, Tung Nguyen, Mobeen Mahmood, Fiona Feng, Steven Y. Feng, Haoran Zhao, Michael Yu, Varun Gangal, Chelsea Zou, Zihan Wang, Jessica P. Wang, Pawan Kumar, Oleksandr Pokutnyi, Robert Gerbicz, Serguei Popov, John-Clark Levin, Mstyslav Kazakov, Johannes Schmitt, Geoff Galgon, Alvaro Sanchez, Yongki Lee, Will Yeadon, Scott Sauers, Marc Roth, Chidozie Agu, Søren Riis, Fabian Giska, Saiteja Utpala, Zachary Giboney, Gashaw M. Goshu, Joan of Arc Xavier, Sarah-Jane Crowson, Mohinder Maheshbhai Naiya, Noah Burns, Lennart Finke, Zerui Cheng, Hyunwoo Park, Francesco Fournier-Facio, John Wydallis, Mark Nandor, Ankit Singh, Tim Gehrunger, Jiaqi Cai, Ben McCarty, Darling Duclosel, Jungbae Nam, Jennifer Zampese, Ryan G. Hoerr, Aras Bacho, Gautier Abou Loume, Abdallah Galal, Hangrui Cao, Alexis C. Garretson, Damien Sileo, Qiuyu Ren, Doru Cojoc, Pavel Arkhipov, Usman Qazi, Lianghui Li, Sumeet Motwani, Christian Schröder de Witt, Edwin Taylor, Johannes Veith, Eric Singer, Taylor D. Hartman, Paolo Rissone, Jaehyeok Jin, Jack Wei Lun Shi, Chris G. Willcocks, Joshua Robinson, Aleksandar Mikov, Ameya Prabhu, Longke Tang, Xavier Alapont, Justine Leon Uro, Kevin Zhou, Emily de Oliveira Santos, Andrey Pupasov Maksimov, Edward Vendrow, Kengo Zenitani, Julien Guillod, Yuqi Li, Joshua Vendrow, Vladyslav

Kuchkin, and Ng Ze-An. Humanity's last exam. *CoRR*, abs/2501.14249, 2025. 1.1, 1.3, 2.1, 2.2, 3.2, 3.3, 6.1, 6.3

Chandan K. Reddy and Parshin Shojaee. Towards scientific discovery with generative AI: progress, opportunities, and challenges. In *AAAI*, pages 28601–28609. AAAI Press, 2025. 1.2

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. *CoRR*, abs/2311.12022, 2023. 1.3, 2.2, 3, 3.2, 3.3

M.-A-P. Team, Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, Kang Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, Kaixin Deng, Shian Jia, Sichao Jiang, Yiyan Liao, Rui Li, Qinrui Li, Sirun Li, Yizhi Li, Yunwen Li, Dehua Ma, Yuansheng Ni, Haoran Que, Qiyao Wang, Zhoufutu Wen, Siwei Wu, Tianshun Xing, Ming Xu, Zhenzhu Yang, Zekun Moore Wang, Jun Zhou, Yuelin Bai, Xingyuan Bu, Chenglin Cai, Liang Chen, Yifan Chen, Chengtuo Cheng, Tianhao Cheng, Keyi Ding, Siming Huang, Yun Huang, Yaoru Li, Yizhe Li, Zhaoqun Li, Tianhao Liang, Chengdong Lin, Hongquan Lin, Yinghao Ma, Tianyang Pang, Zhongyuan Peng, Zifan Peng, Qige Qi, Shi Qiu, Xingwei Qu, Shanghaoran Quan, Yizhou Tan, Zili Wang, Chenqing Wang, Hao Wang, Yiya Wang, Yubo Wang, Jiajun Xu, Kexin Yang, Ruibin Yuan, Yuanhao Yue, Tianyang Zhan, Chun Zhang, Jinyang Zhang, Xiyue Zhang, Xingjian Zhang, Yue Zhang, Yongchi Zhao, Xiangyu Zheng, Chenghua Zhong, Yang Gao, Zhoujun Li, Dayiheng Liu, Qian Liu, Tianyu Liu, Shiwen Ni, Junran Peng, Yujia Qin, Wenbo Su, Guoyin Wang, Shi Wang, Jian Yang, Min Yang, Meng Cao, Xiang Yue, Zhaoxiang Zhang, Wangchunshu Zhou, Jiaheng Liu, Qunshu Lin, Wenhao Huang, and Ge Zhang. Supergpqa: Scaling LLM evaluation across 285 graduate disciplines. *CoRR*, abs/2502.14739, 2025. 1.3, 3

