# ASTROMATHBENCH: EVALUATING THE MATHEMATICAL REASONING CAPABILITIES OF LLMS FOR ASTRONOMICAL PROBLEMS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Large language models (LLMs) have demonstrated remarkable capabilities in advanced mathematical reasoning tasks. However, computational reasoning in astronomy with LLMs presents unique challenges yet has received much less attention. Existing benchmarks often fail to capture LLMs' cross-disciplinary reasoning abilities and are prone to data contamination. To address these problems, we propose AstroMathBench, a benchmark for evaluating LLMs' computational reasoning capabilities in astronomy. AstroMathBench covers five task types in astronomical reasoning: formula calculation, formula derivation, unimodal computation, multimodal computation, and code-based calculation. Beyond task-level evaluation, we decompose the reasoning process into five cognitive dimensions, enabling fine-grained assessment of LLMs' capability boundaries. Moreover, we design a cross-validation-based pipeline to mitigate data contamination by separating private and public data and employing isomorphic variant generation. The resulting dataset contains 554 seed problems that have been manually verified and checked for contamination, with 90.1% fully code-formalized. Our evaluation of 10 leading LLMs finds that the highest accuracy reaches only 76%, highlighting the significant challenges of computational reasoning in astronomy and the need for LLMs with stronger cross-disciplinary reasoning abilities. In general, we hope that AstroMathBench will drive advances in AI for astronomy.

## 1 INTRODUCTION

Large language models (LLMs) have demonstrated remarkable capabilities across diverse domains, such as dialog (Ou et al., 2024), tool usage (Qin et al., 2024), and code generation (Jiang et al., 2026). Among these capabilities, mathematical reasoning serves as a principal measure of LLMs' problem-solving skills (Ahn et al., 2024). As such, several benchmarks have been proposed to perform fine-grained assessments of LLMs' mathematical abilities. For example, GSM8K (Cobbe et al., 2021) focuses on elementary arithmetic reasoning and serves as an entry-level test of mathematical problem solving; MATH (Hendrycks et al., 2021) extends the evaluation to high school and Olympiad-level problems across algebra, geometry, combinatorics, and calculus that require more complex multi-step reasoning. Moreover, several benchmarks (Amini et al., 2019; Gao et al., 2025; Xu et al., 2024) examine different dimensions of mathematical reasoning, ranging from language understanding and symbolic manipulation to structured problem solving. Collectively, they have driven methodological advances, including Chain-of-Thought (CoT) prompting (Wei et al., 2022), self-consistency decoding (Wang et al., 2023), and verifier-guided reasoning (Li et al., 2023), thereby improving LLM performance in mathematics.

Despite these advances, a systematic evaluation of computational reasoning in astronomy remains largely underexplored. Existing benchmarks are insufficient for assessing LLMs' capabilities in astronomy-related problem-solving because of the fundamental differences between astronomical computation and standard mathematical reasoning. First, astronomical problems rely on domain-specific knowledge and require integrating mathematical operations with astronomical constants, coordinate systems, physical models, and celestial mechanics. Second, astronomical reasoning chains span tightly coupled, multi-dimensional physical processes. Typical tasks involve interdependent formulas, numerical methods, and unit or time-system conversions, whose interactions cannot



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

be fully captured by purely abstract mathematical logic. Third, astronomical results carry explicit physical meaning and application constraints. Unlike symbolic answers, astronomical computations yield physically grounded quantities, where even minor computational errors can render physically invalid results. Fourth, astronomical tasks are sensitive to numerical precision and require rigorous control over units and accuracy thresholds; yet, these requirements are absent from existing benchmarks.

Consequently, computational reasoning in astronomy presents unique challenges for LLMs. Recently, several studies have evaluated LLMs in astronomy-related contexts, such as question answering (Li et al., 2025), multimodal understanding (Shi et al., 2025), and visualization (Joseph et al., 2025). However, these benchmarks primarily emphasize conceptual understanding and auxiliary skills, rather than computational reasoning. Typical tasks, such as time-system conversion, coordinate transformation, celestial position prediction, and photometric calculation, have not yet been covered by them. In addition, they often suffer from data contamination (Deng et al., 2024).

To bridge these gaps, we propose AstroMathBench, a new benchmark for evaluating computational reasoning in astronomy with LLMs. AstroMathBench targets the fundamental requirements for astronomical computation by organizing problems into five task categories: formula calculation, formula derivation, unimodal (text-only) computation, multimodal (text + image) computation, and code-based calculation. Beyond overall performance measures, it probes the LLM's boundaries across five cognitive dimensions: problem comprehension, condition robustness, knowledge consistency, reverse reasoning, and anti-interference. These dimensions reveal deeper limitations in reasoning that are often obscured by accuracy-based evaluations. Moreover, we adopt a cross-validation-based pipeline that separates private and public data and employs isomorphic variant generation to mitigate data contamination and memorization effects.

Finally, we conduct a comprehensive evaluation of ten state-of-the-art LLMs on AstroMathBench, providing fine-grained analyses of reasoning trajectories, failure modes, and model behaviors. The results show that the performance of the evaluated LLMs is inferior (with the highest accuracy of only 76%), highlighting significant challenges for computationally intensive astronomical tasks that require complex multi-step cross-disciplinary reasoning. We hope that AstroMathBench will offer actionable insights to advance computational reasoning with LLMs in scientific contexts.

## 2 The AstroMathBench Framework

As illustrated in Figure 1, we propose a pipeline to construct a computational reasoning benchmark in astronomy, covering all stages from data preparation to evaluation. The pipeline consists of three stages: *i) data collection and preprocessing*, *ii) data pollution prevention*, and *iii) random derivation and testing*. Next, we will describe each stage in more detail.

### 2.1 Data Collection and Preprocessing

**Data Sources** The original questions in AstroMathBench are collected from four main sources: *i)* existing benchmarks, *ii)* astronomical Olympiads, *iii)* professional books, and *iv)* human experts. We confirm that the built dataset is free to use and distribute under the CC-BY-4.0 license.

**Task Types** To assess the performance of LLMs for computational reasoning in astronomy, we cover five task types across different dimensions:

- **Formula Calculation (FC)**: This type of question asks an LLM to apply established astronomical or physical formulae to compute specific numerical results. It tests the LLM's ability to identify relevant equations, substitute appropriate parameters, handle units, and perform accurate mathematical calculations. Basic mathematical skills and attention to numerical precision are essential for this type of question.
- **Formula Derivation (FD)**: This type of question asks an LLM to derive a mathematical or physical formula related to astronomy from fundamental principles, known laws, and given conditions. It requires the LLM to understand astronomical concepts, physical assumptions, and logical steps in the derivation process.
- **Unimodal Computation (UC)**: This type of question evaluates an LLM's ability to perform multi-step reasoning that involves quantitative analysis in astronomical contexts.

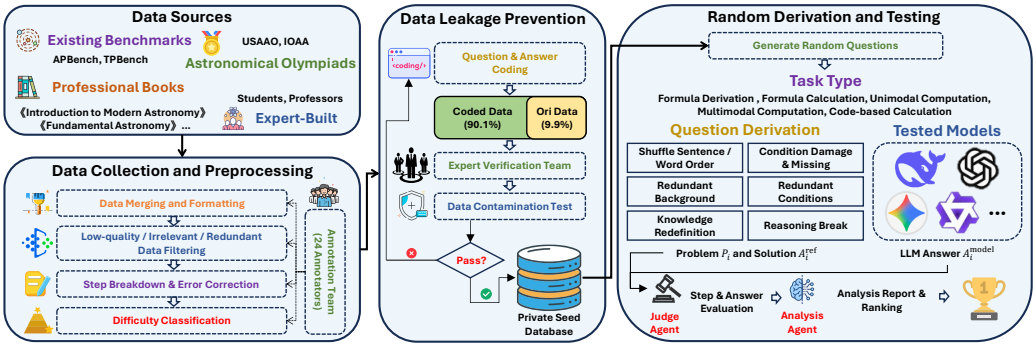



Figure 1: An overview of the AstroBench framework. The proposed framework consists of four stages: data sources, data collection and preprocessing, data leakage prevention, and random derivation and testing. First, data are collected from multiple sources. The collected samples undergo unified cleaning, low-quality filtering, annotation, and difficulty classification. Next, the curated data are transformed into executable code representations and verified by domain experts. After contamination testing, qualified samples are securely stored in a private database. During evaluation, the target model receives randomly derived test variants from the database, which are assessed by a Judge Agent that automatically analyzes model outputs and generates evaluation reports.

Rather than focusing solely on the final results, it emphasizes the reasoning process that connects physical concepts, assumptions, and intermediate calculations. The LLM should decompose the problem into subproblems, justify each step, and apply structured computational thinking to reach a correct conclusion.

- **Multimodal Computation (MC)**: This type of question asks an LLM to combine visual information (e.g., images, diagrams, or plots) with textual descriptions to perform calculations or quantitative reasoning. It requires the LLM to correctly interpret visual data, extract measurable or symbolic information, and integrate it with astronomical knowledge and mathematical operations via multimodal reasoning.
- **Code-based Calculation (CC)**: This type of question requires an LLM to solve computational problems by coding. The LLM should translate astronomical models or equations into algorithms, implement numerical methods, and execute code.

**Data Distribution and Preprocessing** The task-type and difficulty-level distributions of AstroMathBench are provided in Appendix D. AstroMathBench is constructed through an expert workflow involving data merging, filtering, annotation correction, and formula-based difficulty classification. The red-highlighted categories (CC and MC) are derived from base tasks through category expansion and processing rather than collected independently. Detailed preprocessing procedures, difficulty coefficients, role definitions, and task examples are provided in Appendices B and L.

### 2.2 Data Leakage Prevention

To mitigate the impact of data contamination and prevent future data leakage, we introduce a code-based mutation framework.

**Code-Based Mutation** Traditional data augmentation methods often rely on simple rule-based transformations or random sampling, which may not effectively generate truly novel data instances. In contrast, our proposed code-based mutation approach leverages programming perturbation variables and constants to construct a function variant generator. The generator is designed to create new data by manipulating the underlying structure of the original problems. The generator operates at the code level, enabling it to generate an infinite number of problem instances and solution processes that are at the same difficulty level as the original data but have not been seen before. For example, we introduce controlled variable assignments into the problem-solving code, yielding a diverse set of new problems that preserve the core semantic and structural characteristics of the original ones. Specifically, given an original problem $Q$ with ground-truth answer $A$, Algorithm 1 generates a perturbed problem-answer pair $(Q_i, A_i)$ by modifying numerical constants while preserving physical

validity and reasoning structure. *i) Setup:* Let $\mathcal{C} = \{c_1, \ldots, c_n\}$ denote the numerical constants in $Q$. The solution process is represented as an ordered set of formulas $\mathcal{F} = \{f_1, \ldots, f_k\}$, where intermediate results are stored in $\mathcal{R}$. *ii) Perturbation:* Each constant $c_j$ is perturbed by sampling $c_j^{(i)}$ from the uniform distribution $\mathcal{U}\big([\alpha_j c_j, \beta_j c_j]\big)$. The sampling process is repeated until all domain-specific and physical constraints are satisfied, yielding the perturbed constant set $\mathcal{C}^{(i)}$. *iii) Recomputation:* Using $\mathcal{C}^{(i)}$, it re-evaluates $\mathcal{F}$ sequentially. The output of the final formula is taken as the new answer $A_i$. *iv) Output:* The new problem statement $Q_i$ is obtained by replacing $\mathcal{C}$ with $\mathcal{C}^{(i)}$ in $Q$. The algorithm finally outputs $(Q_i, A_i)$, which preserves the original reasoning structure with controlled numerical variation. Finally, among the original 554 basic problems, 90.1% can successfully generate function variants, while the remaining 9.9% cannot and thus remain in their original form. The high success rate of variant generation confirms the effectiveness of the code-based mutation mechanism.

**Private Seed Database Construction** We subject both the generated function variants and the original data to expert validation and data contamination testing. Problems that pass contamination testing are compiled in the private seed database. This database serves as a clean data source for LLMs. Each time AstroMathBench participates in a test, the data it uses is generated from the seed database. This effectively alleviates data contamination caused by directly using the original problems. Using data from the seed database ensures that LLMs are tested on unseen, uncontaminated data, thereby providing a more accurate assessment of their performance and generalization.

**Memory Contamination Testing for Pre-existing Data** In addition, to prevent data contamination, we follow the memory-contamination testing method for tabular data described in (Bordt et al., 2023) to test our data. We have made several changes to the testing process to better align with the characteristics of our problem-based data. Specifically, we set up the following tests: *i) Random Test:* We present consecutive content from a random position within the problem and require the model to complete the next sentence. This test aims to evaluate whether the model has memorized specific sentence-level patterns in the data. By using a random position, we increase the test's difficulty and improve its ability to detect potential contamination. The similarity metric measures how closely the model's completion matches the expected next sentence, providing a quantitative assessment of contamination. *ii) First Word Test:* We provide consecutive content from a random position in the problem and ask the model to complete the next word. This test focuses on detecting word-level memorization. The exact match rate is used to assess whether the model accurately predicts the next word, providing a finer-grained measure of contamination. *iii) Specific Variable Value and Name Test:* We extract specific variable names and their corresponding values from the problem, erase them from the original problem, and then request the model to provide the values and names of the specified variables. This test is designed to detect whether the model has memorized specific variable-related information in the problems. The use of both exact- and partial-match rates allows us to capture different levels of contamination. Based on these tests, we can assess whether the pre-trained models have already been exposed to the corresponding data.

### 2.3 Cognitive Derivation

Inspired by the evaluation from the lens of human cognition in (Liu et al., 2025), we derive new assessment questions and evaluation dimensions at cognitive levels. This will help test LLMs' problem-solving more comprehensively.

To evaluate the computational reasoning abilities of LLMs, we design a set of five evaluation dimensions inspired by fundamental processes in human cognition. Human problem-solving is not a monolithic capability but rather a coordinated process involving *problem understanding*, *condition modeling*, *structured reasoning*, *knowledge consistency*, and *interference suppression*. Drawing on the cognitive perspective, we propose an evaluation framework that decomposes LLM reasoning performance across five axes, enabling a finer-grained, process-aware assessment.

**(1) Problem Comprehension:** We first evaluate whether the model constructs a semantic representation of the problem independent of its linguistic form. In human cognition, comprehension precedes reasoning: individuals can solve the same problem even when sentences are reordered differently, as long as the underlying meaning remains intact. On the contrary, when the word or-

der is disrupted, the semantics and coherence of the sentence become unreadable to individuals, and therefore, the problem becomes unsolvable. To mirror this capability, we assess models under perturbations such as sentence-order shuffling and word-order variation. Performance degradation under these conditions indicates reliance on sentence- and word-order patterns rather than on semantic understanding, thereby directly limiting computational reasoning.

**(2) Condition Robustness:** We then focus on the model's ability to assess the validity of problem conditions. Human reasoners do not attempt to 'repair' a problem when essential conditions are missing or when key values are corrupted; instead, they recognize that the problem is under-specified or invalid. Accordingly, we evaluate LLMs under settings of condition omission and value damage. In cases of missing or damaged conditions, the correct cognitive response is to identify the problem as unsolvable.

**(3) Knowledge Consistency:** This reflects humans' ability to reason under altered or context-dependent rules. Humans can flexibly revise prior knowledge when new definitions are explicitly provided. Accordingly, we evaluate LLMs under knowledge redefinition settings to test whether models can adhere to task-specific rules when the problem remains well-defined.

**(4) Reverse Reasoning:** This captures the structured and multi-stage nature of human reasoning. Many cognitive tasks cannot be solved through direct forward inference alone; instead, humans frequently reason backward from goals, construct intermediate hypotheses, or bridge gaps in incomplete reasoning chains. To reflect this process, we introduce reasoning breaks that require forward, intermediate, and backward reasoning. This dimension assesses whether LLMs possess flexible reasoning strategies beyond linear CoT generation, a key requirement for computational tasks.

**(5) Anti-Interference:** We evaluate a model's ability to suppress irrelevant or misleading information, a hallmark of human attentional control. Real-world reasoning rarely occurs in clean environments; humans selectively attend to task-relevant information while filtering out redundant background and extraneous conditions. We simulate such scenarios by injecting irrelevant context and redundant constraints to test whether LLMs can maintain focus on causally relevant information. Robust performance under interference indicates stronger robustness in reasoning.

To sum up, these five cognitive dimensions provide an evaluation that aligns LLM reasoning assessment with human cognition. Rather than measuring correctness alone, our framework aims to diagnose how models reason, providing deeper insights into their strengths and limitations in computational reasoning.

### 2.4 Random Derivation and Automated Testing

In the AstroMathBench evaluation protocol, test samples are obtained by querying the private seed database, which randomly generates variant instances across five task types. The difficulty level of the generated instances remains consistent across different runs, ensuring fairness and reproducibility. These variant instances constitute the core benchmark dataset and are primarily used to assess a model's computational and reasoning capabilities in astronomy. Building on this base set, additional variants are generated by problem-derivation code to further evaluate LLMs' cognitive abilities from five dimensions. Each final evaluation instance consists of three components: a variant problem $P_i$, a reference solution $A_i^{\text{ref}}$, and the solution produced by the evaluated model $A_i^{\text{model}}$. The assessment is conducted using an LLM-as-a-Judge method, which performs both process-level and result-level evaluations. The resulting evaluation data are subsequently processed by an analysis agent that automatically generates multidimensional analytical reports and provides insights into the model's performance across different evaluation axes. The evaluation prompts are provided in Appendix N.

## 3 Experiments

### 3.1 Experimental Setup

We evaluated ten LLMs, consisting of five closed-source models, i.e., GPT-5.6-Luna (OpenAI Team, 2026), Qwen3-Max (Qwen Team, 2025b), Gemini-2.5-Flash (Gemini Team, 2025a), As-

| Task | Metric | Mode | GPT-5.6-Luna | AstroOne | Qwen3-Max | Gemini-2.5-Flash | Claude-haiku-4.5 | AstroSage3-8B | Llama-3.1-8B | Llama-3.1-70B | Qwen2.5-Math-72B | Qwen3-VL-32B |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **FD** | Score | *Ori* | 9.10 | **9.13** | 9.04 | 8.68 | 7.71 | 2.63 | 4.03 | 5.24 | 6.90 | 8.31 |
| | | *+Think* | 9.41 | 6.35 | 9.15 | 8.52 | 8.03 | – | – | – | – | 7.14 |
| | | *+CoT* | 9.38 | 8.78 | 9.03 | 8.54 | 7.96 | 2.97 | 3.74 | 5.21 | 7.09 | 8.25 |
| | | *-System* | 9.18 | 8.08 | 8.85 | 8.75 | 7.26 | 3.00 | 4.11 | 5.18 | 7.00 | 8.57 |
| **UC** | Score | *Ori* | **8.19** | 6.68 | 7.97 | 4.04 | 6.89 | 2.65 | 4.12 | 4.17 | 6.29 | 4.42 |
| | | *+Think* | 8.16 | 3.16 | 8.12 | 3.82 | 7.04 | – | – | – | – | 6.33 |
| | | *+CoT* | 8.13 | 6.40 | 8.09 | 4.51 | 6.91 | 2.77 | 4.16 | 4.39 | 6.44 | 4.59 |
| | | *-System* | 8.18 | 6.55 | 8.03 | 3.51 | 6.92 | 2.59 | 3.89 | 3.51 | 6.50 | 4.61 |
| | Acc | *Ori* | **0.74** | 0.54 | 0.66 | 0.32 | 0.47 | 0.04 | 0.06 | 0.12 | 0.34 | 0.31 |
| | | *+Think* | 0.75 | 0.32 | 0.67 | 0.30 | 0.47 | – | – | – | – | 0.52 |
| | | *+CoT* | 0.73 | 0.53 | 0.68 | 0.35 | 0.49 | 0.03 | 0.07 | 0.16 | 0.34 | 0.37 |
| | | *-System* | 0.76 | 0.54 | 0.68 | 0.27 | 0.46 | 0.03 | 0.07 | 0.14 | 0.34 | 0.34 |
| **FC** | Score | *Ori* | **7.45** | 6.66 | 7.11 | 5.90 | 6.79 | 2.81 | 4.20 | 4.60 | 6.21 | 4.86 |
| | | *+Think* | 7.57 | 4.68 | 7.34 | 6.01 | 6.52 | – | – | – | – | 6.25 |
| | | *+CoT* | 7.77 | 6.74 | 7.49 | 5.32 | 6.82 | 2.99 | 4.20 | 3.52 | 6.12 | 3.83 |
| | | *-System* | 7.46 | 6.22 | 7.26 | 5.93 | 6.46 | 2.65 | 3.90 | 4.66 | 6.27 | 4.13 |
| | Acc | *Ori* | **0.70** | 0.58 | 0.61 | 0.49 | 0.39 | 0.04 | 0.13 | 0.16 | 0.30 | 0.30 |
| | | *+Think* | 0.72 | 0.36 | 0.58 | 0.42 | 0.36 | – | – | – | – | 0.51 |
| | | *+CoT* | 0.75 | 0.59 | 0.64 | 0.39 | 0.45 | 0.00 | 0.16 | 0.16 | 0.33 | 0.22 |
| | | *-System* | 0.71 | 0.54 | 0.59 | 0.46 | 0.39 | 0.01 | 0.13 | 0.19 | 0.26 | 0.26 |

Table 1: Overall performance of each tested model on FD, UC, and FC tasks. "Ori" denotes the original base model; "+Think" applies a slow-thinking evaluation protocol, "+CoT" uses CoT prompting, and "-System" removes the system prompt. The values in red indicate anomalous results due to erroneous outputs. Bold font marks the best performance under the "Ori" setting.

troOne (Zhejiang Lab, 2024), and Claude-haiku-4.5 (Anthropic, 2025), and five open-source models, i.e., Llama-3.1-8B, Llama-3.1-70B (Touvron et al., 2023), Qwen2.5-Math-72B (Qwen Team, 2024), Qwen3-VL-32B (Qwen Team, 2025a), and AstroSage3-8B (de Haan et al., 2025). During evaluation, we used DeepSeek-V3.2 (DeepSeek-AI, 2025) as the judge model. Implementation details are described in Appendix E.

We employ a combination of process- and results-oriented measures to assess model performance. The process-oriented evaluation includes five indicators, each on a scale from 0 to 10: *correctness*, *completeness*, *relevance*, *clarity & coherence*, and *precision & detail*. These five measures, from general to specific, assess the consistency between the model's problem-solving process and the ground truth across different dimensions. We then calculate the *overall score* as the average score for the five indicators. The results-oriented evaluation uses *accuracy* to assess the ratio of each LLM's answers that align with the ground truth.

### 3.2 Overall Results

We first compare the overall performance of each LLM for different task types. The results for FD, UC, and FC tasks are shown in Table 1.

For the FD task, when reasoning strategies are not used, AstroOne, a model specific to astronomy, achieves the best performance. This indicates that AstroOne exhibits strong domain knowledge representation and recall capabilities relative to models of similar scale, showing the effectiveness of domain-focused pretraining. Among open-source models, Qwen3-VL-32B achieves performance comparable to closed-source models, highlighting its potential in scientific domains.

For the UC task, which is more difficult than FD, AstroOne demonstrates clear limitations in computational reasoning, underperforming leading closed-source models. GPT-5.6-Luna achieves the best performance on this task; however, it, as well as all other tested models, exhibits noticeable performance degradation. This suggests that existing models still lack sufficient capability to perform complex scientific reasoning. Even the top-performing model, GPT-5.6-Luna, attains only 74% accuracy on astronomical computations, which remains inadequate for real-world applications. The finding implies that, despite substantial progress in mathematical reasoning, existing LLMs still struggle with cross-disciplinary scientific computation, in which mathematical reasoning must be coupled with domain knowledge.

For the FC task, GPT-5.6-Luna again shows the best performance. Although its difficulty is slightly lower than that of UC, FC's overall performance decreases further than that of UC. This trend raises an important concern: existing models tend to prioritize textual understanding and generic mathematical skills, while largely neglecting the structured, symbolic nature of formulas. Such a limitation is likely to constrain future progress in the application of LLMs in astronomy and, more broadly, in scientific reasoning tasks that rely heavily on formal symbolic representations.

