ASTROALERTBENCH: EVALUATING THE ACCURACY, REASONING, AND SELF-ASSESSMENT OF MULTIMODAL LLMS IN ASTRONOMICAL CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern astronomical observatories generate a massive volume of multimodal data, creating a critical bottleneck for expert human review. While multimodal Large Language Models (LLMs) have shown promise in interpreting complex visual and textual inputs, their ability to perform specialized scientific classification while providing interpretable reasoning remains understudied. We introduce AstroAlertBench, a comprehensive multimodal benchmark designed to evaluate LLMs' performance in astronomical event review along a three-stage logical chain: metadata grounding, scientific reasoning, and hierarchical classification over five categories. We utilize a pilot sample of 1,500 real-world alerts from the Zwicky Transient Facility (ZTF), a wide-field survey that scans the northern sky to detect transient astronomical events. On this dataset, we benchmark 13 model configurations of frontier closed-source and open-weight LLMs supporting visual input. Our results reveal that high accuracy does not always align with model rationale self-assessment, impacting its reliability as a real-world assistant. Together, *AstroAlertBench* provides a framework for developing calibrated and interpretable astronomical assistants.

1 INTRODUCTION

The landscape of artificial intelligence has been fundamentally altered by the emergence of Large Language Models (LLMs) based on the Transformer architecture (Vaswani et al., 2017). These models, including proprietary series such as GPT (Brown et al., 2020; Achiam et al., 2023) and high-performance open-weight families like Qwen (Bai et al., 2023; Yang et al., 2025), have demonstrated proficiency across a broad spectrum of zero-shot reasoning tasks (Wei et al., 2022). The scaling of these architectures has enabled capabilities in multi-step reasoning and the integration of information from diverse sources, typically facilitated by extensive pre-training on diverse corpora followed by instruction tuning and alignment (Bai et al., 2023). Furthermore, the integration of visual encoders has birthed a new generation of Multimodal LLMs (MLLMs) capable of processing interleaved visual and textual data (Team et al., 2024). By mapping disparate modalities into a shared feature space, these models can address complex, open-ended problems that require the simultaneous interpretation of both semantic and structural visual cues (Radford et al., 2021; Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023).

This progress has prompted a transition toward utilizing LLMs as specialized scientific assistants (Taylor et al., 2022; Romera-Paredes et al., 2024). Research is increasingly moving beyond static information retrieval toward the development of autonomous LLM agents designed to execute end-to-end research workflows (Boiko et al., 2023; Bran et al., 2023). However, as scientific inquiries increase in complexity, general-purpose models encounter a dual bottleneck. First, high costs for commercial, closed-source APIs often prohibit large-scale deployment on high-volume scientific data streams (Pan et al., 2024). Second, general open-weight models frequently struggle with the long-tail domain knowledge and reasoning required for specialized research tasks (Pan et al., 2024; Guo et al., 2024). These limitations create a critical need for comprehensive frameworks to evaluate whether these models possess the underlying scientific logic necessary for reliable deployment.

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.