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, Anima Anandkumar, Karianne Bergen, Carla P. Gomes, Shirley Ho, Pushmeet Kohli, Joan Lasenby, Jure Leskovec, Tie-Yan Liu, Arjun Manrai, Debora S. Marks, Bharath Ramsundar, Le Song, Jimeng Sun, Jian Tang, Petar Velickovic, Max Welling, Linfeng Zhang, Connor W. Coley, Yoshua Bengio, and Marinka Zitnik. Scientific discovery in the age of artificial intelligence. *Nat.*, 620(7972):47–60, 2023. 1.2

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In *NeurIPS*, 2024. 2.2

xAI. Grok-4. https://docs.x.ai/docs/models/grok-4-0709, 2025. 5.1

Qiming Xie, Zengzhi Wang, Yi Feng, and Rui Xia. Ask again, then fail: Large language models' vacillations in judgement. *CoRR*, abs/2310.02174, 2023. 2.4

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jian Yang, Jiaxi Yang, Jingren Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025. 5.1

Hengjie Yu and Yaochu Jin. Unlocking the potential of AI researchers in scientific discovery: What is missing? *CoRR*, abs/2503.05822, 2025. 1.2

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, Yean Cheng, Yifan An, Yilin Niu, Yuanhao Wen, Yushi Bai, Zhengxiao Du, Zihan Wang, Zilin Zhu, Bohan Zhang, Bosi Wen, Bowen Wu, Bowen Xu, Can Huang, Casey Zhao, Changpeng Cai, Chao Yu, Chen Li, Chendi Ge, Chenghua Huang, Chenhui Zhang, Chenxi Xu, Chenzheng Zhu, Chuang Li, Congfeng Yin, Daoyan Lin, Dayong Yang, Dazhi Jiang, Ding Ai, Erle Zhu, Fei Wang, Gengzheng Pan, Guo Wang, Hailong Sun, Haitao Li, Haiyang Li, Haiyi Hu, Hanyu Zhang, Hao Peng, Hao Tai, Haoke Zhang, Haoran Wang, Haoyu Yang, He Liu, He Zhao, Hongwei Liu, Hongxi Yan, Huan Liu, Huilong Chen, Ji Li, Jiajing Zhao, Jiamin Ren, Jian Jiao, Jiani Zhao, Jianyang Yan, Jiaqi Wang, Jiayi Gui, Jiayue Zhao, Jie Liu, Jijie Li, Jing Li, Jing Lu, Jingsen Wang, Jingwei Yuan, Jingxuan Li, Jingzhao Du, Jinhua Du, Jinxin Liu, Junkai Zhi, Junli Gao, Ke Wang, Lekang Yang, Liang Xu, Lin Fan, Lindong Wu, Lintao Ding, Lu Wang, Man Zhang, Minghao Li, Minghuan Xu, Mingming Zhao, Mingshu Zhai, Pengfan Du, Qian Dong, Shangde Lei, Shangqing Tu, Shangtong Yang, Shaoyou Lu, Shijie Li, Shuang Li, Shuang-Li, Shuxun Yang, Sibo Yi, Tianshu Yu, Wei Tian, Weihan Wang, Wenbo Yu, Weng Lam Tam, Wenjie Liang, Wentao Liu, Xiao Wang, Xiaohan Jia, Xiaotao Gu, Xiaoying Ling, Xin Wang, Xing Fan, Xingru Pan, Xinyuan Zhang, Xinze Zhang, Xiuqing Fu, Xunkai Zhang, Yabo Xu, Yandong Wu, Yida Lu, Yidong Wang, Yilin Zhou, Yiming Pan, Ying Zhang, Yingli Wang, Yingru Li, Yinpei Su, Yipeng Geng, Yitong Zhu, Yongkun Yang, Yuhang Li, Yuhao Wu, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yuxuan Zhang, Zezhen Liu, Zhen Yang, Zhengda Zhou, Zhongpei Qiao, Zhuoer Feng, Zhuorui Liu, Zichen Zhang, Zihan Wang, Zijun Yao, Zikang Wang, Ziqiang Liu, Ziwei Chai, Zixuan Li, Zuodong Zhao, Wenguang Chen, Jidong Zhai, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *CoRR*, abs/2508.06471, 2025. 5.1

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*, 2023. 1.3, 2.4, 5.1

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark W. Barrett, and Ying Sheng. Sglang: Efficient execution of structured language model programs. In *NeurIPS*, 2024. 5.1

Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. From automation to autonomy: A survey on large language models in scientific discovery. *CoRR*, abs/2505.13259, 2025. 2.3

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models. In *NAACL-HLT (Findings)*, pages 2299–2314. Association for Computational Linguistics, 2024. 1.3, 2.1

A. ATLAS Details

A.1. ATLAS Subjects

We show all the subjects and sub-fields of ATLAS in Table 8.