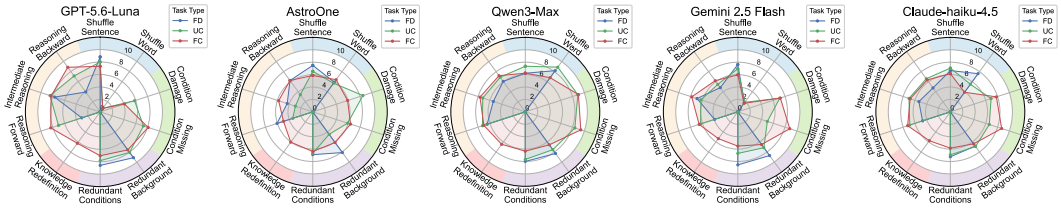



Figure 2: Performance of the five closed-source models in each cognitive dimension. The intervals highlighted in blue correspond to "Problem Comprehension"; the intervals shown in green represent "Condition Robustness"; the intervals marked in purple relate to "Anti-Interference"; the red intervals correspond to "Knowledge Consistency"; and the intervals displayed in orange represent "Reverse Reasoning". Note that the performance of the five open-source models is presented in Figure 5 in the appendix due to space limitations.

| Model | CC | | MC | | |
|---|---|---|---|---|---|
| | UC-Accuracy | FC-Accuracy | FD-Score | UC-Score | UC-Accuracy |
| GPT-5.6-Luna | **0.67** | 0.55 | 7.70 | 8.43 | **0.61** |
| AstroOne | 0.60 | 0.52 | - | - | - |
| Qwen3-Max | 0.63 | 0.52 | - | - | - |
| Gemini-2.5-Flash | **0.65** | 0.55 | **8.45** | **8.53** | **0.61** |
| Claude-haiku-4.5 | 0.58 | 0.49 | 7.60 | 6.42 | 0.33 |
| AstroSage3-8B | 0.00 | 0.04 | - | - | - |
| Llama-3.1-8B | 0.17 | 0.22 | - | - | - |
| Llama-3.1-70B | 0.31 | 0.36 | - | - | - |
| Qwen2.5-Math-72B | 0.06 | 0.22 | - | - | - |
| Qwen3-VL-32B | 0.56 | 0.48 | 7.90 | 5.28 | 0.33 |

Table 2: Performance of each tested model on CC and MC tasks. Values in red indicate anomalous results due to erroneous outputs. Bold font marks the best performance among tested models.

Across all three tasks, slow thinking degrades AstroOne's performance, indicating it does not benefit from extended reasoning. The causes will be analyzed in error analysis (§ 3.7). In contrast, the Qwen series, GPT-5.6-Luna, and Claude-Haiku-4.5 benefit from reasoning strategies, including slow thinking and CoT prompting. These results suggest that their pretraining successfully internalizes reasoning-oriented behaviors. The system prompt analysis indicates that removing explicit astronomical guidance generally degrades performance. This implies that existing models lack sufficiently robust internal representations of astronomical knowledge. The emphasis on general-purpose abilities appears to come at the expense of reduced effectiveness in solving domain-specific problems.

Finally, as shown in Table 2, code-based calculation generally improves the accuracy of most models. This suggests that code execution effectively augments reasoning capabilities. However, for Qwen2.5-Math-72B, which emphasizes pretrained mathematical reasoning capabilities, code-based calculation degrades its performance. For multimodal computation, we only evaluated the four multimodal LLMs: GPT-5.6-Luna , Gemini-2.5-Flash, Claude-haiku-4.5, and Qwen3-VL-32B. When visual information is added, we observe a decline in performance on FD tasks. In contrast, a performance gain is observed in text-only unimodal reasoning when visual input is added. These findings suggest that text-image integration is critical for astronomical reasoning.

We further validate the robustness of the evaluation protocol using multiple evaluator models; the results are provided in Appendix H. Additional experimental results and example model analysis reports are provided in Appendices M.1 and O, respectively.

### 3.3 Cognitive Dimension Evaluation

The results for cognitive dimension evaluation are shown in Figures 2 and 5. Specifically, we systematically evaluate each model's behavior, with particular focus on its robustness to input perturbations, condition variations, and reasoning consistency.

| Model | Sentence | Word | Condition & Variable | | | |
|---|---|---|---|---|---|---|
| | SIM | Acc | #PEM | #PAM | #PEMS | #PAMS |
| GPT-5.6-Luna | 0.564 | 6.7% | 0.2% | 28.2% | 0.4% | 35.6% |
| DeepSeek-V3.2 | 0.619 | 1% | 0.2% | 28.8% | 0.4% | 36.3% |
| GPT-5.1 | 0.639 | 7.2% | 0 | 14.4% | 0 | 24.2% |
| Qwen3-Max | 0.641 | 12.9% | 0 | 9.2% | 0 | 14.0% |

Table 3: Results of dataset contamination tests. "Sentence", "Word", and "Condition & Variable" represent different testing levels. SIM denotes similarity computed using bge-large-en-v1.5. #PEM and #PAM denote perfect and partial match accuracy, while #PEMS and #PAMS apply the same criteria with similarity $\geq 0.9$ regarded as a successful match.

| Task | Eval. Type | Spearman | ICC(2,1) | $\Delta$-Score |
|---|---|---|---|---|
| UC | Process | 0.950 | 0.626 | 1.090 |
| | Result | 0.729 | 0.613 | 0.120 |
| FC | Process | 0.900 | 0.945 | 0.008 |
| | Result | 1.000 | 1.000 | 0 |
| MC | Process | 0.790 | 0.750 | 0.820 |
| | Result | 0.866 | 0.865 | 0.070 |
| Cog | Process | 0.675 | 0.665 | 0.173 |
| | Result | 0.710 | 0.713 | 0.019 |
| FD | Process | 0.898 | 0.670 | 0.352 |
| CC | Result | 0.937 | 0.938 | 0.010 |

Table 4: Correlation between human and model scoring.

We observe that all models exhibit a performance degradation on the problem comprehension dimension. This indicates that, under the shuffle sentence setting, the tested models cannot align with human cognition. Under the shuffle word setting, Qwen3-Max achieves the best overall performance, showing robustness to word-order perturbations and effectiveness in rejecting ill-posed inputs. AstroOne shows a noticeably weaker rejection capability when encountering problems inconsistent with human cognition, although it still outperforms other models except for Qwen3-Max. In contrast, GPT-5.6-Luna exhibits almost no ability to reject such inputs. Although optimized for mathematical reasoning, Qwen2.5-Math-72B fails to assess cognitive validity. In terms of condition robustness, Qwen3-Max continues to exhibit superior performance, showing its effectiveness in identifying and modeling fine-grained conditions and making accurate judgments accordingly. Nevertheless, most other models suffer from insufficient sensitivity to conditional information. In the anti-interference evaluation, Qwen3-Max and GPT-5.6-Luna demonstrate strong resistance to background noise and irrelevant conditions, maintaining stable performance, although they show degradation relative to their performance in a noise-free setting. Other models achieve only moderate levels of anti-interference capability. On the knowledge consistency dimension, leading models exhibit similar performance, remaining moderate overall. This suggests that they have limited capability to incorporate new astronomical definitions and instead tend to rely on memorized knowledge that closely aligns with their training corpora. In the reverse reasoning test, GPT-5.6-Luna achieves the highest performance, whereas other models exhibit instability. The performance of open-source models is consistently poor across all cognitive dimensions. Nevertheless, Qwen2.5-Math-72B performs well on tasks of mathematical reasoning reproduction. Overall, AstroOne ranks in the lower-middle tier in cognitive evaluation. AstroSage3-8B performs poorly across all aspects. Additional results for cognitive dimension evaluation can also be found in Appendix M.2.

### 3.4 Data Pollution Testing

We conduct a dataset contamination analysis at the sentence, word, and condition & variable levels. The results are presented in Table 3. To establish a baseline, we first evaluate the embedding of unrelated sentence pairs drawn from five distinct domains. Over 10,000 sentence pairs, the bge-large-en-v1.5 model yields an average similarity of 0.4246 with a standard deviation of 0.0582. The 95th percentile similarity is about 0.52, suggesting that the embedding space is centered around a similarity of about 0.4 for semantically unrelated inputs. The sentence-level contamination tests show similarity scores in the range of 0.62–0.64. While these values are higher than those of unrelated sentences, they remain below the threshold used for higher-confidence semantic overlap ($\geq$ 0.8), and thus do not constitute strong evidence of memorization or direct data leakage. At the word level, Qwen3-Max achieves the highest accuracy of only 12.9%, well below the 50% threshold that indicates significant overlap. This suggests that lexical-level memorization or contamination is negligible. For condition & variable matching, the contamination signals remain weak. Using exact matching, the highest #PEM is about 0.2%, while the highest #PAM reaches 28.8%, both far below 50%. Similarly, under embedding-based matching, #PEMS peaks at 0.4%, and #PAMS at 36.3%. All of them remain below the threshold that indicates dataset leakage. Given the embedding model's inherently high baseline similarity, we conclude that the benchmark exhibits no notable data leakage.

| Model | Score (mean ± var) | Acc (mean ± var) | Score Range | Acc Range | Max △ Score | Max △ Acc |
|---|---|---|---|---|---|---|
| DeepSeek V3.2 | 7.79 ± 0.027 | 0.59 ± 0.00013 | 7.60–8.00 | 0.58–0.61 | 0.40 | 0.03 |
| Gemini 2.5 Flash | 7.43 ± 0.017 | 0.50 ± 0.00388 | 7.29–7.63 | 0.42–0.56 | 0.34 | 0.14 |
| Kimi K2.5 | 7.29 ± 0.022 | 0.61 ± 0.00083 | 7.07–7.46 | 0.56–0.63 | 0.39 | 0.07 |
| Qwen3-Max | 7.68 ± 0.008 | 0.58 ± 0.00048 | 7.60–7.83 | 0.54–0.60 | 0.23 | 0.06 |
| AstroOne | 7.48 ± 0.029 | 0.57 ± 0.00182 | 7.31–7.65 | 0.53–0.60 | 0.36 | 0.07 |

Table 5: Evaluation stability of models across different question variants.

### 3.5 Human Random Testing

We randomly sampled reported results from five task types (FD, UC, MC, CC, and FC) and the cognitive dimension (Cog), followed by a manual evaluation. Considering the high cost of manual annotation, we maximized the number of randomly sampled instances in each category whenever possible. Since some task types contained only a limited number of reported results, the final sample sizes differed across categories, resulting in 301 manually evaluated instances, including 25 FD, 50 UC, 21 MC, 100 CC, 25 FC, and 80 Cog samples. Based on the human evaluation results, we further assess the reliability of the LLM-as-Judge paradigm. The comparison between LLM-based and human evaluations is shown in Table 4. LLM-based scoring exhibits consistently strong rank correlation with human judgment, with Spearman's $\rho$ ranging from 0.675 to 1.000 across all settings ($p < 0.001$), indicating high reliability in preserving relative model ordering. Absolute agreement, measured by $ICC(2,1)$, shows moderate to high consistency, with lower agreement on Cog and UC process-level evaluation (≈0.62–0.67) and substantially higher agreement on FC tasks, particularly at the result level (1.000). Score deviation analysis further reveals that discrepancies between model and human scoring are more pronounced for process-level assessments, whereas result-level evaluations yield minimal deviation ($\Delta \leq 0.12$). Overall, these results suggest that model-based evaluation provides a reliable proxy for human judgment, especially for result-oriented and structurally constrained tasks, while process-level scoring remains comparatively more subjective.

### 3.6 Evaluation Stability on Different Question Variants

We conducted an additional experiment to test whether the generated variant sets are truly equivalent in validity. To address this issue, we conducted an additional experiment. Specifically, we randomly sampled 100 seed instances and generated five test variants for each instance, while keeping the question type and difficulty level aligned with the original seed data. We then evaluated four subject models on each of the five variant sets. As shown in Table 5, the results show that the overall performance trends of all four models remain consistent across the five test sets. Meanwhile, the score variations stay within an acceptable range for open-ended evaluation: the maximum fluctuation in Score is below 0.4, and the maximum fluctuation in Accuracy is below 0.14. These findings suggest that the variant data constructed from the seed instances maintain good consistency in difficulty and can provide equivalent and stable assessments of model capability.

### 3.7 Error Analysis

Current models show recurring errors: i) reasoning instability, where think-style prompting can induce repetitive loops and reasoning drift; ii) poor unit awareness and scale consistency, leading to physically meaningless or numerically incorrect results; iii) capability imbalance, where models optimized for QA or mathematics fail on numerical reasoning or executable code generation; and iv) unreliable structured generation, including malformed JSON, syntax errors, repetitive outputs, and non-executable code. These issues substantially hinder automated evaluation and deployment. More detailed analyses are provided in Appendix G.

## 4 Conclusion

In this paper, we present AstroBench, a benchmark for evaluating astronomical computational reasoning in LLMs across diverse tasks and cognitive settings, with an emphasis on contamination robustness and long-term validity. Extensive experiments reveal major limitations in computational stability, formula understanding, code-based reasoning, unit conversion, and magnitude control, highlighting a significant gap between current LLMs and the demands of astronomical reasoning.

## AI USE STATEMENT

AI tools were used in this study solely to aid and polish writing and to assist in evaluating the process-based experimental results presented in this paper. They were not involved in research design, data analysis, or the formation of conclusions. Details are described in the paper.

## REPRODUCIBILITY STATEMENT

We have made our experimental code, along with the associated evaluation code and variant data, publicly available on an anonymous GitHub repository (including intermediate experimental results reported in the paper). The seed datasets involved in the paper are not publicly released. Our code and data are publicly available at `https://anonymous.4open.science/r/AstroBench-705D`.

## REFERENCES


Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. Large language models for mathematical reasoning: Progresses and challenges. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pp. 225–237, 2024. doi: 10.18653/v1/2024.eacl-srw.17.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2357–2367, 2019. doi: https://doi.org/10.18653/v1/N19-1245.

Anthropic. Claude haiku 4.5: Our fastest model, a lightweight version of our most powerful AI, at a more affordable price. `https://www.anthropic.com/claude/haiku`, 2025.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690, 2024. doi: 10.1609/aaai.v38i16.29720.

Sebastian Bordt, Harsha Nori, and Rich Caruana. Elephants never forget: Testing language models for memorization of tabular data. In *NeurIPS 2023 Second Table Representation Learning Workshop*, 2023. URL `https://openreview.net/forum?id=l1u7jA60wT`.

Sébastien Bubeck, Christian Coester, Ronen Eldan, Timothy Gowers, Yin Tat Lee, Alexandru Lupsasca, Mehtaab Sawhney, Robert Scherrer, Mark Sellke, Brian K. Spears, Derya Unutmaz, Kevin Weil, Steven Yin, and Nikita Zhivotovskiy. Early science acceleration experiments with GPT-5. *arXiv:2511.16072*, 2025. URL `https://arxiv.org/abs/2511.16072`.

Daniel J. H. Chung, Zhiqi Gao, Yurii Kvasiuk, Tianyi Li, Moritz Münchmeyer, Maja Rudolph, Frederic Sala, and Sai Chaitanya Tadepalli. Theoretical physics benchmark (TPBench) - a dataset and study of AI reasoning capabilities in theoretical physics. *Machine Learning: Science and Technology*, 6(2):30505, 2025. doi: 10.1088/2632-2153/adfcb0.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv:2110.14168*, 2021. URL `https://arxiv.org/abs/2110.14168`.

Tijmen de Haan, Yuan-Sen Ting, Tirthankar Ghosal, Tuan Dung Nguyen, Alberto Accomazzi, Azton Wells, Nesar Ramachandra, Rui Pan, and Zechang Sun. Achieving GPT-4o level performance in astronomy with a specialized 8b-parameter large language model. *Scientific Reports*, 15:13751, 2025. doi: 10.1038/s41598-025-97131-y.

DeepSeek-AI. Deepseek-v3.2: Pushing the frontier of open large language models. *arXiv:2512.02556*, 2025. URL `https://arxiv.org/abs/2512.02556`.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. Investigating data contamination in modern benchmarks for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8706–8719, 2024. doi: 10.18653/v1/2024.naacl-long.482.

Aniket Didolkar, Anirudh Goyal, Nan Rosemary Ke, Siyuan Guo, Michal Valko, Timothy Lillicrap, Danilo Rezende, Yoshua Bengio, Michael Mozer, and Sanjeev Arora. Metacognitive capabilities of LLMs: An exploration in mathematical problem solving. *Advances in Neural Information Processing Systems*, 37:19783–19812, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2318d75a06437eaa257737a5cf3ab83c-Abstract-Conference.html.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, Zhengyang Tang, Benyou Wang, Daoguang Zan, Shanghaoran Quan, Ge Zhang, Lei Sha, Yichang Zhang, Xuancheng Ren, Tianyu Liu, and Baobao Chang. Omni-MATH: A universal olympiad level mathematic benchmark for large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=yaqPf0KAlN.

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv:2507.06261*, 2025a. URL https://arxiv.org/abs/2507.06261.

Gemini Team. Gemini 3 Pro – Best for complex tasks and bringing creative concepts to life. https://deepmind.google/models/gemini/pro, 2025b.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, 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, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, 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, Mingxu Zhou, 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, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, 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, 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 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z.

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 *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1 (NeurIPS Datasets and Benchmarks 2021)*, 2021. URL https://openreview.net/forum?id=7Bywt2mQsCe.

Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. A survey on large language models for code generation. *ACM Transactions on Software Engineering and Methodology*, 35 (2):58:1–58:72, 2026. doi: 10.1145/3747588.

Mingyu Jin, Weidi Luo, Sitao Cheng, Xinyi Wang, Wenyue Hua, Ruixiang Tang, William Yang Wang, and Yongfeng Zhang. Disentangling memory and reasoning ability in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1681–1701, 2025. doi: 10.18653/v1/2025.acl-long.84.

Sebastian Antony Joseph, Syed Murtaza Husain, Stella S. R. Offner, Stéphanie Juneau, Paul Torrey, Adam S. Bolton, Juan P. Farias, Niall Gaffney, Greg Durrett, and Junyi Jessy Li. AstroVisBench: A code benchmark for scientific computing and visualization in astronomy. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=qXiTFAgEx4.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Jie Li, Fuyong Zhao, Panfeng Chen, Jiafu Xie, Xiangrui Zhang, Hui Li, Mei Chen, Yanhao Wang, and Ming Zhu. An astronomical question answering dataset for evaluating large language models. *Scientific Data*, 12:447, 2025. doi: 10.1038/s41597-025-04613-9.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5315–5333, 2023. doi: 10.18653/v1/2023.acl-long.291.

Jiayu Liu, Zhenya Huang, Wei Dai, Cheng Cheng, Jinze Wu, Jing Sha, Song Li, Qi Liu, Shijin Wang, and Enhong Chen. CogMath: Assessing LLMs' authentic mathematical ability from a human cognitive perspective. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 38692–38707, 2025. URL https://proceedings.mlr.press/v267/liu25ab.html.

Zimu Lu, Aojun Zhou, Houxing Ren, Ke Wang, Weikang Shi, Junting Pan, Mingjie Zhan, and Hongsheng Li. MathGenie: Generating synthetic data with question back-translation for enhancing mathematical reasoning of LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2732–2747, 2024. URL 10.18653/v1/2024.acl-long.151.

OpenAI. GPT-4 technical report. *arXiv:2303.08774*, 2023. URL https://arxiv.org/abs/2303.08774.

OpenAI Team. Gpt-5.6: Frontier intelligence that scales with your ambition. https://openai.com/index/gpt-5-6/, 2026.

Jiao Ou, Junda Lu, Che Liu, Yihong Tang, Fuzheng Zhang, Di Zhang, and Kun Gai. DialogBench: Evaluating LLMs as human-like dialogue systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6137–6170, 2024. doi: 10.18653/v1/2024.naacl-long.341.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802, 2024. doi: 10.52202/079017-2197.

Lucas Carrit Delgado Pinheiro, Ziru Chen, Bruno Caixeta Piazza, Ness Shroff, Yingbin Liang, Yuan-Sen Ting, and Huan Sun. Large language models achieve gold medal performance at the international olympiad on astronomy & astrophysics (IOAA). *arXiv:2510.05016*, 2025. URL https://arxiv.org/abs/2510.05016.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=dHng2O0Jjr.

Qwen Team. Qwen2.5-Math technical report: Toward mathematical expert model via self-improvement. *arXiv:2409.12122*, 2024. URL https://arxiv.org/abs/2409.12122.

Qwen Team. Qwen3-VL technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Qwen Team. Qwen3 technical report. *arXiv:2505.09388*, 2025b. URL https://arxiv.org/abs/2505.09388.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Jinghang Shi, Xiaoyu Tang, Yang Huang, Yuyang Li, Xiao Kong, Yanxia Zhang, and Caizhan Yue. AstroMMBench: A benchmark for evaluating multimodal large language models capabilities in astronomy. *arXiv:2510.00063*, 2025. URL https://arxiv.org/abs/2510.00063.

Y.-S. Ting, Tuan Dung Nguyen, Tirthankar Ghosal, Rui Pan, Hardik Arora, Zechang Sun, Tijmen de Haan, Nesar Ramachandra, Azton Wells, Sandeep Madireddy, and Alberto Accomazzi. AstroMLab 1: Who wins astronomy jeopardy!? *Astronomy and Computing*, 51:100893, 2025. doi: 10.1016/j.ascom.2024.100893.

Yuan-Sen Ting and Teaghan O'Briain. Teaching astronomy with large language models. *The Open Journal of Astrophysics*, 9, 2026. doi: 10.33232/001c.160220.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv:2302.13971*, 2023. URL https://arxiv.org/abs/2302.13971.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Yutong Wang, Yunxiang Xiao, Yonglin Tian, Junyong Li, Jing Wang, and Yisheng Lv. Astroreview: An llm-driven multi-agent framework for telescope proposal peer review and refinement. *arXiv:2512.24754*, 2025. URL https://arxiv.org/abs/2512.24754.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. URL https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Liang Xu, Hang Xue, Lei Zhu, and Kangkang Zhao. SuperCLUE-Math6: Graded multi-step math reasoning benchmark for LLMs in Chinese. *arXiv:2401.11819*, 2024. URL https://arxiv.org/abs/2401.11819.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36:11809–11822, 2023. URL https://proceedings.neurips.cc/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract.html.

Yuhang Zhang, Bohan Jiao, and Zhaohui Dang. A large language model-based approach to spacecraft formation mission planning. *IFAC-PapersOnLine*, 59(20):2166–2170, 2025. doi: 10.1016/j.ifacol.2025.11.480.

Fuyong Zhao, Yuyang Li, Zhenyu Liu, Panfeng Chen, Cunshi Wang, Jifeng Liu, Hui Li, and Yanhao Wang. Unveiling the power of multimodal large language models for radio astronomical image understanding and question answering. *Machine Learning: Science and Technology*, 6(4):045005, 2025. doi: 10.1088/2632-2153/ae0c56.

Zhejiang Lab. AstroOne, a large language model for astronomy launched. `https://en.zhejianglab.com/news/202501/t20250127_4158.shtml`, 2024.

Bingke Zhu, Xiaoxiao Wang, Minghui Jia, Yihan Tao, Xiao Kong, Ali Luo, Yingying Chen, Ming Tang, and Jinqiao Wang. FLARE: A framework for stellar flare forecasting using stellar physical properties and historical records. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 9501–9509, 2025. doi: 10.24963/ijcai.2025/1056.

# A Extended Related Work

## A.1 LLM Benchmarks in Astronomy

There are numerous benchmarks for evaluating LLMs' capabilities in solving astronomical problems. Astro-QA (Li et al., 2025) was a question-answering benchmark to evaluate factual knowledge in astronomy. Then, AstroMLab 1 (Ting et al., 2025) assessed LLMs' understanding about frontier scientific consensus through literature review. The IOAA benchmark (Pinheiro et al., 2025) used Olympiad-level problems to evaluate LLMs' performance in astronomy & astrophysics. AstroMMBench (Shi et al., 2025) presented a multimodal benchmark to evaluate LLMs' astrophysical image understanding through expert-validated visual question answering. AstroVisBench (Joseph et al., 2025) evaluated LLMs' abilities to transform observational data into visual plots by code generation. TPBench (Chung et al., 2025) assessed the LLM reasoning for high-difficulty problems in theoretical physics, where cosmological problems were also included. Despite significant progress, existing benchmarks still lack a unified evaluation of computational reasoning tailored to astronomical problems and are vulnerable to data contamination.

## A.2 Astronomical Foundation Models and LLMs

Significant effort has been devoted to developing astronomy-specific foundation models and LLMs. AstroSage (de Haan et al., 2025) showed that continued pre-training on large-scale astronomical corpora enabled an 8B-parameter model to achieve astronomical semantic understanding comparable to the GPT-4o level. The FLARE model (Zhu et al., 2025) introduced a hybrid reasoning framework that integrates physical attributes and evolutionary histories via soft prompts and residual modules for stellar flare forecasting. RadioAstroVQA (Zhao et al., 2025) constructed a visual question dataset for radio observations to fine-tune multimodal LLMs (MLLMs) for radio source classification, surpassing the performance of existing deep learning models. In addition, AstroReview (Wang et al., 2025), AstroPlan (Zhang et al., 2025), and AstroTutor (Ting & O'Briain, 2026) explore the use of LLMs for telescope operation assistance, spacecraft formation planning, and interactive astronomy education, respectively. The above developments have highlighted the applications of LLMs in astronomy from different aspects. However, these methods do not consider how to improve LLMs' computational reasoning capabilities.

