ATMOSSCI-BENCH: Evaluating the Recent Advance of Large Language Model for Atmospheric Science

Chenyue Li[†], Wen Deng[†], Mengqian Lu[†], Binhang Yuan[†]

[†]HKUST

Abstract

The rapid advancements in large language models (LLMs), particularly in their reasoning capabilities, hold transformative potential for addressing complex challenges in atmospheric science. However, leveraging LLMs effectively in this domain requires a robust and comprehensive evaluation benchmark. To address this need, we present ATMOSSCI-BENCH, a novel benchmark designed to systematically assess LLM performance across five core categories of atmospheric science problems: hydrology, atmospheric dynamics, atmospheric physics, geophysics, and physical oceanography. We employ a template-based question generation framework, enabling scalable and diverse multiple-choice questions curated from graduate-level atmospheric science problems. We conduct a comprehensive evaluation of representative LLMs, categorized into four groups: instruction-tuned models, advanced reasoning models, math-augmented models, and domain-specific climate models. Our analysis provides some interesting insights into the reasoning and problem-solving capabilities of LLMs in atmospheric science. We believe ATMOSSCI-BENCH can serve as a critical step toward advancing LLM applications in climate service by offering a standard and rigorous evaluation framework. Our source codes are currently available at [`https://github.com/Relaxed-System-Lab/AtmosSci-Bench`].

1 Introduction

The rapid advancement of large language models (LLMs) [1], especially in their reasoning capabilities, offers transformative potential for addressing complex challenges in atmospheric science [2, 3, 4, 5]. However, the development of reliable and effective LLM-based applications for climate-related tasks requires *a robust and comprehensive evaluation framework. Such a benchmark is essential to systematically assess the performance of LLMs across a diverse array of atmospheric science problems*, ensuring their utility, accuracy, and robustness in this critical domain.

Constructing a comprehensive benchmark for atmospheric science is crucial to harness the recent advancements in large language models (LLMs) for diverse climate service applications [3]. Note that atmospheric science presents unique and complex challenges, ranging from micro-scale processes like cloud dynamics to global-scale climate systems. To ensure that LLMs can effectively contribute to solving these real-world problems, it is essential to establish a benchmark that evaluates their performance, especially their reasoning and interpretative abilities — Such a well-designed benchmark will not only foster innovation but also provide a standardized framework for assessing the utility, accuracy, and robustness of LLMs in this field.

Atmospheric science problems differ significantly from the mathematical and physical problems commonly found in existing LLM benchmarks [6, 7]. This field is inherently interdisciplinary, requiring the integration of theoretical knowledge with real-world phenomena. Atmospheric science involves analyzing and synthesizing heterogeneous data types, such as spatial coordinates, temperatures, wind patterns, and empirical estimates, which are often presented in varied formats and units. Furthermore, solving these problems necessitates the selection of appropriate physical models and mathematical methods to ensure accuracy, adding layers of complexity beyond traditional benchmarks. As such, constructing a benchmark tailored to atmospheric science is a necessary complement to existing evaluations, enabling a more comprehensive assessment of LLMs' reasoning capabilities.

To address this need, we present ATMOSSCI-BENCH, a novel benchmark to comprehensively evaluate the recent advance of LLMs for atmospheric science. Concretely, we summarize our key contributions:

Contribution 1. We construct ATMOSSCI-BENCH, a multiple-choice question benchmark designed to evaluate LLM performance across five core categories of atmospheric science: (i) *hydrology*, (ii) *atmospheric dynamics*, (iii) *atmospheric physics*, (vi) *geophysics*, and (v) *physical oceanography*. The benchmark is carefully curated from graduate-level atmospheric science problems, ensuring a high standard of relevance and complexity. Technically, ATMOSSCI-BENCH employs a template-based question generation framework. Using a manually implemented, rule-based mechanism, each question template can be systematically expanded into a desired number of concrete questions through effective symbolic extensions. This approach ensures both scalability and diversity in the question set, providing a robust tool for assessing LLM capabilities in reasoning and problem-solving within the domain of atmospheric science.

Contribution 2. We conduct a comprehensive evaluation that includes a wide range of representative open-source and proprietary LLMs, which can be concretely categorized into four classes: (i) *instruction models* that have been fine-tuned for instruction following; (ii) *reasoning models* that have been aligned with advanced reasoning abilities; (iii) *math models* that have been augmented with more mathematical skills; and (vi) *domain-specific climate models* that have been continuously pre-trained with climate-relevant corpus.

Contribution 3. We carefully analyze the evaluation results and summarize the following interesting findings:

- ***Finding 1.*** *Reasoning models (e.g., Deepseek-R1) outperform instruction, math, and domain-specific models, demonstrating the superior significance of advanced reasoning ability in atmospheric science tasks.*
- ***Finding 2.*** *The inference time scaling introduces interesting quality-efficiency tradeoffs for reasoning models—Increasing reasoning token length enhances model accuracy up to* 16*K tokens, while further expansion yields diminishing returns.*
- ***Finding 3.*** *While reasoning models show better robustness when handling arithmetic tasks with higher numerical precision when compared with other model categories, they still relatively struggle with symbolic perturbation.*

2 Related Work

LLM advances. LLMs, such as OPT [8], LLAMA [9], GPT [10], GEMINI [11], CLAUDE [12], and MIXTRAL[13], have demonstrated remarkable performance across a wide range of applications. While general-purpose LLMs exhibit strong adaptability, domain-specific models have also been developed to enhance performance in specialized fields. In the context of atmospheric science, climate-focused LLMs such as CLIMATEBERT [14], CLIMATEGPT [4], and CLIMAX [15] are designed to address the unique challenges of climate modeling and analysis, which illustrates a promising paradigm different from traditional approaches that designing a specific model for some particular task [16, 17, 18, 19, 20]. More recently, reasoning models, including GPT-O1 [21], GEMINI-2.0-FLASH-THINKING [22], QWQ [23], and DEEPSEEK-R1 [24], have emerged, highlighting advancements in mathematical and scientific problem-solving. These models leverage sophisticated reasoning techniques, presenting exciting opportunities for tackling complex challenges in atmospheric science.