A.2. ATLAS Question Type Distribution

We also analyze the question type of ATLAS in Figure 8 with the definition in Table 9.

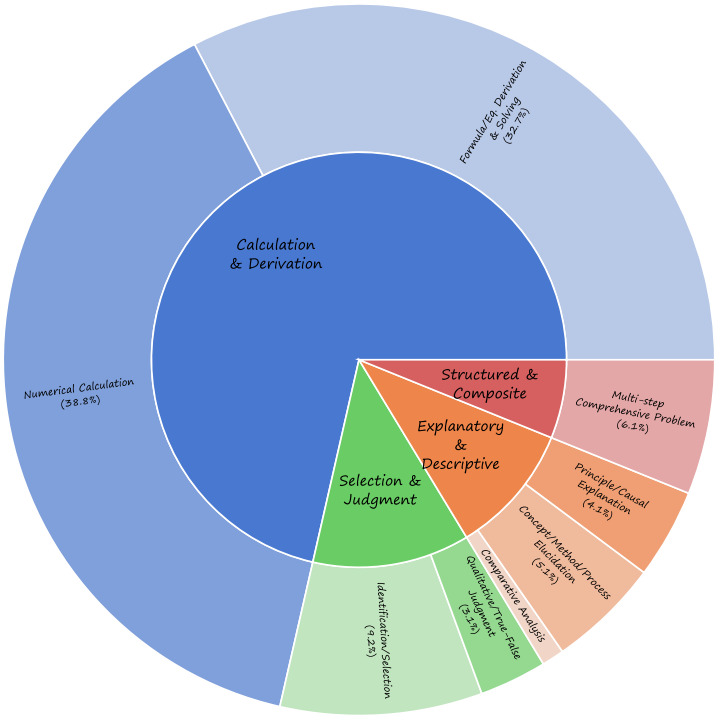

Figure 8: Hierarchical Distribution of Question Types in ATLAS

B. Question Contributors

We have collaborated with scholars from over 25 different universities or research organizations to contribute to ATLAS, and the statistics of these institutions are shown in Figure 9.

C. Expert Review

C.1. Peer Review Stage

We employ a structured peer review process governed by a formal scoring rubric, exemplified in Table 10 for the mathematics domain. Each problem is independently evaluated by two to three domain experts who score its quality across three core dimensions: (1) Content & Format, (2) Scientific Value, and (3) Difficulty. A problem is advanced from the peer review stage only if it achieves an average score of at least 3.0 across all reviews.

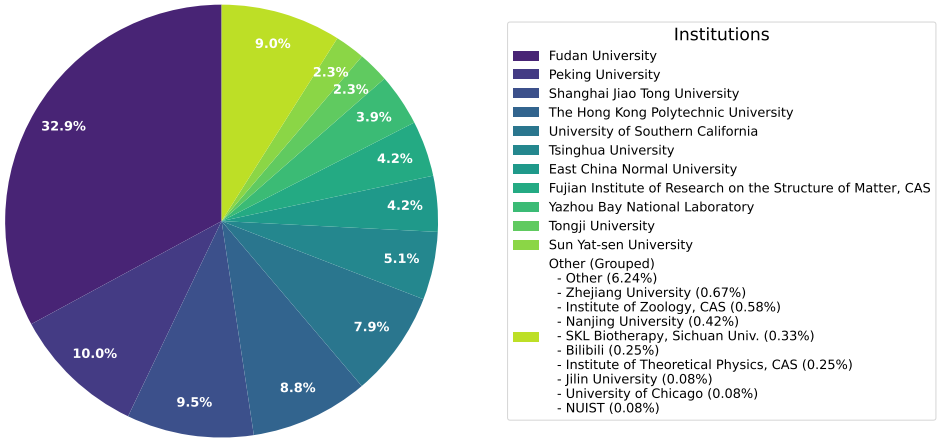

Figure 9: Distribution of Question Contributing Institutions.

C.2. Meta Review Stage

The meta review stage we invite experts to give a "Meta Review" for every question passed in peer review stage, we sample 50 questions failed in this stage and summarize the reason in Table 11.

D. Expert Review

E. Prompts for Evaluation

Prompt E.1 and Prompt E.2 demonstrates the details instructions leveraged for evaluation.