## A.3 Reasoning with Large Language Models

LLMs, such as GPT-5 (Bubeck et al., 2025), DeepSeek (Guo et al., 2025), and Gemini 2.5 (Gemini Team, 2025a), have demonstrated advanced reasoning capabilities through pre-training on massive, diverse corpora. To further enhance their reasoning abilities, prior work introduced structured inference paradigms. For example, Chain-of-Thought (CoT) prompting (Wei et al., 2022) enabled step-by-step problem decomposition; Tree-of-Thoughts (ToT) (Yao et al., 2023) and Graph-of-Thoughts (GoT) (Besta et al., 2024) generalized CoT by exploring branching and graph-structured reasoning

paths. These methods have shown strong performance on multi-step and combinatorial reasoning tasks. Beyond prompting strategies, external data augmentation, e.g., Retrieval-Augmented Generation (RAG) (Lewis et al., 2020), and training-time optimization, e.g., Group Relative Policy Optimization (GRPO) (Shao et al., 2024), were also used to enhance the reasoning process. Recent studies have also introduced novel strategies to enhance LLMs' mathematical reasoning abilities. For instance, Didolkar et al. (2024) adopted a metacognition-inspired reasoning approach to improving the ability of LLMs in mathematical problem-solving. Jin et al. (2025) proposed to disentangle memory retrieval and reasoning strategies into two explicit steps to boost the performance of LLMs in challenging reasoning tasks. Lu et al. (2024) enhanced the mathematical reasoning of LLMs by generating diverse real-solution problems from seed data, thereby augmenting the corpus. Finally, multimodal LLMs, such as GPT-4V (OpenAI, 2023) and Gemini 3 Pro (Gemini Team, 2025b), showed improved reasoning skills in cross-modal scenarios.

## B Details of Dataset Collection and Processing

**Data Preprocessing** The preprocessing of AstroMathBench was divided into four main steps: data merging and formatting, data filtering, data annotation and correction, and difficulty classification. The annotation team comprises 24 researchers from our group, as well as other groups with whom we collaborate. Their roles and responsibilities are listed below.

- Two Ph.D. students in the cross-disciplinary field of astronomy and computer science are first-level workers.
- Two graduate students with at least two years of experience in the intersection of LLMs and astronomy are second-level workers.
- The remaining graduate students in the field of computer science are third-level workers.

In data merging and formatting, first-level workers are responsible for classifying data into different task types and merging data from various sources into a unified data table. The second-level worker is responsible for formatting the merged data. In data filtering, first-level and second-level workers are responsible for filtering out the following categories of data: i) samples with ill-defined problems, ambiguous questions, or answers that are unclear or overly generic (e.g., a question that asks for a general scientific explanation without a specific reasoning goal); ii) data irrelevant to astronomical computation, such as tasks involving image plotting, true or false judgments, or commonsense trivia that require no domain-specific inference; iii) redundant entries, including near-duplicate questions or repeated content across different sources. In data annotation and correction, third-level workers are responsible for the initial annotation of solution steps and final answers. Second-level workers review the annotations and correct inconsistencies. First-level workers adjudicate disputed cases and perform the final verification. Each worker is presented with a question and its answer at a time and asked to search web pages, review reference books, and discuss with other workers to write an explanation of how the answer is reached. In difficulty classification, we use the following equation to assign the level to each question:

$$\mathcal{D} = \mathcal{D}_{\text{base}} \times (1 + S) \times (1 + I) \times T, \tag{1}$$

where $\mathcal{D}_{\text{base}}$ denotes the base difficulty coefficient, $S$ represents the step complexity, $I$ indicates whether image comprehension is required, and $T$ is the question-type coefficient.

**Annotation Protocol and Quality Control** *i) Compensation Status:* All annotators were members of collaborating academic research groups and participated as part of their regular research activities. All annotators participated as research assistants within the project and received monthly stipends in accordance with institutional policies ($ 220 for Ph.D. students and $ 117 for Master's students). *ii) Annotation Guideline:* Before annotation, all annotators participated in a standardized online training session based on a written annotation guideline. The training covered the annotation objectives, labeling criteria, annotation workflow, and quality requirements to ensure consistent understanding and implementation of the annotation protocol across all annotators. *iii) Quality Control:* The 100% of the dataset was initially annotated by a third-level annotator and reviewed by a designated second-level annotator. All reviewed samples were subsequently submitted to the first-level annotators for final verification. Samples with annotation disagreements, accounting for approximately 10% of the dataset, were jointly adjudicated by two first-level annotators before the

**Algorithm 1** Problem Generation by Parameter Perturbation

**Input:** Original problem $Q$ and its answer $A$
**Output:** Variant problem $Q_i$ and its answer $A_i$
Extract numerical values $\mathcal{C} \leftarrow \{c_1, \ldots, c_n\}$ from $Q$
Infer solution formulas $\mathcal{F} \leftarrow \{f_1, \ldots, f_k\}$ from $(Q, A)$
**repeat**
  **for** each constant $c_j \in \mathcal{C}$ **do**
    Define perturbation interval $\Delta_j = [\alpha_j c_j, \beta_j c_j]$
    Sample perturbed value $c_j^{(i)} \sim \mathcal{U}(\Delta_j)$
  **end for**
  Form perturbed constant set $\mathcal{C}^{(i)} = \{c_1^{(i)}, \ldots, c_n^{(i)}\}$
**until** all physical constraints are satisfied
Initialize intermediate results $\mathcal{R} \leftarrow \emptyset$
**for** $j = 1$ to $k$ **do**
  Compute $r_j \leftarrow f_j(\mathcal{C}^{(i)}, \mathcal{R})$
  Update intermediate results $\mathcal{R} \leftarrow \mathcal{R} \cup \{r_j\}$
**end for**
Set final answer $A_i \leftarrow r_k$
Generate new problem $Q_i \leftarrow Q$ with $\mathcal{C}$ replaced by $\mathcal{C}^{(i)}$

final annotation was confirmed. *iv) Multi-level Workflow:* Each sample was independently annotated by one third-level annotators supervised by a designated second-level annotator. The second-level annotator then reviewed all annotations to ensure consistency and compliance with the annotation guidelines. Samples with annotation disagreements were escalated to the first-level annotators, who resolved the disagreements through discussion and adjudication. For samples without annotation disagreements, the reviewed annotations were also submitted to the first-level annotators for final verification. The verified or adjudicated annotation was adopted as the final label.

**Omitted Algorithm for Problem Generation** We present the procedure of generating a perturbed problem-answer pair by modifying numerical constants while preserving physical validity and reasoning structure in Algorithm 1. A detailed description can be found in Section 2.2.

**Difficulty Coefficient Assignment** To quantify problem difficulty, we adopt a unified scoring method that integrates source reliability, modality, question type, and reasoning complexity. Specifically, each problem is first assigned a base difficulty score $\mathcal{D}_{\text{base}}$ according to its source, as reviewed by first-level annotators: problems from APBench and *Fundamental Astronomy* are assigned $\mathcal{D}_{\text{base}} = 0.5$; those from *Introduction to Modern Astronomy*, *Fundamentals of Astrometry and Celestial Mechanics*, and expert-built datasets are assigned $0.6$; problems from USAAAO are assigned $0.7$; and those from IOAA/IOOA and TPBench are assigned $0.8$. To further account for additional challenges, we introduce a modality coefficient $I = 0.5$ for image-based problems and a question-type coefficient $T$, where reasoning or formula-derivation problems are assigned $T = 0.7$, and direct formula-based computation problems are assigned $T = 0.5$. Finally, we incorporate a step-complexity coefficient $S$ based on the number of reasoning steps required, with $S = 0$ for problems solvable in 0–1 step, $S = 0.3$ for 1–3 steps, $S = 0.5$ for 4–5 steps, and $S = 0.7$ for problems requiring six or more steps.

**Variable-Perturbation Problem Generation** For the process of generating variable-perturbation problems in cognitive experiments, the specific implementation is as follows: given an original question and answer pair $(Q, A)$, Algorithm 2 generates a set of aligned variants by selectively perturbing placeholders in the question. The algorithm first extracts all placeholders $\mathcal{P}$ from $Q$ and samples a subset $\mathcal{P}' \subseteq \mathcal{P}$ with $|\mathcal{P}'| = \lfloor |\mathcal{P}|/2 \rfloor$. The same subset $\mathcal{P}'$ is used across all perturbation types to ensure structural alignment. For each perturbation type $i \in \mathcal{I}$, a perturbation operator $\phi_i$ is applied to the selected placeholders, producing a variant $(Q_i, A_i)$. The output of the algorithm is a set of aligned question-answer pairs $\{(Q_i, A_i)\}_{i \in \mathcal{I}}$, which differ only in localized placeholder substitutions. The perturbation operator is defined as:

$$\phi_i(p) = \begin{cases} p, & i = \text{ori}, \\ \texttt{<Replace>}, & i = \text{rep}, \\ u \sim \mathcal{U}, & i = \text{uns}, \end{cases} \tag{2}$$

**Algorithm 2** Structured Question Variant Generation

**Input:** Original question-answer pair $(Q, A)$
**Output:** A set of perturbed variants $\{(Q_i, A_i)\}_{i \in \mathcal{I}}$
Extract placeholder set $\mathcal{P} = \{p_1, \ldots, p_n\}$ from $Q$
Sample a subset $\mathcal{P}' \subseteq \mathcal{P}$ with $|\mathcal{P}'| = \lfloor n/2 \rfloor$
**for all** perturbation types $i \in \mathcal{I}$ **do**
    **if** $i =$ original **then**
        $Q_i \leftarrow Q$ and $A_i \leftarrow A$
    **else**
        Apply perturbation operator $\phi_i$ on $\mathcal{P}'$
        $Q_i \leftarrow \phi_i(Q)$ and $A_i \leftarrow \phi_i(A)$
    **end if**
**end for**

where $\mathcal{U}$ denotes a predefined set of mathematically unsolvable expressions.

Finally, the data distribution and the distribution of question difficulty coefficients are shown in Figure 3.

## C Details of Code-Based Mutation

To make the code-based mutation process easier to understand, we add a concrete example of a code-formalized altitude–azimuth problem in Fig. 6 and Fig. 7. The original problem (*'Given the declination of Perseus $\alpha$ star as $\delta = 49°.8$ and the hour angle $t = 324°$, calculate the altitude $h$ and azimuth $A$ of this star when observed at latitude $\phi = 39°.1$'*) asks for the altitude $h$ and azimuth $A$ of a star given the observer latitude $\phi$, stellar declination $\delta$, and hour angle $t$. This problem can be naturally formalized as a Python function that takes $(\phi, \delta, t)$ as input and computes $h$ and $A$ using the standard spherical trigonometric relations. During mutation, the generated question differs from the original one only in the sampled parameter values, while the physical setting, computational procedure, and reasoning logic are kept unchanged. This allows us to obtain a sufficiently different numerical problem with a different final answer, without changing the underlying difficulty level. We explicitly check the physical constraints of perturbable variables before accepting a generated function variant. In this example, the valid ranges are $-90° \leq \phi \leq 90°$ for geographic latitude, $-90° \leq \delta \leq 90°$ for stellar declination, and $0° \leq t \leq 360°$ for hour angle. After identifying such constraints, the system samples candidate values within the proposed ranges and executes the generated function to verify numerical validity. In our implementation, each candidate range is tested using 20 random samples. A range is accepted only if all executions are completed without runtime errors or invalid numerical results. Otherwise, the range is rejected or adjusted. Not all problems admit code-based variants. For symbolic derivation problems, such as deriving the differential equations of the two-body problem *'The two-body problem concerns how two celestial bodies move under mutual gravitational attraction when only these two bodies are considered (neglecting perturbations from all others). Let there be only the Sun $\odot$ and a planet $P$. Establish an arbitrary spatial coordinate system $O - xyz$(As shown in the figure¡image¿GSTD-Fig-1¡/image¿). $r_\odot$ is the position vector of the Sun, $r_P$ is the position vector of the planet, and $r$ is the position vector of the planet relative to the Sun. Denote the masses of the Sun and the planet by $M$ and $m$. Based on the given information, derive the differential equations for the two-body problem.'*, there may be no independent numerical parameters that can be safely perturbed. Modifying the assumptions, coordinate definitions, or symbolic quantities in such cases may change the logical structure of the problem itself rather than produce a controlled variant. Therefore, no function variant is generated for these cases.

## D Dataset Information

We introduce the distribution of the AstroMathBench dataset and present representative instances for each task type, accompanied by detailed explanations of both their explicit objectives and implicit evaluation targets.

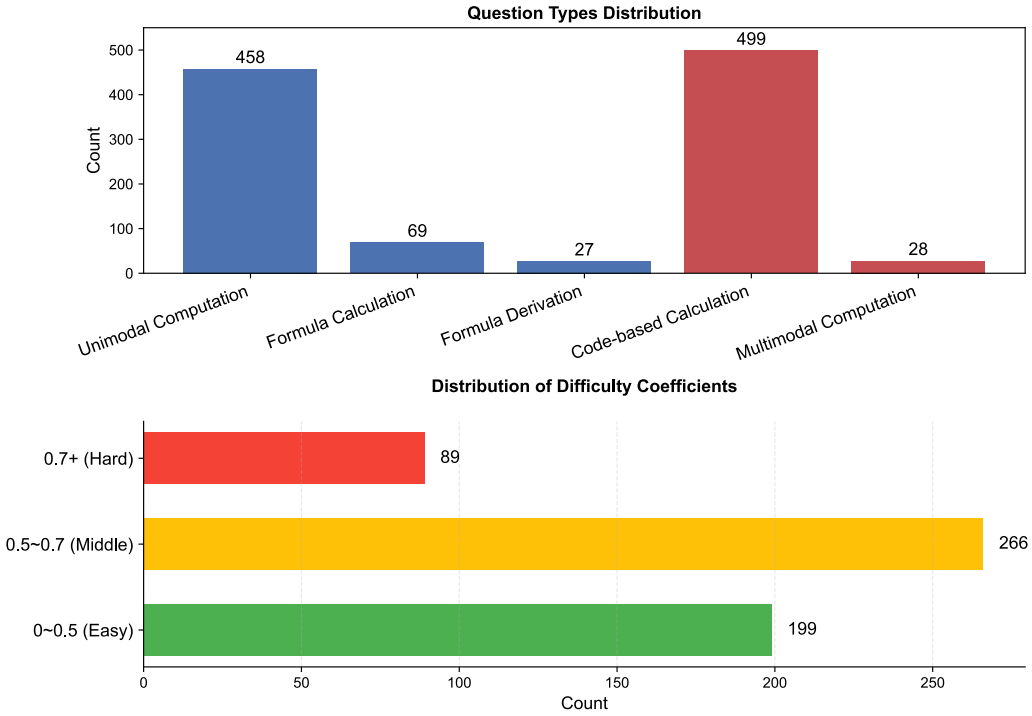



Figure 3: Distribution of problems in the AstroMathBench dataset by task type and difficulty level.

We have now confirmed the following coverage statistics, noting that some questions may span 2–3 astronomy subfields: celestial mechanics (364), astrodynamics (278), astronautics (206), astrophysics (195), astrometry (82), astronomical techniques and methods (53), cosmology (29), high-energy theory (23), space engineering (8), and history of astronomy (2).

While each task is framed around a concrete astrophysics problem, their deeper purpose is to systematically probe LLMs' capabilities in physical understanding, multi-step reasoning, mathematical formalization, and cross-modal integration.

Figure 3 illustrates the distribution problems in the AstroMathBench dataset by task type and difficulty level. Note that the categories highlighted in red (i.e., Code-based Calculation and Multimodal Computation) are generated through processing and expanding the base categories, rather than being independently collected.

## E Model Configuration

| Model | #Params | #Tokens | Window Size | Think | Image | Open | Access |
|---|---|---|---|---|---|---|---|
| AstroOne | 70B | 32B | – | ✓ | × | × | API |
| GPT-5.6-Luna | – | – | 400K | ✓ | ✓ | ✓ | API |
| Qwen3-Max | > 1000B | 36T | 256K | ✓ | × | × | API |
| Gemini 2.5 Flash | – | – | 1M | ✓ | ✓ | × | API |
| Claude-haiku-4.5 | – | – | 200K | ✓ | ✓ | × | API |
| DeepSeek-V3.2 | 685B | – | 128K | ✓ | × | ✓ | API |
| AstroSage-8B | 8B | 5.3B | – | × | × | ✓ | Weights |
| Llama3.1-8B | 8B | > 15T | 128k | × | × | ✓ | Weights |
| Llama3.1-70B | 70B | > 15T | 128k | × | × | ✓ | Weights |
| Qwen2.5-Math-72B | 72B | – | 4K | × | × | ✓ | Weights |
| Qwen3-VL-32B | 32B | – | 128K | ✓ | ✓ | ✓ | Weights |

Table 6: Detailed information of the models evaluated in the experiments.

The experiments with open-source models were conducted on a server with four NVIDIA RTX A6000 GPUs. For closed-source models, we used their official APIs. An overview of the models tested is presented in Table 6, where "#Params" indicates their respective parameters, "#Tokens" reveals the amount of pre-training data used by the model, "Window Size" reveals the size of the model context window, "Think" indicates whether the model supports slow-thinking capabilities, "Image" indicates whether the model supports multimodal analysis, "Open" indicates whether the model is open-source or closed-source, "Access" indicates how we test the model, either through web interfaces and APIs or through their weights, and a dashed line ('–') represents an unknown value.

## F Details of Human Random Testing

In human random testing, the manually sampled results were independently reviewed and annotated by two Ph.D. students with interdisciplinary expertise in astronomy and computer science, both of whom are members of the research group. The annotators assessed whether the reasoning process generated by each model was consistent with the reference solution and whether the final answer was correct. Since the evaluation was performed against authoritative reference answers, these judgments were straightforward and objective. To ensure a fair evaluation, the annotators were blind to the identity of the model that generated each result throughout the annotation process.

## G Details of Error Analysis

**Reasoning Failures** In knowledge reasoning, a representative phenomenon is that AstroOne exhibits significant degradation when think strategies are enabled. An inspection of the analysis logs reveals that AstroOne is prone to repetitive reasoning loops, leading to the reasoning trajectory to progressively drift away from the original problem and, in some cases, to be prematurely truncated. This highlights the necessity of incorporating thought-termination mechanisms during training and of improving the quality control of think-style training data to prevent degenerate reasoning.

**Unit Awareness and Scale Consistency** A pervasive issue across models is the neglect of units and order-of-magnitude consistency during intermediate reasoning steps. This issue is particularly critical in astronomy, where calculations often span multiple disciplines. Any order-of-magnitude error can lead to dramatically incorrect results, whereas ignoring units renders outputs physically meaningless. Notably, this becomes especially severe when models rely on code-based computation, in which unit omissions or scale confusion propagate directly into numerical outputs without corrections.

**Performance Collapse** Performance collapses are observed in AstroSage3-8B and LLaMA-3.1-8B. These models appear to be pretrained on question-answering data. As a result, they exhibit a complete loss of capability when confronted with numerical reasoning tasks. Conversely, Qwen2.5-Math-72B, which prioritizes mathematical computation ability, demonstrates an opposite failure mode: during evaluations that require code-based execution, it hardly generates executable code. These cases indicate that astronomical computation cannot rely on isolated capabilities (e.g., purely mathematical or symbolic) but instead requires the integration of mathematical, physical, and programmatic competencies.

**Structured Output and Executability** Regarding output format compliance, our evaluations reveal substantial deficiencies in models' structured generation and executability. Several models struggle to produce valid JSON outputs, rendering their results difficult to parse or to use directly. AstroSage3-8B, in particular, not only generates malformed JSON files but also frequently produces excessive repetitive strings. While DeepSeek-V3.2 and Qwen3-VL-32B achieve higher output quality, they still occasionally violate strict JSON syntax, necessitating regex-based post-hoc extraction. In contrast, Gemini-2.5-Flash and Claude-Haiku-4.5 rely primarily on regex-based extraction to recover usable JSON, indicating limited stability in generating structured output. The problem is more severe for LLaMA-3.1-8B, whose generated JSON often contains fundamental syntactic errors (e.g., missing closing braces), requiring extensive automated repair before parsing becomes possible.

| Evaluation Model | Model | FD Score | UC Score | UC Acc | FC Score | FC Acc |
|---|---|---|---|---|---|---|
| DeepSeek-V3.2 | Qwen3-Max | 9.04 | 7.97 | 0.66 | 7.11 | 0.61 |
| Qwen3-Max | Qwen3-Max | 8.75 | 7.76 | 0.63 | 7.51 | 0.57 |
| GPT-5.1 | Qwen3-Max | 8.85 | 7.79 | 0.72 | 7.10 | 0.59 |
| RMAD | - | 0.01 | 0.01 | 0.05 | 0.02 | 0.02 |
| DeepSeek-V3.2 | DeepSeek-V3.2 | 8.94 | 7.32 | 0.60 | 6.99 | 0.55 |
| Qwen3-Max | DeepSeek-V3.2 | 8.72 | 7.49 | 0.61 | 7.05 | 0.57 |
| GPT-5.1 | DeepSeek-V3.2 | 9.03 | 7.50 | 0.67 | 6.87 | 0.62 |
| RMAD | - | 0.01 | 0.01 | 0.05 | 0.01 | 0.05 |
| DeepSeek-V3.2 | Gemini-2.5-Flash | 8.68 | 4.04 | 0.32 | 5.90 | 0.49 |
| Qwen3-Max | Gemini-2.5-Flash | 8.41 | 5.07 | 0.38 | 6.48 | 0.43 |
| GPT-5.1 | Gemini-2.5-Flash | 8.78 | 4.56 | 0.37 | 6.34 | 0.52 |
| RMAD | - | 0.02 | 0.08 | 0.07 | 0.04 | 0.07 |

Table 7: Evaluation results across different models.

**Code Generation** In code generation tasks, AstroSage3-8B often produces snippets that can be identified via regex but fail to execute due to inconsistent or incorrect indentation. AstroOne, when operating under the think strategy, often directly triggers runtime errors that interrupt the reasoning process. Even when execution succeeds, the outputs may contain unexpected artifacts (e.g., special characters), severely compromising output consistency.

**Hidden Formatting Errors** An additional, subtler yet impactful issue arises when models generate JSON strings that appear syntactically correct but contain invalid internal field values that violate JSON constraints. This issue is particularly frequent in LLaMA-3.1-8B, further undermining its reliability in automated evaluation.

In summary, we summarize and analyze representative failure modes observed during evaluation. These findings reveal that existing models suffer from critical capability gaps and the neglect of essential competencies. In particular, deficiencies in output format and syntax remain prominent. Such weaknesses pose substantial challenges for automated evaluation and deployment, particularly in scenarios involving JSON and executable code generation.

## H Model Evaluation Bias

The evaluation depends on an LLM-as-a-judge setup, with DeepSeek-V3.2 used for process-level scoring. That makes the benchmark easier to scale, but it also introduces another model's bias into the assessment of reasoning quality. The LLM-as-a-judge evaluations may introduce bias from the evaluator model itself (Panickssery et al., 2024). However, we would like to clarify the following points: Our use of LLM-as-a-judge for process-level evaluation follows common practice in the current literature. LLM-based evaluation has been widely adopted in tasks that require strong semantic understanding, including scientific reasoning benchmarks in medicine, mathematics, and astronomy, as well as literature mining and scientific reviewing. In contrast, traditional automatic metrics such as BLEU, ROUGE, or embedding-based similarity mainly rely on surface-form matching, and therefore are not well suited to assessing reasoning quality or semantic correctness in open-ended responses. This limitation is especially pronounced in our setting, where the focus is on process-level reasoning quality.

We took several steps to reduce evaluation bias as much as possible: i) We used clearly specified and detailed scoring criteria in the evaluation prompts to reduce subjectivity (see Appendix, Experiment Evaluation Prompt Setting 8,9,10). ii) We emphasized process-level scoring, requiring the evaluator to focus on the correctness and consistency of intermediate reasoning steps rather than only the final answer. This helps reduce the influence of stylistic or wording differences.

The results in Table 4 provide empirical support for the reliability of our evaluation setup. We observe high correlation and strong agreement between human expert scores and model-based scores, with relatively small score deviations. This suggests that evaluator-induced bias is limited in our setting.

Finally, to address this concern, we conduct an additional experiment in which multiple LLM judges from different sources independently score the same set of model outputs. We then analyze the

| Evaluation Model | Model | FD Score | UC Score | UC Acc | FC Score | FC Acc |
|---|---|---|---|---|---|---|
| DeepSeek-V3.2 | GPT-5.6-Luna | 9.10 | 8.19 | 0.74 | 7.45 | 0.70 |
| Qwen3-Max | GPT-5.6-Luna | 9.35 | 8.06 | 0.70 | 7.73 | 0.68 |
| ChatGLM-5.3 | GPT-5.6-Luna | 9.38 | 8.07 | 0.74 | 7.42 | 0.71 |
| Ensemble Score | - | 9.28 ± 0.13 | 8.11 ± 0.06 | 0.73 ± 0.02 | 7.53 ± 0.14 | 0.70 ± 0.01 |

Table 8: Evaluation results from Ensemble Model.

| Model | FD Score | UC Score | UC Acc |
|---|---|---|---|
| Gemini-2.5-Flash | 8.45 | 8.53 | 0.61 |
| Claude-haiku-4.5 | 7.60 | 6.42 | 0.33 |
| Qwen3-VL-32B | 7.90 | 5.28 | 0.33 |
| GPT-5.1 | 9.00 | 7.68 | 0.67 |
| GLM-4.6V | 7.70 | 5.64 | 0.22 |
| Kimi K2.5 | 8.57 | 8.22 | 0.72 |

Table 9: Performance comparison on the MC task.

average variation across judges using Relative Mean Absolute Deviation (RMAD). The results are summarized in the table below.

