

Rethinking Domain Specialization for Open-Ended Scientific Reasoning in Astronomy Language Models

Vanessa Lama*, Sanjay Das*, Emily Herron*, Yuan-Sen Ting^{†§}, Tijmen de Haan[‡], Junqi Yin*, Tirthankar Ghosal*, Feiyi Wang*

*Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA

[†]The Ohio State University, Columbus, Ohio, USA

[‡]QUP/IPNS, High Energy Accelerator Research Organization (KEK), Tsukuba, Ibaraki, Japan

[§]Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany

Email: {lamav,dass3,ghosalt}@ornl.gov

Abstract—Domain-specialized language models are widely used for scientific question answering, but stronger general-purpose systems raise a sharper question: when does domain-specific fine-tuning remain valuable for open-ended scientific reasoning? We study this in astronomy with a curated QA benchmark from publicly available 2017–2026 Olympiad-style materials. The free-response subset contains 300 questions, including 204 text-only and 96 image-linked examples. We compare open-weight and API-served general-purpose, multimodal, and astronomy-specialized models using judge-based correctness and complementary reference metrics. Strong general-purpose models establish the highest correctness baseline in this testbed, while analyses of metric agreement, judge sensitivity, benchmark composition, and modality reveal variation not captured by a single leaderboard. These results motivate treating domain specialization as a task- and deployment-dependent property and highlight the role of domain-specific evaluation in determining which models, capabilities, and evaluation criteria are appropriate for scientific workflows.

Index Terms—AI for Science, astronomy, scientific reasoning, domain specialization, question answering, benchmarking, multimodal evaluation, LLM-as-judge

I. INTRODUCTION

Large language models (LLMs) and multimodal foundation models are increasingly used for scientific question answering, literature understanding, and knowledge-intensive reasoning. Astronomy provides a concrete testbed for broader AI-for-science evaluation because its problems combine specialized terminology, quantitative reasoning, diagrams, and physical assumptions. Rigorous evaluation remains difficult because open-ended problems are poorly served by exact match or shallow overlap metrics.

A central question for scientific language modeling is not whether domain specialization is useful, but when it provides measurable advantages over strong general-purpose models. Astronomy-specialized models have improved over earlier general baselines, while open or locally served models raise deployment questions not captured by accuracy alone. At the same time, general-purpose models have advanced rapidly on mathematical, scientific, expert-level, and olympiad-style

reasoning benchmarks [1]–[4]. Domain fine-tuning should therefore be compared with current general-purpose systems, not only smaller or older baselines. Most astronomy-focused evaluations emphasize multiple-choice QA or similar datasets, where mapping response to a definitive solution can provide context on model performance. However, we have yet to fully capture free-response reasoning with explanations, intermediate steps, and quantitative or conceptual conclusions.

To examine this gap, we study a curated astronomy QA benchmark derived from Olympiad-style materials spanning 2017–2026. We focus on the free-response subset, including text-only and image-linked questions, while a multiple-choice track is reserved for future expansion of our study. Our evaluation complements broader scientific and multimodal benchmarks such as MMLU, ScienceQA, GPQA, SciBench, MMMU, and OlympiadBench [1]–[6] by focusing on open-ended astronomy problems with reference solutions, multi-part reasoning structure, and image-linked scientific context. We evaluate open-weight and API-served models with general-purpose, multimodal, reasoning, and astronomy-specialized capabilities.

We use this benchmark to characterize scientific reasoning across model types, question properties, modalities, and evaluation methods.

The contributions of this paper are:

- a structured 2017–2026 free-response astronomy QA corpus and conversion pipeline preserving subpart context, reference solutions, and associated figures;
- a unified benchmarking framework for comparing general-purpose, multimodal, and astronomy-specialized language models on open-ended astrophysics reasoning tasks;
- a systematic characterization of model behavior across question properties, evaluation metrics, judge selection, modality, and benchmark robustness.

Fig. 1 summarizes the experimental pipeline from benchmark construction through model response generation, evaluation, and exploratory analysis.

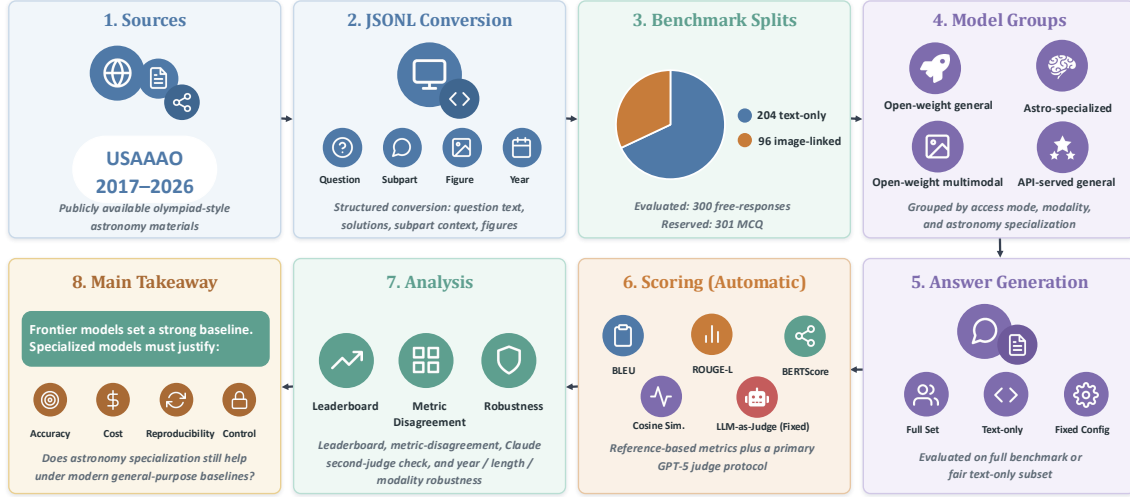


Fig. 1. Experimental pipeline for constructing the benchmark, generating model responses, evaluating scientific answer quality, and analyzing model behavior across performance, metric-agreement, robustness, and modality views.

