Auditable Multi-View Hosted Multimodal Judging for AstroCLIMB

Kewen Zhu
Independent

Junyuan Tan
University of California, Berkeley

Abstract

We study AstroCLIMB, a four-way relationship-classification task over pairs of astronomy figure/caption endpoints. Starting from a leakage-audited lexical and low-level image-statistics baseline (M7), we evaluate a hosted multimodal judge under a fixed, auditable protocol. GPT-5.6 Luna receives only the two endpoint contents and produces integer class probabilities for three views in both endpoint directions. On 2,000 frozen validation pairs, all 12,000 calls succeeded. Equal-weight fusion reaches macro-F1 0.67, versus M7's 0.40 (paired component-bootstrap delta $+0.27$, 95% CI $[+0.24, +0.29]$); Luna-only reaches 0.68. These are conditional local-validation results, not an official Luna leaderboard score. We therefore report an auditable system study without claiming paper-level generalization.

1 Introduction

AstroCLIMB (Grezes et al., 2026; AstroCLIMB Organizers, 2026) asks a system to classify a pair of astronomy figure/caption objects as *same figure*, *same paper*, *related papers*, or *unrelated papers*. Pairs may be image–image, image–caption, or caption–caption. The task tests whether systems can combine heterogeneous scientific endpoints while avoiding shortcuts from identifiers and metadata. The benchmark's *related papers* target is citation-based, while the original judging prompt describes substantive relatedness; throughout, we treat the latter as a semantic proxy rather than citation-edge ground truth.

Our M7 system uses symmetric word-level TF–IDF summaries, endpoint availability, and deterministic low-level image statistics. It obtains 0.40 macro-F1 on the frozen observable-content component split and 0.40 on the official public run (submission 56179217). Our research question is: under a fixed, auditable protocol and the same frozen split, how much does a hosted multimodal judge add to this content-only baseline? The resulting system, AstroLex-Luna, emphasizes provenance and failure visibility: anonymous cache keys, digest-bearing receipts, no stored payloads, and an all-six-or-fallback contract.

Our contributions are a documented content-only baseline and request protocol; an auditable report of one cached hosted-judge run, with conditional uncertainty and explicit failure handling; and an error analysis that separates benchmark semantics from the prompt's relatedness proxy. The contribution is an evaluation procedure and empirical analysis, not a new model architecture or a new aggregation algorithm.

This work is a systems evaluation, not a claim that a proprietary endpoint is universally superior. Related scientific multimodal resources include SciMMIR (Wu et al., 2024), MISTI (Taechoyotin and Acuna, 2024), and Multimodal ArXiv (Li et al., 2024); graph-based representations such as SPECTER (Cohan et al., 2020) use information excluded here.

2 Data and leakage audit

The official training and test files each contain 10,000 pair rows. Separate DOI, paper-title, author, citation-metadata, and image-identifier fields are excluded from prediction features. Identifiers and content hashes support splitting and traceability. Class labels supervise M7 training and score validation predictions; they are never supplied to the hosted judge. Contaminated sample/solution values are likewise excluded from modeling and selection.

We project permitted content to decoded captions, image bytes, modality, and endpoint SHA-256 digests. The fixed component-greedy-v2 split (seed 20260911) has 8,000 training and 2,000 validation rows, 8,671 observable-

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 34–37
November 10, 2026 ©2026 Association for Computational Linguistics

content components, exact validation class counts 200/600/600/600, and zero endpoint or pair overlap. The term "leakage-audited" below means observable-content component disjoint only: paper_disjoint=false because paper identities are unavailable for every endpoint. One endpoint is shared between distributed train and test files, so we do not claim a fully independent official test set. This content-overlap audit cannot detect whether the hosted model encountered the papers during pretraining.

3 Method

M7 is our frozen lexical and image-statistics baseline. Its 328 features combine symmetric word TF–IDF pair summaries, availability, and 64 low-level image statistics (color, brightness, gradient, and spatial summaries). A class-balanced multinomial logistic regression is fit using training-only TF–IDF, standardization, $C = 1$, and at most 2,000 iterations.

Hosted judge. For each pair, GPT-5.6 Luna (medium reasoning effort) is queried for direct, counterfactual, and boundary/ambiguity views, each in A/B and B/A directions. Inputs contain only raw captions and preprocessed images (EXIF transpose, RGB, longest edge at most 1536, JPEG quality 90); no tools are enabled, temperature is unset, and API storage is disabled (`store=false`). The six responses are aggregated as specified below.