LLM benchmarks. Assessing LLMs is crucial for ensuring their effectiveness in deployment across various domains [25]. Traditional benchmarks like `GSM8K` [26] and `MATH` [6] have become less effective as state-of-the-art models achieve near-perfect scores, necessitating more challenging benchmarks to evaluate reasoning capabilities accurately. Recent benchmarks target specialized fields, such as `GPQA-Diamond` [27] for expert-level science, `AIME2024` [28] for advanced mathematics, and `SCIBENCH` [7] for collegiate science problems. However, a comprehensive LLM benchmark for atmospheric science remains underrepresented, where `CLIMAQA` [29] only offers basic definition-based assessments, lacking depth in evaluating complex problem-solving abilities. Designing a good LLM benchmark requires principled guidance to ensure robust, accurate, and meaningful evaluation. For example, A notable advancement is the introduction of symbolic extensions in benchmarking, as seen in `GSM-Symbolic` [30], `VarBench` [31], and `MM-PhyQA`. These benchmarks introduce question variants by altering numerical values or modifying problem structures, improving robustness, and mitigating contamination risks. Notably, `GSM-Symbolic` highlights that even minor perturbations can significantly impact model performance, revealing fragilities in LLM reasoning. Additionally, numerical reasoning plays a fundamental role in evaluating LLMs, especially for scientific applications. Papers like NumberCookbook [32] and NumeroLogic [33] uncover weaknesses in LLMs' ability to process numerical information accurately, emphasizing that tokenization strategies and internal number representation significantly affect arithmetic performance [34]. Despite advancements in benchmarking, a rigorous climate-focused evaluation framework is still missing.

3 Benchmark Construction

3.1 Benchmark Overview

We introduce a comprehensive multiple-choice question (MCQ) benchmark, ATMOSSCI-BENCH, specifically designed for atmospheric science to enable more effective evaluation of LLMs. Unlike traditional metrics such as exact match, BLEU, or F1 scores, which primarily assess superficial similarity, MCQs offer well-defined answer choices, reducing ambiguity and enabling a more precise assessment of model comprehension and logical inference [35]. This structured format ensures a more robust evaluation of LLMs' capabilities in tackling atmospheric science challenges.

Design principles: To ensure a rigorous evaluation of LLMs in atmospheric science, we adhere to a set of well-defined principles that emphasize reasoning and interpretative abilities:

- *Deep understanding of essential physical equations:* Atmospheric science is governed by fundamental physical equations, and a meaningful evaluation requires that LLMs not only recall these principles but also apply them appropriately in the corresponding contexts. Thus, the questions should be designed to assess both conceptual comprehension and the ability to use these equations in problem-solving, ensuring the benchmark measures true scientific reasoning rather than mere memorization.
- *Complex reasoning and multi-step logic:* Many real-world atmospheric problems require synthesizing information from multiple sources, integrating equations, and applying multi-step logical reasoning. To reflect these challenges, benchmark questions should be crafted to go beyond simple recall, testing the model's ability to handle intricate reasoning and dynamic problem-solving scenarios inherent to the field.
- *Appropriate numerical arithmetic processing:* Accurate numerical computation is essential for scientific disciplines, where correct reasoning leads to fixed, verifiable answers. By incorporating numerical problems, we provide a structured and objective evaluation framework, eliminating ambiguities in assessment. This approach also enables seamless integration of reasoning tasks, extending the benchmark's scope to evaluate mathematical intuition and computational fluency.

3.2 Data Source and Preprocessing

To ensure the rigor and relevance of the benchmark, we curated questions from authoritative graduate-level textbooks and exam materials widely recognized in atmospheric science education. These sources provide high-quality, well-established content that aligns with the complexity and depth required for evaluating LLMs in this domain.

We leverage Mathpix OCR [36] to extract both questions and their corresponding explanations from the collected materials. For multi-part problems or sequential questions where solving one step is necessary to proceed to the next, we consolidated them into single questions to enhance the complexity and depth of reasoning required. This approach preserves the logical progression of problem-solving, ensuring a comprehensive assessment of model capabilities. The benchmark covers distinct sub-fields of atmospheric science, each representing a key subject:

- *Hydrology* examines the distribution, movement, and properties of water on Earth, including the water cycle, precipitation, rivers, lakes, and groundwater dynamics.
- *Atmospheric dynamics* focuses on the motion of the atmosphere, including large-scale weather systems, wind patterns, and governing forces of atmospheric circulation.
- *Atmospheric physics* covers physical processes such as radiation, thermodynamics, cloud formation, and energy transfer within the atmosphere.
- *Geophysics* encompasses the physical processes of the Earth, including its magnetic and gravitational fields, seismic activity, and internal structure.
- *Physical oceanography* investigates the physical properties and dynamics of ocean water, including currents, waves, tides, and ocean-atmosphere interactions.