II. BACKGROUND AND RELATED WORK

Scientific question answering has evolved from broad knowledge-based benchmarks toward increasingly specialized evaluations of domain knowledge, complex reasoning, and multimodal understanding. Yet evaluating language models remains a moving target as models, data, capabilities, and scientific use cases rapidly evolve. As AI becomes more integrated into scientific practice, evaluation must likewise mature toward systematic frameworks that identify which measures of model capability and reliability are most meaningful for a given scientific workflow.

A. Scientific Question Answering and Benchmarking

Scientific QA benchmarks test reasoning over technical material rather than surface retrieval. MMLU, ScienceQA, GPQA, and SciBench cover broad academic knowledge, multimodal science reasoning, expert questions, and college-level problem solving [1], [2], [5], [6]. MMMU and Olympiad-Bench emphasize expert-level visual and olympiad-style reasoning [3], [4]. Particularly in astronomy, AstroMLab 1 introduced a multiple-choice benchmark [7]. As an extension, our benchmark focuses on open-ended problems requiring physical modeling, derivation, numerical estimation, and explanation in an attempt to evaluate a separate aspect of scientific QA.

B. Evaluation of Open-Ended Model Outputs

Open-ended scientific responses are difficult to evaluate with exact match or multiple-choice accuracy. BLEU and ROUGE-L provide inexpensive and reproducible measures of response similarity, but primarily capture lexical overlap rather than scientific correctness [8], [9]. BERTScore and sentence-embedding cosine similarity capture semantic similarity be-

yond lexical overlap, but may still assign high similarity to scientifically incorrect responses [10], [11]. LLM-as-judge evaluation provides a practical approach for assessing open-ended responses, but its reliability depends on factors such as judge selection, evaluation criteria, and potential biases [12]–[14]. SciTrust and EAIRA similarly argue for multi-faceted scientific evaluation [15], [16]. We therefore complement judge-based evaluation with reference metrics and robustness analyses to provide a more comprehensive assessment of model performance.

C. Domain-Specialized and Multimodal Models for Science

Domain-adapted language models seek to improve scientific capabilities through continued pretraining or instruction tuning on specialized corpora. In astronomy, AstroLLaMA introduced models trained on astronomy literature [17], while subsequent AstroMLab efforts expanded domain-specific benchmarking and model development through the AstroLLaMA and AstroSage families [18]–[21]. AstroSage further represents domain adaptation at scale, with its 70B-scale model trained using leadership-class computing resources on Frontier [21]. These efforts motivate evaluating specialization not as an assumed advantage, but alongside general-purpose models and practical considerations such as open-weight access, local deployment, and data governance. We also consider multimodal capability, as astronomy problems frequently incorporate diagrams, plots, and sky maps.

III. METHODOLOGY

A. Dataset Source and Construction

The dataset is derived from publicly available USA Astronomy and Astrophysics Organization (USAAAO) Olympiad materials spanning 2017–2026. Annual L^AT_EX question and

solution files are processed through a structured conversion pipeline to produce JSONL records that preserve question text, reference solutions, subpart organization, year identifiers, and associated figures. The resulting free-response dataset contains 300 examples.

B. Record Construction

Each JSONL record represents a standalone problem or individually evaluated subpart and includes the year, problem identifier, question text, reference answer, length category, and associated image paths when applicable. The core schema is `{year, id, question, reference_answer, image_paths, length_bucket}`, with additional metadata used to preserve parent-problem relationships, subpart structure, source information, and figures. Length categories are assigned heuristically based on problem structure, number of sub-parts, and expected reasoning depth rather than fixed token thresholds, and are used only for robustness analysis. For multi-part problems, shared context is retained to preserve the interpretation of equations, constants, diagrams, and coordinate conventions. Representative records are provided in the Appendix.

C. Evaluation Splits

The curated free-response dataset contains 204 text-only and 96 image-linked examples, with their distribution across years shown in Fig. 2. We define two evaluation settings based on model modality. The *text-only* setting includes the 204 examples that do not require visual input, providing a common evaluation set for all models. The *full* setting includes all 300 examples and is used to evaluate models capable of processing both textual and visual inputs. This separation enables comparisons across model classes without disadvantaging text-only models while retaining the multimodal characteristics of the benchmark. First-round multiple-choice materials are reserved for future development as a complementary accuracy-based evaluation track.

IV. EXPERIMENTAL SETUP

A. Dataset

Fig. 2 summarizes the overall dataset composition, including question format, year and modality distributions, and question-length categories.

B. Models

The evaluated systems span general-purpose, multimodal, and astronomy-specialized language models under both local and API-served configurations (Fig. 1). Open-weight models include Llama and GPT-OSS general-purpose models, Gemma multimodal models, and the astronomy-adapted AstroLLaMA and AstroSage families. API-served models include GPT-5.5, Claude Sonnet 4.6, and GPT-OSS 120B.

Evaluation follows model input capability and serving support. Text-only models use the shared 204-example subset, while models supporting image inputs use the full 300-example benchmark.

TABLE I
COMPUTE AND SERVING CONFIGURATIONS USED FOR BENCHMARK INFERENCE.

Model group	Serving	Compute configuration
Open 1B–20B text models	Local	A100 80GB; typically single-GPU inference
Open 70B text models	Local	A100 80GB; multi-GPU inference as required
Open multimodal models	Local	A100 80GB; image-enabled inference on the full benchmark
GPT-5 judge and GPT-5.5	API	ORNL Azure OpenAI; provider-managed compute
Claude Sonnet 4.6 and GPT-OSS 120B	API	OLCF/AmSC OpenAI-compatible endpoint; provider-managed compute

C. Inference Configuration

All models use a common generation and evaluation pipeline where supported. Each response is generated once and evaluated against its reference solution using automatic metrics and a fixed GPT-5 judge, which assigns scores of correct (1.0), partially correct (0.5), or incorrect (0.0). GPT-5.5 is evaluated without additional reasoning effort, while GPT-OSS 120B uses the provider-required low-reasoning setting. Higher-reasoning configurations are reserved for future ablation studies.