Table 8: Statistics of the ATLAS dataset, detailing the number of questions per sub-field.

Subject	Sub-field	Count
Biology	Molecular Biology and Biotechnology	11
	Genetics and Bioinformatics	59
	Immunology	2
	Physiology and Integrative Biology	3
	Neuroscience and Psychology	3
	Ecology	5
	Biophysics and Biochemistry	15
	Cell Biology	4
Mathematics	Analysis	18
	Statistics and Operations Research	9
	Algebra and Geometry	51
	Differential Equations and Dynamical Systems	9
	Computational Mathematics	3
	Interdisciplinary Mathematics	4
Chemistry	Physical Chemistry	21
	Inorganic Chemistry	69
	Organic Chemistry	8
	Analytical Chemistry	11
	Chemical Engineering and Technology	2
	Theoretical and Computational Chemistry	6
Physics	Relativity	11
	Astrophysics	5
	Thermodynamics and Statistical Physics	22
	Electrodynamics	50
	Quantum Mechanics	33
	Classical Mechanics	48
	Fluid Mechanics	6
CS	Computer Science and Technology Fundamentals	15
	Computer Architecture	27
	Artificial Intelligence	16
	Computer Software	3
Earth Sci.	Geography	19
	Geodesy	19
	Space Physics	9
	Atmospheric Chemistry	31
	Solid Earth Geophysics	7
	Marine Science	5
	Hydrology	11
	Geochemistry	3
	Geology	5
Materials Sci.	Material Synthesis and Processing Technology	15
	Metal Materials	13
	Fundamental Materials Science	7
	Material Testing and Analysis Technology	11
	Composite Materials	64
	Organic Polymer Materials	23
	Materials Surface and Interface	7
	Total	**798**

Table 9: A Typology of Question Forms Based on Structural Analysis. This table categorizes questions not by their subject matter, but by their formal structure and expected answer type.

Primary Category	**Secondary Category**	**Features & Description**
Calculation & Derivation	Single-Value Calculation	Requires calculating a specific numerical result based on given conditions and data.
	Formula Derivation	Requires deriving the mathematical relationship between variables; the answer is a formula or equation.
Explanatory & Descriptive	Principle/Causal Explanation	Typically asks "Why," requiring an explanation of the scientific principles behind a phenomenon or conclusion.
	Process/Method Description	Typically asks "How," requiring a description of operational steps or mechanisms.
	Concept/Feature Elucidation	Requires defining terms, or listing and describing the types and features of an object or concept.
	Comparative Analysis	Requires comparing the similarities and differences between two or more items.
Selection & Judgment	Multiple Choice / Select	Presents multiple statements, requiring the selection of all correct or qualifying items.
	True/False Judgment	Presents a single statement, requiring a judgment of its correctness (true or false).
Specific Info. Retrieval	Fill-in-the-Blank (Cloze)	A descriptive text with blanks that need to be filled with the correct technical terms or information.
	Direct Information Recall	Requires writing specific information like chemical formulas or equations directly from the prompt.
Structured Multi-Part Problem	Comprehensive Analysis	Revolves around a central scenario, with multiple, logically connected sub-questions.

Table 10: ATLAS Expert Peer Review Scoring Rubric (Math Domain as a Tempalte).

Dimension	Score	Standard Description
Content & Format	5	Content is rigorous, format is perfect, expression is precise.
(0-5)	3	Content is reasonable, with minor areas for improvement.
	1	Multiple clear errors, requires major revision.
Scientific Value	5	Extremely high value, tests cross-domain integration and top-tier thinking.
(0-5)	3	High value, tests application of multiple knowledge points.
	1	Limited value, tests simple memorization or mechanical operation.
Difficulty Rating	5	Extremely difficult (e.g., IMO/Putnam level for Math).
(0-5)	3	High difficulty (e.g., AIME/CMO level for Math).
	1	Basic level (e.g., High school foundation).