3.3 Question Generation Framework

To rigorously evaluate the reasoning and problem-solving capabilities of LLMs in atmospheric science, we employ symbolic MCQ generation techniques inspired by the `GSM-Symbolic` framework [30], enhanced with a rule-based mechanism. This approach enables the creation of scalable and diverse question sets while ensuring logical coherence and alignment with real-world physical laws. Instead of fixed numerical values, we also design a template-based question perturbation mechanism with placeholder variables, which can be systematically instantiated through symbolic extensions. This ensures that models are tested on genuine reasoning ability rather than pattern matching from the potentially contaminated training data. Figure 1 illustrates the question construction pipeline as we enumerate below.

- *Question template construction*: We invite domain experts in atmospheric science to systematically transform selected questions (OCR extracted) into reusable templates. The experts manually identify numerical values within each question and replace them with variable placeholders, ensuring flexibility for symbolic instantiation. These variable placeholders, highlighted in different colors in Figure 1, allow for systematic variation while preserving the original scientific integrity of the problem.
- *Numerical assignment in question template*: We design a rule-based mechanism for valid numerical assignments in each question template. Note that many variables in atmospheric science problems are interdependent, meaning that the inappropriate assignment of some value(s) could lead to unrealistic or invalid physical scenarios. To fulfill this requirement, we ask the experts for each question template to define: (i) a valid numerical range (*min*, *max*) for each variable to ensure scientifically plausible values; (ii) a granularity parameter (i.e., the smallest step size between values) to control precision, allowing for variation in significant digits — this variation affects numerical representation, potentially influencing LLM arithmetic performance [33, 32, 34]; and (iii) a set of rule-based constraints that are manually implemented to enforce logical dependencies (e.g., in Figure 1, ensuring $t_1 < t_2$). We believe these manual configurations ensure that all generated instances remain scientifically valid while allowing systematic variation in numerical representation.
- *Automatic problem solver to support value perturbation*: For each question, we utilize GPT-4o to generate an initial Python implementation based on the corresponding explanatory materials (e.g., textbook solutions). This synthesized solution is then manually reviewed, verified, and refined by experts to ensure correctness and adherence to the intended problem-solving methodology. Once validated, the solver can automatically compute the correct answer for any given set of valid input variables, ensuring consistency and scalability in question generation. Note that to ensure consistency, accuracy, and alignment with real-world scientific standards, we also manually assign appropriate units and define significant digits for rounding the final answer in each automatic problem solver. This standardization maintains numerical precision while preventing inconsistencies in representation, ensuring that generated answers adhere to established atmospheric science conventions.
- *Incorrect option generation*: To effectively assess LLM reasoning, multiple-choice questions require plausible but incorrect distracting options that challenge the model's understanding while avoiding trivial elimination strategies [37]. We design the following mechanisms to generate incorrect options: (i) producing an incorrect answer by randomly swapping two variables in the computation; (ii) altering a single variable in the equation to generate a close but incorrect result; (iii) randomly assigning all variables within their predefined