D. Computational Platform

Local open-weight inference was performed on an 8-GPU NVIDIA A100 system using a CUDA-enabled PyTorch environment. The benchmark is inference-only; no training or fine-tuning was performed as part of this study. Models were allocated one or multiple GPUs according to memory requirements, with independent evaluations executed concurrently when resources permitted. Per-example predictions, metrics, judge decisions, and run metadata were retained, with resumable execution used to limit repeated inference following interruptions.

API-served models were accessed through Oak Ridge National Laboratory (ORNL) Azure OpenAI and Oak Ridge Leadership Computing Facility (OLCF)/American Science Cloud (AmSC) OpenAI-compatible endpoints. Model identifiers, access dates, and run metadata were retained for reproducibility, while data loading, metric computation, serialization, and judge evaluation were handled through the same benchmark pipeline used for local models. Provider-side GPU utilization and energy measurements were not available; local evaluations retained wall-clock and per-example execution records.

E. Evaluation Protocol and Metrics

Models are evaluated on the full or text-only benchmark according to their input capabilities, with comparisons made primarily within the same setting. Free-response answers are assessed using a primary LLM-as-judge score alongside complementary reference-based metrics. Let y denote the reference answer and $\hat{y}$ the generated response. Then,

USAAAO_QA data overview

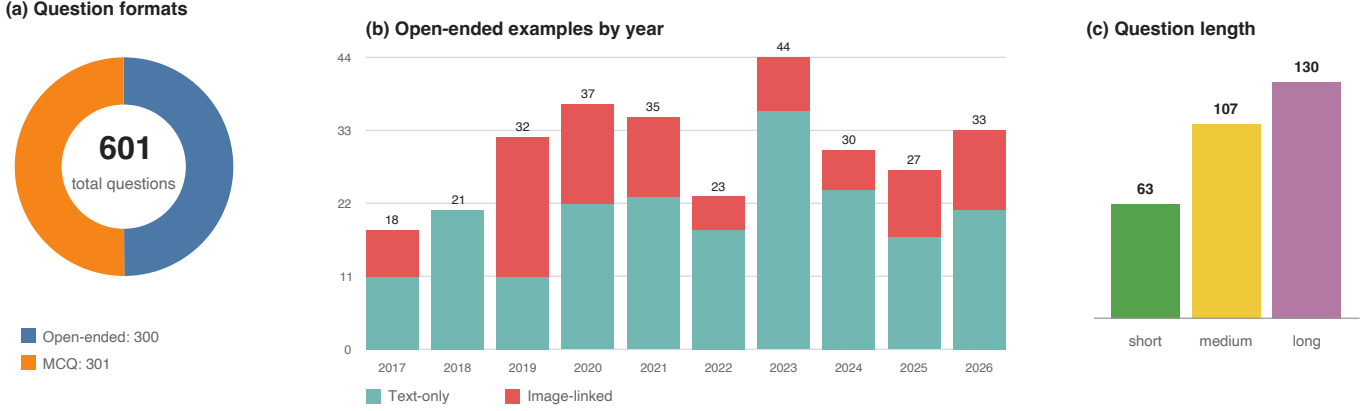


Fig. 2. Overview of the USAAAO_QA corpus. (a) Distribution of the 601 questions between free-response and multiple-choice formats. (b) Distribution of the 300 free-response examples across years and text-only or image-linked modalities. (c) Distribution of the free-response examples across short, medium, and long length categories. The 301 multiple-choice questions are reserved for a future benchmark track.

a) Judge score.: The primary correctness metric is a fixed GPT-5 judge score,

$$s_{\text{judge}}(\hat{y}, y) \in \{0, 0.5, 1\},$$

corresponding to incorrect, partially correct, and correct responses, respectively. We report the mean judge score

$$\bar{s}_{\text{judge}} = \frac{1}{N} \sum_{i=1}^N s_{\text{judge}}(\hat{y}_i, y_i).$$

b) Reference metrics.: We complement judge-based correctness with BLEU, ROUGE-L, BERTScore, and sentence-embedding cosine similarity. BLEU measures modified n -gram precision with a brevity penalty,

$$\text{BLEU} = \text{BP} \exp \left(\sum_{n=1}^4 w_n \log p_n \right),$$

where p_n denotes modified n -gram precision and w_n are uniform weights. ROUGE-L measures longest-common-subsequence overlap between the generated and reference responses. BERTScore instead compares contextual token embeddings through similarity-based matching; we report baseline-rescaled F1 scores, which may take negative values. Sentence-level semantic similarity is computed as

$$\text{CosSim}(\hat{y}, y) = \frac{e(\hat{y}) \cdot e(y)}{\|e(\hat{y})\| \|e(y)\|},$$

where $e(\cdot)$ denotes the sentence-embedding model. We treat these reference-based metrics as complementary diagnostic measures rather than direct measures of scientific correctness.

F. Score Uncertainty and Robustness

Because benchmark composition varies by year, modality, and length category, we quantify uncertainty in aggregate model scores using nonparametric bootstrap resampling. For

each replicate b , we sample N examples with replacement and compute

$$\bar{s}_{\text{judge}}^{(b)} = \frac{1}{N} \sum_{i=1}^N s_i^{(b)}.$$

Confidence intervals are obtained from the empirical distribution of the bootstrap means.

We further examine robustness across benchmark characteristics. Year-level variation is summarized as

$$\sigma_{\text{year}}^2 = \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} (\bar{s}_y - \bar{s})^2,$$

where $\bar{s}_y$ and $\bar{s}$ denote the year-specific and overall mean judge scores, respectively. For models evaluated on the full benchmark, the modality gap is

$$\Delta_{\text{modality}} = \bar{s}_{\text{text}} - \bar{s}_{\text{image}},$$

where positive values indicate lower performance on image-linked questions. Together, these analyses assess the stability of aggregate model comparisons across benchmark composition. Bootstrap intervals capture variation from example resampling only, as each model generates one response per question.

V. RESULTS

We evaluate domain specialization relative to strong general-purpose baselines through complementary views of aggregate performance, modality, metric agreement, judge sensitivity, and robustness.