Table 11: Analysis of Rejection Reasons for Meta Question Review

Main Category	Subcategory	Percentage
Content & Logical Flaws	***Subtotal***	**46%**
	Incorrect Answer or Fact	16%
	Calculation or Derivation Error	14%
	Oversimplified or Missing Premise	8%
	Flawed Logic / Violates Principles	6%
	Ignores Provided Context/Data	2%
Difficulty & Scope	***Subtotal***	**38%**
	Difficulty Too Low	24%
	Out of Scope / Uncollected Type	8%
	Low Value / Rote Memorization	6%
Content Quality & Formatting	***Subtotal***	**16%**
	Formatting or Rendering Error	8%
	Missing or Unclear Content	4%
	Mismatched or Hard-to-Verify Content	4%

Prompt E.1: Prompt for Prediction

Problem:

```
{problem}
```

Instructions:

- Solve the problem step by step. If the problem contains multiple sub-questions, make sure to solve each one individually.
- At the end, output **only** the final answers in the following format:

````
```json
{
  "answers": [
    "answer to sub-question 1",
    "answer to sub-question 2",
    ...
  ]
}
```
````

- Each item in the list should be the **final answer** to a sub-question.
- If there is only one question, return a list with a single item.
- **Do not** include any explanation, reasoning steps, or additional text outside the JSON list.
- **Do** put the JSON list in the block of ```` ```json ... ``` ````

Prompt E.2: Prompt for Judgement

You are an expert answer grader. Your task is to evaluate whether the candidate's **final answer** matches the **provided standard answer**. Follow the grading protocol strictly and **do not generate or modify answers**. Only compare the candidate's response to the given standard.

Evaluation Guidelines

1. Reference Standard

- The **standard answer is always correct** — never question its validity.
- The **question itself is valid** — do not critique or reinterpret it.
- Do **not** regenerate, fix, or complete the candidate's answer — only **evaluate** what is provided.

2. Comparison Strategy

- Carefully analyze the **question type** and **standard answer format**:
  - Determine whether an **exact match** is required, or whether **partial correctness** is acceptable (e.g., for multi-component or expression-based answers).
  - This judgment should be based on the **question's phrasing and answer structure**.
- Evaluate **only the candidate's final answer**, ignoring reasoning or explanation.

- Ignore differences in **formatting, style**, or **variable naming**, as long as the content is equivalent.
- For **mathematical expressions**, check **step-by-step equivalence** (e.g., by simplifying both expressions and comparing results).
- For **multiple-choice questions**, only the **final selected option** and its **associated content** matter.
- For **decimal or fraction comparisons**, consider the answers equivalent if the relative error is $\leq$ **±0.1**.

3. Multi-part Answers

- If the question requires **multiple components or selections**, all parts must match the standard answer exactly.
- Compare each component individually.
- **Partial correctness is not acceptable** — label as incorrect if any part is wrong.

4. Validity Check Immediately reject the candidate's answer if it meets **any of the following criteria**:

- **INCOMPLETE**: Final sentence is cut off or the answer is clearly unfinished.
- **REPETITIVE**: Contains repeated phrases or outputs in a loop.
- **REFUSAL**: Explicitly states inability to answer (e.g., "I cannot answer this question").
- Use label **C**.

Grading Scale

Grade	Label	Description
A	CORRECT	Exact or semantically equivalent match; includes numerically equivalent results (within ±0.0001)
B	INCORRECT	Any deviation from the standard answer; includes partial matches
C	INVALID	Answer is INCOMPLETE, REPETITIVE, or a REFUSAL

Evaluation Procedure & Output Format

1. **Check for Validity First**:
    - If the answer is incomplete, repetitive, or a refusal, **immediately assign label C** with the reason and stop further evaluation.
2. **If Valid, Compare Content**:
    - Analyze the question type: Are strict matches required (e.g., order, format, completeness)?
    - Apply tolerances: Accept allowed variations (e.g., unformatted but equivalent math, missing labels in MCQs).
    - Carefully compare final answers for:
        - Semantic or mathematical equivalence
        - Relative error tolerance (± 0.1)
        - Expression format flexibility
3. **Produce a Final Judgment**:
    - For each sub-question, return:

```
```json
{
  "label": "A" / "B" / "C",
  "explanation": "Brief justification here"
}
```
```

    - At the end, return a list of these JSON objects for each sub-question.

```
```json
{
  "judgements": [
    {
        "label": "A" / "B" / "C" for sub-question 1,
        "explanation": "Brief justification here for sub-question 1"
    },
    {
        "label": "A" / "B" / "C" for sub-question 2,
        "explanation": "Brief justification here for sub-question 2"
    },
    ...
  ]
}
```
```

    - If there is only one question, return a list with a single item.
    - **Do** put the JSON list in the block of `json`.

Task Input

```
{problem}


{answer}


{prediction}
```

Begin Evaluation Below:
Analyze the candidate's answer step by step, then provide a **final structured judgment**.

Table 13: The performance of various LLMs on the test set of ATLAS, as judged by GPT-OSS-120B, is sorted by average accuracy. Each LLM is prompted to generate four predictions, and we report the average accuracy as well as the mG-Pass@$\{2, 4\}$ scores. A high mG-Pass score indicates a high level of stability across multiple predictions.