Response and aggregation protocol. We use model alias `gpt-5.6-luna`. Each response must contain exactly four integer fields, in class order SF/SP/RP/UP, each in $[0, 100]$ and summing to 100; malformed responses count as failures. The view instructions are "View: direct. Classify the relation.", "View: counterfactual. Ignore superficial wording/layout and classify.", and "View: boundary. Classify conservatively at the nearest class boundary." Each follows the same label definitions and is issued with both endpoint orders. Appendix A reproduces the common prompt; the accompanying `PROTOCOL.md` also records the response schema and settings.

For class c, we compute

$$g_c = \left[\prod_{k=1}^{6} \frac{\max(p_{kc}, 10^{-9})}{100} \right]^{1/6}, \qquad (1)$$

System	Macro	SF	SP	RP	UP
M7 baseline	0.40	0.39	0.49	0.29	0.45
Luna-only	0.68	0.82	0.70	0.39	0.82
M7+Luna fusion	0.67	0.80	0.71	0.37	0.81

Table 1: Frozen validation results from one cached API run. SF, SP, RP, and UP denote same figure, same paper, related papers, and unrelated papers.

where p_{kc} is the returned integer. We normalize g, set SF to zero for image–image and text–text pairs, and renormalize. The same mask is applied to M7 before equal-weight probability fusion; prediction is by argmax, with ties resolved in the stated class order. Table 1 reports the original unmasked M7 baseline; its hard predictions and F1 are unchanged by masking in this run. Pooling both orders symmetrizes the cached ensemble; it does not guarantee identical responses on a new API run.

Fusion, failure handling, and uncertainty. The primary prediction is a fixed 0.5/0.5 fusion of masked M7 and Luna. A pair is aggregated only if all six calls succeed; any failure discards the partial Luna result and falls back to masked M7. This all-six-or-fallback behavior is implemented and tested; the reported run had no failures. We use 2,000 component-bootstrap replicates. The confidence interval is a fixed-run conditional component-bootstrap interval: it describes sampling variation conditional on these cached responses and does not include API randomness, model-alias changes, or serving drift. The live validation run required an explicit external-API flag and used an append-only response cache. The documented split and code specify how to rerun M7 with authorized data. Cached Luna responses permit reaggregation of this run; neither the protocol nor access to the service guarantees identical fresh responses.

4 Results

On the frozen validation set, M7 scores 0.40 macro-F1, Luna-only 0.68, and the fixed fusion 0.67. The fusion improves over M7 by $+0.27$; its paired component-bootstrap 95% CI is $[+0.24, +0.29]$. Luna-only has conditional bootstrap CI $[0.66, 0.70]$. The fused per-class F1 values are 0.80 (same figure), 0.71 (same paper), 0.37 (related papers), and 0.81 (unrelated papers); related papers remains the hardest class. Values shown here and in Table 1 are two-significant-digit

renderings; exact values remain in the evidence artifacts.

Fusion as a negative result. Fixed fusion is 0.011 macro-F1 below Luna-only. It raises same-paper F1 from 0.70 to 0.71, but lowers the other three class scores (Table 1); deltas are computed from unrounded values. We retain it as the prespecified comparison, not as the best-performing predictor. These results support the value of the hosted judge relative to M7, but do not support an overall macro-F1 benefit from adding M7 to Luna. No fusion weight was retuned on these outcomes, and we do not claim statistical significance for the Luna-only versus fusion difference.

Descriptive modality slices show fused macro-F1 of 0.52 for image–image ($n = 571$), 0.58 for image–caption ($n = 838$), and 0.51 for text–text ($n = 591$). Each slice averages over all four labels; SF has no support and contributes zero F1 in image–image and text–text. These availability-confounded scores should not be directly ranked as evidence of modality-specific understanding.

Error analysis. Table 2 shows the fused confusion matrix. Of 600 related-paper pairs, 263 (43.8%) are predicted as same paper and 165 (27.5%) as unrelated; only 142 (23.7%) are recovered. The reverse same-paper-to-related error occurs in 16 of 600 pairs (2.7%), showing an asymmetric boundary error. Luna-only has the same dominant failure: 276 related pairs become same paper, while 153 are correctly recovered. Fusion reduces the former count to 263 but increases related-to-unrelated errors from 147 to 165, leaving related-paper F1 lower overall.

The original prompt describes related papers as "objects from different but substantively related papers", whereas the task label requires a citation link. Content similarity is therefore an imperfect proxy for the target relation. This mismatch is a plausible contributor to the errors, but the confusion counts alone do not establish its cause. We preserve the original prompt and cached predictions rather than retroactively relabeling this experiment.

Exploratory call-count comparison. Using immutable cached responses, direct A/B (one call) scores 0.65, direct bidirectional (two) 0.64, all three A/B views (three) 0.66, and the full bidirectional three-view protocol (six) 0.67 under the same fusion. The six-versus-three difference (0.0081) is post-hoc and exploratory; it does not establish a causal contribution of any prompt view. Counts and cost do not by themselves demonstrate a benefit. Establishing a visual reasoning mechanism would require controlled image/caption interventions, which we did not run. We also cannot prove absence of pretraining contamination from this audit.