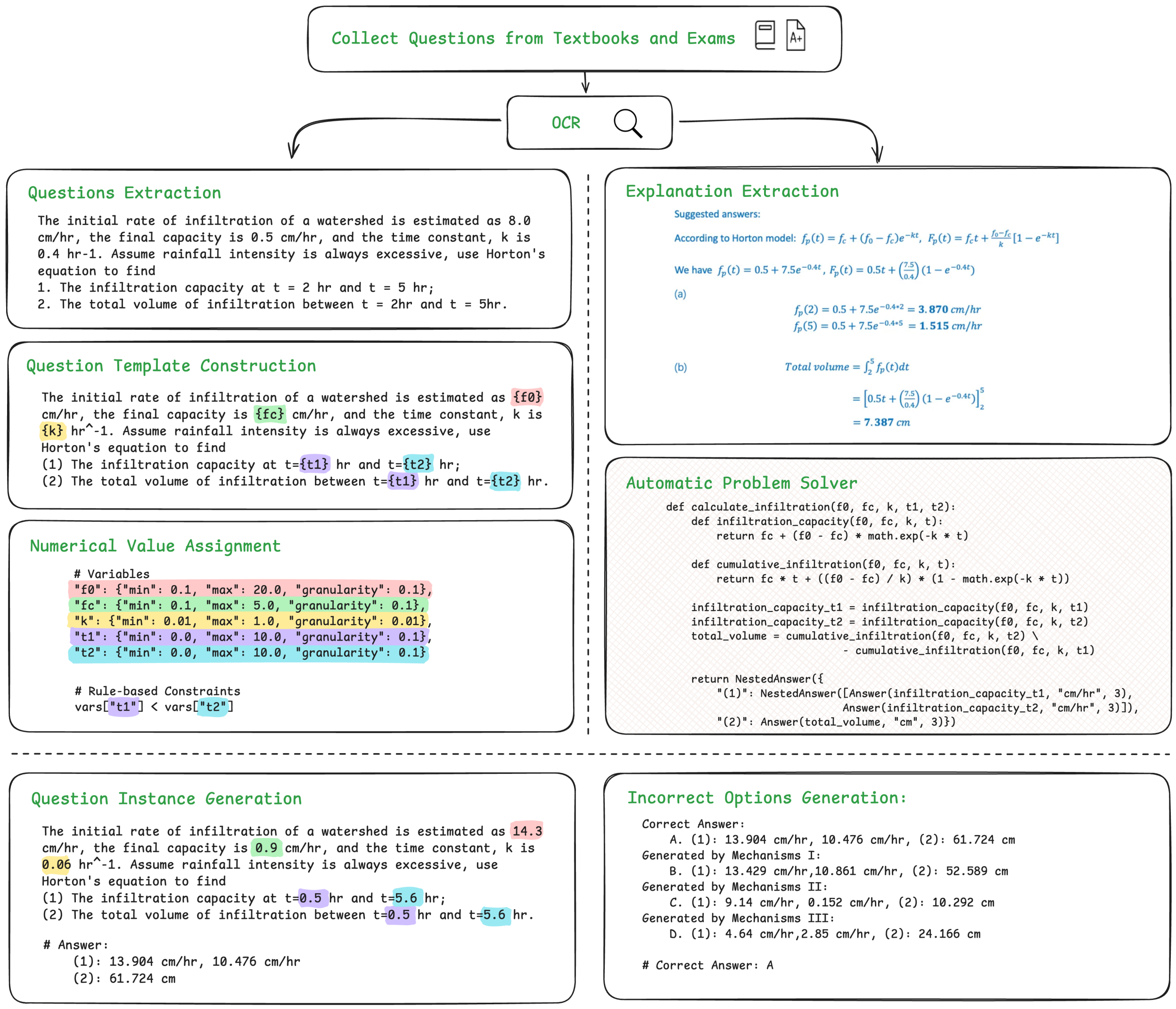

Figure 1: Construction pipeline of our template-based question generation framework. Blocks on the middle left represent the question generation process, where variables are highlighted in different colors. Blocks on the middle right depict the automatic problem solver, which derives the answer from given variables. Bottom blocks illustrate an example of a generated question and its corresponding options.

constraints, ensuring adherence to the rule-based mechanism; and (vi) if above three methods fail to generate valid incorrect options (i.e., those satisfying the scientific constraints of the rule-based mechanism), we use a default strategy, where incorrect options are generated as scaled multiples of the correct answer (e.g., $\times 2, \times 3, \times 4$).

4 Evaluation Setup

We design three main experiments to assess LLM performance on our benchmark, focusing on comprehensive performance comparison among various LLMs (*Q1*), reasoning ability variations for the tasks (*Q2*), and robustness of the benchmark results for real-world deployment (*Q3*). We enumerate these concrete questions below:

- *Q1. How do various state-of-the-art LLMs (i.e., falling into different categories of instruction, math, reasoning,* *and domain-specific models) comprehensively perform for the proposed atmospheric science benchmark?*

- *Q2. How do the models specialized in reasoning perform during inference time scaling, i.e., how can we improve the model's test accuracy by increasing the length of reasoning tokens?*
- *Q3. How robust are the benchmark results, especially when we variate the scientific numerical precision and the degree of perturbation introduced by symbolic variation?*

4.1 Constructed Benchmark Dataset

To answer the above three questions and systematically evaluate LLMs on atmospheric science tasks, we leverage our question generation framework consisting of 67 question templates to construct a concrete benchmark dataset. As we mentioned in Section 3.3, each template supports multiple variations, ensuring a diverse and scalable question set. We consider three levels of significant digits—*Low*, *Standard*, and *High*—to analyze the impact of numerical precision on LLM performance. To answer *Q1* (assessing the overall performance of various LLM categories) and *Q2* (inference scaling of reasoning models), we construct `ATMOSSCI-BENCH10`, where 10 question test sets are generated, with each test set being constructed from all question templates while maintaining predefined significant digits (Standard). To investigate model robustness *Q3*, we construct additional test sets: (i) to investigate the influence of scientific numerical precision, we generate two additional `ATMOSSCI-BENCH10`, each varying the level of significant digits (*Low*, *High*) along with *Standard* level to measure the effect of numerical precision; (ii) to evaluate the robustness under symbolic variation, we generate `ATMOSSCI-BENCH30`, which consists of 30 test sets for each question template, with controlled symbolic variations to analyze sensitivity to numerical perturbations.

4.2 Benchmark Models

To comprehensively assess LLM performance in atmospheric science, we include state-of-the-art LLMs falling into four categories: (i) instruction models, (ii) reasoning models, (iii) math models, and (iv) domain-specific models. This categorization enables a structured comparison of general-purpose, specialized, and domain-adapted models.

Instruction models. Instruction-tuned models serve as strong general-purpose baselines, optimized for following prompts and single-step inference tasks, where we include:

- GPT-4o, GPT-4o-mini [10]: OpenAI's instruction-tuned models.
- Qwen2.5-Instruct (3B, 7B, 32B, 72B) [38]: Instruction-tuned Qwen models with enhanced abilities.
- Gemma-2-9B-it, Gemma-2-27B-it [39]: Google's open-weight instruction models; along with Gemini-2.0-Flash-Exp [40], the powerful Gemini model optimized for efficiency.
- Llama-3.3-70B-Instruct, Llama-3.1-405B-Instruct-Turbo [41]: Meta's widely used instruction models.
- DeepSeek-V3 [42]: Deepseek's latest MoE-based instruction model for general tasks.

Math models. Mathematical LLMs specialize in problem-solving, computational reasoning, and theorem proving — such ability is essential for atmospheric problems. Towards this end, we include:

- DeepSeek-Math-7B-Instruct and DeepSeek-Math-7B-RL [43]: Deepseek's math-focused models trained for theorem proving.
- Qwen2.5-Math (1.5B, 7B, 72B) [44]: Qwen's recent models optimized for mathematics.