Model	#Tokens	Accuracy (%) ↑	mG-Pass@2 (%) ↑	mG-Pass@4 (%) ↑
OpenAI GPT-5-High	32k	43.8	34.2	33.5
Gemini-2.5-Pro	32k	39.9	30.6	28.8
OpenAI o3-High	32k	37.4	25.9	23.8
Grok-4	32k	35.4	26.2	24.4
Qwen3-235B-A22B-2507	32k	29.7	21.6	19.8
DeepSeek-V3.1	32k	29.5	20.2	18.5
Doubao-Seed-1.6-thinking	32k	28.8	20.1	18.5
DeepSeek-R1-0528	32k	26.1	18.4	16.6
OpenAI o4-mini	32k	24.1	15.2	13.6
GPT-OSS-120B-High	32k	23.3	14.6	12.8

F. Performance on the Test Set of ATLAS

Overall Performance. Table 13 presents the performance of all LLMs evaluated on the test set of ATLAS, as judged by GPT-OSS-120B, and ordered by average accuracy. OpenAI GPT-5-High ranks highest with an accuracy of 43.8%, followed by Gemini-2.5-Pro at 39.9%, OpenAI o3-High at 37.4%, and Grok-4 at 35.4%. The lower-performing models include Qwen3-235B-A22B-2507 at 39.6%, Doubao-Seed-1.6-thinking at 28.8%, DeepSeek-R1-0528 at 26.1%, OpenAI o4-mini at 24.1%, and GPT-OSS-120B at 23.3%. The mG-Pass@2 and mG-Pass@4 scores, which indicate stability across multiple predictions, exhibit a similar pattern, with OpenAI GPT-5-High achieving the highest scores of 34.2% and 33.5%, respectively, while GPT-OSS-120B scores the lowest, at 14.6% and 12.8%. In comparison to Table 3, where OpenAI GPT-5-High leads with an accuracy of 42.9%, followed closely by Gemini-2.5-Pro at 35.3%, the overall ranking remains consistent, though OpenAI o3 models show competitive performance in the mid range. Furthermore, the accuracy of OpenAI o4-mini shows only a slight variation, from 24.1% in Table 3 to 22.4% in Table 13, suggesting relative consistency. Other models also demonstrate minor fluctuations.

Subject Performance. Figure 10 illustrates the performance of all LLMs across different subjects in ATLAS's test set. OpenAI GPT-5 consistently achieves the highest accuracy and mG-Pass scores across all subjects, standing out as the clear leader. Gemini-2.5-Pro also delivers competitive results, particularly in Chemistry, Physics, and Biology. Grok-4 demonstrates notable strength in Computer Science, achieving the best scores in this domain. In contrast, Qwen3-235B-A22B and Qwen3-235B-A22B-2507 generally show weaker performance across most subjects, while DeepSeek-R1-0528 and OpenAI o4-mini remain in the lower tier with moderate results. Doubao-Seed-1.6-thinking and DeepSeek-V3.1 produce mixed outcomes, performing well in some subjects but lagging in others. For Specific Subjects:

- **Chemistry**: OpenAI GPT-5 leads by a large margin, followed by Gemini-2.5-Pro, with Grok-4 and Doubao-Seed-1.6-thinking showing moderate results.
- **Computer Science**: Grok-4 achieves the best overall performance, with GPT-5, o3, and Doubao-Seed-1.6-thinking trailing behind.
- **Earth Science**: GPT-5 ranks highest, while Gemini-2.5-Pro and o3 achieve competitive performance.