Gold \ Pred.	SF	SP	RP	UP
SF	200	0	0	0
SP	60	489	16	35
RP	30	263	142	165
UP	12	23	17	548

Table 2: Fused confusion counts on the 2,000 validation pairs. Rows are gold labels; columns are predictions. Abbreviations follow Table 1.

Auditability. The receipt records model, protocol, cache and prediction digests, SDK version, external-API status, and 12,000 successful calls. Summing the cache usage records gives 20,232,306 input and 3,783,736 output tokens; at the rates recorded for this run, USD 0.20/1.20 per million input/output tokens, estimated cost is USD 8.59. The accompanying protocol and `EVIDENCE.json` map the reported aggregates to hashed result and prediction artifacts. These records bind aggregates to local artifacts; hashes alone do not provide independent access to those artifacts. Raw endpoint payloads and per-pair predictions are not redistributed in this bundle. A fresh evaluation requires authorized competition data and access to the hosted endpoint, and remains sensitive to service drift.

5 Conclusion

The official system result in this paper is the M7 public run (score 0.40, submission 56179217). Luna is a locally evaluated hosted component, not an official system run. Under the observable-content component split and this fixed cached run, the constrained hosted judge substantially improves local validation over M7. This conclusion is conditional on one fixed split and cached service run; it does not establish paper-level generalization or a causal mechanism.

6 Limitations

The split is not paper-disjoint, so the results do not establish paper-level generalization. Luna is

proprietary, its serving behavior may drift, and the conditional bootstrap intervals do not capture API randomness. We cannot prove that the hosted model was free of pretraining contamination. Raw endpoint content can also contain visible titles or citations even though separate metadata fields are excluded. The six-view protocol does not establish causal visual or cross-modal understanding; controlled image/caption interventions would be needed for mechanism attribution. The fixed fusion was specified before reading the reported validation outcome, whereas Luna-only and call-count comparisons are exploratory. The result is neither a new architecture nor an official Luna leaderboard score. The prompt uses substantive relatedness as a proxy for the citation-edge definition of related papers; the cached results therefore also reflect this semantic mismatch, rather than an exact operationalization of citation linking.

Acknowledgements

Generative AI use. Generative AI tools, including OpenAI Codex, assisted with code implementation, experimental analysis, and manuscript drafting and revision. The authors are responsible for the scientific claims, references, and final content. The evaluated GPT-5.6 Luna predictions were produced under the protocol described in Section 3.

References

AstroCLIMB Organizers. 2026. AstroCLIMB: Astronomy citation linking from illustrations. WASP 2026 shared task and dataset documentation. Accessed 2026-09-11.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282. Association for Computational Linguistics.

Felix Grezes, Kelly Lockhart, Alberto Accomazzi, Atilla Alkan, Ethan Seefried, and Tirthankar Ghosal. 2026. AstroCLIMB: Astronomy citation linking from illustrations - a multimodal benchmark (overview of the WASP 2026 shared task). In *Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14387. Association for Computational Linguistics.

Pawin Taechoyotin and Daniel Acuna. 2024. Misti: Metadata-informed scientific text and image representation through contrastive learning. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 155–164. Association for Computational Linguistics.

Siwei Wu, Yizhi Li, Kang Zhu, Ge Zhang, Yiming Liang, Kaijing Ma, Chenghao Xiao, Haoran Zhang, Bohao Yang, Wenhu Chen, Wenhao Huang, Noura Al Moubayed, Jie Fu, and Chenghua Lin. 2024. SciMMIR: Benchmarking scientific multi-modal information retrieval. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12560–12574. Association for Computational Linguistics.

A Frozen judging instructions

The following common prefix is reproduced from the evaluated protocol, with line wrapping only:

> Labels: same_figure=exactly an image and its own caption; same_paper=distinct objects from one paper; related_papers=objects from different but substantively related papers; unrelated_papers=no substantive relation. Return integer probabilities summing to 100.

Each request appends one of the three view instructions in Section 3 and the two endpoint content blocks. Each view is called once with A/B and once with B/A. The four JSON keys are the label names above; values must be integers between 0 and 100, with no additional fields and a total of 100. Application-level validation enforces the total as well as schema compliance.

The original prompt equates relatedness with substantive content similarity, including in its description of the unrelated class. A pair can be topically similar without a citation link, and a citation can connect objects with dissimilar content. Accordingly, this prompt does not exactly operationalize the benchmark's citation-based target. A citation-aligned prompt comparison would require new predictions and evaluation; no such comparison was performed for this paper. We retain the original prompt and scores to keep that distinction inspectable.