As shown in the Table 7, under our benchmark design and prompt constraints, although different evaluator models produce slightly different absolute scores, they remain highly consistent in overall trends and model rankings (RMAD ¡ 8%). This suggests that the impact of single-judge bias is mitigated to a meaningful extent in our setup. Meanwhile, we also agree that an ensemble of judges would be a more ideal solution, although it comes with substantially higher evaluation cost. We view this as an important direction for future work.

To examine the robustness of our evaluation protocol, we additionally conduct a small-scale study using an ensemble of three independent evaluator models, including DeepSeek-V3.2, Qwen3-Max, and ChatGLM-5.3. As shown in Table 8, the ensemble scores are computed as the mean across the three evaluators, with the standard deviation reported to characterize evaluator-level variation. The relatively small deviations across all metrics (e.g., 9.28 ± 0.13 for FD and 7.53 ± 0.14 for FC) indicate that the evaluation results are generally consistent across different evaluator models. This provides supplementary evidence that the conclusions drawn from a single evaluator are unlikely to be dominated by model-specific evaluation bias. For simplicity and consistency, we therefore use a single evaluator in the main experiments, while the ensemble-based evaluation is provided as a robustness check in the appendix.

## I Multimodal Task Evaluation Extension

We selected the current set of models mainly based on two considerations: *i) Representativeness:* The selected models cover different types of multimodal architectures and training paradigms, including frontier proprietary models, open-source vision-language models, and instruction-tuned models. This allows them to reflect the major current technical directions to a reasonable extent. *ii) Practical cost constraints:* Systematic evaluation of large multimodal models is expensive in both computation and monetary cost, especially on our benchmark, which involves complex astronomical reasoning across multiple dimensions.

At the same time, we would like to emphasize that the primary goal of our work is not to provide an exhaustive ranking of all existing multimodal models. Rather, our goal is to introduce a challenging and scalable benchmark for multimodal astronomical reasoning, and to use representative models to reveal the current capability boundaries and failure modes of existing methods.

As shown in Table 9, to better address it while maintaining a reasonable cost budget, we have additionally included results for three mainstream multimodal models on the multimodal subtask MC. Overall, on multimodal tasks, Kimi K2.5 shows relatively strong performance.

## J Explanation of Difficulty Coefficient Definition

We provide a clearer explanation of the motivation behind our difficulty modeling design, together with additional empirical validation. We would like to further clarify that the coefficients in Eq.

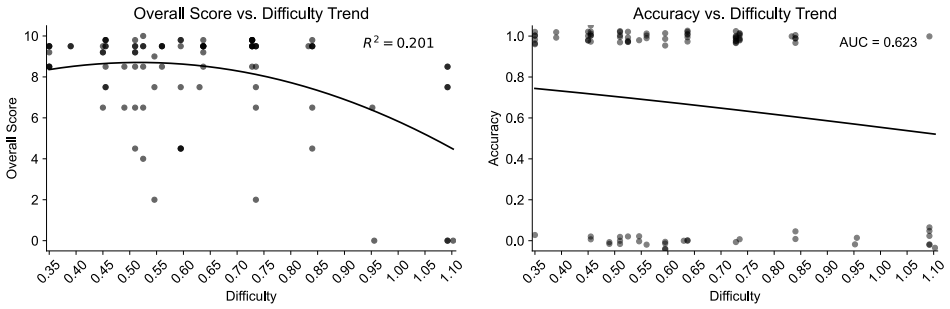



Figure 4: Model scores and accuracy across different difficulty levels.

(1) were not set at random. Instead, they were designed based on commonly recognized factors that affect problem difficulty in educational assessment and problem-solving research, including the following aspects:

First, the data source is used to characterize differences in dataset selection criteria and overall difficulty level. For example, compared with Olympiad-style problems, questions drawn from textbooks or introductory astronomy materials are generally easier for non-astronomy students to solve after a short period of study; accordingly, we assign them a base difficulty coefficient of 0.5. Problems that students with some astronomy background can solve within a reasonable amount of time are assigned a coefficient of 0.6. In contrast, competition problems such as those from USAAAO or IOAA typically require systematic training and longer preparation, and are therefore assigned higher base coefficients of 0.7–0.8. Second, modality introduces additional cognitive burden for both humans and models. For instance, image-based questions usually require extra visual interpretation, which increases problem complexity. We therefore assign a higher difficulty weight (0.5) to multimodal problems. Third, problem type distinguishes between questions that can be solved through direct formula application and those that require reasoning or formula derivation. The latter usually involve more complex cognitive processes and are thus assigned higher difficulty coefficients. Fourth, reasoning steps are a common indicator of problem complexity. In general, problems requiring more solution steps tend to be more difficult. We therefore set the corresponding coefficients according to the number of reasoning steps needed to solve each problem.

In addition, we adopt a multiplicative formulation to combine these factors, with the goal of capturing their interactions. For example, when multi-step reasoning and multimodal information appear together, the problem difficulty often increases substantially; such an interaction is difficult to reflect through a simple linear sum.

To further address the concern, we also include additional empirical analysis in Figure 4. Specifically, we randomly sampled 93 questions spanning different difficulty ranges and conducted the following analyses: i) logistic regression to model how the discrete metric Accuracy varies with Difficulty; and ii) polynomial regression to fit the relationship between the continuous metric Overall Score and Difficulty. The results suggest that the current difficulty coefficient design is reasonably well aligned with the actual problem difficulty: both Accuracy and Overall Score show an overall decreasing trend as difficulty increases. Meanwhile, we also acknowledge that human and model cognition may differ. For example, even with incomplete knowledge coverage, a model may still answer some highly difficult questions correctly. This observation suggests that developing more precise ways to characterize and measure "difficulty" remains an important direction for future research.

## K Release Plan Explanation

Specifically, upon publication, we will release the following resources:

**Benchmark Construction Resources** We will release the complete codebase of the mutation-based problem generation pipeline, including the code generator, perturbation strategies, generation

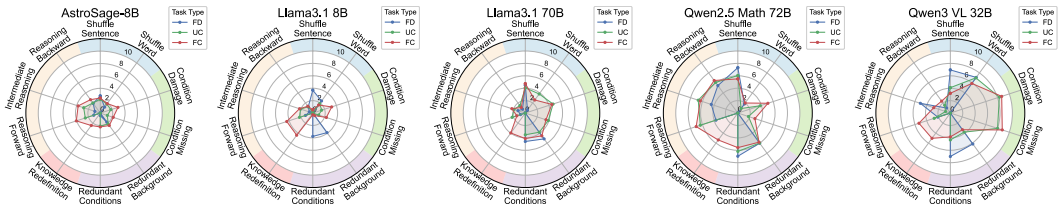



Figure 5: Performance of the five open-source models in each cognitive dimension. The intervals highlighted in blue correspond to "Problem Comprehension"; the intervals shown in green represent "Condition Robustness"; the intervals marked in purple relate to "Anti-Interference"; the red intervals correspond to "Knowledge Consistency"; and the intervals displayed in orange represent "Reverse Reasoning".

prompts, and constraint checking modules. These resources enable researchers to reproduce the proposed problem generation process and construct new variants following the same principles.

**Evaluation Resources** We will release the evaluation scripts, configuration files, and the processed benchmark instances used in our experiments. The generated problem variants, associated images, and metadata required for model evaluation will also be publicly available.

**Data Release and Hidden Evaluation Protocol** The original seed database contains privately collected expert-designed problems and serves as the foundation for generating diverse variants. To preserve benchmark integrity, prevent potential contamination, and maintain a reliable future evaluation environment, we will not release the complete seed database and corresponding ground-truth answers. Instead, these resources will be maintained as a hidden evaluation set.

To support broader access and continuous evaluation, we plan to establish an online evaluation platform where researchers can submit their models and obtain evaluation results on newly generated variants. The platform will distinguish the main leaderboard from additional diagnostic leaderboards based on different evaluation dimensions. During the transition period, researchers may also request access to additional evaluation data through the contact information provided in our repository.

**Human Annotation Resources** The human annotations are used for benchmark construction and quality verification. Since they are directly associated with the hidden evaluation set, we will not release the complete annotation records to avoid exposing test information. Instead, we will provide detailed annotation protocols and quality-control procedures in the paper and supplementary materials to ensure transparency and reproducibility.

## L Examples for Different Question Types

**Formula Derivation (Text-only):** As illustrated in Table 10, the formula derivation Example explicitly requires the model to derive the Virial theorem from given premises. Beyond this surface requirement, the task implicitly evaluates a model's capability for deep interdisciplinary reasoning. Successfully solving the problem demands an accurate understanding of the physical meaning and mathematical structure of the virial theorem across classical mechanics, statistical physics, and astrophysics. Moreover, the model must abstract key physical quantities from natural language descriptions, perform multi-step symbolic reasoning grounded in Newton's second law, and present a logically coherent and standardized derivation. The task further probes consistency in concept definitions, continuity of logical exposition, and sensitivity to the historical and theoretical context of the theorem, rather than reliance on memorized conclusions or superficial paraphrasing.

**Formula Derivation (Image + Text):** As illustrated in Table 11, we present a multimodal formula derivation example, where the model is explicitly asked to derive the differential equations governing the two-body problem by leveraging both textual descriptions and a coordinate system depicted in an image. Implicitly, this task places stronger emphasis on image–text joint understanding. The model must correctly reconstruct the physical scenario from heterogeneous representations, identify

```
keywordstyle
def generate_alt_az_problem(phi: float, delta: float, t: float) -> Dict[str, str]:
    """ Generate an astronomy computational reasoning problem and its solution steps.

    Parameters
    ----------
    phi : float. Observer's latitude in degrees.
    delta : float. Declination of the star in degrees.
    t : float. Hour angle of the star in degrees.

    Returns
    -------
    Dict[str, str]. A dictionary containing the question, solution steps, and final answer. ""
        "

    def deg_sin(x: float) -> float: return math.sin(math.radians(x))

    def deg_cos(x: float) -> float: return math.cos(math.radians(x))

    def deg_asin(x: float) -> float: return math.degrees(math.asin(x))

    # Step 1: Altitude
    sin_h = ( deg_sin(phi) * deg_sin(delta) + deg_cos(phi) * deg_cos(delta) * deg_cos(t))

    h = deg_asin(sin_h)

    # Step 2: Azimuth from sine formula
    sin_A = (deg_cos(delta) * deg_sin(t)) / deg_cos(h)

    # Step 3: Azimuth from cosine formula
    cos_A = (deg_sin(phi) * deg_sin(h) - deg_sin(delta)) / (deg_cos(phi) * deg_cos(h))

    # Step 4: Determine unique azimuth using quadrant
    A_rad = math.atan2(sin_A, cos_A)
    A = math.degrees(A_rad)

    question = (
        f"Given the declination of a star as δ = {delta:.1f}° and the hour angle "
        f"t = {t:.1f}°, calculate the altitude h and azimuth A of this star when "
        f"observed at latitude φ = {phi:.1f}°.")

    solution = f""" Step 1: Using the spherical trigonometric formula: sin h = sin φ sin δ +
        cos φ cos δ cos t

sin h = sin({phi:.1f}°)·sin({delta:.1f}°) + cos({phi:.1f}°)·cos({delta:.1f}°)·cos({t:.1f}°) =
    {sin_h:.4f}

Thus, h = {h:.1f}°

Step 2: Using the spherical trigonometric formula: sin A = (cos δ sin t) / cos h

sin A = (cos({delta:.1f}°)·sin({t:.1f}°)) / cos({h:.1f}°) = {sin_A:.4f}

Step 3: Using the spherical trigonometric formula: cos A = (sin φ sin h - sin δ) / (cos φ cos
    h)

cos A = (sin({phi:.1f}°)·sin({h:.1f}°) - sin({delta:.1f}°)) / (cos({phi:.1f}°)·cos({h:.1f}°))
    = {cos_A:.4f}

Step 4: Combining sin A and cos A to determine the correct quadrant: A = {A:.1f}° """

    final_answer = f"h = {h:.1f}°, A = {A:.1f}°"

    return {"question": question, "solution_steps": solution.strip(), "final_answer":
        final_answer}
```

Figure 6: Function Definition Point for Code-Formalized Problems

hidden assumptions (e.g., absence of external forces, choice of inertial frame), and apply foundational laws of classical mechanics—Newton's second law and universal gravitation—to formulate rigorous vector-valued differential equations. This example also evaluates the model's ability to manage multiple symbolic variables, reason about reference frames and relative motion, and transition from conceptual descriptions to formal mathematical expressions, which are central indicators of physics-style theoretical reasoning.

keywordstyle

```python
def main() -> None:
    """
    Program entry point. Randomly generate input parameters and
    produce a complete astronomy problem with solution.
    """
    phi = random.uniform(20.0, 50.0)       # latitude
    delta = random.uniform(-10.0, 70.0)    # declination
    t = random.uniform(0.0, 360.0)         # hour angle

    problem = generate_alt_az_problem(phi, delta, t)

    print(problem["question"])
    print(problem["solution_steps"])
    print(problem["final_answer"])
```

Figure 7: Invocation Entry Point: Example of Code-Formalized Problems

| | |
|---|---|
| **Task** | Formula Derivation |
| **Difficulty** | 0.546 |
| **Image** | None |
| **Question** | A class of dynamical methods for determining the mass of galaxies is the application of the virial theorem. The virial theorem was proposed by the German physicist Rudolf Julius Clausius in the 19th century. For a thermodynamic system composed of many molecules, let the mass of a single molecule be $m$, its position vector relative to an arbitrarily chosen origin be $\mathbf{r}$, and the resultant force of all forces acting on it be $\mathbf{F}$. Consider the inner product $\mathbf{r} \cdot \mathbf{F}$. The sum of this inner product over all molecules in the system, $\sum \mathbf{r} \cdot \mathbf{F}$, is called the virial. Now, please derive the formula of the virial theorem based on Newton's second law. |
| **Solution** | **According to Newton's second law:**<br>$$\mathbf{F} = m\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2}, \quad \mathbf{r} \cdot \mathbf{F} = m\,\mathbf{r} \cdot \frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2}$$<br>**Considering that:**<br>$$m\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathbf{r} \cdot \frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}\right) = m\left(\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}\right)^2 + m\,\mathbf{r} \cdot \frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2} = m\left(\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}\right)^2 + \mathbf{r} \cdot \mathbf{F}$$<br>Summing over the entire system:<br>$$\sum m\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathbf{r} \cdot \frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}\right) = \sum m\left(\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}\right)^2 + \sum \mathbf{r} \cdot \mathbf{F}$$<br>If only gravitational forces exist among the molecules, the entire system is a gravitational system, and the forces acting on it are all central forces. Then, for a stable system governed by central forces, we conclude:<br>$$\sum m\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathbf{r} \cdot \frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}\right) = 0$$<br>**According to:** All central forces are conservative forces, so $\sum \mathbf{r} \cdot \mathbf{F}$ is equivalent to the potential energy $U$ of a conservative system, which may also be written as:<br>$$U = \sum \mathbf{r} \cdot \mathbf{F}$$<br>The total kinetic energy of all molecules in the system is:<br>$$T = \frac{1}{2}\sum mv^2 = \frac{1}{2}\sum m\left(\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}\right)^2$$<br>**Conclude:** The differential form of the virial theorem can be written as:<br>$$2T + U = 0$$<br>where $T$ is the total kinetic energy of the system, and $U$ is the total gravitational potential energy of the entire system. |
| **Final Answer** | None |

Table 10: Formula Derivation Example: Derivation of the Virial Theorem

**Unimodal Computation (Text-only):** As illustrated in Table 12, the unimodal computation example shows the model is explicitly tasked with estimating the mass of a globular cluster. Implicitly, this problem assesses a broader spectrum of scientific reasoning abilities. These include invoking relevant astrophysical background knowledge (e.g., globular clusters, velocity dispersion, the virial theorem), maintaining consistency across units and dimensions (e.g., converting angular scales to physical distances), and constructing reasonable physical assumptions given incomplete

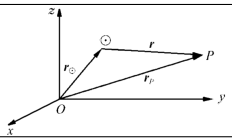
| | |
|---|---|
| **Task** | Formula Derivation |
| **Difficulty** | 0.819 |
| **Image** |   |
| **Question** | The two-body problem concerns how two celestial bodies move under mutual gravitational attraction when only these two bodies are considered (neglecting perturbations from all others). Let there be only the Sun $\odot$ and a planet $P$. Establish an arbitrary spatial coordinate system $O-xyz$(As shown in the figure<image>GSTD-Fig-1</image>). $r_\odot$ is the position vector of the Sun, $r_P$ is the position vector of the planet, and $r$ is the position vector of the planet relative to the Sun. Denote the masses of the Sun and the planet by $M$ and $m$. Based on the given information, derive the differential equations for the two-body problem. |
| **Solution** | **According to Newton's second law:** $$\mathbf{F}=m\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2},\quad \mathbf{r}\cdot\mathbf{F}=m\,\mathbf{r}\cdot\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2}$$ **Considering that:** According to: The Sun is subject to the gravitational force exerted by the planet $$\mathbf{F}_\odot=k^2\frac{Mm}{r^2}\frac{\mathbf{r}}{|\mathbf{r}|}=k^2\frac{Mm}{r^3}\mathbf{r}.$$ Likewise, the planet is subject to the gravitational force exerted by the Sun. $$\mathbf{F}_P=k^2\frac{Mm}{r^2}\frac{-\mathbf{r}}{|\mathbf{r}|}=-k^2\frac{Mm}{r^3}\mathbf{r}.$$ From Newton's second law, both the Sun and the planet accelerate under the action of force $$\mathbf{F}=\frac{\mathrm{d}(m\mathbf{v})}{\mathrm{d}t}=m\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2}.$$ Conclude: $$\mathbf{F}_\odot=M\frac{\mathrm{d}^2\mathbf{r}_\odot}{\mathrm{d}t^2}=k^2\frac{Mm}{r^3}\mathbf{r},$$ $$\mathbf{F}_P=m\frac{\mathrm{d}^2\mathbf{r}_P}{\mathrm{d}t^2}=-k^2\frac{Mm}{r^3}\mathbf{r}.$$ According to: $$\mathbf{r}=\mathbf{r}_P-\mathbf{r}_\odot.$$ Conclude: $$\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2}=\frac{\mathrm{d}^2}{\mathrm{d}t^2}(\mathbf{r}_P-\mathbf{r}_\odot)=-k^2\frac{M+m}{r^3}\mathbf{r}=-\frac{\mu}{r^3}\mathbf{r}.$$ According to: $$\mu=k^2(M+m).$$ Establish a Cartesian coordinate system $x$–$y$–$z$ with the Sun as the origin. Conclude: The differential equations of the two-body problem are $$\begin{cases}\frac{\mathrm{d}^2x}{\mathrm{d}t^2}=-\frac{\mu}{r^3}x,\\ \frac{\mathrm{d}^2y}{\mathrm{d}t^2}=-\frac{\mu}{r^3}y,\\ \frac{\mathrm{d}^2z}{\mathrm{d}t^2}=-\frac{\mu}{r^3}z,\end{cases}$$ where $r=\sqrt{x^2+y^2+z^2}$. |
| **Final Answer** | None |

Table 11: Formula Derivation Example: Derivation of the differential equations for the two-body problem

information. Crucially, the task combines multi-step logical reasoning with numerical estimation, thereby testing whether the model possesses physics-intuition-driven quantitative reasoning, rather than merely executing symbolic or arithmetic computations.

**Multimodal Computation:** As illustrated in Table 13, we present a multimodal computation example, where the explicit objective is the calculation of an oscillation frequency using both visual and textual inputs. Implicitly, the task probes the model's ability to translate complex natural language scenarios into a formal physical model, with particular emphasis on Keplerian orbital mechan-

| | |
|---|---|
| **Task** | Unimodal Computation |
| **Difficulty** | 0.546 |
| **Image** | None |
| **Question** | Estimate the mass of a globular cluster with a radial velocity dispersion $\sigma_r = 16.2$ km/s. The cluster has an angular diameter of $\theta = 3.56'$ and is a distance $d = 9630$ pc away from us. |
| **Solution** | **Step 1:** In order to estimate the mass of the globular cluster, we use the Virial theorem, which relates the time-averaged values of the kinetic and potential energy of a gravitationally bound system.<br>$$\langle T \rangle = -\frac{1}{2}\langle U \rangle$$<br>Note that $\sigma^2 = \sigma_x^2 + \sigma_y^2 + \sigma_z^2$ and $\sigma_r^2 = \sigma_x^2 = \sigma_y^2 = \sigma_z^2$ (assuming velocity dispersions are randomly distributed in a globular cluster). Thus,<br>$$\sigma^2 = 3\sigma_r^2.$$<br>**Step 2:** Additionally, we use the expression<br>$$\langle U \rangle = -\frac{3}{5}\frac{GM^2}{R}$$<br>for the gravitational potential energy of a uniform sphere.<br>$$\frac{1}{2}M\sigma^2 = -\frac{1}{2}\left(-\frac{3}{5}\frac{GM^2}{R}\right)$$<br>Substituting $\sigma^2 = 3\sigma_r^2$,<br>$$\frac{1}{2}M\left(3\sigma_r^2\right) = \frac{3}{10}\frac{GM^2}{R}.$$<br>**Step 3:** We also note that:<br>$$R = \frac{D}{2} = \frac{d\theta}{2}$$<br>where $\theta$ needs to be converted from arcminutes to radians before being substituted. Solving for $M$, replacing $R$ with the expression from above, and substituting numerical values gives<br>$$M = \frac{5\sigma_r^2 R}{G} = \frac{5\sigma_r^2 d\theta}{2G} \approx 3.03 \times 10^{36}\,\text{kg}.$$ |
| **Final Answer** | $$M = 3.03 \times 10^{36}\,\text{kg}$$ |

Table 12: Unimodal Computation Example: Mass Estimation of a Globular Cluster

ics, including orbital periods, eccentricity, perihelion/aphelion properties, and velocity directions. Additionally, the example evaluates the model's capability for optimization-based temporal reasoning under multiple constraints, precise handling of implicit physical conditions (such as changes in velocity sign), robustness to missing or weakly specified visual information, and holistic integration of knowledge across astronomy, mathematics, and logical inference.

**Formula Calculation:** As illustrated in Table 14, we present a formula calculation example, where the explicit objective is the calculation of an oscillation frequency from given parameters. Implicitly, this task demands domain knowledge in astrophysics and asteroseismology, including an understanding of the physical meaning of large frequency separation, phase terms, angular degree, and radial order in asymptotic relations. The model must correctly map natural language descriptions to appropriate mathematical formulas, perform accurate symbolic substitution and numerical computation, and distinguish between core physical parameters and correction terms. Furthermore, in the absence of explicitly provided formulas, the task tests the model's ability to engage in knowledge-retrieval-style reasoning and structured computation, which constitutes a primary objective of this formula-centric task category.