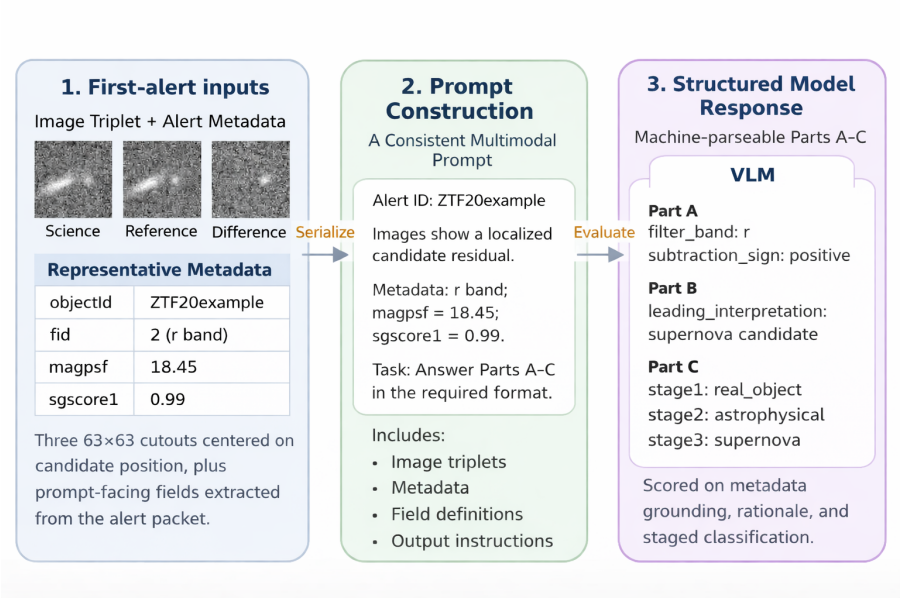

Figure 1: The AstroAlertBench pipeline combines ZTF science, reference, and difference cutouts with alert metadata in a multimodal prompt. Models return structured responses covering Part A metadata grounding, Part B scientific rationale, and Part C staged classification.

Addressing these challenges requires a shift from outcome-based metrics toward a staged evaluation logic (Guo et al., 2024). Scientific problem-solving is not merely a task of pattern recognition; it requires a verifiable logical chain that connects perception to reasoning and final application. For LLMs to be integrated into expert discovery pipelines, they must provide more than correct classification labels (Stoppa et al., 2025; Lightman et al., 2023). They must offer interpretable reasoning and accurate self-evaluations regarding their internal logic to ensure reliability in scientific environments (Guo et al., 2017; Kadavath et al., 2022; Lightman et al., 2023).

Astronomy provides a representative environment for evaluating these capabilities. Time-domain surveys, such as the Zwicky Transient Facility (ZTF) (Bellm et al., 2019; Graham et al., 2019; Masci et al., 2019), generate millions of nightly "alerts" signaling potential astrophysical discoveries (Masci et al., 2019). To process this data, the astronomy community relies on machine learning pipelines for automated categorization (Duev et al., 2019; Mahabal et al., 2019; Förster et al., 2021). However, as noted by Stoppa et al. (2025), these models primarily function as black boxes, lacking the human-readable rationales necessary for experts to prioritize follow-up resources. While Multimodal LLMs offer a path toward interpretable review, they face a significant domain gap in specialized astronomical logic, exacerbated by a lack of domain-specific benchmarks (Pan et al., 2024).

In this work, we introduce *AstroAlertBench*[1], a multimodal benchmark designed to evaluate LLMs in astronomical event review. We design the evaluation as a three-stage logical chain: (1) **Metadata Grounding**, assessing the model's ability to accurately read and restate verifiable observational parameters from the alert metadata; (2) **Scientific Rationale**, assessing the ability to propose astrophysical hypotheses and provide grounded reasoning; and (3) **Staged Classification**, testing a hierarchical decision process covering artifact rejection, physical origin, and final astrophysical categorization across five operationally meaningful categories. This staged approach allows for the isolation of failures in perception from errors in high-level reasoning (Guo et al., 2024).

A central component of this benchmark is evaluating **model self-assessment and self-correction**. We evaluate this through three lenses: (i) whether average rationale self-ratings are associated with classification accuracy across models, (ii) whether per-alert self-ratings distinguish correct

[1]The project website, dataset, and code are available at https://astroalertbench.com, https://huggingface.co/datasets/AnonymousUser16384/AstroAlertBench, and https://github.com/AnonymousUser16384/AstroAlertBench respectively.

from incorrect classifications, and (iii) a behavioral "second-rollout" test of whether models can successfully fix mistakes on alerts with low first-pass self-ratings. Across these dimensions, we benchmark 13 model configurations from the GPT, Claude, Gemini, Qwen, and Kimi families. Together, AstroAlertBench provides a framework for evaluating the scientific reasoning and reliability of multimodal LLMs as astronomical assistants.

2 Related Work

2.1 Astronomical Alert Triage and Broker Systems

The operational demands of modern time-domain surveys, most notably the Zwicky Transient Facility (ZTF) (Bellm et al., 2019; Graham et al., 2019; Masci et al., 2019), have necessitated the development of high-throughput alert distribution and filtering systems (Patterson et al., 2019). These surveys generate millions of real-time alerts that must be triaged to separate instrumental or subtraction artifacts from genuine astrophysical phenomena (Duev et al., 2019; Mahabal et al., 2019). Community broker systems, such as the Automatic Learning for the Rapid Classification of Events (ALeRCE), utilize specialized machine-learning pipelines to discriminate between classes such as supernovae, variable stars, and active galactic nuclei (AGN) (Carrasco-Davis et al., 2021; Förster et al., 2021). Similarly, the Bright Transient Survey (BTS) utilizes deep-learning frameworks like BTSbot to automate the identification and follow-up of bright transients (Rehemtulla et al., 2024). While these traditional models achieve high classification accuracy, they predominantly function as "black boxes," providing scores or labels without the human-readable scientific rationales required for expert review and resource prioritization.