Reasoning models. Reasoning ability is the core technique to improve LLMs' performance over complicated tasks. We include the recent advanced reasoning models focus on deep logical reasoning and multi-step problem-solving:

- GPT-o1 [21]: OpenAI's reasoning-optimized model.
- QwQ-32B-Preview [23]: Reasoning model based on Qwen2.5-32B.
- Gemini-2.0-Flash-Thinking-Exp (01-21) [22]: Extended Gemini-2.0-Flash-Exp for enhanced reasoning.
- DeepSeek-R1 [24]: Deepseek's RL-trained model for complex problem-solving.

Domain-specific models. We also include some models that are specially tailored for climate-related and atmospheric science tasks by supervised fine-tuning or continuous pre-training:

- ClimateGPT-7B, ClimateGPT-70B [4]: Climate models pre-trained on domain-specific data.

5 Evaluation Results and Discussion

5.1 End-to-end Evaluation Results

Experimental setup. To comprehensively evaluate the performance of four categories of LLMs on atmospheric science tasks and assess whether AtmosSci-Bench provides a sufficiently challenging and discriminative evaluation framework, we conduct a systematic performance comparison using our `ATMOSSCI-BENCH10` benchmark across four representative LLM categories introduced in Section 4. We standardize experimental settings for each category as: (i) Reasoning models use 32K max context length, including the reasoning tokens; (ii) Instruction and math models use 8K max output tokens, balancing response quality and efficiency; (iii) Domain-specific models are set to 4K context length, the maximum capacity they support. By controlling these variables, we ensure that performance differences reflect genuine capability gaps rather than confounding factors, allowing us to validate whether AtmosSci-Bench effectively differentiates model performance and highlights reasoning proficiency.

Results and analysis. We present accuracy across different atmospheric science tasks, along with an overall performance comparison in Table 1 with three key observations:

Table 1: Comparison across four LLMs categories in terms of accuracy (%) and symbolic standard deviation for Hydrology (Hydro), Atmospheric Dynamics (AtmDyn), Atmospheric Physics (AtmosPhy), Geophysics (GeoPhy), and Physical Oceanography (PhyOcean).

Category	Model	Hydro	AtmDyn	AtmosPhy	GeoPhy	PhyOcean	Overall Acc	SymStd.
Instruction Models	Gemma-2-9B-it	24.0	13.78	15.71	11.43	20.0	15.08	4.07
	Gemma-2-27B-it	56.0	29.73	45.71	41.43	37.5	36.72	5.94
	Qwen2.5-3B-Instruct	46.0	29.19	34.28	30.0	37.5	32.09	7.71
	Qwen2.5-7B-Instruct	62.0	38.92	51.43	51.43	42.5	44.78	5.12
	Qwen2.5-32B-Instruct	58.0	47.3	64.28	62.86	55.0	53.73	6.05
	Qwen2.5-72B-Instruct-Turbo	72.0	50.0	77.86	44.29	62.5	57.61	4.73
	Llama-3.3-70B-Instruct	78.0	42.94	67.14	51.91	52.5	52.11	3.73
	Llama-3.1-405B-Instruct-Turbo	72.0	48.92	64.29	57.14	62.5	55.52	6.17
	GPT-4o-mini	48.0	42.16	58.57	40.0	40.0	45.67	4.78
	GPT-4o	68.0	51.08	74.28	60.0	55.0	58.36	5.19
	Gemini-2.0-Flash-Exp	90.0	58.11	68.57	77.14	62.5	64.93	4.29
	Deepseek-V3	94.0	57.3	73.57	64.28	62.5	64.48	6.28
Reasoning Models	QwQ-32B-Preview	88.0	60.27	87.86	74.28	60.0	69.55	4.57
	Gemini-2.0-Flash-Thinking-Exp (01-21)	100.0	78.11	85.0	91.43	80.0	82.69	3.87
	GPT-o1	100.0	82.7	92.14	92.86	87.5	87.31	3.32
	Deepseek-R1	98.0	85.68	95.0	95.71	82.5	89.4	3.48
Math Models	Deepseek-Math-7B-RL	22.0	20.54	27.86	24.29	37.5	23.58	4.44
	Deepseek-Math-7B-Instruct	36.0	28.38	33.57	30.0	40.0	30.9	4.04
	Qwen2.5-Math-1.5B-Instruct	50.0	30.0	22.86	34.29	30.0	30.45	3.24
	Qwen2.5-Math-7B-Instruct	54.0	31.35	39.28	35.71	32.5	35.22	6.14
	Qwen2.5-Math-72B-Instruct	70.0	54.87	73.57	62.86	40.0	59.85	6.27
Domain-Specific Models	ClimateGPT-7B	26.0	18.65	21.43	11.43	30.0	19.7	5.25
	ClimateGPT-70B	24.0	25.4	30.0	40.0	27.5	27.91	4.45

- AtmosSci-Bench *effectively differentiates LLM performance across categories, with reasoning models demonstrating the highest proficiency.* The results confirm that our benchmark successfully distinguishes LLM performance, particularly in assessing reasoning proficiency. Reasoning models (69.55% - 89.4%) significantly outperform instruction models (15.08% - 64.93%), demonstrating superior consistency with lower symbolic reasoning standard deviation (SymStd) [30]. DeepSeek-R1, the best-performing reasoning model, achieves 89.4% accuracy, while the top instruction model, Gemini-2.0-Flash-Exp, only reaches 64.93%, a