**Code-Based Calculation:** As illustrated in Table 15, we present a code-based calculation example, which evaluates the model's end-to-end capability to formalize a natural language physics problem and solve it computationally. This includes understanding concepts from astrodynamics and orbital mechanics (e.g., low Earth orbit, Hohmann transfer, semi-major axis, $\Delta v$, gravitational parameters), identifying and modeling implicit assumptions (such as the two-body approximation and patched-conic methods), and translating mathematical expressions into executable Python code. The task further assesses rigor in unit consistency, numerical accuracy, and result interpretation. At its

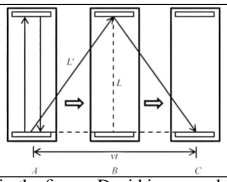
| | |
|---|---|
| **Task** | Multimodal Computation |
| **Difficulty** | 1.1025 |
| **Image** |  |
| **Question** | <image>JSTL-Fig-2</image>As shown in the figure, David is researching an Earth-like exoplanet near the constellation of Bootes, orbiting a star twice the mass of our Sun. On March 14, 2025, the planet was observed at 3 AU from its parent star. After some time, David observes it again and finds that it has already passed its farthest point from the star and is now at 4 AU. If the planet is known to have a period of 5 years and an eccentricity of 0.5, what is the shortest amount of time that could have passed since the first observation on March 14, 2025? Assume that David views the orbit face-on and that the exoplanet's velocity component along the semimajor axis has opposite sign at the two observation points. |
| **Solution** | **Step 1:** First, we need to find the planet's semi-major axis through Kepler's third law.<br>We have:<br>$$\frac{a^3}{T^2} = M$$<br>$$a = \sqrt[3]{MT^2} = \sqrt[3]{2 \times 5^2} \approx 3.684 \text{ AU}$$<br>**Step 2:** Next, let's find the true anomaly at 3 AU and at 4 AU. This gives the position of the planet in the elliptical orbit. Note that when the planet is at 3 AU, it is before passing its farthest distance from its parent star, i.e., its true anomaly must be between $0^\circ$ and $180^\circ$. This also means that the true anomaly when the planet is at 4 AU must be between $180^\circ$ and $360^\circ$ since it has passed its farthest distance.<br>We have:<br>$$r = \frac{a(1-e^2)}{1+e\cos f}, \cos f = \frac{1}{e}\left(\frac{a(1-e^2)}{r} - 1\right)$$<br>$$f_{3\text{AU}} = 99^\circ 5' 27'', f_{4\text{ AU}} = 231^\circ 47' 36''$$<br>**Step 3:** From the true anomaly, we convert it to eccentric anomaly. Similarly, the eccentric anomaly when the planet is at 3 AU must be between $0^\circ$ and $180^\circ$, and the eccentric anomaly when the planet is at 4 AU must be between $180^\circ$ and $360^\circ$.<br>We have:<br>$$\cos E = \frac{e + \cos f}{1 + e\cos f}$$<br>$$E_{3\text{ AU}} = 68^\circ 12' 7.2'', E_{4\text{ AU}} = 260^\circ 7' 18.9''$$<br>**Step 4:** From the eccentric anomaly, we convert it to the mean anomaly. Similarly, the mean anomaly when the planet is at 3 AU must be between 0 and $\pi$, and the mean anomaly when the planet is at 4 AU must be between $\pi$ and $2\pi$. (The calculation below is in radians.)<br>We have:<br>$$M = E - e\sin E$$<br>$$M_{3\text{ AU}} = 0.7261 \text{ rad}, M_{4\text{ AU}} = 5.0326 \text{ rad}$$<br>**Step 5:** From the mean anomaly, we can find the elapsed time since March 14, 2025.<br>We have:<br>$$M = \frac{2\pi}{T} \times (t - t_0), \Delta M = \frac{2\pi}{T}\Delta t$$<br>$$\Delta t = \Delta M \times \frac{T}{2\pi} = (M_{4\text{ AU}} - M_{3\text{ AU}}) \times \frac{5 \text{ yrs}}{2\pi}$$<br>$$\Delta t \approx 3.427 \text{ years}$$ |
| **Final Answer** | $$\Delta t \approx 3.427 \text{ years}$$ |

Table 13: Multimodal Computation Example: Calculation of Minimum Exoplanet Orbital Transit Time

core, this example measures whether the model can perform end-to-end scientific reasoning, spanning natural language understanding, physical modeling, mathematical derivation, programmatic implementation, and quantitative conclusion.

| | |
|---|---|
| **Task** | Formula Calculation |
| **Difficulty** | 0.5 |
| **Image** | None |
| **Question** | Calculate the stellar oscillation frequency with the following parameters: Large frequency separation $\Delta\nu = 287.74\,\mu$Hz Phase offset $\epsilon = 1.41$ Angular degree $l = 2$ Radial order $n = 24$ Offset term $d_{02} = 4.41\,\mu$Hz |
| **Solution** | **Step 1:** Confirm the input parameters:<br>$$\Delta\nu = 287.74\mu\text{Hz}, \epsilon = 1.41, l = 2, n = 24, d_{02} = 4.41\mu\text{Hz}.$$<br>**Step 2:** Use the oscillation frequency formula:<br>$$\nu_{n,l} \approx \Delta\nu \left(n + \frac{l}{2} + \epsilon\right) - d_{0l}.$$<br>**Step 3:** Compute the term in parentheses:<br>$$24 + \frac{2}{2} + 1.41 = 26.41.$$<br>**Step 4:** Compute the frequency:<br>$$287.74 \times 26.41 - 4.41 = 7595.07\,\mu\text{Hz}.$$ |
| **Final Answer** | The oscillation frequency is $\nu_{24,2} = 7595.07\,\mu$Hz |

Table 14: Formula Calculation Example: Calculation of Oscillation Frequency

| | |
|---|---|
| **Task** | Code-based Calculation |
| **Difficulty** | 0.5 |
| **Image** | None |
| **Question** | Suppose a spacecraft was orbiting in a low Earth orbit at an altitude of 400 km. The spacecraft makes a single orbital maneuver to place it into a Mars transfer orbit. Delta-v ($\Delta v$) refers to the change in velocity during an orbital maneuver. What is the $\Delta v$ required for this trans-Mars injection? The semimajor axes of the orbits of Earth and Mars are $1.496 \times 10^8$ km and $2.279 \times 10^8$ km, respectively. |
| **Code** | Python Code<br>```python<br>def generate_trans_mars_injection_problem(<br>    leo_altitude_km: float,<br>    a_earth_sun_km: float,<br>    a_mars_sun_km: float,<br>    mu_sun_km3_s2: float = 1.32712440018e11,<br>    mu_earth_km3_s2: float = 3.986004418e5,<br>    r_earth_km: float = 6371.0,<br>) -> Dict[str, str]:<br><br>    if leo_altitude_km <= 0:<br>        raise ValueError("leo_altitude_km must be positive.")<br>    if a_earth_sun_km <= 0 or a_mars_sun_km <= 0:<br>        raise ValueError("Semimajor axes must be positive.")<br>    if a_mars_sun_km <= a_earth_sun_km:<br>        raise ValueError("For a Mars transfer, Mars' semimajor axis should exceed Earth's.")<br><br>    # --- Step 1: Hohmann transfer semimajor axis (km) ---<br>    a_transfer_km = 0.5 * (a_earth_sun_km + a_mars_sun_km)<br><br>    # --- Steps 2-3: Heliocentric speeds and hyperbolic excess speed (km/s) ---<br>    r_earth_heliocentric_km = a_earth_sun_km<br>    v1_km_s = math.sqrt(mu_sun_km3_s2 * (2.0 / r_earth_heliocentric_km - 1.0 / a_transfer_km))<br>    v_earth_km_s = math.sqrt(mu_sun_km3_s2 / r_earth_heliocentric_km)<br>    v_inf_km_s = v1_km_s - v_earth_km_s<br><br>    # --- Step 4: LEO orbital speed and escape speed (km/s) ---<br>    r_p_km = r_earth_km + leo_altitude_km<br>    v_orb_km_s = math.sqrt(mu_earth_km3_s2 / r_p_km)<br>    v_esc_km_s = math.sqrt(2.0) * v_orb_km_s<br><br>    # --- Step 5: Perigee speed on departure hyperbola and delta-v (km/s) ---<br>    v_p_km_s = math.sqrt(v_esc_km_s**2 + v_inf_km_s**2)<br>    delta_v_km_s = v_p_km_s - v_orb_km_s<br><br>    final_answer = f"\\boxed{{\\Delta v = {delta_v_km_s:.3f}\\ \\text{{km/s}}}}"<br><br>    return final_answer<br>``` |
| **Final Answer** | $\Delta v = 3.569 km/s$ |

Table 15: Code-based Calculation Example

# M Additional Experimental Results

## M.1 More Results on Different Task Types

This section provides additional experimental results on different task types. We report multi-dimensional evaluation scores for each model, including correctness, completeness, relevance, coherence, and detail, to assess the quality of the reasoning process. As shown in Table 16, we present a fine-grained comparison of different models on FD. Table 17 reports detailed performance results of different models on UC. In addition, Table 18 reports the evaluation results on FC.

| Model | Mode | Overall | Correctness | Completeness | Relevance | Coherence | Detail |
|---|---|---|---|---|---|---|---|
| GPT-5.6-Luna | *Ori* | 9.10 (5.95) | 9.50 (6.21) | 8.79 (5.76) | 9.82 (6.44) | **9.79** (6.43) | **9.50** (6.22) |
| | *+Think* | 9.41 (6.17) | 9.82 (6.43) | 9.00 (5.89) | 9.94 (6.54) | 9.85 (6.47) | 9.79 (6.43) |
| | *+CoT* | 9.38 (6.14) | 9.82 (6.43) | 9.00 (5.89) | 9.88 (6.48) | 9.79 (6.41) | 9.79 (6.41) |
| | *-System* | 9.18 (6.01) | 9.62 (6.29) | 8.82 (5.78) | 9.88 (6.48) | 9.91 (6.51) | 9.79 (6.42) |
| | *+Image* | 7.70 (6.94) | 7.90 (7.10) | 7.30 (6.58) | 9.00 (8.17) | 8.70 (7.90) | 7.20 (6.51) |
| AstroOne | *Ori* | **9.13** (5.87) | **9.56** (6.15) | **9.06** (5.82) | **9.88** (6.38) | 9.31 (6.01) | 8.88 (5.69) |
| | *+Think* | 6.35 (4.07) | 6.71 (4.29) | 6.03 (3.86) | 6.88 (4.41) | 6.74 (4.31) | 6.06 (3.88) |
| | *+CoT* | 8.78 (5.67) | 9.18 (5.92) | 8.53 (5.51) | 9.26 (5.98) | 8.94 (5.77) | 8.59 (5.55) |
| | *-System* | 8.08 (5.20) | 8.35 (5.37) | 8.00 (5.15) | 8.50 (5.48) | 8.32 (5.37) | 7.91 (5.10) |
| Qwen3-Max | *Ori* | 9.04 (5.91) | 9.47 (6.19) | 8.82 (5.78) | 9.79 (6.42) | 9.32 (6.09) | 8.82 (5.77) |
| | *+Think* | 9.15 (5.98) | 9.59 (6.27) | 9.00 (5.88) | 9.76 (6.40) | 9.41 (6.16) | 8.88 (5.82) |
| | *+CoT* | 9.03 (5.90) | 9.32 (6.09) | 9.03 (5.90) | 9.65 (6.31) | 9.24 (6.03) | 8.74 (5.70) |
| | *-System* | 8.85 (5.78) | 9.18 (5.99) | 8.74 (5.72) | 9.44 (6.18) | 9.44 (6.18) | 8.56 (5.60) |
| DeepSeek-V3.2 | *Ori* | 8.94 (5.86) | 9.24 (6.04) | 8.85 (5.80) | 9.76 (6.41) | 9.32 (6.12) | 8.97 (5.89) |
| | *+Think* | 8.75 (5.71) | 9.12 (5.94) | 8.35 (5.45) | 9.71 (6.37) | 9.38 (6.13) | 8.44 (5.51) |
| | *+CoT* | 9.18 (6.00) | 9.62 (6.29) | 8.91 (5.82) | 9.82 (6.44) | 9.65 (6.32) | 8.88 (5.81) |
| | *-System* | 9.04 (5.91) | 9.50 (6.21) | 8.76 (5.73) | 9.82 (6.43) | 9.38 (6.13) | 8.82 (5.78) |
| Gemini-2.5-Flash | *Ori* | 8.68 (5.65) | 9.03 (5.88) | 8.65 (5.63) | 9.56 (6.25) | 9.21 (6.01) | 8.71 (5.67) |
| | *+Think* | 8.52 (5.54) | 8.91 (5.80) | 8.53 (5.54) | 9.85 (6.46) | 9.18 (6.01) | 8.59 (5.61) |
| | *+CoT* | 8.54 (5.48) | 8.76 (5.62) | 8.56 (5.49) | 9.56 (6.23) | 9.15 (5.97) | 8.53 (5.50) |
| | *-System* | 8.75 (5.67) | 9.12 (5.89) | 8.68 (5.61) | 9.47 (6.19) | 9.29 (6.07) | 8.68 (5.63) |
| | *+Image* | 8.45 (7.75) | 9.05 (8.28) | 8.30 (7.60) | 9.25 (8.45) | 9.10 (8.33) | 8.05 (7.38) |
| Claude-haiku-4.5 | *Ori* | 7.71 (5.00) | 7.91 (5.13) | 7.44 (4.83) | 9.32 (6.10) | 8.85 (5.78) | 7.65 (4.97) |
| | *+Think* | 8.03 (5.21) | 8.38 (5.42) | 7.62 (4.95) | 9.38 (6.13) | 8.91 (5.81) | 7.65 (4.97) |
| | *+CoT* | 7.96 (5.20) | 8.24 (5.36) | 7.59 (4.97) | 9.44 (6.19) | 8.88 (5.81) | 7.59 (4.96) |
| | *-System* | 7.26 (4.72) | 7.41 (4.80) | 6.94 (4.50) | 9.15 (5.99) | 8.59 (5.60) | 6.91 (4.48) |
| | *+Image* | 7.60 (6.86) | 7.85 (7.07) | 7.35 (6.65) | 8.95 (8.14) | 8.50 (7.71) | 7.45 (6.73) |
| AstroSage3-8B | *Ori* | 2.63 (1.73) | 2.19 (1.45) | 2.44 (1.61) | 4.63 (3.01) | 3.63 (2.41) | 2.25 (1.49) |
| | *+CoT* | 2.97 (1.94) | 2.47 (1.63) | 2.88 (1.88) | 5.29 (3.43) | 4.26 (2.79) | 2.56 (1.66) |
| | *-System* | 3.00 (1.95) | 2.68 (1.72) | 2.74 (1.78) | 4.76 (3.15) | 3.94 (2.58) | 2.59 (1.67) |
| Llama-3.1-8B | *Ori* | 4.03 (2.54) | 3.62 (2.26) | 3.88 (2.44) | 6.94 (4.45) | 6.29 (4.04) | 3.79 (2.40) |
| | *+CoT* | 3.74 (2.41) | 3.29 (2.13) | 3.41 (2.19) | 6.47 (4.19) | 5.35 (3.45) | 3.35 (2.15) |
| | *-System* | 4.11 (2.66) | 3.76 (2.43) | 4.00 (2.58) | 6.47 (4.22) | 6.09 (3.97) | 3.32 (2.17) |
| Llama-3.1-70B | *Ori* | 5.24 (3.33) | 5.26 (3.33) | 4.94 (3.13) | 7.65 (4.96) | 6.68 (4.31) | 4.56 (2.87) |
| | *+CoT* | 5.21 (3.32) | 5.18 (3.28) | 4.82 (3.06) | 7.53 (4.88) | 6.82 (4.40) | 4.82 (3.05) |
| | *-System* | 5.18 (3.31) | 5.18 (3.29) | 4.94 (3.16) | 7.59 (4.94) | 7.00 (4.54) | 4.71 (2.99) |
| Qwen2.5-Math-72B | *Ori* | 6.90 (4.42) | 6.85 (4.37) | 6.47 (4.13) | 8.82 (5.73) | 7.91 (5.11) | 6.71 (4.29) |
| | *+CoT* | 7.09 (4.53) | 7.32 (4.67) | 6.74 (4.27) | 9.00 (5.85) | 8.12 (5.26) | 6.74 (4.30) |
| | *-System* | 7.00 (4.49) | 7.18 (4.60) | 6.74 (4.30) | 9.03 (5.88) | 8.15 (5.29) | 6.94 (4.46) |
| Qwen3-VL-32B | *Ori* | 8.31 (5.38) | 8.62 (5.57) | 8.32 (5.39) | 9.50 (6.21) | 8.74 (5.68) | 8.21 (5.33) |
| | *+Think* | 7.14 (4.56) | 7.41 (4.72) | 7.00 (4.45) | 8.09 (5.17) | 7.56 (4.83) | 7.03 (4.48) |
| | *+CoT* | 8.25 (5.32) | 8.62 (5.56) | 8.29 (5.34) | 9.41 (6.14) | 8.65 (5.61) | 8.03 (5.19) |
| | *-System* | 8.57 (5.55) | 8.91 (5.77) | 8.47 (5.48) | 9.47 (6.17) | 8.79 (5.70) | 8.38 (5.42) |
| | *+Image* | 7.90 (7.17) | 8.50 (7.70) | 7.50 (6.79) | 8.80 (8.01) | 8.50 (7.73) | 7.60 (6.89) |

Table 16: Performance of different LLMs for the Formula Derivation (FD) task. Parenthetical scores report results after applying the weighted difficulty coefficient $\mathcal{D}$. "Ori" denotes the original model; "+Think" indicates that the model performs reasoning under a slow-thinking evaluation protocol; "+CoT" denotes the use of a chain-of-thought prompting strategy; "-System" removes the system prompt; "+Code" indicates code-based reasoning, corresponding to the CC task only; and "+Image" indicates text-image reasoning (if supported), corresponding to the MC task only. Values in red indicate anomalous results due to erroneous outputs. Bold font marks the best performance under the "Ori" setting.

## M.2 Additional Results for Cognitive Derivation Testing

This section presents additional results from the cognitive derivation experiments. We consider ten fine-grained dimensions: Shuffle Sentence, Shuffle Word Order, Redundant Background, Redundant Conditions, Condition Damage, Condition Missing, Knowledge Redefinition, Reasoning Forward,

| Model | Mode | Overall | Correctness | Completeness | Relevance | Coherence | Detail | Accuracy |
|---|---|---|---|---|---|---|---|---|
| GPT-5.6-Luna | *Ori* | **8.19** (4.40) | **8.46** (4.55) | **7.89** (4.24) | **9.48** (5.09) | **9.05** (4.87) | **8.09** (4.35) | **0.74** |
| | *+Think* | 8.16 (4.40) | 8.44 (4.56) | 7.86 (4.24) | 9.46 (5.09) | 8.99 (4.84) | 8.16 (4.40) | 0.75 |
| | *+CoT* | 8.13 (4.36) | 8.40 (4.51) | 7.84 (4.21) | 9.45 (5.07) | 8.98 (4.82) | 8.06 (4.33) | 0.73 |
| | *-System* | 8.18 (4.40) | 8.45 (4.55) | 7.87 (4.23) | 9.45 (5.07) | 8.96 (4.82) | 8.11 (4.36) | 0.76 |
| | *+Code* | - | - | - | - | - | - | 0.67 |
| | *+Image* | 8.43 (7.66) | 8.75 (7.94) | 8.14 (7.41) | 9.72 (8.90) | 9.39 (8.59) | 8.31 (7.55) | 0.61 |
| AstroOne | *Ori* | 6.68 (3.59) | 6.84 (3.68) | 6.61 (3.56) | 7.48 (4.01) | 7.08 (3.80) | 6.57 (3.53) | 0.54 |
| | *+Think* | 3.16 (1.98) | 3.74 (2.06) | 3.50 (1.92) | 3.99 (2.19) | 3.77 (2.07) | 3.58 (1.96) | 0.32 |
| | *+CoT* | 6.40 (3.42) | 6.57 (3.51) | 6.37 (3.40) | 7.21 (3.85) | 6.84 (3.65) | 6.28 (3.37) | 0.53 |
| | *-System* | 6.55 (3.52) | 6.73 (3.62) | 6.53 (3.51) | 7.31 (3.93) | 6.94 (3.72) | 6.51 (3.50) | 0.54 |
| | *+Code* | - | - | - | - | - | - | 0.60 |
| Qwen3-Max | *Ori* | 7.97 (4.29) | 8.19 (4.41) | **7.89** (4.24) | 8.82 (4.74) | 8.43 (4.53) | 7.94 (4.27) | 0.66 |
| | *+Think* | 8.12 (4.38) | 8.36 (4.51) | 8.08 (4.35) | 8.94 (4.81) | 8.57 (4.61) | 7.99 (4.31) | 0.67 |
| | *+CoT* | 8.09 (4.35) | 8.31 (4.48) | 8.01 (4.31) | 8.99 (4.84) | 8.58 (4.62) | 8.10 (4.36) | 0.68 |
| | *-System* | 8.03 (4.34) | 8.25 (4.46) | 7.95 (4.29) | 8.91 (4.80) | 8.53 (4.60) | 7.92 (4.28) | 0.68 |
| | *+Code* | - | - | - | - | - | - | 0.63 |
| DeepSeek-V3.2 | *Ori* | 7.32 (3.91) | 7.55 (4.03) | 7.20 (3.84) | 8.29 (4.42) | 7.76 (4.14) | 7.24 (3.87) | 0.60 |
| | *+Think* | 7.41 (3.99) | 7.59 (4.09) | 7.32 (3.95) | 8.44 (4.55) | 7.97 (4.29) | 7.30 (3.93) | 0.62 |
| | *+CoT* | 7.23 (3.90) | 7.45 (4.03) | 7.09 (3.82) | 8.26 (4.46) | 7.60 (4.10) | 7.13 (3.85) | 0.60 |
| | *-System* | 7.56 (4.08) | 7.81 (4.22) | 7.18 (3.88) | 8.76 (4.72) | 7.83 (4.23) | 7.36 (3.97) | 0.60 |
| | *+Code* | - | - | - | - | - | - | 0.61 |
| Gemini-2.5-Flash | *Ori* | 4.04 (2.14) | 4.17 (2.21) | 4.07 (2.15) | 4.70 (2.48) | 4.50 (2.37) | 4.00 (2.12) | 0.32 |
| | *+Think* | 3.82 (2.01) | 3.92 (2.06) | 3.85 (2.02) | 4.47 (2.35) | 4.28 (2.24) | 3.84 (2.02) | 0.30 |
| | *+CoT* | 4.51 (2.38) | 4.63 (2.44) | 4.51 (2.38) | 5.21 (2.74) | 5.02 (2.64) | 4.53 (2.39) | 0.35 |
| | *-System* | 3.51 (1.84) | 3.61 (1.90) | 3.54 (1.86) | 4.16 (2.18) | 4.00 (2.09) | 3.52 (1.84) | 0.27 |
| | *+Code* | - | - | - | - | - | - | 0.65 |
| | *+Image* | 8.53 (7.83) | 8.75 (8.03) | 8.56 (7.88) | 9.67 (8.87) | 9.31 (8.54) | 8.50 (7.80) | 0.61 |
| Claude-haiku-4.5 | *Ori* | 6.89 (3.69) | 7.03 (3.76) | 6.67 (3.56) | 8.56 (4.59) | 7.71 (4.13) | 6.60 (3.52) | 0.47 |
| | *+Think* | 7.04 (3.74) | 7.16 (3.81) | 6.81 (3.61) | 8.72 (4.66) | 7.83 (4.17) | 6.70 (3.55) | 0.47 |
| | *+CoT* | 6.91 (3.69) | 7.05 (3.78) | 6.71 (3.57) | 8.64 (4.64) | 7.73 (4.14) | 6.60 (3.51) | 0.49 |
| | *-System* | 6.92 (3.69) | 7.01 (3.74) | 6.84 (3.64) | 8.61(4.61) | 7.88 (4.21) | 6.73 (3.58) | 0.46 |
| | *+Code* | - | - | - | - | - | - | 0.58 |
| | *+Image* | 6.42 (5.91) | 6.42 (5.94) | 6.08 (5.62) | 8.89 (8.21) | 8.00 (7.33) | 6.25 (5.76) | 0.33 |
| AstroSage3-8B | *Ori* | 2.65 (1.41) | 2.12 (1.13) | 2.87 (1.53) | 5.09 (2.74) | 3.71 (1.99) | 2.44 (1.30) | 0.04 |
| | *+CoT* | 2.77 (1.49) | 2.29 (1.23) | 2.92 (1.56) | 5.37 (2.89) | 4.18 (2.24) | 2.57 (1.37) | 0.03 |
| | *-System* | 2.59 (1.38) | 2.14 (1.14) | 2.78 (1.48) | 5.02 (2.69) | 3.77 (2.02) | 2.41 (1.28) | 0.03 |
| | *+Code* | - | - | - | - | - | - | 0.00 |
| Llama-3.1-8B | *Ori* | 4.12 (2.18) | 3.68 (1.93) | 4.17 (2.21) | 6.83 (3.65) | 5.86 (3.12) | 3.92 (2.06) | 0.06 |
| | *+CoT* | 4.16 (2.20) | 3.86 (2.04) | 4.11 (2.17) | 6.63 (3.54) | 5.74 (3.06) | 3.88 (2.05) | 0.07 |
| | *-System* | 3.89 (2.07) | 3.52 (1.87) | 3.86 (2.05) | 6.47 (3.46) | 5.56 (2.96) | 3.57 (1.90) | 0.07 |
| | *+Code* | - | - | - | - | - | - | 0.17 |
| Llama-3.1-70B | *Ori* | 4.17 (2.20) | 4.02 (2.12) | 4.21 (2.22) | 6.25 (3.33) | 5.57 (2.97) | 3.87 (2.04) | 0.12 |
| | *+CoT* | 4.39 (2.33) | 4.32 (2.29) | 4.33 (2.30) | 6.25 (3.33) | 5.72 (3.06) | 4.05 (2.15) | 0.16 |
| | *-System* | 3.51 (1.87) | 3.49 (1.86) | 3.39 (1.80) | 5.15 (2.77) | 4.62 (2.48) | 3.16 (1.68) | 0.14 |
| | *+Code* | - | - | - | - | - | - | 0.31 |
| Qwen2.5-Math-72B | *Ori* | 6.29 (3.35) | 6.24 (3.32) | 6.33 (3.36) | 8.72 (4.67) | 8.01 (4.29) | 6.11 (3.25) | 0.34 |
| | *+CoT* | 6.44 (3.44) | 6.43 (3.43) | 6.43 (3.42) | 8.76 (4.70) | 8.10 (4.34) | 6.30 (3.36) | 0.34 |
| | *-System* | 6.50 (3.47) | 6.51 (3.47) | 6.50 (3.47) | 8.83 (4.74) | 8.15 (4.37) | 6.39 (3.41) | 0.34 |
| | *+Code* | - | - | - | - | - | - | 0.06 |
| Qwen3-VL-32B | *Ori* | 4.42 (2.39) | 4.49 (2.43) | 4.44 (2.40) | 5.27 (2.85) | 4.98 (2.70) | 4.37 (2.37) | 0.31 |
| | *+Think* | 6.33 (3.40) | 6.51 (3.50) | 6.26 (3.36) | 7.29 (3.91) | 7.00 (3.75) | 6.20 (3.33) | 0.52 |
| | *+CoT* | 4.59 (2.45) | 4.70 (2.51) | 4.58 (2.45) | 5.38 (2.88) | 5.08 (2.71) | 4.59 (2.46) | 0.37 |
| | *-System* | 4.61 (2.49) | 4.68 (2.54) | 4.61 (2.49) | 5.45 (2.94) | 5.10 (2.75) | 4.53 (2.45) | 0.34 |
| | *+Code* | - | - | - | - | - | - | 0.56 |
| | *+Image* | 5.28 (5.06) | 5.39 (5.15) | 5.22 (5.00) | 6.67 (6.39) | 6.19 (5.91) | 5.14 (4.96) | 0.33 |

Table 17: Performance of different LLMs for the Unimodal Computation (UC) task. Parenthetical scores report results after applying the weighted difficulty coefficient $\mathcal{D}$. "Ori" denotes the original model; "+Think" indicates that the model performs reasoning under a slow-thinking evaluation protocol; "+CoT" denotes the use of a chain-of-thought prompting strategy; "-System" removes the system prompt; "+Code" indicates code-based reasoning, corresponding to the CC task only. Values in red indicate anomalous results due to erroneous outputs. Bold font marks the best performance under the "Ori" setting.