2.2 Multimodal Reasoning and Scientific Benchmarks

The evaluation of Vision-Language Models (VLMs) has shifted from basic object recognition toward assessing complex, multi-discipline reasoning capabilities (Radford et al., 2021; Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023). Large-scale benchmarks such as Yue et al. (2024) and Lu et al. (2023) have established baseline performance for frontier models like GPT-4 (Achiam et al., 2023), Gemini (Team et al., 2024), and Qwen3 (Yang et al., 2025) across college-level academic tasks. However, as scientific inquiries move into specialized domains, general-purpose benchmarks often fail to capture the long-tail knowledge and cascading complexity required for research-grade problem solving (Taylor et al., 2022; Romera-Paredes et al., 2024). To address this, recent work has introduced domain-specific benchmarks in fields like chemistry and laboratory automation (Boiko et al., 2023; Bran et al., 2023). Specifically, Guo et al. (2024) developed staged reasoning protocols for molecular structure elucidation, while recent efforts in scientific discovery benchmarking have begun to decompose tractable tasks into hierarchical sub-goals, such as inspiration retrieval and hypothesis composition (Liu et al., 2025). More broadly, complementary work examines how reliably LLMs assess their own outputs and report confidence (Kadavath et al., 2022; Geng et al., 2024).

Within the astronomical domain, early efforts to evaluate Large Language Models have remained largely unimodal and restricted in their classification scope. For instance, Pan et al. (2024) introduced a suite to assess textual domain expertise and fact-retrieval through multiple-choice questions, while Stoppa et al. (2025) utilized foundation models for the textual interpretation of binary real-versus-bogus image classifications. These approaches, however, are fundamentally limited by their reliance on single-modality inputs and simplified decision tasks—either focusing exclusively on textual scientific knowledge retrieval (Pan et al., 2024) or providing binary interpretations based solely on visual stamps without integrating the accompanying numerical alert metadata (Stoppa et al., 2025). **In this work, we present *AstroAlertBench* as an advancement that bridges and expands upon these efforts** by introducing a multimodal, three-stage evaluation framework that addresses the complex requirements of multi-class alert triage. By requiring models to ground high-level astrophysical rationales in both raw image triplets and serialized alert parameters across a five-category taxonomy, our benchmark establishes a more comprehensive standard for evaluating the reliability, scientific logic, and self-assessment of LLM agents in time-domain astronomy.



Figure 2: **AstroAlertBench taxonomy.** The inner ring shows the five benchmark classes with equal weight: *bogus*, *asteroid*, *supernova*, *AGN*, and *variable star*. The outer ring illustrates representative subtype categories; this is not an exhaustive classification tree. Wedge sizes are schematic and do not encode subtype counts.

3 The AstroAlertBench Dataset

This section details the construction and composition of **AstroAlertBench**, focusing on the transition from raw astronomical data to a structured multimodal benchmark. We decompose the dataset generation into data acquisition, automated visual preprocessing, and taxonomic design.

3.1 Data Acquisition and Multimodal Content

We utilize the fully public ALeRCE real-time stamp classifier dataset (Carrasco-Davis et al., 2021) as the authoritative source for ground-truth labels and alert-level inputs. This dataset is derived from the Zwicky Transient Facility (ZTF) alert stream (Patterson et al., 2019), which issues nightly notifications of transient, variable, and moving objects. Each benchmark example is constructed from the **first-detection alert** of a unique alert source to evaluate rapid triage capabilities without reliance on multi-epoch light curves. Class-specific label provenance is detailed in Appendix B.1.1.

Each datapoint consists of two primary modalities:

- **Image Triplet:** Three 63×63 pixel cutouts centered on the candidate position: the *Science* image (current observation), the *Reference* image (historical baseline), and the *Difference* image (Science minus Reference).
- **Alert Metadata:** A subset of 19 prompt-facing features extracted from the ZTF AVRO packet, including photometric attributes (e.g., PSF magnitude), spatial and morphological context (e.g., distance to the nearest reference source), and historical detection counts.