A. Aggregate Model Performance

Table II gives aggregate performance across the 2017–2026 free-response benchmark. API-served general-purpose models set the highest baseline: GPT-5.5 reaches 0.6700 on the full benchmark, while Claude Sonnet 4.6 and GPT-OSS 120B reach 0.6814 and 0.6544 on the text-only subset. Among open-weight multimodal models, Gemma 4 31B IT is strongest at

TABLE II

FREE-RESPONSE BENCHMARK RESULTS FOR 2017–2026. *Judge* REPORTS PERFORMANCE ON EACH MODEL’S EVALUATION SETTING, WHILE *Judge*_{text} REPORTS PERFORMANCE ON THE SHARED 204-EXAMPLE TEXT-ONLY SUBSET. BOLD VALUES INDICATE THE HIGHEST FULL-BENCHMARK AND SHARED TEXT-ONLY JUDGE SCORES.

Model	Setting	<i>N</i>	Judge	<i>Judge</i> _{text}	BLEU	ROUGE-L	BERTScore F1	Cosine Sim.
Gemma 3 4B IT	Full	300	0.0967	0.0931	0.0143	0.1757	-0.0336	0.5860
Gemma 4 26B A4B IT	Full	300	0.3567	0.3897	0.0161	0.1828	-0.0311	0.5980
Gemma 4 31B IT	Full	300	0.3933	0.4314	0.0155	0.1872	-0.0237	0.6064
GPT-5.5	Full	300	0.6700	0.7083	0.0127	0.2168	-0.0592	0.6072
Claude Sonnet 4.6	Text-only	204	0.6814	0.6814	0.0115	0.1908	-0.0649	0.6077
GPT-OSS 120B	Text-only	204	0.6544	0.6544	0.0120	0.1899	-0.1162	0.5910
Llama 3.2 1B Instruct	Text-only	204	0.0343	0.0343	0.0175	0.1724	-0.0447	0.5725
Llama 3.1 8B	Text-only	204	0.0172	0.0172	0.0187	0.1312	-0.0800	0.5162
Llama 3 8B Instruct	Text-only	204	0.1201	0.1201	0.0209	0.1758	-0.0736	0.5894
Llama 3.1 70B Instruct	Text-only	204	0.2598	0.2598	0.0224	0.2058	-0.0023	0.6193
GPT-OSS 20B	Text-only	204	0.3284	0.3284	0.0114	0.1478	-0.1476	0.5988
AstroLLaMA 3 8B AIC	Text-only	204	0.0833	0.0833	0.0208	0.1403	-0.0607	0.5220
AstroSage 8B	Text-only	204	0.1005	0.1005	0.0167	0.1280	-0.0680	0.5307
AstroSage 70B (20251009)	Text-only	204	0.3333	0.3333	0.0204	0.1966	0.0034	0.5997

0.3933. Because this view mixes input settings, we next isolate the shared text-only subset.

B. Text-Only Comparison

To control for modality, we compare all models on the shared 204-example text-only subset. GPT-5.5, Claude Sonnet 4.6, and GPT-OSS 120B achieve judge scores of 0.7083, 0.6814, and 0.6544, respectively. Among locally evaluated open-weight models, AstroSage 70B and GPT-OSS 20B obtain similar scores of 0.3333 and 0.3284. This comparison provides a common setting for examining the effect of domain specialization relative to general-purpose baselines. Appendix Table V further reports specialization deltas against comparable general models. We next examine whether reference-based metrics reflect the same performance patterns observed under judge-based evaluation.

C. Metric Disagreement

Fig. 3 shows substantial variation across metrics on the shared text-only subset. ROUGE-L and sentence-level cosine similarity show moderate positive correlations with judge score ($r = 0.68$ and $r = 0.66$, respectively), whereas BERTScore shows little association ($r = -0.11$). BLEU is negatively correlated with judge score ($r = -0.60$), with scores remaining low even among models receiving substantially higher correctness judgments.

These differences are particularly relevant when comparing domain-specialized and general-purpose models. Domain-specific terminology may increase lexical or semantic similarity to reference solutions without necessarily indicating a correct solution, while correct responses may follow alternative wording or derivations. We therefore treat reference-based metrics as complementary diagnostics and use judge-based scoring as the primary correctness measure. Given these differences across metrics, we assess the sensitivity of the evaluation to judge selection.

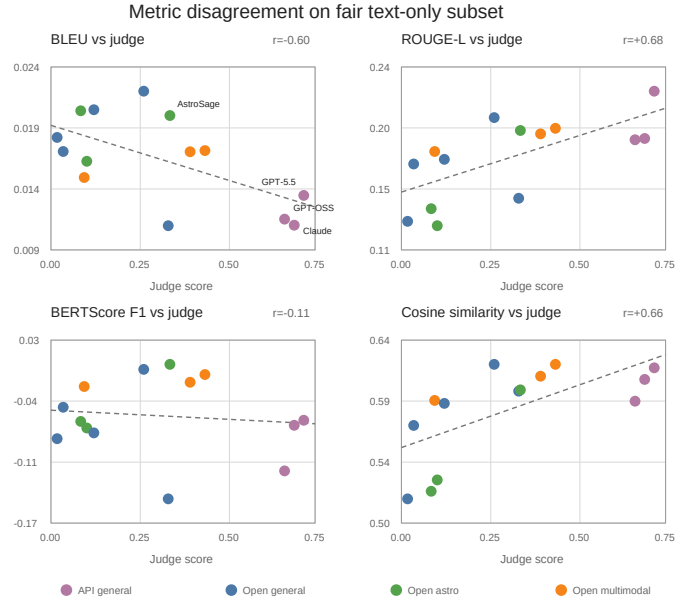


Fig. 3. Relationship between judge scores and reference-based metrics on the shared text-only subset. Correlations vary considerably across BLEU, ROUGE-L, BERTScore F1, and sentence-level cosine similarity, illustrating that reference similarity does not consistently track judged correctness.

D. Judge Sensitivity

To assess whether model comparisons depend on evaluator choice, we re-score the five highest-performing eligible models under the primary GPT-5 judge on the shared 204-example text-only benchmark using Claude Sonnet 4.6 as a second judge. Claude Sonnet 4.6 is excluded to avoid self-evaluation.