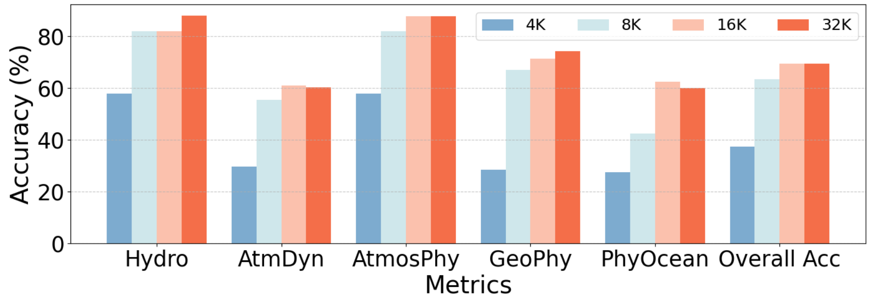

Figure 2: **Model Reasoning Step Comparison** in terms of accuracy(%) on QwQ-32B-Preview models ranging from 4K up to 32K.

substantial 24.47% gap. This clear performance variance underscores ATMOSSCI-BENCH 's ability to challenge advanced LLMs, ensuring that strong reasoning skills translate into measurable performance gains.

- *Math models do not show a clear advantage over instruction models.* Despite their specialization, math models do not significantly outperform instruction models, suggesting that mathematical optimization alone is insufficient for solving atmospheric science challenges.
- *Domain-specific models underperform despite climate specialization, indicating a need for reasoning-augmented approaches.* Domain-specific models perform poorly despite their climate specialization, with CLIMATEGPT-7B and CLIMATEGPT-70B achieving only 19.7% and 27.91% accuracy, respectively. Their inability suggests that specialized training alone may not compensate for weak reasoning capabilities. This highlights there is a need for a reasoning model in this atmospheric science. ATMOSSCI-BENCH provides a rigorous evaluation framework to guide the development of such reasoning-augmented domain models, addressing the limitations of existing approaches.

In conclusion, to answer *Q1* regarding the overall performance of various LLM categories, our evaluation reveals that *reasoning models significantly outperform instruction, math, and domain-specific models in atmospheric science tasks, highlighting their superior adaptability to advanced reasoning challenges, while domain-specific models struggle despite specialized training.*

5.2 Inference Scaling for Reasoning Models

Experimental setup. To answer *Q2*, i.e., whether increasing the length of reasoning tokens improves the performance of reasoning models, we conduct an inference time scaling evaluation on ATMOSSCI-BENCH10 using the QWQ-32B-PREVIEW model, varying its reasoning token limits from 4K up to 32K. By systematically increasing the token limit, we aim to determine whether a longer inference process leads to higher accuracy and whether there exists an optimal threshold beyond which additional tokens provide minimal benefit.

Results and analysis. As shown in Figure 2, increasing the reasoning token limit generally improves model accuracy, but the gains diminish beyond a certain threshold. Across all evaluated metrics, including overall accuracy, performance is consistently lower at 4K tokens, improves significantly at 8K and 16K tokens, and then plateaus beyond 16K tokens, with 32K tokens offering only marginal improvement. This trend suggests that while extending reasoning length enhances model performance up to a certain point, it further increases yield, diminishing returns without proportional accuracy gains. Thus, our answer to *Q2* is that *increasing the length of reasoning tokens improves model accuracy up to 16K tokens, beyond which performance gains diminish, indicating an optimal threshold for inference time scaling.*

5.3 Robustness of ATMOSSCI-BENCH

To evaluate the robustness of ATMOSSCI-BENCH (*Q3*), we conduct experiments to assess: (i) robustness against variations in numerical precision and (ii) robustness to different degrees of perturbation introduced by symbolic variation.

Table 2: Performance Comparison Among Different Significant digits in terms of accuracy (%) and symbolic standard deviation.

Model	Sig. Digits	Overall Acc	SymStd.
Qwen2.5-7B-Instruct	Low	43.58	3.84
	Standard	44.78	5.12
	High	38.51	6.56
Qwen2.5-Math-7B-Instruct	Low	36.12	4.03
	Standard	35.22	6.14
	High	35.82	4.1
QwQ-32B-Preview	Low	68.51	4.42
	Standard	69.55	4.57
	High	71.34	4.03

Experiments for numerical precision. (*Setup*). We hypothesize that increasing the significant digits in numerical variables may increase the difficulty of test sets or deteriorate the performance due to the fragility of the arithmetic ability of LLMs [32, 33, 34]. To test this, we conduct our experiment across three configurations of significant digits, assessing the QWEN-class models from three different LLM categories, including QWEN2.5-7B-INSTRUCT, QWEN2.5-MATH-7B-INSTRUCT and QWQ-32B-PREVIEW. We variate three degrees of scientific numerical precision: (i) *Low* significant digit, where digits are 0–3 digits shorter than the "Standard" configuration; (ii) *Standard* significant digit that predefined in our template-based question generation framework, which aims to introduce diversity while maintaining realistic and valid values; and (iii) *High* significant digit that extends the significant digits by an additional 10 digits beyond the "Standard" configuration to facilitate a more rigorous comparison.