3.2 Automated Image Preprocessing and Serialization

An automated visual pipeline prepares 32-bit astronomical data for 8-bit vision-language encoders. For each alert, the three cutouts are tiled horizontally and labeled Science, Reference, and Difference, with the candidate at the center of each panel (Figure 1). Metadata is serialized into a compact human-readable format with explicit field definitions to provide observational context (Appendix B.2.1).

3.3 Benchmark Taxonomy and Evaluation Set Construction

Mirroring the real-time filtering standards used by major astronomical data brokers (Carrasco-Davis et al., 2021; Rehemtulla et al., 2024), AstroAlertBench assesses LLMs across five primary categories: *bogus*, *asteroid*, *supernova*, *AGN*, and *variable star* (Figure 2). This structure evaluates a model's ability to reject artifacts and classify astronomical objects. We construct a balanced evaluation set of 1,500 examples (300 per class) to ensure that benchmark scores reflect multimodal reasoning rather than majority-class priors, as the natural ZTF alert stream is highly imbalanced (Rehemtulla et al., 2024). Details in dataset construction are in Appendix B.1.

4 The AstroAlertBench Discovery Chain and Results

4.1 Protocol Overview

AstroAlertBench evaluates model performance through a three-stage discovery chain designed to isolate failures in data perception from errors in high-level reasoning: **Metadata Grounding** (Perception), **Scientific Rationale** (Reasoning), and **Staged Classification** (Decision). We benchmark 13 model configurations across various architectures and reasoning modes; specific model configurations and hyperparameters are detailed in Appendix A.2. We evaluate all models without fine-tuning on the benchmark. We also initialize a Zooniverse project[2] as a precursor to future community-scale human-in-the-loop participation.

4.2 Part A: Metadata Grounding

Part A evaluates metadata perception to ensure that reasoning failures are not caused by grounding errors. Models must map raw numerical parameters to their scientific definitions using a mapping key provided in the prompt. Full details of these definitions and the prompt template are provided in Appendix B.2.1. For example, given the raw metadata `fid: 1`, the model must cross-reference the provided information table to correctly translate the numerical filter ID (`fid`) into its scientific name, mapping `1` to the *g*-band filter (Patterson et al., 2019). We evaluate this interpretive mapping across six objective fields: `fid` (filter ID), `isdiffpos` (difference-residual sign), `magpsf` (PSF magnitude), `sigmapsf` (magnitude error), `ndethist` (historical detections), and `ncovhist` (historical coverages). Success is measured by a match with the ground-truth alert packet. **All 13 evaluated model configurations achieved 100% accuracy on Part A across all six metadata fields**, confirming correct performance on the tested metadata mappings. The full per-field accuracy results for each model are reported in Appendix C.1.

4.3 Part B: Scientific Rationale

In Part B, models must produce a structured scientific justification across three dimensions: **key evidence**, **leading interpretation**, and **alternative analysis**. Following each rationale, the model assigns itself a quality score (0–5) for each dimension based on the rubric in Stoppa et al. (2025). We summarize these responses using the *mean self-reasoning score* and the *self pass rate* (the proportion of alerts averaging ≥ 4 on the rubric).

Detailed results for all 13 runs are provided in Appendix C.2. `Claude Opus 4.7 think` achieves the highest end-to-end five-class accuracy (60.60%), while both `Opus 4.7` configurations report the lowest mean rationale self-ratings (both 3.99). In comparison, `Gemini 2.5 Pro` reports the highest mean self-rating among all 13 configurations (4.89), despite its lower classification accuracy (41.93%). Section 5 examines the association between these reasoning-quality ratings and classification correctness. A complementary evaluation of classification confidence is provided in Appendix E.6.