Table III shows that judge selection affects score magnitude more strongly for some models than others, while the highest-performing models remain stable across both evaluators. Agreement ranges from 0.608 for GPT-OSS 20B to 0.882 for GPT-OSS 120B, with score differences ranging from -0.0147 to $+0.1593$. Because the primary GPT-5 judge shares a model family with some evaluated systems, same-family

TABLE III

SECOND-JUDGE SENSITIVITY ON THE SHARED TEXT-ONLY SUBSET. Δ DENOTES THE CLAUDE SCORE MINUS THE GPT-5 SCORE; AGREEMENT IS THE FRACTION OF EXAMPLES RECEIVING IDENTICAL 0/0.5/1 SCORES.

Model	N	GPT-5	Claude	Δ	Agree
GPT-5.5	204	0.7083	0.6936	-0.0147	0.868
GPT-OSS 120B	204	0.6544	0.6691	+0.0147	0.882
Gemma 4 31B IT	204	0.4314	0.5319	+0.1005	0.623
AstroSage 70B	204	0.3333	0.3627	+0.0294	0.868
GPT-OSS 20B	204	0.3284	0.4877	+0.1593	0.608

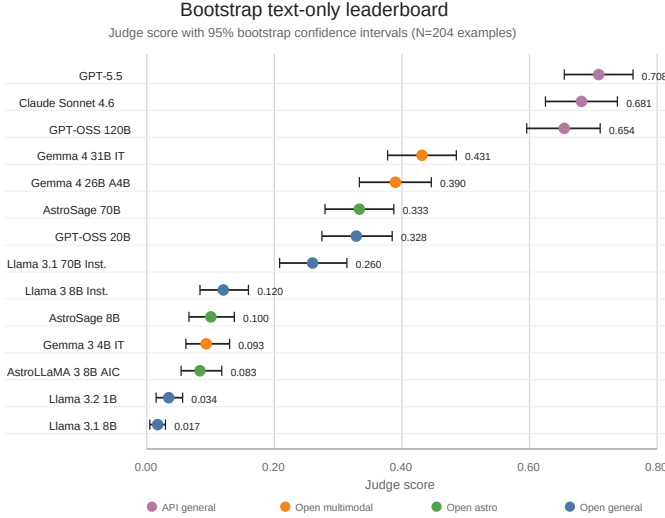


Fig. 4. Bootstrap estimates of model performance on the shared text-only subset ($N = 204$). Points denote mean judge scores and intervals show 95% nonparametric bootstrap confidence intervals.

evaluator preference is a potential confound [14]. However, the observed score shifts do not indicate a uniform advantage for models from the same family. These results reinforce the need to interpret small score differences cautiously and motivate examining whether model comparisons remain stable across variations in benchmark composition.

E. Performance Robustness

We first assess whether differences in aggregate performance exceed variation induced by benchmark composition. Fig. 4 shows 95% bootstrap confidence intervals for judge scores on the shared text-only subset. Large performance differences remain well separated under resampling, whereas intervals overlap for closer comparisons, including GPT-5.5 and Claude Sonnet 4.6 and AstroSage 70B and GPT-OSS 20B. These results suggest that small differences in aggregate scores should not be interpreted as definitive model rankings.

We further examine variation across exam year and question-length category. Fig. 5 shows that performance varies across years, indicating sensitivity to changes in benchmark composition and problem content. Despite this variation, broader differences among model groups are generally preserved.

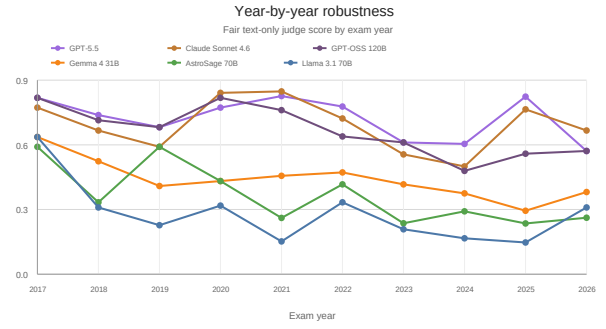


Fig. 5. Year-level performance on the shared text-only subset. Lines show mean judge scores by exam year for selected models.

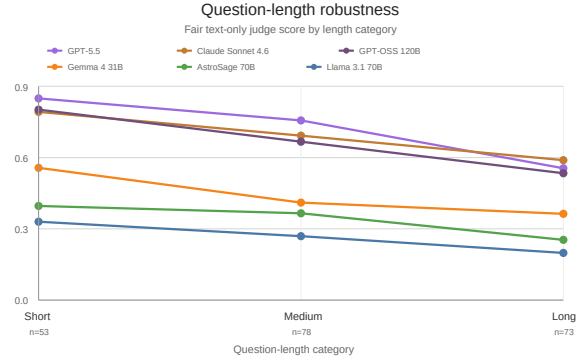


Fig. 6. Performance across question-length categories on the shared text-only subset. Lines show mean judge scores for short, medium, and long problems ($N = 53, 78$, and 73 , respectively).

Question length produces a more consistent pattern. As shown in Fig. 6, judge scores decline from short to medium and long problems across all selected models. Because the length categories also reflect problem structure and expected reasoning depth, this trend provides a coarse indication of increasing difficulty for longer, more involved problems.

F. Modality Effects

Furthermore, we examine whether performance differs between text-only and image-linked questions for models evaluated on the full benchmark. Fig. 7 shows lower judge scores on image-linked questions for GPT-5.5 and the larger Gemma variants. GPT-5.5 decreases from 0.708 on text-only questions to 0.589 on image-linked questions, while Gemma 4 31B and Gemma 4 26B decrease from 0.431 to 0.312 and from 0.390 to 0.286, respectively. Gemma 3 4B remains substantially lower in both settings.

The consistent gap among the higher-performing full-setting models suggests that image-linked problems introduce additional challenges beyond those observed on the text-only subset. However, modality is confounded with problem content and difficulty, so these differences should not be attributed solely to visual reasoning capability. A controlled multimodal comparison across a broader set of vision-capable models would be needed to isolate this effect.