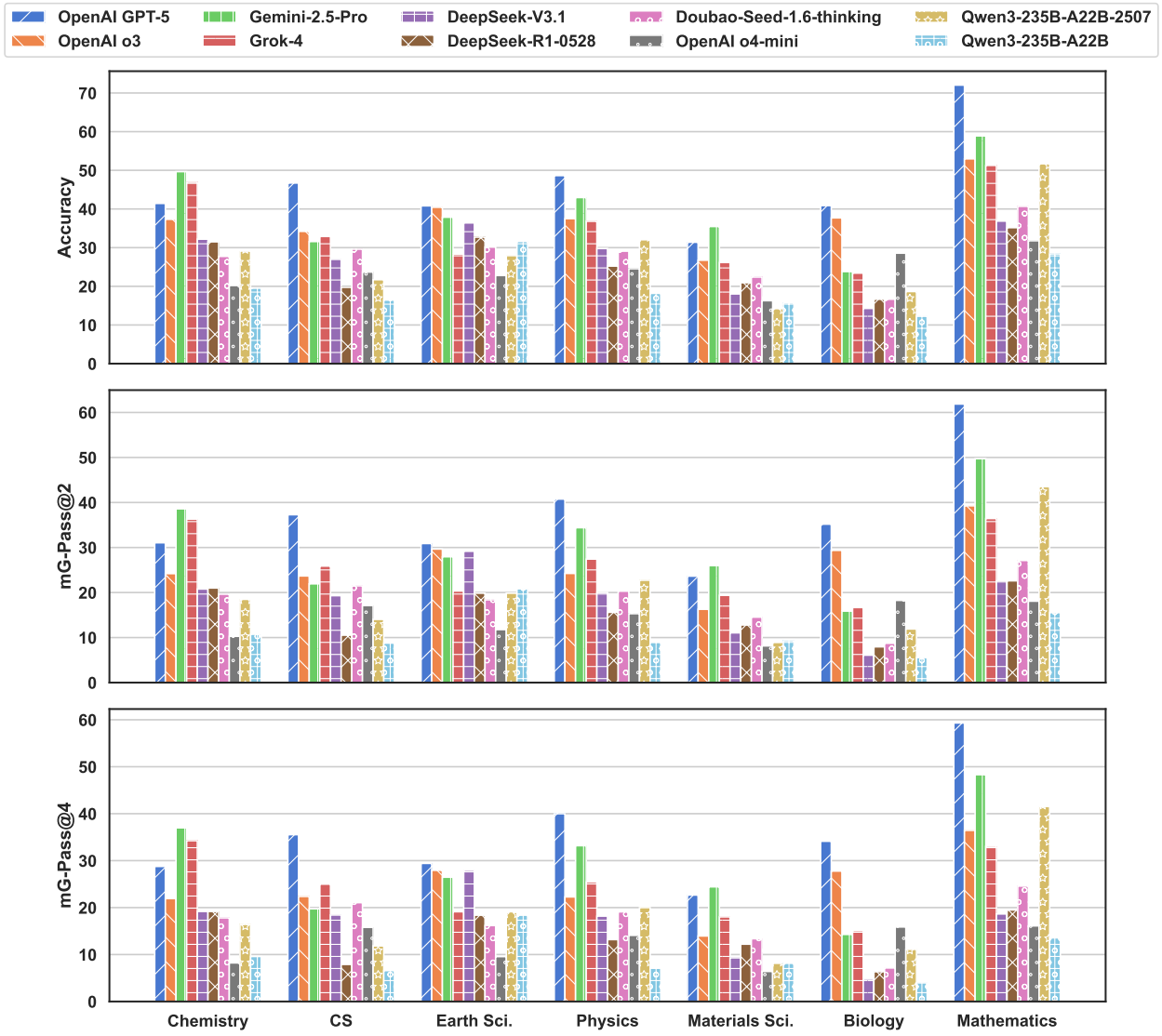

Figure 10: The performance of different LLMs across different subjects of ATLAS's test set.

- **Physics**: GPT-5 dominates, with Gemini-2.5-Pro also performing strongly.
- **Materials Science**: GPT-5 again leads, followed by Gemini-2.5-Pro and o3 as the next tier of models.
- **Biology**: GPT-5 significantly surpasses all other models, while Gemini-2.5-Pro and o3 achieve moderate accuracy and stability.
- **Mathematics**: GPT-5 shows overwhelming dominance, with Qwen3-235B-A22B-2507 and Gemini-2.5-Pro forming the second tier of performance.

By comparing these performances with those on the validation set (Figure 10), we can assess the consistency of the subject-specific outcomes. For example, GPT-5 consistently dominates across both datasets, while Grok-4 maintains its strength in Computer Science. Such consistency highlights the inherent strengths and weaknesses of the models across knowledge domains.

G. Contributors

Our team is composed of researchers with diverse technical backgrounds, each of whom contributed in different ways to the success of this project. The core contributors were responsible for all stages of the work, including data collection strategy, data quality screening, evaluation design, result analysis, and manuscript preparation. Project contributors, who come from multiple research communities, coordinated data collection efforts and oversaw data quality control. Data contributors provided realistic and challenging questions, bringing their domain expertise to strengthen the dataset. The corresponding authors initiated and supervised the project and secured the resources necessary to complete this work.