4.4 Part C: Staged Classification

Part C decomposes the five-class decision into a hierarchical cascade: **Stage-1** distinguishes instrumental *artifacts* from *real_objects*, **Stage-2** routes real objects to *solar_system* or *astrophysical*,

[2] https://www.zooniverse.org/projects/AstroAlert

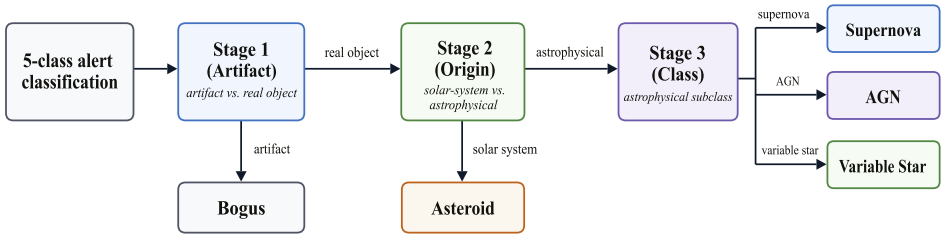

Figure 3: **Hierarchical classification cascade.** The five benchmark categories are mapped through three sequential decision stages: artifact rejection (Stage-1), physical origin (Stage-2), and astrophysical subclass assignment (Stage-3).

and **Stage-3** assigns the specific astrophysical subclass among *supernova*, *variable_star*, and *AGN*. Figure 3 maps the five benchmark classes to this cascade. We measure per-stage accuracy, conditional accuracy at each decision point (restricted to where the previous stages were correct), and end-to-end 5-class accuracy. Table 1 shows the stage-wise results; Figure 4 shows the per-class accuracy heatmap.

#	Run	Stage-1	Stage-2	Stage-3	Stage-3 cond.	5-class (E2E)	S3 macro-F1
1	Claude Opus 4.7 think	$\underline{86.60 \pm 0.88}$	$\mathbf{81.60 \pm 1.00}$	$\mathbf{66.80 \pm 1.22}$	$\mathbf{60.78 \pm 1.26}$	$\mathbf{60.60 \pm 1.26}$	**0.558**
2	GPT-5.4 high-think	83.93 ± 0.95	$\underline{80.00 \pm 1.03}$	59.40 ± 1.27	47.11 ± 1.29	$\underline{51.07 \pm 1.29}$	0.435
3	Kimi K2.5 think	81.36 ± 1.01	71.94 ± 1.16	$\underline{60.19 \pm 1.27}$	53.45 ± 1.29	49.33 ± 1.29	0.500
4	Claude Opus 4.7 nothink	85.20 ± 0.92	70.93 ± 1.17	54.47 ± 1.29	$\underline{60.44 \pm 1.26}$	48.87 ± 1.29	$\underline{0.548}$
5	Qwen3.5-397B-A17B think	80.87 ± 1.02	72.80 ± 1.15	52.80 ± 1.29	32.44 ± 1.21	44.27 ± 1.28	0.310
6	GPT-5.4 no-think	77.47 ± 1.08	71.53 ± 1.17	57.00 ± 1.28	37.89 ± 1.25	43.67 ± 1.28	0.329
7	Gemini 2.5 Pro high-think	82.27 ± 0.99	63.60 ± 1.24	45.53 ± 1.29	52.33 ± 1.29	41.93 ± 1.27	0.509
8	Gemini 2.5 Flash no-think	83.47 ± 0.96	63.73 ± 1.24	37.80 ± 1.25	52.00 ± 1.29	36.27 ± 1.24	0.456
9	Qwen3.5-397B-A17B nothink	79.17 ± 1.05	60.95 ± 1.26	40.39 ± 1.27	45.61 ± 1.29	34.93 ± 1.23	0.430
10	Qwen3.5-35B-A3B think	77.15 ± 1.35	62.98 ± 1.55	50.98 ± 1.61	32.12 ± 1.50	26.73 ± 1.14	0.271
11	Qwen3.5-35B-A3B nothink	80.40 ± 1.03	60.88 ± 1.27	31.65 ± 1.21	37.51 ± 1.26	25.50 ± 1.13	0.273
12	Qwen3.5-4B nothink	79.30 ± 1.10	54.21 ± 1.35	28.68 ± 1.22	38.67 ± 1.32	22.32 ± 1.08	0.307
13	Qwen3.5-4B think	$\mathbf{87.07 \pm 1.88}$	78.86 ± 2.29	42.59 ± 2.78	45.60 ± 2.80	8.41 ± 0.72	0.241

Table 1: **Part C staged classification performance.** Metrics include artifact rejection (Stage-1), origin identification (Stage-2), subclass assignment (Stage-3), conditional subclass accuracy (Stage-3 cond.), joint end-to-end accuracy (5-class), and macro-F1 for astrophysical subclasses (S3 macro-F1). **Bold** and underline denote the first and second best performance in each column, respectively. Results are reported over $1,500$ alerts.