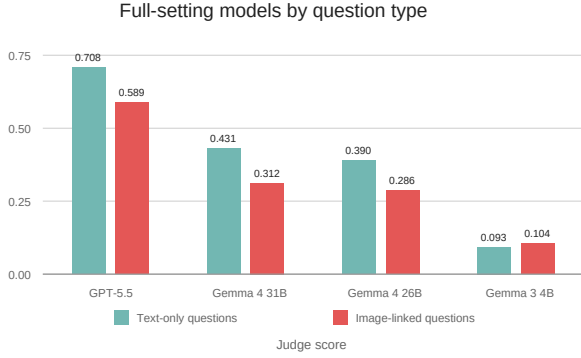


Fig. 7. Performance by question modality for models evaluated on the full benchmark. Bars show mean judge scores on the 204 text-only and 96 image-linked questions.

TABLE IV
FAILURE PATTERNS IDENTIFIED THROUGH TARGETED ANNOTATION OF
MODEL RESPONSES.

Failure type	Description
Incomplete derivation	Correct concept or relation is identified, but the reasoning or calculation is not completed.
Numerical / unit error	Reasoning is plausible, but errors occur in algebra, units, constants, or numerical scale.
Visual grounding	Required information from a figure, plot, or diagram is ignored or incorrectly interpreted.
Context fragmentation	A subpart is answered without necessary context or results established earlier in the problem.

VI. ERROR ANALYSIS

We conduct a targeted qualitative error analysis on a subset of model responses to characterize failures not fully captured by aggregate correctness or reference-based similarity metrics. The annotated responses reveal four recurring categories: incomplete derivations, numerical or unit errors, visual grounding failures, and context fragmentation. Table IV summarizes these failure patterns.

These failure patterns provide additional context for the metric disagreement observed in Fig. 3. A response may use appropriate domain terminology or begin from the correct physical relation while still producing an incomplete derivation or incorrect numerical result. Similarly, errors involving visual evidence or shared problem context may not be adequately captured by lexical or embedding-based similarity. These observations reinforce the use of reference-based metrics as complementary diagnostics rather than stand-alone measures of scientific correctness.

Because the error annotation covers only a subset of responses, we treat this analysis as qualitative rather than as an estimate of failure prevalence across the benchmark. Extending the annotation across models and question categories

would enable a more systematic comparison of how failure modes vary with model specialization, modality, and problem characteristics.

VII. DISCUSSION

Our results suggest that the value of domain specialization should be established through task- and deployment-specific evaluation rather than assumed from domain adaptation alone. In this astronomy benchmark, strong API-served general-purpose models establish a high correctness baseline, while specialized open-weight models offer different properties related to local deployment, reproducibility, controllability, and data governance. As reflected in Tables II and I, these systems also operate under different computational and serving environments. Model selection for scientific workflows therefore involves both capability and deployment requirements that are not represented by aggregate accuracy alone.

This comparison should not be interpreted as a definitive test of domain adaptation itself. The astronomy-specialized models evaluated here are not domain-adapted versions of the strongest contemporary general-purpose models, and differences in base model capability, scale, training data, and adaptation strategy remain confounding factors. Rather, the results raise a more targeted question: whether adapting stronger contemporary open models can improve domain-specific reasoning while retaining the access and deployment properties that motivate specialization in the first place.

Our evaluation also demonstrates that the answer depends on how scientific capability is measured. Reference-based metrics show varying agreement with judge-based correctness, evaluator choice affects some model scores, and performance varies across benchmark composition, question length, and modality. These differences argue against reducing open-ended scientific evaluation to a single metric or ranking. Instead, complementary evaluation views can help identify where models succeed, where they fail, and whether observed differences are robust to the characteristics of the scientific task.

More broadly, this study illustrates a growing challenge for AI for Science: model capabilities, architectures, and deployment options are evolving faster than any fixed benchmark can fully characterize them. Domain-specific evaluation can therefore serve not only to rank models, but to identify which capabilities, evaluation criteria, and deployment properties matter for a particular scientific workflow. Repeating such characterization across scientific domains can help establish where general-purpose models are sufficient, where specialization or additional scaffolding provides measurable value, and which evaluation signals are appropriate for scientific use.

Astronomy provides one controlled setting for this analysis, but the broader methodology extends naturally to domains involving specialized knowledge, quantitative reasoning, multimodal evidence, and domain-specific computational constraints. As AI becomes more deeply integrated into scientific workflows, such evaluations can provide empirical guidance for choosing among general models, domain adaptation, retrieval, tool use, and other forms of scientific AI augmentation

rather than assuming that any one approach is universally preferred.

VIII. LIMITATIONS

This study focuses on the free-response benchmark; the multiple-choice track is reserved for future evaluation. Text-only and full-setting results are based on different subsets and therefore should not be interpreted as direct comparisons. Similarly, comparisons between API-served and open-weight models characterize task performance rather than the full set of deployment tradeoffs.

Because the source materials are publicly available, potential training-data contamination cannot be ruled out, particularly for broadly web-trained models [22]. Variation across exam years provides a coarse temporal diagnostic but cannot establish whether individual problems were encountered during training or distinguish contamination from changes in question content and difficulty.

The LLM-as-judge evaluation is also sensitive to judge selection and rubric design. Our second-judge analysis provides an initial robustness check but does not replace expert adjudication or a broader multi-judge evaluation. Likewise, the targeted error analysis covers only a subset of responses and should not be interpreted as an estimate of failure-mode prevalence across the full benchmark.

Finally, local and API-served models differ in serving infrastructure, context handling, generation policies, and available hardware information. Provider-side resource utilization is unavailable, and this study does not provide a complete comparison of inference cost, GPU-hours, or energy consumption across serving environments.

IX. FUTURE WORK

Future work will extend the benchmark across task formats, model adaptation strategies, and measures of scientific reliability. Incorporating the reserved multiple-choice track will enable comparison between constrained-answer accuracy and open-ended reasoning within the same astronomy testbed. We also plan to expand the multimodal evaluation with additional vision-capable models to better isolate visual reasoning from differences in problem content and difficulty.