Table 14: List of Contributors

Contribution Type	Contributors
Core Contributor	Hongwei Liu[1], Junnan Liu[1], Shudong Liu[1]
Project Contributor	Haodong Duan[1], Yuqiang Li[1], Mao Su[1], Xiaohong Liu[2], Guangtao Zhai[2], Xinyu Fang[3,1], Qianhong Ma[2,1], Taolin Zhang[4,1], Zihan Ma[5,1], Yufeng Zhao[4,1], Peiheng Zhou[1], Linchen Xiao[1], Wenlong Zhang[1], Shijie Zhou[6], Xingjian Ma[6], Siqi Sun[6], Jiaye Ge[1], Meng Li[1], Yuhong Liu[1], Jianxin Dong[1], Jiaying Li[1], Hui Wu[1], Hanwen Liang[1], Jintai Lin[15], Yanting Wang[17], Jie Dong[2], Tong Zhu[16], Tianfan Fu[20], Conghui He[1], Qi Zhang[6]
Corresponding Author	Lei Bai[1], Kai Chen[1], Songyang Zhang[1]
Data Contributors	Yuqiang Li[2], Ben Gao[7], Mao Su[1], Shengdu Chai[1,6], Xuefeng Wei[8], Zicheng Zhang[2], Chunyi Li[2], Yiheng Wang[2,1], Weijia Li[9], Fenghua Ling[1], Zhou Yuhao[10,1], Xu Wanghan[2,1], He Xuming[3,1], Liu Yidi[11], Jiaqi Wei[3,1], Zhiqian Huang[6], Rui Hua[6], Pinxian Bie[6], Wenhui Qiu[6], Peng Guo[6], Junli Sun[6], Qizheng You[6], Na Wei[6], Xinyuan Zhang[6], Yurong Mou[6], Mingfeng Xie[6], Zhexuan Yu[6], Yundi Chen[6], Feng Cui[6], Kunhua Li[6], Xueting Cao[6], Liming Rao[6], Xujing Wang[6], Zichao Wang[6], Yuanhao Li[6], Zhiyuan Chen[6], Yunke Jin[6], Ruizhi Xue[6], Yibai Zhang[6], Xiao Zhou[6], Chenqing Fan[6], Zhenhao Guo[6], Junhua Liu[6], Ziqing Zhu[6], Yehao Zhang[6], Shaorong Chen[6], Tao Jin[6], Hushui Chen[6], Yidan Liu[6], Haixing Gong[6], Yifu Zhang[6], Zhibo Yu[6], Bin Wang[6], Jun You[6], Zhe Zhao[6], Lujie Yuan[6], Xiaofei Chen[6], Lin Zhang[6], Congyuan Yue[6], Zhengjie Yu[6], Tianyi Shen[6], Yutian Hou[6], Zhengyang Liu[6], Yunwen Guo[6], Shuang Li[6], Shutong Yue[2], Chi Shu[12], Yunzhang Li[6], Zhiwei He[2], Jushi Kai[2], Hailong Li[6], Yuchen He[2], Jiarong Jin[2], Jie Zhang[6], Fulin Wang[2], Xingyuan Yan[9], Haifeng Wang[13], Yuting Li[2], Yuncong Hu[2], Yadong Wu[2], Zhenghong Guo[2], Hongqiang Xiong[14], Jintai Lin[15], Yanting Wang[17], Ning Shen[3], Wang Chen[6], Kaipeng Zheng[2], Zhiwen Xue[15], Tong Liu[18], Shizhen Zhao[2], Jiye Wu[19], Zixuan Chen[2], Xiangying Shen[9], Yan Yu[15], Jieru Zhao[2], Zhezhi He[2], Qiu Yang[15], Ying Zhang[6], Zhe-Ning Chen[8], Juepeng Zheng[9], Jiuke Wang[9], Xiang Zhang[9], Xingyuan Yan[9], Meng Yang[9], Zhen Pan[2]

Main Affiliations
[1] Shanghai AI Lab
[2] Shanghai Jiao Tong University
[3] Zhejiang University
[4] Tsinghua University
[5] Xian Jiaotong University
[6] Fudan University
[7] Wuhan University
[8] Fujian Institute of Research on the Structure of Matter, Chinese Academy of Sciences
[9] Sun Yat-sen University
[10] Sichuan University
[11] University of Science and Technology of China
[12] University of Chicago
[13] Yazhouwan National Laboratory
[14] Jilin University
[15] Peking University
[16] East China Normal University
[17] Institute of Theoretical Physics, Chinese Academy of Sciences
[18] The Hong Kong Polytechnic University
[19] Nanjing University of Information Science and Technology
[20] Nanjing University,