Five-class classification performance. As shown in Table 1, `Claude Opus 4.7 think` leads the benchmark with $60.60 \pm 1.26\%$ end-to-end 5-class accuracy, representing a significant 9.53 percentage point lead over the second-ranked `GPT-5.4 high-think`. The three highest-ranking models—`Opus 4.7 think`, `GPT-5.4 high-think`, and `Kimi K2.5 think`—all utilize reasoning-enabled modes. Notably, `Kimi K2.5 think` achieves the highest accuracy among the open-weight models, outperforming several frontier closed-source configurations.

Staged decision results. **Stage-1** (real-vs-artifact) exhibits consistent high performance across all model scales, with an accuracy spread of 77.15–87.07%. **Stage-2** (physical origin) begins to separate the closed-source frontier and reasoning-enabled open-weight models from standard configurations. Leading results from `Claude Opus 4.7 think` (81.60%) and `GPT-5.4 high-think` (80.00%) are joined by strong open-weight performances, such as `Kimi K2.5 think` (71.94%). **Stage-3 conditional accuracy** reveals the most significant performance gap. Its spread of 32.12–60.78% is among the largest in Table 1, indicating that while most models effectively clear the initial artifact filter, few maintain high accuracy in fine-grained astrophysical categorization.

Class-level failure modes. Figure 4 reveals three universal patterns: (i) **AGN-VS Confusion**: AGN accuracy remains below 8% for all models, with sources overwhelmingly misclassified as variable stars. Beyond image cutouts, the alert metadata provide physical evidence to distinguish AGNs from variable stars through spatial and host context, g, r, i, z photometry, and detection

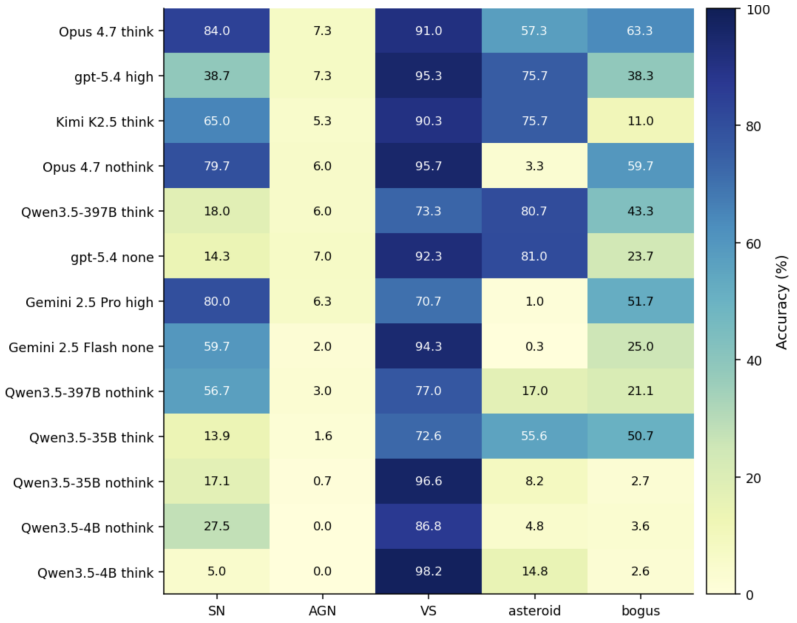

Figure 4: **Per-class accuracy heatmap across 13 runs.** Rows are the 13 evaluated model configurations, sorted top-to-bottom by absolute 5-class accuracy; columns are the five benchmark classes (*supernova*, *AGN*, *variable star*, *asteroid*, *bogus*).

history. ALeRCE's high stamp-classification accuracy suggests that these inputs contain sufficient discriminative information, so the observed AGN collapse primarily reflects a limitation of the evaluated MLLMs (Carrasco-Davis et al., 2021; Förster et al., 2021). (ii) **Gemini's asteroid collapse**: The `Gemini 2.5` family exhibits a localized blind spot on asteroids ($\leq 1.0\%$), despite `2.5 Pro` otherwise being competitive on the supernova class at 80.0%. (iii) **Qwen bogus collapse**: Most smaller `Qwen` variants fail to identify the *bogus* class ($< 4\%$), driven by a strong majority-class bias in Stage-1 that predicts *real_object* for nearly all inputs.