(*Results and analysis*). Table 2 reveals distinct performance trends among the three models when varying the number of significant digits. Among the three models, the reasoning model QWQ-32B-PREVIEW indicates performance improvement with increasing numerical precision, indicating its rigorousness in scientific tasks—where real analytics could be enabled instead of simple pattern matching. The math-specialized model, i.e., QWEN2.5-MATH-7B-INSTRUCT, remains stable across different levels of significant digits, with minimal variations in accuracy, potentially reflecting its specialization in numerical processing. In contrast, the standard instruction model, i.e., QWEN2.5-7B-INSTRUCT, suffers from a significant drop in accuracy as numerical precision increases, suggesting its risk of reliance on pattern matching rather than desired numerical reasoning, making it more fragile with more precise scientific numerical computation. Conclusively, in terms of robustness raised in *Q3*, we believe that the *reasoning model demonstrates higher resilience to extended numerical sequences, while the instruction model exhibits significant sensitivity to variations in numerical precision.*

Experiments for symbolic variation. (*Setup*). Inspired by `GSM-Symbolic` [30], which demonstrates that modifying numerical variables in the `GSM8K` dataset led to significant performance drops, suggesting that LLMs may rely on pattern matching rather than genuine logical reasoning. We aim to assess the robustness of advanced reasoning models under varying degrees of symbolic perturbation. To examine this, we evaluate three reasoning models—DEEPSEEK-R1, GEMINI-2.0-FLASH-THINKING-EXP (01-21), and QWQ-32B-PREVIEW—on `ASTMOSSCI-BENCH30`, consisting of 30 question test sets that vary in numerical variables. We systematically modify numerical variables within a scientifically reasonable range, introducing controlled variations to assess whether performance remains stable or degrades significantly with perturbation.

(*Results and analysis*). Figure 3 illustrates the empirical performance distribution of reasoning models on `ASTMOSSCI-BENCH30`. We observe that for both DEEPSEEK-R1 and QWQ-32B-PREVIEW, the accuracy of the original question set (dashed line in Figure 3) is approximately one standard deviation away from the mean accuracy across perturbed instances, indicating a notable shift. In contrast, GEMINI-2.0-FLASH-THINKING-EXP (01-21) exhibits a more minor deviation, with accuracy within one standard deviation but skewed toward the right side of the distribution. As the degree of numerical perturbation increases, we observe a consistent downward trend in model performance, reinforcing the notion that LLMs, even those specialized in reasoning, could still struggle with symbolic variation. To answer <u>*Q3*</u> w.r.t symbolic variation, the results indicate that *the reasoning models evaluated in our benchmark could still be under the risk of insufficient robustness under symbolic perturbation, as increasing the degree of variation leads to significant and often unpredictable drops in accuracy.* This suggests that our benchmark effectively reveals weaknesses in reasoning models' ability to generalize beyond pattern-matching strategies. Furthermore, these findings could imply that the tested reasoning models are likely trained on in-distribution data sources, such as standard textbooks in atmospheric science. Their performance may thus be heavily influenced by pattern-matching within familiar distributions rather than true logical reasoning. The observed performance degradation under perturbation further highlights the need for robust evaluation frameworks that test models beyond their training distributions.

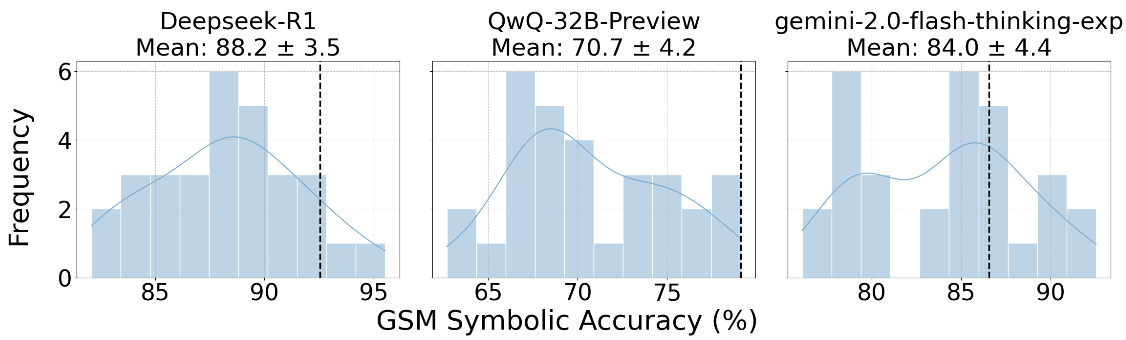

Figure 3: Performance distribution among reasoning LLMs (DEEPSEEK-R1, QWQ-32B-PREVIEW and GEMINI-2.0-FLASH-THINKING-EXP) on ASTMOSSCI-BENCH30. The Y-axis represents the frequency of the symbolic test sets achieving the accuracy shown on the X-axis. The black vertical dash lines denote the accuracy of the original question set. GEMINI-2.0-FLASH-THINKING-EXP refers to the EXP-01-21 version.

6 Conclusion

In this paper, we introduced ATMOSSCI-BENCH, a novel benchmark designed to systematically evaluate the reasoning and problem-solving capabilities of LLMs in atmospheric science. Our benchmark covers five core categories—hydrology, atmospheric dynamics, atmospheric physics, geophysics, and physical oceanography—through a scalable, template-based question generation framework that ensures diversity and complexity in multiple-choice question assessments. By conducting a comprehensive evaluation across four distinct model categories—instruction-tuned models, advanced reasoning models, math-augmented models, and domain-specific climate models—we provide key insights into the strengths and limitations of LLMs in addressing atmospheric science problems. Our findings highlight that reasoning models outperform other categories, demonstrating stronger problem-solving and reasoning capabilities in the domain of atmospheric science. This also underscores the benchmark's effectiveness in differentiating models. We believe that ATMOSSCI-BENCH (where all the implementations are fully open-sourced) can serve as an essential step toward advancing the application of LLMs in climate-related decision-making by offering a standardized and rigorous evaluation framework for future research.