A key direction is evaluating astronomy adaptation using stronger contemporary open-weight base models. This will help separate the effects of domain specialization from differences in model scale and underlying capability. Beyond aggregate correctness, future evaluations will examine targeted scientific properties such as topic robustness, physical consistency, multi-step reasoning, and uncertainty calibration, including whether model confidence reliably tracks correctness across problem difficulty.

We further plan to evaluate retrieval- and agent-based approaches that combine language-model reasoning with astronomy resources, task decomposition, scientific tools, and numerical or unit verification. These experiments can help determine when domain adaptation, retrieval, tool use, or agentic scaffolding provides measurable benefits over stronger

general-purpose baselines, and whether those benefits justify additional training, computational, and deployment costs.

X. CONCLUSION

We presented a benchmark for open-ended astronomy question answering using Olympiad-style problems from 2017–2026 and used it to characterize general-purpose and astronomy-specialized language models under a common scientific reasoning task. The evaluation considers aggregate correctness, modality, metric agreement, judge sensitivity, and robustness across benchmark characteristics, providing a broader view of model behavior than a single leaderboard.

Strong general-purpose models establish a high correctness baseline in this testbed, while the value of domain specialization remains dependent on model capability, task requirements, and deployment constraints. Our analyses also show that reference-based similarity does not consistently reflect judged scientific correctness and that small performance differences can be sensitive to evaluator choice and benchmark composition.

More broadly, these results motivate domain-specific evaluation as a means of identifying which models, capabilities, and evaluation criteria are appropriate for scientific workflows. As AI systems continue to evolve, such characterization can help determine where general-purpose models are sufficient and where domain adaptation or additional scientific scaffolding provides measurable value.

XI. LLM USAGE DISCLOSURE

Large language models assisted with code generation, data-format validation, figure generation, and drafting. The authors reviewed conversion scripts, JSONL records, result tables, figures, and manuscript text. Benchmark predictions were generated by the evaluated systems and scored with the fixed pipeline described above. The authors remain responsible for the final content, analysis, and conclusions.

XII. ARTIFACT DESCRIPTION

Code and analysis workflows are maintained at <https://github.com/vanessalama09/astrobench>. The artifact supports benchmark construction, evaluation, and reproduction of tables and figures from per-example results. Model regeneration requires appropriate checkpoints, compute resources, or API access. The structured USAAO dataset will be released following documentation and licensing review.

ACKNOWLEDGMENT

This research used resources of the Oak Ridge Leadership Computing Facility (OLCF), which is a DOE Office of Science User Facility at the Oak Ridge National Laboratory supported by the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. Tirthankar Ghosal was supported by the U.S. Department of Energy, Advanced Scientific Computing Research, through the SciDAC-RAPIDS3 institute.

REFERENCES

- [1] D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman, “GPQA: A graduate-level google-proof Q&A benchmark,” 2023. [Online]. Available: <https://arxiv.org/abs/2311.12022>
- [2] X. Wang, Z. Hu, P. Lu, Y. Zhu, J. Zhang, S. Subramaniam, A. R. Lomaba, S. Zhang, Y. Sun, and W. Wang, “SciBench: Evaluating college-level scientific problem-solving abilities of large language models,” 2024. [Online]. Available: <https://arxiv.org/abs/2307.10635>
- [3] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su, and W. Chen, “MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI,” 2024. [Online]. Available: <https://arxiv.org/abs/2311.16502>
- [4] C. He, R. Luo, Y. Bai, S. Hu, Z. L. Thai, J. Shen, J. Hu, X. Han, Y. Huang, Y. Zhang, J. Liu, L. Qi, Z. Liu, and M. Sun, “OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems,” 2024. [Online]. Available: <https://arxiv.org/abs/2402.14008>
- [5] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, “Measuring massive multitask language understanding,” 2021. [Online]. Available: <https://arxiv.org/abs/2009.03300>
- [6] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan, “Learn to explain: Multimodal reasoning via thought chains for science question answering,” 2022. [Online]. Available: <https://arxiv.org/abs/2209.09513>
- [7] Y.-S. Ting, T. D. Nguyen, T. Ghosal, R. Pan, H. Arora, Z. Sun, T. de Haan, N. Ramachandra, A. Wells, S. Madireddy, and A. Accomazzi, “AstroMLab 1: Who wins astronomy jeopardy!?” *Astronomy and Computing*, vol. 51, p. 100893, 2025. [Online]. Available: <https://arxiv.org/abs/2407.11194>
- [8] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: A method for automatic evaluation of machine translation,” in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, P. Isabelle, E. Charniak, and D. Lin, Eds. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: <https://aclanthology.org/P02-1040/>
- [9] C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: <https://aclanthology.org/W04-1013/>
- [10] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTScore: Evaluating text generation with BERT,” 2020. [Online]. Available: <https://arxiv.org/abs/1904.09675>
- [11] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using siamese BERT-networks,” 2019. [Online]. Available: <https://arxiv.org/abs/1908.10084>
- [12] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, “Judging LLM-as-a-judge with MT-Bench and chatbot arena,” 2023. [Online]. Available: <https://arxiv.org/abs/2306.05685>
- [13] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, “G-Eval: NLG evaluation using GPT-4 with better human alignment,” 2023. [Online]. Available: <https://arxiv.org/abs/2303.16634>
- [14] K. Wataoka, T. Takahashi, and R. Ri, “Self-preference bias in LLM-as-a-judge,” 2024. [Online]. Available: <https://arxiv.org/abs/2410.21819>
- [15] E. Herron, J. Yin, and F. Wang, “SciTrust: Evaluating the trustworthiness of large language models for science,” in *SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, Nov. 2024, p. 17. [Online]. Available: <https://doi.org/10.1109/SCW63240.2024.00017>
- [16] F. Cappello, S. Madireddy, R. Underwood, N. Getty, N. L.-P. Chia, N. Ramachandra, J. Nguyen, M. Keceli, T. Mallick, Z. Li, M. Ngom, C. Zhang, A. Yanguas-Gil, E. Antoniuk, B. Kailkhura, M. Tian, Y. Du, Y.-S. Ting, A. Wells, B. Nicolae, A. Maurya, M. M. Rafique, E. Huerta, B. Li, I. Foster, and R. Stevens, “EAIIRA: Establishing a methodology for evaluating AI models as scientific research assistants,” 2025. [Online]. Available: <https://arxiv.org/abs/2502.20309>
- [17] T. D. Nguyen, Y.-S. Ting, I. Ciuca, C. O’Neill, Z.-C. Sun, M. Jabłońska, S. Kruk, E. Perkowski, J. Miller, J. J. Li, J. Peek, K. Iyer, T. Rozanski, P. Khetarpal, S. Zaman, D. Brodrick, S. J. Rodriguez Mendez, T. Bui, A. Goodman, A. Accomazzi, J. Naiman, J. Cranney, K. Schawinski, and R. Raileanu, “AstroLLaMA: Towards specialized foundation models in astronomy,” in *Proceedings of the Second Workshop on Information Extraction from Scientific Publications*, T. Ghosal, F. Grezes, T. Allen, K. Lockhart, A. Accomazzi, and S. Blanco-Cuaresma, Eds. Bali, Indonesia: Association for Computational Linguistics, Nov. 2023, pp. 49–55. [Online]. Available: <https://aclanthology.org/2023.wiesp-1.7/>
- [18] R. Pan, T. D. Nguyen, H. Arora, A. Accomazzi, T. Ghosal, and Y.-S. Ting, “AstroMLab 2: AstroLLaMA-2-70b model and benchmarking specialised LLMs for astronomy,” in *Proceedings of SC 2024-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2024, pp. 87–96.
- [19] T. de Haan, Y.-S. Ting, T. Ghosal, T. D. Nguyen, A. Accomazzi, A. Wells, N. Ramachandra, R. Pan, and Z. Sun, “AstroMLab 3: Achieving GPT-4o level performance in astronomy with a specialized 8b-parameter large language model,” *Scientific Reports*, vol. 15, no. 1, p. 13751, Apr. 2025. [Online]. Available: <https://doi.org/10.1038/s41598-025-97131-y>
- [20] T. de Haan, Y.-S. Ting, T. Ghosal, T. D. Nguyen, A. Accomazzi, E. Herron, V. Lama, R. Pan, A. Wells, and N. Ramachandra, “AstroMLab 4: Benchmark-topping performance in astronomy Q&A with a 70b-parameter domain-specialized reasoning model,” 2026. [Online]. Available: <https://arxiv.org/abs/2505.17592>
- [21] —, “AstroSage: Leading performance in astronomy Q&A with a 70b-parameter domain-specialized model,” in *ICML 2025: Forty-Second International Conference on Machine Learning*, 2025. [Online]. Available: <https://www.alcf.anl.gov/publications/astrosage-leading-performance-astronomy-qa-70b-parameter-domain-specialized-model>
- [22] C. Deng, Y. Zhao, X. Tang, M. Gerstein, and A. 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*, 2024, pp. 8706–8719. [Online]. Available: <https://arxiv.org/abs/2311.09783>