Intermediate Reasoning, and Reasoning Backward. Tables 19 and 20 report the cognitive testing performance on FD. Tables 21 and 22 present the results on UC. Tables 23 and 24 report the performance on FC. The omitted visualizations of open-source models for cognitive dimension evaluation can be found in Figure 5, and a detailed description can be found in Section 3.3.

| Model | Mode | Overall | Correctness | Completeness | Relevance | Coherence | Detail | Accuracy |
|---|---|---|---|---|---|---|---|---|
| GPT-5.6-Luna | *Ori* | **7.45** (3.41) | **7.71** (3.54) | **7.02** (3.22) | **9.01** (4.13) | **8.84** (4.06) | **7.62** (3.50) | **0.70** |
| | *+Think* | 7.57 (3.47) | 7.88 (3.62) | 7.06 (3.24) | 9.03 (4.14) | 8.77 (4.03) | 7.53 (3.45) | 0.72 |
| | *+CoT* | 7.77 (3.57) | 8.10 (3.72) | 7.24 (3.32) | 9.28 (4.26) | 9.09 (4.18) | 7.79 (3.58) | 0.75 |
| | *-System* | 7.46 (3.43) | 7.74 (3.56) | 7.01 (3.22) | 9.20 (4.23) | 8.96 (4.12) | 7.67 (3.53) | 0.71 |
| | *+Code* | - | - | - | - | - | - | 0.55 |
| AstroOne | *Ori* | 6.66 (3.06) | 6.93 (3.18) | 6.38 (2.93) | 7.60 (3.50) | 7.39 (3.40) | 6.47 (2.97) | 0.58 |
| | *+Think* | 4.68 (2.13) | 4.86 (2.21) | 4.50 (2.04) | 5.75 (2.61) | 5.22 (2.37) | 4.68 (2.13) | 0.36 |
| | *+CoT* | 6.74 (3.09) | 6.95 (3.18) | 6.53 (2.99) | 8.01 (3.68) | 7.78 (3.56) | 6.74 (3.09) | 0.59 |
| | *-System* | 6.22 (2.86) | 6.33 (2.91) | 6.03 (2.77) | 7.39 (3.40) | 7.19 (3.31) | 6.27 (2.89) | 0.54 |
| | *+Code* | - | - | - | - | - | - | 0.52 |
| Qwen3-Max | *Ori* | 7.11 (3.26) | 7.32 (3.36) | 6.88 (3.16) | 8.30 (3.81) | 8.05 (3.70) | 6.97 (3.20) | 0.61 |
| | *+Think* | 7.34 (3.36) | 7.54 (3.46) | 7.04 (3.23) | 8.70 (3.99) | 8.48 (3.89) | 7.32 (3.35) | 0.58 |
| | *+CoT* | 7.49 (3.44) | 7.6 (3.5) | 7.23 (3.32) | 8.81 (4.04) | 8.60 (3.95) | 7.33 (3.37) | 0.64 |
| | *-System* | 7.26 (3.34) | 7.38 (3.39) | 7.11 (3.27) | 8.83 (4.05) | 8.73 (4.01) | 7.48 (3.43) | 0.59 |
| | *+Code* | - | - | - | - | - | - | 0.52 |
| DeepSeek-V3.2 | *Ori* | 6.99 (3.20) | 7.24 (3.31) | 6.72 (3.07) | 8.45 (3.87) | 8.02 (3.67) | 6.88 (3.14) | 0.55 |
| | *+Think* | 5.69 (2.62) | 5.83 (2.68) | 5.43 (2.50) | 6.96 (3.19) | 6.72 (3.09) | 5.71 (2.63) | 0.46 |
| | *+CoT* | 7.18 (3.30) | 7.38 (3.39) | 6.89 (3.17) | 8.63 (3.98) | 8.18 (3.77) | 7.09 (3.26) | 0.58 |
| | *-System* | 6.91 (3.16) | 7.13 (3.27) | 6.37 (2.92) | 8.49 (3.89) | 7.67 (3.52) | 6.64 (3.04) | 0.58 |
| | *+Code* | - | - | - | - | - | - | 0.59 |
| Gemini-2.5-Flash | *Ori* | 5.90 (2.71) | 6.06 (2.78) | 5.65 (2.60) | 7.20 (3.32) | 7.05 (3.25) | 5.95 (2.74) | 0.49 |
| | *+Think* | 6.01 (2.75) | 6.19 (2.83) | 5.83 (2.66) | 7.34 (3.36) | 7.25 (3.32) | 6.12 (2.80) | 0.42 |
| | *+CoT* | 5.32 (2.43) | 5.49 (2.50) | 5.27 (2.41) | 6.51 (2.98) | 6.36 (2.91) | 5.37 (2.45) | 0.39 |
| | *-System* | 5.93 (2.71) | 6.12 (2.80) | 5.80 (2.66) | 7.16 (3.28) | 6.81 (3.12) | 5.98 (2.74) | 0.46 |
| | *+Code* | - | - | - | - | - | - | 0.55 |
| Claude-haiku-4.5 | *Ori* | 6.79 (3.10) | 6.78 (3.09) | 6.58 (3.01) | 9.04 (4.15) | 8.31 (3.81) | 6.67 (3.05) | 0.39 |
| | *+Think* | 6.52 (2.97) | 6.54 (2.98) | 6.39 (2.91) | 8.67 (3.97) | 8.12 (3.72) | 6.42 (2.92) | 0.36 |
| | *+CoT* | 6.82 (3.11) | 6.78 (3.09) | 6.67 (3.05) | 9.07 (4.16) | 8.51 (3.91) | 6.79 (3.11) | 0.45 |
| | *-System* | 6.46 (2.95) | 6.46 (2.95) | 6.41 (2.93) | 8.70 (4.00) | 8.21 (3.77) | 6.36 (2.91) | 0.39 |
| | *+Code* | - | - | - | - | - | - | 0.49 |
| AstroSage3-8B | *Ori* | 2.81 (1.27) | 2.29 (1.03) | 3.00 (1.37) | 5.46 (2.50) | 4.07 (1.86) | 2.65 (1.20) | 0.04 |
| | *+CoT* | 2.99 (1.36) | 2.55 (1.16) | 3.17 (1.45) | 5.67 (2.60) | 4.66 (2.14) | 2.70 (1.23) | 0.00 |
| | *-System* | 2.65 (1.20) | 2.21 (0.99) | 2.76 (1.25) | 5.20 (2.37) | 3.94 (1.79) | 2.34 (1.05) | 0.01 |
| | *+Code* | - | - | - | - | - | - | 0.04 |
| Llama-3.1-8B | *Ori* | 4.20 (1.93) | 4.00 (1.83) | 4.12 (1.89) | 7.10 (3.26) | 6.72 (3.09) | 3.91 (1.80) | 0.13 |
| | *+CoT* | 4.20 (1.90) | 3.94 (1.78) | 4.07 (1.84) | 7.06 (3.23) | 6.65 (3.05) | 3.93 (1.78) | 0.16 |
| | *-System* | 3.90 (1.78) | 3.82 (1.74) | 3.62 (1.66) | 6.51 (2.99) | 6.33 (2.90) | 3.63 (1.65) | 0.13 |
| | *+Code* | - | - | - | - | - | - | 0.22 |
| Llama-3.1-70B | *Ori* | 4.60 (2.09) | 4.74 (2.15) | 4.37 (1.99) | 6.99 (3.20) | 6.44 (2.95) | 4.30 (1.96) | 0.16 |
| | *+CoT* | 3.52 (1.61) | 3.49 (1.60) | 3.26 (1.50) | 5.30 (2.44) | 4.91 (2.26) | 3.36 (1.54) | 0.16 |
| | *-System* | 4.66 (2.12) | 4.67 (2.12) | 4.51 (2.05) | 6.93 (3.16) | 6.43 (2.93) | 4.31 (1.96) | 0.19 |
| | *+Code* | - | - | - | - | - | - | 0.36 |
| Qwen2.5-Math-72B | *Ori* | 6.21 (2.83) | 6.18 (2.81) | 6.14 (2.80) | 8.81 (4.04) | 8.41 (3.86) | 6.12 (2.78) | 0.30 |
| | *+CoT* | 6.12 (2.79) | 6.00 (2.73) | 6.09 (2.78) | 8.88 (4.08) | 8.43 (3.87) | 6.04 (2.76) | 0.33 |
| | *-System* | 6.27 (2.85) | 6.36 (2.90) | 6.18 (2.82) | 8.82 (4.05) | 8.46 (3.88) | 5.94 (2.70) | 0.26 |
| | *+Code* | - | - | - | - | - | - | 0.22 |
| Qwen3-VL-32B | *Ori* | 4.86 (2.22) | 4.85 (2.21) | 4.86 (2.22) | 6.52 (2.99) | 6.23 (2.86) | 4.91 (2.24) | 0.30 |
| | *+Think* | 6.25 (2.87) | 6.40 (2.94) | 6.04 (2.76) | 7.58 (3.48) | 7.30 (3.35) | 6.13 (2.81) | 0.51 |
| | *+CoT* | 3.83 (1.74) | 3.80 (1.73) | 3.78 (1.72) | 5.27 (2.40) | 5.02 (2.29) | 3.89 (1.77) | 0.22 |
| | *-System* | 4.13 (1.90) | 4.02 (1.85) | 4.14 (1.90) | 5.51 (2.54) | 5.29 (2.44) | 4.08 (1.88) | 0.26 |
| | *+Code* | - | - | - | - | - | - | 0.48 |

Table 18: Performance of different LLMs for the Formula Calculation (FC) task. Parenthetical scores report results after applying the weighted difficulty coefficient $\mathcal{D}$. "Ori" denotes the original model; "+Think" indicates that the model performs reasoning under a slow-thinking evaluation protocol; "+CoT" denotes the use of a chain-of-thought prompting strategy; "-System" removes the system prompt; "+Code" indicates code-based reasoning, corresponding to the CC task only; and "+Image" indicates text-image reasoning (if supported), corresponding to the MC task only. Values in red indicate anomalous results due to erroneous outputs. Bold font marks the best performance under the "Ori" setting.

| Model | Perturbation | Overall | Correctness | Completeness | Relevance | Coherence | Detail | Accuracy |
|---|---|---|---|---|---|---|---|---|
| GPT-5.6-Luna | *Shuffle Sentence* | 8.82 (5.72) | 9.18 (5.93) | 8.47 (5.50) | 9.68 (6.33) | 9.62 (6.27) | 9.44 (6.13) | - |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.00 |
| | *Redundant Background* | 9.09 (5.92) | 9.41 (6.13) | 8.71 (5.68) | 9.76 (6.40) | 9.65 (6.31) | 9.44 (6.16) | - |
| | *Redundant Conditions* | 8.59 (5.67) | 8.97 (5.91) | 8.25 (5.46) | 9.66 (6.43) | 9.53 (6.33) | 9.06 (5.99) | - |
| | *Condition Damage* | - | - | - | - | - | - | - |
| | *Condition Missing* | - | - | - | - | - | - | - |
| | *Knowledge Redefinition* | - | - | - | - | - | - | - |
| | *Reasoning Forward* | 3.12 (2.04) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 7.67 (5.37) | - | - | - | - | - | - |
| | *Reasoning Backward* | 3.94 (2.49) | - | - | - | - | - | - |
| AstroOne | *Shuffle Sentence* | 7.49 (4.76) | 7.85 (5.01) | 7.35 (4.68) | 7.82 (5.00) | 7.62 (4.84) | 7.29 (4.64) | - |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.59 |
| | *Redundant Background* | 8.11 (5.29) | 8.47 (5.52) | 8.21 (5.35) | 8.68 (5.68) | 8.12 (5.29) | 7.97 (5.19) | - |
| | *Redundant Conditions* | 6.90 (4.42) | 7.24 (4.64) | 6.76 (4.33) | 7.26 (4.69) | 7.24 (4.65) | 6.68 (4.29) | - |
| | *Condition Damage* | - | - | - | - | - | - | - |
| | *Condition Missing* | - | - | - | - | - | - | - |
| | *Knowledge Redefinition* | - | - | - | - | - | - | - |
| | *Reasoning Forward* | 6.06 (4.07) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 4.33 (3.07) | - | - | - | - | - | - |
| | *Reasoning Backward* | 6.24 (4.23) | - | - | - | - | - | - |
| Qwen3-Max | *Shuffle Sentence* | 6.22 (3.94) | 6.53 (4.15) | 6.06 (3.82) | 6.50 (4.13) | 6.56 (4.19) | 6.03 (3.82) | - |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.82 |
| | *Redundant Background* | 8.30 (5.31) | 8.71 (5.58) | 7.97 (5.09) | 8.82 (5.65) | 8.62 (5.53) | 7.94 (5.08) | - |
| | *Redundant Conditions* | 7.99 (5.25) | 8.38 (5.50) | 7.68 (5.05) | 8.44 (5.56) | 8.41 (5.53) | 7.68 (5.05) | - |
| | *Condition Damage* | - | - | - | - | - | - | - |
| | *Condition Missing* | - | - | - | - | - | - | - |
| | *Knowledge Redefinition* | - | - | - | - | - | - | - |
| | *Reasoning Forward* | 6.47 (4.42) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 5.42 (3.86) | - | - | - | - | - | - |
| | *Reasoning Backward* | 6.06 (4.00) | - | - | - | - | - | - |
| DeepSeek-V3.2 | *Shuffle Sentence* | 7.94 (5.16) | 8.38 (5.44) | 7.68 (4.98) | 8.74 (5.69) | 8.91 (5.82) | 7.82 (5.10) | - |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.12 |
| | *Redundant Background* | 9.04 (5.90) | 9.53 (6.23) | 8.76 (5.73) | 9.82 (6.43) | 9.47 (6.20) | 8.71 (5.70) | - |
| | *Redundant Conditions* | 8.49 (5.56) | 8.97 (5.88) | 8.21 (5.37) | 9.03 (5.93) | 8.91 (5.84) | 8.24 (5.40) | - |
| | *Condition Damage* | - | - | - | - | - | - | - |
| | *Condition Missing* | - | - | - | - | - | - | - |
| | *Knowledge Redefinition* | - | - | - | - | - | - | - |
| | *Reasoning Forward* | 5.88 (3.96) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 5.42 (3.81) | - | - | - | - | - | - |
| | *Reasoning Backward* | 7.29 (4.95) | - | - | - | - | - | - |
| Gemini-2.5-Flash | *Shuffle Sentence* | 7.62 (5.06) | 7.94 (5.25) | 7.41 (4.92) | 8.41 (5.60) | 8.12 (5.41) | 7.53 (5.01) | - |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.18 |
| | *Redundant Background* | 8.71 (5.66) | 9.09 (5.90) | 8.65 (5.63) | 9.68 (6.33) | 9.21 (6.01) | 8.76 (5.73) | - |
| | *Redundant Conditions* | 8.55 (5.57) | 8.97 (5.84) | 8.65 (5.62) | 9.18 (6.01) | 8.94 (5.84) | 8.35 (5.47) | - |
| | *Condition Damage* | - | - | - | - | - | - | - |
| | *Condition Missing* | - | - | - | - | - | - | - |
| | *Knowledge Redefinition* | - | - | - | - | - | - | - |
| | *Reasoning Forward* | 3.94 (2.72) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 6.92 (4.89) | - | - | - | - | - | - |
| | *Reasoning Backward* | 4.82 (3.23) | - | - | - | - | - | - |
| Claude-haiku-4.5 | *Shuffle Sentence* | 6.65 (4.24) | 6.91 (4.38) | 6.26 (3.98) | 7.94 (5.13) | 7.56 (4.87) | 6.29 (4.00) | - |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.76 |
| | *Redundant Background* | 6.82 (4.41) | 7.03 (4.52) | 6.59 (4.27) | 8.38 (5.44) | 8.15 (5.29) | 6.38 (4.13) | - |
| | *Redundant Conditions* | 7.24 (4.65) | 7.50 (4.80) | 6.76 (4.34) | 8.59 (5.56) | 8.21 (5.32) | 7.03 (4.51) | - |
| | *Condition Damage* | - | - | - | - | - | - | - |
| | *Condition Missing* | - | - | - | - | - | - | - |
| | *Knowledge Redefinition* | - | - | - | - | - | - | - |
| | *Reasoning Forward* | 4.71 (3.10) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 5.33 (3.73) | - | - | - | - | - | - |
| | *Reasoning Backward* | 4.76 (3.19) | - | - | - | - | - | - |

Table 19: Cognitive evaluation performance of different LLMs for the Formula Derivation (FD) task.

| Model | Perturbation | Overall | Correctness | Completeness | Relevance | Coherence | Detail | Accuracy |
|---|---|---|---|---|---|---|---|---|
| AstroSage-8B | *Shuffle Sentence* | 2.78 (1.75) | 2.50 (1.58) | 2.59 (1.62) | 4.21 (2.72) | 3.65 (2.37) | 2.47 (1.56) | - |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.12 |
| | *Redundant Background* | 1.79 (1.09) | 1.71 (1.03) | 1.71 (1.02) | 2.53 (1.57) | 2.35 (1.45) | 1.59 (0.95) | - |
| | *Redundant Conditions* | 0.41 (0.27) | 0.35 (0.23) | 0.41 (0.27) | 1.00 (0.65) | 1.00 (0.66) | 0.35 (0.24) | - |
| | *Condition Damage* | - | - | - | - | - | - | - |
| | *Condition Missing* | - | - | - | - | - | - | - |
| | *Knowledge Redefinition* | - | - | - | - | - | - | - |
| | *Reasoning Forward* | 1.12 (0.76) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 0.92 (0.68) | - | - | - | - | - | - |
| | *Reasoning Backward* | 1.94 (1.28) | - | - | - | - | - | - |
| Llama3.1-8B | *Shuffle Sentence* | 3.68 (2.35) | 3.29 (2.09) | 3.29 (2.10) | 6.18 (3.98) | 5.71 (3.71) | 3.24 (2.04) | - |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.24 |
| | *Redundant Background* | 3.90 (2.49) | 3.53 (2.24) | 3.71 (2.36) | 6.41 (4.14) | 5.56 (3.58) | 3.50 (2.22) | - |
| | *Redundant Conditions* | 3.76 (2.39) | 3.35 (2.12) | 3.71 (2.35) | 6.00 (3.84) | 5.76 (3.74) | 3.18 (2.02) | - |
| | *Condition Damage* | - | - | - | - | - | - | - |
| | *Condition Missing* | - | - | - | - | - | - | - |
| | *Knowledge Redefinition* | - | - | - | - | - | - | - |
| | *Reasoning Forward* | 1.24 (0.82) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 0.58 (0.41) | - | - | - | - | - | - |
| | *Reasoning Backward* | 1.24 (0.85) | - | - | - | - | - | - |
| Llama3.1-70B | *Shuffle Sentence* | 4.50 (2.84) | 4.35 (2.76) | 4.29 (2.69) | 6.18 (3.88) | 5.94 (3.76) | 3.71 (2.32) | - |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.06 |
| | *Redundant Background* | 5.15 (3.29) | 5.00 (3.17) | 4.65 (2.96) | 7.59 (4.95) | 6.71 (4.35) | 4.47 (2.84) | - |
| | *Redundant Conditions* | 4.58 (2.97) | 4.65 (3.03) | 4.24 (2.73) | 6.35 (4.14) | 5.94 (3.88) | 4.12 (2.64) | - |
| | *Condition Damage* | - | - | - | - | - | - | - |
| | *Condition Missing* | - | - | - | - | - | - | - |
| | *Knowledge Redefinition* | - | - | - | - | - | - | - |
| | *Reasoning Forward* | 0.82 (0.59) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 1.25 (0.91) | - | - | - | - | - | - |
| | *Reasoning Backward* | 0.76 (0.54) | - | - | - | - | - | - |
| Qwen2.5-Math-72B | *Shuffle Sentence* | 7.31 (4.72) | 7.56 (4.87) | 6.97 (4.49) | 8.71 (5.67) | 8.29 (5.41) | 6.91 (4.46) | - |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.18 |
| | *Redundant Background* | 5.99 (3.82) | 6.09 (3.88) | 5.82 (3.70) | 8.12 (5.26) | 7.38 (4.76) | 5.68 (3.61) | - |
| | *Redundant Conditions* | 6.96 (4.49) | 7.03 (4.54) | 6.59 (4.22) | 8.94 (5.83) | 8.26 (5.35) | 6.53 (4.19) | - |
| | *Condition Damage* | - | - | - | - | - | - | - |
| | *Condition Missing* | - | - | - | - | - | - | - |
| | *Knowledge Redefinition* | - | - | - | - | - | - | - |
| | *Reasoning Forward* | 3.71 (2.53) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 4.50 (3.17) | - | - | - | - | - | - |
| | *Reasoning Backward* | 5.47 (3.67) | - | - | - | - | - | - |
| Qwen3-VL-32B | *Shuffle Sentence* | 6.95 (4.39) | 7.26 (4.60) | 6.62 (4.16) | 7.65 (4.86) | 7.38 (4.70) | 6.91 (4.37) | - |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.71 |
| | *Redundant Background* | 6.16 (4.06) | 6.26 (4.14) | 5.94 (3.91) | 7.68 (5.08) | 6.82 (4.49) | 6.12 (4.04) | - |
| | *Redundant Conditions* | 7.00 (4.57) | 7.26 (4.73) | 6.79 (4.45) | 8.06 (5.29) | 7.82 (5.11) | 6.91 (4.54) | - |
| | *Condition Damage* | - | - | - | - | - | - | - |
| | *Condition Missing* | - | - | - | - | - | - | - |
| | *Knowledge Redefinition* | - | - | - | - | - | - | - |
| | *Reasoning Forward* | 2.12 (1.45) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 5.08 (3.59) | - | - | - | - | - | - |
| | *Reasoning Backward* | 1.00 (0.62) | - | - | - | - | - | - |

Table 20: Cognitive evaluation performance of different LLMs for the FD task (continued).