References

[1] T Brown, B Mann, N Ryder, M Subbiah, JD Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, et al. Language models are few-shot learners advances in neural information processing systems 33. 2020.

[2] Vincent Nguyen, Sarvnaz Karimi, Willow Hallgren, Ashley Harkin, and Mahesh Prakash. My climate advisor: An application of NLP in climate adaptation for agriculture. In Dominik Stammbach, Jingwei Ni, Tobias Schimanski, Kalyan Dutia, Alok Singh, Julia Bingler, Christophe Christiaen, Neetu Kushwaha, Veruska Muccione, Saeid A. Vaghefi, and Markus Leippold, editors, *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 27–45, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[3] Lujia Zhang, Hanzhe Cui, Yurong Song, Chenyue Li, Binhang Yuan, and Mengqian Lu. On the opportunities of (re)-exploring atmospheric science by foundation models: A case study. *arXiv preprint arXiv:2407.17842*, 2024.

[4] David Thulke, Yingbo Gao, Petrus Pelser, Rein Brune, Rricha Jalota, Floris Fok, Michael Ramos, Ian van Wyk, Abdallah Nasir, Hayden Goldstein, et al. Climategpt: Towards ai synthesizing interdisciplinary research on climate change. *arXiv preprint arXiv:2401.09646*, 2024.

[5] Charles Cao, Jie Zhuang, and Qiang He. LLM-assisted modeling and simulations for public sector decision-making: Bridging climate data and policy insights. In *AAAI-2024 Workshop on Public Sector LLMs: Algorithmic and Sociotechnical Design*, 2024.

[6] 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. *arXiv preprint arXiv:2103.03874*, 2021.

[7] Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*, 2023.

[8] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.

[9] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[10] OpenAI. Openai gpt-4o, 2024.

[11] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv:2403.05530*, 2024.

[12] Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024.

[13] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[14] Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. Climatebert: A pretrained language model for climate-related text. *arXiv preprint arXiv:2110.12010*, 2021.

[15] Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023.

[16] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Graphcast: Learning skillful medium-range global weather forecasting. *arXiv preprint arXiv:2212.12794*, 2022.

[17] Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

[18] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast. *arXiv preprint arXiv:2211.02556*, 2022.

[19] Kang Chen, Tao Han, Junchao Gong, Lei Bai, Fenghua Ling, Jing-Jia Luo, Xi Chen, Leiming Ma, Tianning Zhang, Rui Su, et al. Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *arXiv preprint arXiv:2304.02948*, 2023.

[20] Lei Chen, Xiaohui Zhong, Feng Zhang, Yuan Cheng, Yinghui Xu, Yuan Qi, and Hao Li. Fuxi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science*, 6(1):190, 2023.

[21] OpenAI. Learning to reason with llms. `https://openai.com/index/learning-to-reason-with-llms/`. Accessed: 2025-01-28.

[22] Google DeepMind. Gemini 2.0 flash thinking experimental. `https://deepmind.google/technologies/gemini/flash-thinking/`. Accessed: 2025-01-28.

[23] Qwen Team. Qwq: Reflect deeply on the boundaries of the unknown, November 2024.

[24] 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, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.

[25] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[26] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[27] 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. *arXiv preprint arXiv:2311.12022*, 2023.

[28] MAA. Maa invitational competitions. Mathematical Association of America, February 2024.

[29] Veeramakali Vignesh Manivannan, Yasaman Jafari, Srikar Eranky, Spencer Ho, Rose Yu, Duncan Watson-Parris, Yian Ma, Leon Bergen, and Taylor Berg-Kirkpatrick. Climaqa: An automated evaluation framework for climate foundation models. *arXiv preprint arXiv:2410.16701*, 2024.

[30] Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*, 2024.

[31] Kun Qian, Shunji Wan, Claudia Tang, Youzhi Wang, Xuanming Zhang, Maximillian Chen, and Zhou Yu. Varbench: Robust language model benchmarking through dynamic variable perturbation. *arXiv preprint arXiv:2406.17681*, 2024.

[32] Haotong Yang, Yi Hu, Shijia Kang, Zhouchen Lin, and Muhan Zhang. Number cookbook: Number understanding of language models and how to improve it. *arXiv preprint arXiv:2411.03766*, 2024.

[33] Eli Schwartz, Leshem Choshen, Joseph Shtok, Sivan Doveh, Leonid Karlinsky, and Assaf Arbelle. Numerologic: Number encoding for enhanced llms' numerical reasoning. *arXiv preprint arXiv:2404.00459*, 2024.

[34] Aaditya K Singh and DJ Strouse. Tokenization counts: the impact of tokenization on arithmetic in frontier llms. *arXiv preprint arXiv:2402.14903*, 2024.

[35] Nishant Balepur and Rachel Rudinger. Is your large language model knowledgeable or a choices-only cheater? In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 15–26, 2024.

[36] Mathpix. Mathpix: Ai-powered ocr for math and science, 2024. Accessed: 2024-01-29.

[37] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark, 2024.

[38] Qwen Team. Qwen2.5: A party of foundation models, September 2024.

[39] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric

Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024.

[40] Google DeepMind. Gemini 2.0 flash experimental. `https://deepmind.google/technologies/gemini/flash/`. Accessed: 2025-01-28.

[41] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross

Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked,

Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024.

[42] 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, 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, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2024.

[43] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[44] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.