TABLE V
SPECIALIZATION DELTAS ON THE SHARED TEXT-ONLY SUBSET. Δ_{spec} IS
SPECIALIZED MODEL SCORE MINUS THE NEAREST EVALUATED GENERAL
BASELINE SCORE.

Specialized model	Baseline	Δ_{spec}
AstroSage 70B	Llama 3.1 70B	+0.0735
AstroSage 8B	Llama 3.1 8B	+0.0833
AstroLLaMA 3 8B AIC	Llama 3 8B Inst.	-0.0368

APPENDIX: ADDITIONAL BENCHMARK DETAILS

Specialization Deltas

Specialization should be compared against nearby general baselines, not only against frontier systems. Table V reports text-only judge-score differences between each astronomy-specialized model and the closest evaluated general model at a similar scale.

Prompt Template and Judge Rubric

The answer-generation prompt was:

Answer the astronomy/astrophysics question directly. Provide a concise final answer with essential reasoning only. If the question is quantitative, show the key calculation.
Question: {question}

The judge received the question, reference answer, and model answer, and was instructed to return strict JSON with keys `verdict`, `score`, and `rationale`. Verdicts were restricted to `correct`, `partial`, or `incorrect`; scores were restricted to 1.0, 0.5, or 0.0; and rationales were required to be brief and factual.

Example Benchmark Questions

Fig. 8 shows representative 2026 text-only and image-linked examples. Both follow the JSONL record structure described in the methodology, with question text, reference answer, year, identifier, length bucket, and optional image paths.

(a) Text-only example

1. (10 points) A “gravitational atom”

Bohr’s semiclassical model of the hydrogen atom assumed that the ground state angular momentum of the electron is $\hbar = h/(2\pi)$. While this model is incorrect, the Bohr radius does give the correct lengthscale of the hydrogen atom. By analogy to the Bohr model, consider a bound system consisting of a neutron and an electron interacting purely gravitationally. If the orbital angular momentum is $\hbar$, what would be the radius of this “atom”?

(b) Image-linked example

5. (20 points) Enceladus geysers



Figure 1: Cassini image of geysers from Enceladus. From NASA.

Enceladus is a moon of Saturn which has an outer crust of ice and is believed to have a subsurface liquid water ocean. Here are some useful physical properties:

- Radius of Enceladus (including the ice shell) $R = 250$ km
- Ice shell thickness $H = 25$ km
- Density of ice $\rho_i = 920$ kg/m³
- Surface gravity $g = 0.11$ m/s² (you can assume this is constant throughout the icy shell to the top of the subsurface ocean)
- Semi-major axis of Enceladus’ orbit around Saturn $a = 2.4 \times 10^8$ m
- Mass of Saturn $M = 5.7 \times 10^{26}$ kg

- (a) (10 points) The Cassini spacecraft observed that geysers of water erupt from the surface at a speed of $v = 400$ m/s relative to Enceladus. Calculate whether this water has sufficient speed to escape the gravitational field of Enceladus. If so, will it remain gravitationally bound to Saturn or escape into solar orbit?

Fig. 8. Representative 2026 free-response examples illustrating text-only and image-linked benchmark records.