| Model | Perturbation | Overall | Correctness | Completeness | Relevance | Coherence | Detail | Accuracy |
|---|---|---|---|---|---|---|---|---|
| GPT-5.6-Luna | *Shuffle Sentence* | 8.01 (4.32) | 8.30 (4.48) | 7.69 (4.15) | 9.19 (4.96) | 8.55 (4.61) | 7.89 (4.26) | 0.73 |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.07 |
| | *Redundant Background* | 8.04 (4.33) | 8.33 (4.50) | 7.75 (4.18) | 9.33 (5.02) | 8.62 (4.64) | 7.95 (4.28) | 0.75 |
| | *Redundant Conditions* | 7.83 (4.23) | 8.10 (4.38) | 7.53 (4.07) | 9.02 (4.86) | 8.40 (4.53) | 7.70 (4.16) | 0.70 |
| | *Condition Damage* | - | - | - | - | - | - | 0.58 |
| | *Condition Missing* | - | - | - | - | - | - | 0.73 |
| | *Knowledge Redefinition* | - | - | - | - | - | - | - |
| | *Reasoning Forward* | 7.03 (3.86) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 8.46 (5.42) | - | - | - | - | - | - |
| | *Reasoning Backward* | 7.12 (3.88) | - | - | - | - | - | - |
| AstroOne | *Shuffle Sentence* | 6.53 (3.51) | 6.75 (3.62) | 6.46 (3.47) | 7.65 (4.11) | 7.93 (4.26) | 6.89 (3.70) | 0.53 |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.57 |
| | *Redundant Background* | 4.48 (2.41) | 4.65 (2.51) | 4.38 (2.36) | 5.23 (2.82) | 5.27 (2.84) | 4.73 (2.55) | 0.35 |
| | *Redundant Conditions* | 6.68 (3.58) | 6.95 (3.73) | 6.57 (3.52) | 7.46 (4.01) | 8.07 (4.34) | 6.95 (3.73) | 0.52 |
| | *Condition Damage* | - | - | - | - | - | - | 0.85 |
| | *Condition Missing* | - | - | - | - | - | - | 0.59 |
| | *Knowledge Redefinition* | - | - | - | - | - | - | - |
| | *Reasoning Forward* | 3.39 (2.03) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 2.95 (1.79) | - | - | - | - | - | - |
| | *Reasoning Backward* | 3.42 (2.01) | - | - | - | - | - | - |
| Qwen3-Max | *Shuffle Sentence* | 7.35 (3.93) | 7.63 (4.08) | 7.28 (3.89) | 8.38 (4.49) | 9.05 (4.85) | 8.11 (4.33) | 0.61 |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.89 |
| | *Redundant Background* | 7.54 (4.06) | 7.8 (4.2) | 7.45 (4.01) | 8.70 (4.68) | 9.10 (4.89) | 8.15 (4.38) | 0.63 |
| | *Redundant Conditions* | 7.66 (4.11) | 7.95 (4.27) | 7.51 (4.02) | 8.47 (4.54) | 9.11 (4.89) | 8.16 (4.38) | 0.62 |
| | *Condition Damage* | - | - | - | - | - | - | 0.89 |
| | *Condition Missing* | - | - | - | - | - | - | 0.84 |
| | *Knowledge Redefinition* | - | - | - | - | - | - | - |
| | *Reasoning Forward* | 7.26 (3.98) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 5.42 (3.86) | - | - | - | - | - | - |
| | *Reasoning Backward* | 6.89 (3.76) | - | - | - | - | - | - |
| DeepSeek-V3.2 | *Shuffle Sentence* | 7.21 (3.88) | 7.48 (4.02) | 7.10 (3.81) | 8.25 (4.42) | 8.14 (4.36) | 7.39 (3.96) | 0.58 |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.40 |
| | *Redundant Background* | 7.01 (3.78) | 7.27 (3.92) | 6.94 (3.73) | 7.95 (4.27) | 8.01 (4.30) | 7.29 (3.92) | 0.58 |
| | *Redundant Conditions* | 7.14 (3.84) | 7.44 (4.00) | 7.01 (3.77) | 7.88 (4.23) | 8.09 (4.33) | 7.32 (3.93) | 0.55 |
| | *Condition Damage* | - | - | - | - | - | - | 0.58 |
| | *Condition Missing* | - | - | - | - | - | - | 0.74 |
| | *Knowledge Redefinition* | - | - | - | - | - | - | - |
| | *Reasoning Forward* | 6.12 (3.40) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 6.67 (3.69) | - | - | - | - | - | - |
| | *Reasoning Backward* | 5.66 (3.10) | - | - | - | - | - | - |
| Gemini-2.5-Flash | *Shuffle Sentence* | 6.94 (3.72) | 7.17 (3.85) | 6.97 (3.74) | 8.32 (4.46) | 8.33 (4.47) | 7.42 (3.98) | 0.51 |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.19 |
| | *Redundant Background* | 7.14 (3.87) | 7.38 (4.00) | 7.13 (3.87) | 8.57 (4.63) | 8.69 (4.68) | 7.69 (4.15) | 0.54 |
| | *Redundant Conditions* | 6.70 (3.60) | 6.95 (3.74) | 6.67 (3.58) | 7.68 (4.12) | 7.93 (4.25) | 7.02 (3.76) | 0.46 |
| | *Condition Damage* | - | - | - | - | - | - | 0.71 |
| | *Condition Missing* | - | - | - | - | - | - | 0.50 |
| | *Knowledge Redefinition* | - | - | - | - | - | - | - |
| | *Reasoning Forward* | 6.30 (3.45) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 6.26 (3.49) | - | - | - | - | - | - |
| | *Reasoning Backward* | 5.48 (3.07) | - | - | - | - | - | - |
| Claude-haiku-4.5 | *Shuffle Sentence* | 7.05 (3.75) | 7.23 (3.85) | 6.98 (3.72) | 8.79 (4.71) | 8.47 (4.54) | 7.09 (3.77) | 0.48 |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.55 |
| | *Redundant Background* | 6.91 (3.68) | 7.11 (3.79) | 6.78 (3.61) | 8.72 (4.67) | 8.39 (4.48) | 7.02 (3.73) | 0.48 |
| | *Redundant Conditions* | 6.93 (3.69) | 7.15 (3.80) | 6.78 (3.60) | 8.50 (4.54) | 8.33 (4.45) | 6.97 (3.70) | 0.43 |
| | *Condition Damage* | - | - | - | - | - | - | 0.69 |
| | *Condition Missing* | - | - | - | - | - | - | 0.66 |
| | *Knowledge Redefinition* | - | - | - | - | - | - | - |
| | *Reasoning Forward* | 6.88 (3.72) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 7.01 (3.83) | - | - | - | - | - | - |
| | *Reasoning Backward* | 6.59 (3.58) | - | - | - | - | - | - |

Table 21: Cognitive evaluation performance of different LLMs for the Unimodal Computation (UC) task.

| Model | Perturbation | Overall | Correctness | Completeness | Relevance | Coherence | Detail | Accuracy |
|---|---|---|---|---|---|---|---|---|
| AstroSage-8B | *Shuffle Sentence* | 2.25 (1.21) | 2.05 (1.11) | 2.46 (1.32) | 5.68 (3.05) | 4.66 (2.52) | 2.37 (1.28) | 0.04 |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.18 |
| | *Redundant Background* | 2.19 (1.17) | 2.05 (1.09) | 2.23 (1.19) | 4.65 (2.49) | 4.10 (2.19) | 2.18 (1.16) | 0.04 |
| | *Redundant Conditions* | 2.09 (1.12) | 1.97 (1.05) | 2.11 (1.13) | 4.32 (2.31) | 4.02 (2.14) | 2.10 (1.12) | 0.03 |
| | *Condition Damage* | - | - | - | - | - | - | 0.18 |
| | *Condition Missing* | - | - | - | - | - | - | 0.12 |
| | *Knowledge Redefinition* | - | - | - | - | - | - | - |
| | *Reasoning Forward* | 2.54 (1.38) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 2.81 (1.51) | - | - | - | - | - | - |
| | *Reasoning Backward* | 2.02 (1.07) | - | - | - | - | - | - |
| Llama3.1-8B | *Shuffle Sentence* | 0.13 (0.07) | 0.15 (0.08) | 0.11 (0.06) | 0.27 (0.14) | 0.31 (0.16) | 0.11 (0.06) | 0.00 |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.19 |
| | *Redundant Background* | 0.06 (0.04) | 0.06 (0.04) | 0.04 (0.03) | 0.12 (0.07) | 0.17 (0.11) | 0.04 (0.03) | 0.00 |
| | *Redundant Conditions* | 0.00 (0.00) | 0.00 (0.00) | 0.00 (0.00) | 0.01 (0.01) | 0.03 (0.02) | 0.00 (0.00) | 0.00 |
| | *Condition Damage* | - | - | - | - | - | - | 0.15 |
| | *Condition Missing* | - | - | - | - | - | - | 0.10 |
| | *Knowledge Redefinition* | - | - | - | - | - | - | - |
| | *Reasoning Forward* | 2.26 (1.18) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 2.07 (1.12) | - | - | - | - | - | - |
| | *Reasoning Backward* | 1.25(0.66) | - | - | - | - | - | - |
| Llama3.1-70B | *Shuffle Sentence* | 4.06 (2.16) | 4.11 (2.19) | 4.01 (2.13) | 6.27 (3.35) | 6.11 (3.27) | 4.11 (2.18) | 0.12 |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.38 |
| | *Redundant Background* | 3.82 (2.04) | 3.90 (2.08) | 3.80 (2.02) | 6.06 (3.24) | 5.95 (3.18) | 3.89 (2.06) | 0.10 |
| | *Redundant Conditions* | 3.50 (1.86) | 3.54 (1.88) | 3.45 (1.83) | 5.34 (2.85) | 5.54 (2.96) | 3.53 (1.87) | 0.08 |
| | *Condition Damage* | - | - | - | - | - | - | 0.47 |
| | *Condition Missing* | - | - | - | - | - | - | 0.26 |
| | *Knowledge Redefinition* | - | - | - | - | - | - | - |
| | *Reasoning Forward* | 2.24 (1.20) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 1.91 (1.06) | - | - | - | - | - | - |
| | *Reasoning Backward* | 1.49(0.80) | - | - | - | - | - | - |
| Qwen2.5-Math-72B | *Shuffle Sentence* | 6.04 (3.22) | 6.18 (3.30) | 5.98 (3.18) | 8.47 (4.54) | 8.17 (4.38) | 6.36 (3.39) | 0.32 |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.09 |
| | *Redundant Background* | 5.98 (3.18) | 6.11 (3.25) | 5.93 (3.15) | 8.39 (4.49) | 8.14 (4.35) | 6.23 (3.31) | 0.31 |
| | *Redundant Conditions* | 6.18 (3.29) | 6.36 (3.39) | 6.05 (3.22) | 8.37 (4.47) | 8.24 (4.40) | 6.28 (3.35) | 0.32 |
| | *Condition Damage* | - | - | - | - | - | - | 0.38 |
| | *Condition Missing* | - | - | - | - | - | - | 0.18 |
| | *Knowledge Redefinition* | - | - | - | - | - | - | - |
| | *Reasoning Forward* | 6.35 (3.45) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 6.71 (3.67) | - | - | - | - | - | - |
| | *Reasoning Backward* | 6.10 (3.28) | - | - | - | - | - | - |
| Qwen3-VL-32B | *Shuffle Sentence* | 4.09 (2.22) | 4.22 (2.30) | 4.06 (2.21) | 5.23 (2.85) | 5.57 (3.04) | 4.54 (2.47) | 0.31 |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.70 |
| | *Redundant Background* | 3.80 (2.06) | 3.89 (2.11) | 3.78 (2.05) | 4.60 (2.49) | 4.80 (2.60) | 4.10 (2.22) | 0.27 |
| | *Redundant Conditions* | 4.33 (2.33) | 4.50 (2.42) | 4.26 (2.30) | 5.08 (2.74) | 5.50 (2.97) | 4.52 (2.44) | 0.31 |
| | *Condition Damage* | - | - | - | - | - | - | 0.83 |
| | *Condition Missing* | - | - | - | - | - | - | 0.80 |
| | *Knowledge Redefinition* | - | - | - | - | - | - | - |
| | *Reasoning Forward* | 2.74 (1.58) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 0.81 (0.46) | - | - | - | - | - | - |
| | *Reasoning Backward* | 1.55 (0.92) | - | - | - | - | - | - |

Table 22: Cognitive evaluation performance of different LLMs for the UC task (continued).

| Model | Perturbation | Overall | Correctness | Completeness | Relevance | Coherence | Detail | Accuracy |
|---|---|---|---|---|---|---|---|---|
| GPT-5.6-Luna | *Shuffle Sentence* | 7.32 (3.38) | 7.60 (3.51) | 6.98 (3.22) | 8.79 (4.06) | 8.46 (3.90) | 7.25 (3.34) | 0.70 |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.07 |
| | *Redundant Background* | 7.54 (3.47) | 7.78 (3.58) | 7.14 (3.28) | 9.19 (4.23) | 8.77 (4.04) | 7.56 (3.48) | 0.67 |
| | *Redundant Conditions* | 7.04 (3.24) | 7.27 (3.35) | 6.76 (3.12) | 8.63 (3.97) | 8.24 (3.79) | 7.06 (3.24) | 0.67 |
| | *Condition Damage* | - | - | - | - | - | - | 0.41 |
| | *Condition Missing* | - | - | - | - | - | - | 0.81 |
| | *Knowledge Redefinition* | 6.23(2.86) | 6.36 (2.93) | 6.00 (2.76) | 8.38 (3.86) | 8.61 (3.96) | 7.24 (3.34) | 0.52 |
| | *Reasoning Forward* | 8.33 (3.82) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 8.28 (3.77) | - | - | - | - | - | - |
| | *Reasoning Backward* | 8.84 (4.08) | - | - | - | - | - | - |
| AstroOne | *Shuffle Sentence* | 5.86 (2.68) | 6.18 (2.83) | 5.61 (2.57) | 7.20 (3.30) | 7.36 (3.36) | 6.29 (2.87) | 0.45 |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.64 |
| | *Redundant Background* | 6.11 (2.81) | 6.35 (2.92) | 5.93 (2.73) | 7.36 (3.39) | 7.86 (3.62) | 6.38 (2.93) | 0.48 |
| | *Redundant Conditions* | 6.39 (2.94) | 6.80 (3.12) | 6.14 (2.83) | 7.33 (3.38) | 8.41 (3.88) | 6.86 (3.16) | 0.49 |
| | *Condition Damage* | - | - | - | - | - | - | 0.59 |
| | *Condition Missing* | - | - | - | - | - | - | 0.64 |
| | *Knowledge Redefinition* | 6.04 (2.77) | 5.96 (2.73) | 5.96 (2.74) | 7.86 (3.61) | 8.15 (3.75) | 6.50 (2.99) | 0.48 |
| | *Reasoning Forward* | 4.96 (2.25) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 5.78 (2.66) | - | - | - | - | - | - |
| | *Reasoning Backward* | 6.25 (2.88) | - | - | - | - | - | - |
| Qwen3-Max | *Shuffle Sentence* | 6.23 (2.87) | 6.51 (3.00) | 5.97 (2.75) | 7.80 (3.59) | 8.85 (4.05) | 7.47 (3.43) | 0.52 |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.68 |
| | *Redundant Background* | 6.57 (3.02) | 6.95 (3.20) | 6.36 (2.92) | 8.10 (3.73) | 8.98 (4.14) | 7.70 (3.55) | 0.57 |
| | *Redundant Conditions* | 6.34 (2.92) | 6.67 (3.07) | 6.04 (2.78) | 7.52 (3.45) | 8.86 (4.07) | 7.28 (3.34) | 0.52 |
| | *Condition Damage* | - | - | - | - | - | - | 0.90 |
| | *Condition Missing* | - | - | - | - | - | - | 0.94 |
| | *Knowledge Redefinition* | 6.47 (2.97) | 6.40 (2.93) | 6.33 (2.90) | 8.25 (3.78) | 8.57 (3.93) | 7.29 (3.36) | 0.52 |
| | *Reasoning Forward* | 7.03 (3.21) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 7.68 (3.53) | - | - | - | - | - | - |
| | *Reasoning Backward* | 7.20 (3.31) | - | - | - | - | - | - |
| DeepSeek-V3.2 | *Shuffle Sentence* | 6.05 (2.79) | 6.37 (2.93) | 5.69 (2.62) | 7.96 (3.66) | 8.32 (3.83) | 6.93 (3.19) | 0.46 |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.50 |
| | *Redundant Background* | 6.90 (3.17) | 7.20 (3.31) | 6.67 (3.07) | 8.70 (4.00) | 9.04 (4.15) | 7.38 (3.39) | 0.57 |
| | *Redundant Conditions* | 6.20 (2.86) | 6.61 (3.05) | 5.83 (2.69) | 7.28 (3.36) | 8.56 (3.94) | 6.62 (3.05) | 0.48 |
| | *Condition Damage* | - | - | - | - | - | - | 0.61 |
| | *Condition Missing* | - | - | - | - | - | - | 0.84 |
| | *Knowledge Redefinition* | 6.03 (2.78) | 6.02 (2.77) | 5.97 (2.75) | 8.09 (3.72) | 7.92 (3.64) | 6.45 (2.97) | 0.48 |
| | *Reasoning Forward* | 7.88 (3.60) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 7.54 (3.47) | - | - | - | - | - | - |
| | *Reasoning Backward* | 6.80 (3.13) | - | - | - | - | - | - |
| Gemini-2.5-Flash | *Shuffle Sentence* | 5.97 (2.73) | 6.27 (2.86) | 5.81 (2.66) | 8.41 (3.87) | 8.58 (3.95) | 7.20 (3.30) | 0.48 |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.17 |
| | *Redundant Background* | 6.51 (2.99) | 6.87 (3.16) | 6.29 (2.89) | 8.22 (3.79) | 8.41 (3.88) | 7.17 (3.30) | 0.51 |
| | *Redundant Conditions* | 5.52 (2.53) | 5.88 (2.69) | 5.32 (2.43) | 7.07 (3.25) | 7.99 (3.68) | 6.64 (3.04) | 0.38 |
| | *Condition Damage* | - | - | - | - | - | - | 0.72 |
| | *Condition Missing* | - | - | - | - | - | - | 0.88 |
| | *Knowledge Redefinition* | 5.34 (2.45) | 5.36 (2.46) | 5.36 (2.46) | 7.14 (3.27) | 7.50 (3.44) | 6.20 (2.85) | 0.41 |
| | *Reasoning Forward* | 6.55 (3.00) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 7.81 (3.60) | - | - | - | - | - | - |
| | *Reasoning Backward* | 6.20 (2.86) | - | - | - | - | - | - |
| Claude-haiku-4.5 | *Shuffle Sentence* | 6.16 (2.81) | 6.45 (2.94) | 5.90 (2.69) | 8.72 (4.01) | 8.60 (3.95) | 6.40 (2.92) | 0.42 |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.39 |
| | *Redundant Background* | 6.37 (2.91) | 6.60 (3.02) | 6.21 (2.84) | 8.46 (3.89) | 8.71 (4.00) | 7.04 (3.22) | 0.42 |
| | *Redundant Conditions* | 5.85 (2.68) | 6.04 (2.77) | 5.77 (2.64) | 7.52 (3.44) | 8.65 (3.97) | 6.50 (2.98) | 0.33 |
| | *Condition Damage* | - | - | - | - | - | - | 0.78 |
| | *Condition Missing* | - | - | - | - | - | - | 0.87 |
| | *Knowledge Redefinition* | 5.75 (2.63) | 5.51 (2.52) | 5.79 (2.65) | 8.01 (3.69) | 8.01 (3.68) | 6.12 (2.80) | 0.35 |
| | *Reasoning Forward* | 7.23 (3.30) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 6.81 (3.14) | - | - | - | - | - | - |
| | *Reasoning Backward* | 6.35 (2.91) | - | - | - | - | - | - |

Table 23: Cognitive evaluation performance of different LLMs for the Formula Calculation (FC) task.

| Model | Perturbation | Overall | Correctness | Completeness | Relevance | Coherence | Detail | Accuracy |
|---|---|---|---|---|---|---|---|---|
| AstroSage-8B | *Shuffle Sentence* | 2.43 (1.10) | 2.24 (1.01) | 2.48 (1.12) | 5.67 (2.59) | 4.41 (2.01) | 2.48 (1.12) | 0.06 |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.19 |
| | *Redundant Background* | 2.47 (1.11) | 2.42 (1.08) | 2.43 (1.09) | 4.75 (2.17) | 4.34 (1.97) | 2.47 (1.11) | 0.06 |
| | *Redundant Conditions* | 2.23 (1.00) | 2.17 (0.98) | 2.23 (1.00) | 4.29 (1.94) | 4.22 (1.92) | 2.16 (0.97) | 0.06 |
| | *Condition Damage* | - | - | - | - | - | - | 0.30 |
| | *Condition Missing* | - | - | - | - | - | - | 0.22 |
| | *Knowledge Redefinition* | 2.47 (1.13) | 1.97 (0.90) | 2.75 (1.26) | 5.23 (2.40) | 4.59 (2.11) | 2.45 (1.12) | 0.01 |
| | *Reasoning Forward* | 4.06 (1.87) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 3.75 (1.71) | - | - | - | - | - | - |
| | *Reasoning Backward* | 2.68 (1.20) | - | - | - | - | - | - |
| Llama3.1-8B | *Shuffle Sentence* | 0.52 (0.23) | 0.64 (0.29) | 0.38 (0.17) | 1.03 (0.47) | 0.96 (0.43) | 0.39 (0.18) | 0.01 |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.16 |
| | *Redundant Background* | 0.50 (0.23) | 0.61 (0.28) | 0.36 (0.17) | 0.87 (0.41) | 0.84 (0.39) | 0.38 (0.17) | 0.01 |
| | *Redundant Conditions* | 0.27 (0.13) | 0.29 (0.14) | 0.19 (0.09) | 0.39 (0.18) | 0.43 (0.21) | 0.20 (0.09) | 0.03 |
| | *Condition Damage* | - | - | - | - | - | - | 0.32 |
| | *Condition Missing* | - | - | - | - | - | - | 0.23 |
| | *Knowledge Redefinition* | 4.41 (2.01) | 4.15 (1.89) | 4.37 (2.00) | 6.91 (3.17) | 7.09 (3.25) | 4.31 (1.97) | 0.09 |
| | *Reasoning Forward* | 4.41 (2.10) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 2.25 (1.02) | - | - | - | - | - | - |
| | *Reasoning Backward* | 1.36 (0.62) | - | - | - | - | - | - |
| Llama3.1-70B | *Shuffle Sentence* | 4.74 (2.17) | 4.85 (2.22) | 4.59 (2.10) | 7.71 (3.54) | 7.36 (3.38) | 4.67 (2.13) | 0.12 |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.26 |
| | *Redundant Background* | 4.54 (2.06) | 4.60 (2.08) | 4.38 (1.98) | 6.77 (3.08) | 6.94 (3.15) | 4.59 (2.08) | 0.16 |
| | *Redundant Conditions* | 4.11 (1.88) | 4.22 (1.93) | 3.88 (1.77) | 6.33 (2.91) | 6.68 (3.07) | 4.13 (1.89) | 0.12 |
| | *Condition Damage* | - | - | - | - | - | - | 0.45 |
| | *Condition Missing* | - | - | - | - | - | - | 0.35 |
| | *Knowledge Redefinition* | 3.97 (1.82) | 3.79 (1.74) | 3.93 (1.80) | 6.47 (2.98) | 6.62 (3.04) | 3.94 (1.81) | 0.13 |
| | *Reasoning Forward* | 1.43 (0.66) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 2.28 (1.03) | - | - | - | - | - | - |
| | *Reasoning Backward* | 1.10 (0.51) | - | - | - | - | - | - |
| Qwen2.5-Math-72B | *Shuffle Sentence* | 5.51 (2.51) | 5.75 (2.62) | 5.31 (2.42) | 8.94 (4.11) | 8.72 (4.00) | 5.68 (2.59) | 0.28 |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.20 |
| | *Redundant Background* | 5.80 (2.64) | 5.99 (2.73) | 5.63 (2.57) | 8.78 (4.03) | 8.83 (4.05) | 6.34 (2.90) | 0.30 |
| | *Redundant Conditions* | 5.57 (2.54) | 5.80 (2.65) | 5.36 (2.45) | 8.28 (3.80) | 8.59 (3.95) | 6.03 (2.75) | 0.29 |
| | *Condition Damage* | - | - | - | - | - | - | 0.51 |
| | *Condition Missing* | - | - | - | - | - | - | 0.29 |
| | *Knowledge Redefinition* | 5.36 (2.45) | 5.01 (2.29) | 5.29 (2.42) | 7.68 (3.52) | 8.45 (3.88) | 6.28 (2.88) | 0.22 |
| | *Reasoning Forward* | 7.06 (3.24) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 6.35 (2.92) | - | - | - | - | - | - |
| | *Reasoning Backward* | 6.48 (2.96) | - | - | - | - | - | - |
| Qwen3-VL-32B | *Shuffle Sentence* | 3.19 (1.45) | 3.30 (1.49) | 3.11 (1.41) | 4.96 (2.27) | 5.61 (2.57) | 3.97 (1.81) | 0.19 |
| | *Shuffle Word Order* | - | - | - | - | - | - | 0.59 |
| | *Redundant Background* | 3.32 (1.53) | 3.49 (1.61) | 3.22 (1.48) | 5.41 (2.50) | 5.54 (2.56) | 4.17 (1.92) | 0.20 |
| | *Redundant Conditions* | 3.84 (1.78) | 4.12 (1.91) | 3.65 (1.69) | 5.07 (2.36) | 5.86 (2.73) | 4.04 (1.88) | 0.25 |
| | *Condition Damage* | - | - | - | - | - | - | 0.87 |
| | *Condition Missing* | - | - | - | - | - | - | 0.88 |
| | *Knowledge Redefinition* | 5.03 (2.30) | 4.87 (2.22) | 5.06 (2.31) | 6.51 (2.98) | 6.99 (3.20) | 5.83 (2.67) | 0.36 |
| | *Reasoning Forward* | 5.32 (2.42) | - | - | - | - | - | - |
| | *Intermediate Reasoning* | 2.81 (1.26) | - | - | - | - | - | - |
| | *Reasoning Backward* | 2.38 (1.07) | - | - | - | - | - | - |

Table 24: Cognitive evaluation performance of different LLMs for the FC task (continued).

## N PROMPTS IN ASTROMATHBENCH EVALUATION

This section provides a detailed description of the prompts used in the benchmarking and evaluation protocols. Prompts marked with **[Optional:]** indicate optional components. The term **Image Encoding** specifies whether visual inputs are embedded into the model representations. The **System Prompt** varies according to the task type. For Unimodal Computation, we use *You are an expert in astronomical computational reasoning.*; For Formula Calculation, *You are an expert in astronomical formula calculation.*; And for Formula Derivation, *You are an expert in deriving astronomical formulas.* The **CoT Prompt** indicates whether Chain-of-Thought prompting is employed during inference.

Figures 8–10 correspond to the prompts used in the main experimental evaluation. Figures 11–15 present the prompts designed for cognitive derivative experiments. Figures 16–18 illustrate the prompts used for data contamination analysis. In the evaluation process, we used DeepSeek-V3.2 as the judge model. Figures 19–21 show the prompts used for result evaluation.

**Unimodal Computation and Formula Calculation**

[Optional: Image Encoding, System Prompt]

# Task Definition
Now you are an astronomy expert. Your task is to read the given question and then solve it. You need to solve this problem and provide the problem-solving process and the final result.

[Optional: CoT Prompt]

# Output Format
{
"solution": "Fill in the problem-solving process here",
"final answer": "Only fill in the final result to the question here; no additional steps are needed."
}
Now, here is your question: "{Test_Question}"

Figure 8: Main Experiment Prompt (Setting 1).

**Formula Derivation**

[Optional: Image Encoding, System Prompt]

# Task Definition
Now you are an astronomy expert. Your task is to read the given question and then solve it. You need to solve this problem and provide the problem-solving process.

[Optional: CoT Prompt]

Now, here is your question: "{Test_Question}"

Figure 9: Main Experiment Prompt (Setting 2).

## O EXAMPLES OF GENERATED REPORTS BY ANALYSIS AGENT

To facilitate automated evaluation, we develop an analysis agent for both the base tasks and the cognitively derived tests. The usage of this agent is provided in our code repository. In this section, we show the automated evaluation results of AstroOne (Table 25), Qwen3-Max (Table 26), Gemini 2.5 Flash (Table 27), and AstroSage-8B (Table 28) on three fundamental tasks: FD, UC, and FC.

### Code Calculation

[Optional: System Prompt]

# Task Definition
Now you are an astronomy expert. Your task is to read the given question and then solve it. You need to solve this problem by writing Python code and returning the final result. Python code must be executable.

# Output Format

```
```python
def execute_code():
    # Code to solve the problem
         <codes>
    # final result return
    return result
```
```

Figure 10: Main Experiment Prompt (Setting 3).

### Reverse Reasoning for Unimodal Computation and Formula Calculation

[Optional: System Prompt]

# Task Definition Now you are an astronomy expert. Your task is to read the given problem and its solution, in which some key steps are missing. The missing parts are marked with '<Replace>'. You need to fill in these placeholders based on the context to complete the solution.

# Output Format
{
"filled replace": "List the content filled into the Replace placeholder.",
"complete solution": "Fill the full solution with replace content integrated"
"final answer": "Only fill in the final result to the question here; no additional steps are needed."
}
Now, here is your question: "{Test_Question}".
The solution that needs you to improve is: """{Mask_Solution}"""

Figure 11: Cognitive Evaluation Prompt (Setting 1).

### Reverse Reasoning for Formula Derivation

[Optional: System Prompt]

# Task Definition Now you are an astronomy expert. Your task is to read the given problem and its solution, in which some key steps are missing. The missing parts are marked with '<Replace>'. You need to fill in these placeholders based on the context to complete the solution.

# Output Format
{
"filled replace": "List the content filled into the Replace placeholder.",
"complete solution": "Fill the full solution with replace content integrated"
}
Now, here is your question: "{Test_Question}".
The solution that needs you to improve is: """{Mask_Solution}"""

Figure 12: Cognitive Evaluation Prompt (Setting 2).

### Problem Comprehension, Condition Robustness and Anti-Interference for Unimodal Computation and Formula Calculation

[Optional: System Prompt]

# Task Definition
Now you are an astronomy expert. Your task is to read the given question and then solve it. In principle, you need to solve this problem and provide the problem-solving process and the final result. But, if you think the problem is not unsolvable, just return 'Unsolvable'.

# Output Format
{
"solution": "Fill in the problem-solving process here",
"final answer": "Only fill in the final result to the question here; no additional steps are needed."
}
Now, here is your question: "{Test_Question}"

Figure 13: Cognitive Evaluation Prompt (Setting 3).

### Problem Comprehension, Condition Robustness and Anti-Interference for Formula Derivation

[Optional: System Prompt]

# Task Definition
Now you are an astronomy expert. Your task is to read the given question and then solve it. In principle, you need to solve this problem and provide the problem-solving process. But, if you think the problem is not unsolvable, just return 'Unsolvable'.

Now, here is your question: "{Test_Question}"

Figure 14: Cognitive Evaluation Prompt (Setting 4).

### Knowledge Consistency for Formula Calculation

[Optional: System Prompt]

# Task Definition
Now you are an astronomy expert. Your task is to read the given question and then solve it. You need to solve this problem and provide the problem-solving process and the final result.

# Output Format
{
"solution": "Fill in the problem-solving process here",
"final answer": "Only fill in the final result to the question here; no additional steps are needed."
}
Now, here is your question: "{Test_Question}"

Figure 15: Cognitive Evaluation Prompt (Setting 5).

### Random Test

**System Prompt:** You are a helpful autocomplete bot for astronomical math questions. You will be prompted with a partial question. Your task is to complete only the immediate next sentence, continuing the text naturally and consistently with the dataset. Output only the missing sentence, without quotation marks or additional commentary.

**User Prompt:** Now, start your work. The given sentence is {question}.

Figure 16: Memory Contamination Testing Experiment Prompt (Setting 1).

### First Word Test

**System Prompt:** You are a helpful autocomplete bot for astronomical math questions. You will be prompted with a partial question. Your task is to complete only the immediate next word, continuing the text naturally and consistently with the dataset. Output only the missing word, without quotation marks or additional commentary.

**User Prompt:** Now, start your work. The given sentence is {question}.

Figure 17: Memory Contamination Testing Experiment Prompt (Setting 2).

### Specific Variable Value and Name Test

You are a variable completion assistant. The input will include a question text containing numbered ‘<Replace i>‘ placeholders (e.g., ‘<Replace 1>‘, ‘<Replace 2>‘, ...). For each ‘<Replace i>‘ placeholder, determine the corresponding original content, including:

-Variable name (‘name‘)
-Variable value (‘value‘)

Do not include any explanatory text; only return the completion results.

Input: {masked_question}

Output: [
{
”order”: ”the corresponding placeholder”,
”name”: ”variable name”,
”value”: ”the corresponding value of the variable”
}...]

Figure 18: Memory Contamination Testing Experiment Prompt (Setting 3).

## Process Evaluation Prompt

**System Prompt:** # Task Definition
You are an expert evaluator. You are given: Question, Reference Solution and Candidate Solution. Your task is to evaluate how well the Candidate Solution matches the Reference Solution.

# Evaluation Criteria
Evaluate the Candidate Answer based on the following dimensions:

1. Correctness: Are the core facts, logic, and conclusions consistent with the Reference Solution? Are there any factual errors or incorrect reasoning?

2. Completeness: Does the Candidate Answer cover all key points present in the Reference Solution? Are any important elements missing?

3. Relevance: Is the solution focused on the question? Does it avoid unnecessary or irrelevant information?

4. Clarity & Coherence: Is the solution well-structured and easy to understand? Are explanations clear and logically connected?

5. Precision & Detail: Does the level of detail appropriately match the Reference Solution? Are concepts expressed accurately without being vague or misleading?

# Scoring Instructions
Assign a numerical score from 0 to 10, where:
9–10: Excellent match; nearly identical in meaning and quality
7–9: Minor omissions or slight inaccuracies, but overall correct
6–7: Partially correct; important details missing or unclear
4–6: Significant errors or gaps; only some correct aspects
0–4: Mostly incorrect, irrelevant, or misleading

Use fine-grained judgment. Do not round scores unnecessarily.

# Output Format
Provide your evaluation strictly in the following format:
‘‘‘
Overall Score: <numeric score from 0 to 10>
Evaluation:
- Correctness: <numeric score from 0 to 10>
- Completeness: <numeric score from 0 to 10>
- Relevance: <numeric score from 0 to 10>
- Clarity & Coherence: <numeric score from 0 to 10>
- Precision & Detail: <numeric score from 0 to 10>

Overall Justification:
<Concise explanation of why this score was assigned>
‘‘‘

# Important Notes
Base your evaluation only on the provided Question, Reference Solution, and Candidate Solution.
Do not reward stylistic differences unless they affect clarity or correctness.
If the Candidate Solution introduces new information not present in the Reference Solution:
-Penalize only if it is incorrect or misleading.
-Be objective and consistent across evaluations.

**User Prompt:** Now start your work.
Question: ”{question}”.
Reference Answer: """{Reference_Answer}""".
Candidate Answer: """{Candidate_Answer}""".

Figure 19: Experiment Evaluation Prompt (Setting 1).

## Final Answer Evaluation Prompt

**System Prompt:** You are an expert evaluator. You are given: Question, Reference Answer and Candidate Answer. Your task is to determine whether the Candidate Answer is truly consistent with the Reference Answer.

Consistency means semantic equivalence, not superficial similarity. Please follow these rules carefully:

1. If the Candidate Answer and the Reference Answer express the same meaning, they should be considered consistent, even if:
The wording, structure, or format is different.
One is more concise or more verbose than the other.
One uses synonyms, paraphrases, or different sentence structures.

2. If the answers involve numerical values, they should be considered consistent if:
The values are equal.
Or the difference is within a reasonable or explicitly acceptable error tolerance.
Or one is an approximation of the other with the same underlying meaning.

3. The Candidate Answer should be marked inconsistent if:
It contradicts the Reference Answer.
It misses key information that changes the meaning
It introduces incorrect facts or conclusions.

4. Do not judge based on style, grammar, or formatting alone. Focus only on whether the final expressed meaning and correctness match.

Output only one of the following labels: 'Consistent' or 'Inconsistent'.
Do not provide explanations or additional text.

**User Prompt:** Now start your work.
Question: "{question}".
Reference Answer: """{Reference_Answer}""".
Candidate Answer: """{Candidate_Answer}""".

Figure 20: Experiment Evaluation Prompt (Setting 2).

## Reasoning Completion Evaluation Prompt

**System Prompt:** # Task Definition
You are a rigorous expert in astronomy, astrophysics, and mathematical reasoning.

There is an astronomy-related computational reasoning problem in which some intermediate solution steps were masked. We now provide:
1. The ground-truth answer for the masked steps.
2. A model-generated completion for the masked steps.

Your task is to evaluate how well the model-generated completion matches the ground-truth answer.

# Evaluation guidelines:
- Focus on whether the reasoning logic, mathematical relationships, and physical meaning are consistent.
- Differences in wording, notation, or order of steps are acceptable as long as the reasoning and conclusions are logically equivalent.
- Deduct points if the completion contains incorrect reasoning, incorrect physical concepts, or non-equivalent conclusions.
- Deduct points if key steps present in the ground-truth answer are missing.
- Do NOT penalize valid equivalent derivations or reasonable additional explanations.

# Scoring rules:
- Score range: 0 to 10
- 10: Fully equivalent to the ground-truth answer in logic and conclusions
- 7–9: Mostly correct, with only minor non-critical differences
- 4–6: Partially correct, but with notable omissions or flaws
- 1–3: Mostly incorrect reasoning or inconsistent conclusions
- 0: Completely incorrect or unrelated

# Output format:
- Match Score: <numeric score from 0 to 10>
- Rationale: Brief explanation of the score.

**User Prompt:** Now start your work.
Ground-truth Answer: """{Mask_Answer}"""
Model Completion: """{Candidate_Answer}"""

Figure 21: Experiment Evaluation Prompt (Setting 3).

| Reporting Phase | Analysis of Content |
|---|---|
| **Formula Derivation** | The evaluation results indicate that the model demonstrates strong formula derivation capabilities under the original and chain-of-thought (+CoT) settings, achieving high correctness, completeness, and coherence, suggesting reliable reasoning and alignment with reference solutions. Removing system prompts (-System) slightly reduces performance, highlighting the contribution of contextual guidance, while explicit slow thinking (+Think) appears to disrupt performance, leading to lower scores across all metrics. Overall, the model shows robust ability to derive astronomy-related formulas when guided by structured prompts but is sensitive to prompt framing and additional cognitive interventions, indicating its capability boundaries lie in well-scaffolded derivations and may degrade under atypical or less directed prompting. |
| **Unimodal Computation** | The evaluation results indicate that the model exhibits moderate competency in computational reasoning, with performance consistently lower than in formula derivation tasks. Across settings, structured guidance such as chain-of-thought (+CoT) or the original prompt maintains process-oriented scores near the mid-range, whereas explicit slow thinking (+Think) substantially reduces performance, suggesting that unstructured cognitive prompting can disrupt multi-step quantitative reasoning. Accuracy values are notably low relative to metric scores, indicating that even when reasoning appears complete and coherent, final numerical or logical consistency with reference solutions is limited. Overall, the model is capable of outlining multi-step astronomical computations and maintaining reasonable clarity and relevance, but its ability to reliably produce fully correct quantitative outcomes is constrained, revealing a boundary in precise numerical reasoning and verification under complex multi-step conditions. |
| **Formula Calculation** | The evaluation results show that the model achieves moderate performance in formula calculation, with process-oriented metrics indicating reasonable completeness, clarity, and relevance, while correctness and precision remain slightly lower. Accuracy scores are consistently limited, suggesting that even when the reasoning and procedural steps appear sound, the final numerical results often deviate from reference solutions. Structured prompting such as chain-of-thought (+CoT) maintains or slightly improves overall scores, whereas interventions like slow thinking (+Think) reduce performance across all metrics. These trends indicate that the model can identify relevant formulas, follow calculation procedures, and present solutions coherently, but its ability to reliably produce fully correct numerical outputs is constrained, defining a capability boundary in precise quantitative computation under complex or multi-step calculations. |
| **Summary** | Across the three astronomy-related tasks, the model demonstrates strong conceptual reasoning in formula derivation, moderate competency in multi-step computational reasoning, and moderate performance in formula calculation, with consistent patterns across experimental settings. Structured guidance, particularly chain-of-thought prompts and original contextual framing, generally supports higher process-oriented scores—correctness, completeness, coherence, and relevance—while unstructured interventions like slow thinking tend to degrade performance. A notable cross-task pattern is the divergence between process-oriented metrics and Accuracy: the model often presents logically coherent, complete, and relevant reasoning steps, yet final numerical or logically verified outcomes are less reliable, indicating limitations in precise quantitative execution. Strengths consistently include clarity, coherence, and adherence to relevant concepts, whereas weaknesses appear in precise correctness, detail, and numerical accuracy. Overall, the model's capabilities are bounded by its ability to scaffold reasoning with structured prompts; it reliably outlines reasoning processes and formula derivations but is constrained in producing fully accurate quantitative results under complex, multi-step astronomical calculations. |

Table 25: Example of Analysis Report by AstroOne Evaluation.

| Reporting Phase | Analysis of Content |
|---|---|
| **Formula Derivation** | The evaluation results indicate that Qwen3-Max demonstrates strong formula derivation ability, with consistently high performance across correctness, completeness, relevance, coherence, and detail. Incremental interventions such as engaging in slow thinking (+Think) slightly improve overall scores, particularly in correctness and completeness, suggesting the model benefits from prompting that encourages deliberate reasoning. Chain-of-thought prompts (+CoT) maintain similar overall performance but show a modest trade-off between correctness and clarity, while removing system prompts (-System) slightly reduces performance across most metrics, highlighting the role of structured guidance in supporting derivation accuracy. Overall, the model reliably generates accurate and coherent derivations under standard conditions, with its capability boundaries primarily constrained by the absence of structured prompts or external guidance, indicating that while proficient in applying physical and mathematical principles, it is sensitive to prompt design for maintaining maximal precision and completeness. |
| **Unimodal Computation** | Qwen3-Max demonstrates moderate-to-strong performance in computational reasoning, with process-oriented metrics such as correctness, completeness, relevance, coherence, and detail generally in the 7.9–8.8 range, indicating consistent logical structuring and clarity in multi-step quantitative analysis. Interventions like slow thinking (+Think) and chain-of-thought prompts (+CoT) yield slight improvements across most metrics, suggesting the model benefits from prompts that explicitly encourage stepwise reasoning, while removal of system prompts (-System) produces minimal impact, reflecting relative robustness to prompt structure. Accuracy, however, remains substantially lower ( 0.66–0.68) than the process scores, highlighting a gap between well-articulated reasoning and fully consistent numerical or final-answer correctness. Overall, the model reliably produces logically coherent and detailed reasoning steps but exhibits limitations in achieving precise, fully consistent outcomes, indicating that its capability boundary in computational reasoning is defined more by final-answer fidelity than by reasoning structure. |
| **Formula Calculation** | Qwen3-Max shows moderate performance in formula calculation, with process-oriented metrics generally ranging from 6.9 to 8.8, indicating that it can identify relevant formulas, organize calculations coherently, and maintain clarity and relevance. Interventions like slow thinking (+Think) and chain-of-thought prompts (+CoT) provide incremental gains in completeness and detail, while removal of system prompts (-System) has minimal effect, suggesting the model's reasoning structure is relatively robust to prompt framing. However, Accuracy remains notably lower ( 0.52–0.64), revealing a gap between well-articulated calculation steps and fully correct numerical results, likely due to errors in parameter substitution, arithmetic, or unit handling. Overall, the model is competent at structuring calculations and conveying reasoning but is constrained in reliably producing fully accurate numerical outcomes, defining its capability boundary as stronger in procedural clarity than in final-answer precision. |
| **Summary** | Across formula derivation, computational reasoning, and formula calculation tasks, Qwen3-Max demonstrates consistently strong process-oriented performance, particularly in clarity, coherence, and relevance, indicating robust structuring of reasoning and effective communication of intermediate steps. Metrics such as correctness and completeness remain high in derivation tasks but show moderate declines in calculation-focused tasks, reflecting increasing difficulty in maintaining factual and logical precision when numerical computation is required. Accuracy consistently lags behind process-oriented scores, especially in computational and calculation tasks, revealing a systematic gap between well-articulated reasoning and fully correct final answers, likely due to arithmetic, parameter, or unit-handling errors. Overall, the model excels in constructing logically coherent, detailed, and relevant reasoning chains but is constrained in reliably producing precise numerical outcomes, defining its capability boundary as strong in procedural and conceptual reasoning while limited in final-answer fidelity under astronomy-related computational tasks. |

Table 26: Example of Analysis Report by Qwen3-Max Evaluation.

| Reporting Phase | Analysis of Content |
|---|---|
| **Formula Derivation** | The evaluation results indicate that Gemini 2.5 Flash demonstrates consistently strong performance in formula derivation, with high scores across correctness, completeness, relevance, coherence, and detail. Variations in experimental settings yield only modest differences, suggesting the model's derivation ability is robust to prompt modifications such as slow thinking, chain-of-thought guidance, or system prompt removal. The inclusion of image information slightly lowers completeness and detail, indicating potential limitations in integrating visual context for precise formula derivation. Overall, the model reliably produces accurate, coherent, and relevant derivations from given principles, with its main boundaries appearing in handling additional multimodal input where detailed integration is required. |
| **Unimodal Computation** | The evaluation results show that Gemini 2.5 Flash exhibits limited baseline performance in computational reasoning, with low correctness, completeness, detail, and accuracy under standard text-only settings, indicating challenges in multi-step quantitative analysis. Prompting strategies such as chain-of-thought slightly improve process-oriented metrics and overall reasoning coherence, but without fundamentally raising accuracy, which remains below 40% in most text-only scenarios. Inclusion of code execution or image information produces substantial gains across all metrics and markedly higher accuracy, suggesting that the model's reasoning ability is highly dependent on external or structured support for intermediate calculations. Overall, the model can produce coherent and relevant reasoning steps, but its capability boundary lies in performing reliable multi-step quantitative computations unaided, with significant improvement only when supplemented by code or visual inputs. |
| **Formula Calculation** | The evaluation results indicate that Gemini 2.5 Flash demonstrates moderate capability in formula calculation, with process-oriented metrics such as correctness, completeness, and detail generally in the mid-range and slightly higher scores for relevance and coherence, reflecting that the model can identify appropriate formulae and provide structured reasoning but struggles with fully accurate numerical computation. Accuracy remains below 60% across text-only and chain-of-thought settings, suggesting a persistent gap between the reasoning process and correct final results. Prompt variations and system modifications have limited impact on improving overall performance, while code-assisted execution shows the most substantial gain in accuracy, indicating reliance on structured computational support. Overall, the model is competent in selecting and applying relevant formulas with coherent reasoning but is constrained in reliably producing precise numerical outcomes without external calculation aids. |
| **Summary** | Across astronomy-related reasoning tasks, Gemini 2.5 Flash demonstrates a clear gradient in capability: it performs strongly in formula derivation, moderately in formula calculation, and weakly in multi-step computational reasoning. Process-oriented metrics such as correctness, coherence, and relevance are generally higher than accuracy in tasks requiring numerical computation, indicating that the model can produce logically structured and conceptually sound reasoning even when final results are incorrect. Strengths consistently appear in clarity, coherence, and relevance, while weaknesses emerge in completeness, precision, and accuracy, particularly when computations or multi-step quantitative analysis are involved. Experimental modifications like chain-of-thought prompting or system prompt removal yield modest improvements, whereas structured aids such as code execution or image input substantially boost performance in calculation and reasoning tasks. Overall, the model reliably handles conceptual derivation and reasoning structures but is constrained in producing precise numerical outputs and independent multi-step calculations, with its performance strongly dependent on external computational or visual support for tasks requiring accurate quantitative analysis. |

Table 27: Example of Analysis Report by Gemini 2.5 Flash Evaluation.

| Reporting Phase | Analysis of Content |
|---|---|
| **Formula Derivation** | The evaluation results indicate that AstroSage-8B exhibits generally low performance in the Formula Derivation task, with modest improvements observed under chain-of-thought (+CoT) and system-prompt removal (-System) settings. Across metrics, the model shows relatively better relevance and coherence, suggesting it can maintain focus and structure in its derivations, but struggles substantially with correctness, completeness, and precision, reflecting limited accuracy and detail in applying astronomical principles or deriving formulas from first principles. Enhancements such as step-by-step reasoning (+CoT) yield incremental gains in logical rigor and coverage, while removing system prompts slightly improves factual correctness and completeness, indicating sensitivity to prompt context. Overall, the model demonstrates partial procedural understanding and the ability to organize derivations coherently, but its formula derivation capability remains constrained by shallow reasoning and frequent factual or technical gaps, limiting reliable use for precise astronomical formula generation. |
| **Unimodal Computation** | AstroSage-8B demonstrates limited capability in the Computational Reasoning task, consistently achieving low overall scores across correctness, completeness, and precision metrics, while maintaining higher relevance and moderate coherence, indicating that the model can follow the problem context and structure its responses but struggles with quantitative accuracy and rigorous multi-step reasoning. Notably, the measured Accuracy is near zero across all settings, revealing a stark mismatch between perceived process quality and actual correctness of results, highlighting that even seemingly coherent or detailed solutions rarely produce numerically or logically consistent outcomes. Incremental improvements from chain-of-thought prompts marginally enhance logical flow and coverage but do not meaningfully impact factual correctness or final results. Overall, the model exhibits basic procedural awareness and organization in computational reasoning but is fundamentally constrained in executing reliable multi-step quantitative analysis, making it unsuitable for tasks requiring precise astronomical calculations. |
| **Formula Calculation** | AstroSage-8B shows limited performance in the Formula Calculation task, with moderate relevance and coherence scores indicating that it can structure responses and maintain focus on the question, but consistently low correctness, completeness, and precision reflect frequent errors in formula selection, parameter substitution, unit handling, and arithmetic. Despite occasional detailed reasoning, the measured Accuracy remains near zero across all settings, revealing a pronounced disconnect between the perceived process quality and actual numerical correctness. Chain-of-thought prompts slightly improve logical flow and coverage but do not meaningfully increase the likelihood of producing accurate results. Overall, the model can organize and present calculation steps coherently but lacks reliable quantitative accuracy, constraining its utility for tasks requiring precise application of astronomical formulae. |
| **Summary** | Across Formula Derivation, Computational Reasoning, and Formula Calculation tasks, AstroSage-8B demonstrates a consistent pattern of limited quantitative and logical accuracy, with low correctness, completeness, and precision scores reflecting frequent errors in reasoning, formula application, and arithmetic. Relevance and coherence are relatively stronger across all tasks, indicating the model can maintain focus on questions and structure responses in a logically organized manner, though this does not translate into accurate results. The near-zero measured Accuracy in tasks requiring numerical or formal correctness highlights a persistent disconnect between apparent procedural quality and actual solution validity. Incremental improvements from chain-of-thought prompts or prompt modifications modestly enhance logical flow and coverage but fail to overcome fundamental deficits in multi-step computation and rigorous formula derivation. Overall, the model exhibits basic procedural awareness and coherent exposition but is fundamentally constrained in performing reliable, precise astronomical computations or derivations, limiting its utility to qualitative reasoning or illustrative explanations rather than exact quantitative analysis. |

Table 28: Example of Analysis Report by AstroSage-8B Evaluation.