Impact of reasoning. To quantify the impact of reasoning, we compare model performance across standard and thinking-enabled configurations, measuring the accuracy improvement provided by internal chain-of-thought. For the three strongest models, enabling adaptive or high-thought reasoning yields significant accuracy gains: `Claude Opus 4.7` ($+11.73\%$), `Qwen3.5-397B-A17B` ($+9.34$ percentage points), and `GPT-5.4` ($+7.40\%$). This "think-win" margin scales with overall model capacity, whereas the mid-sized `Qwen3.5-35B-A3B` result is a statistical tie ($+1.23\%$). Notably, the smallest `Qwen3.5-4B` exhibits a performance regression when forced into a thinking mode. Detailed statistical comparisons, including z-scores and standard errors for all pairs, are provided in Appendix D.

5 Evaluating Model Self-Assessment and Scientific Reliability

While high accuracy is a primary benchmark objective, the practical utility of an LLM as a scientific assistant also depends on its calibration. In high-volume fields like time-domain astronomy, where follow-up observations are a scarce resource, a system that cannot accurately signal its own uncertainty poses a significant risk to efficient triage. We therefore evaluate model self-assessment and self-correction through three distinct lenses: the **macro** view (§5.1) investigating associations between average rationale self-ratings and classification accuracy, the **inner** view (§5.2) assessing per-alert error discrimination, and the **behavioral** view (§5.3) testing second-pass error recovery. A complementary experiment (Appendix E.6) evaluates classification confidence using explicitly elicited probabilities. These findings are supplemented by qualitative analysis and expert grading of a representative alert in Appendix G.

5.1 POPULATION-LEVEL TRENDS: PERFORMANCE VERSUS SELF-RATINGS

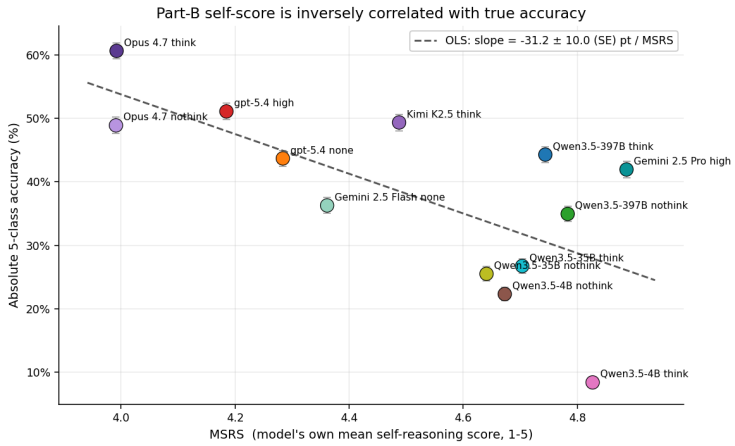

Figure 5: **Inverse correlation between self-assessment and accuracy.** Each point represents one of the 13 evaluated runs ($n = 1,500$). The dashed line indicates an OLS fit with a slope of -31.2 ± 10.0 percentage points of accuracy per unit on the 0–5 self-reasoning scale. Higher empirical accuracy is associated with lower mean rationale self-ratings across these configurations.

Figure 5 reveals a counterintuitive population-level trend: models with higher classification accuracy tend to report lower mean rationale self-ratings. An ordinary least-squares fit through the 13 configurations yields a negative slope of -31.2 ± 10.0 percentage points of accuracy per unit on the 0–5 self-reasoning scale. The two `Claude Opus 4.7` runs occupy the upper-left of the plane, with the thinking configuration achieving the highest accuracy in the batch (60.60%) and both configurations reporting the lowest mean self-ratings (≈ 4.0). Conversely, the smaller `Qwen3.5` model configurations cluster in the lower-right, reporting high mean self-ratings (≥ 4.6) despite lower classification accuracy.

5.2 INSTANCE-LEVEL ANALYSIS: DISTINGUISHING CORRECT CLASSIFICATIONS FROM ERRORS

While more accurate models tend to report lower self-ratings on average, a self-assessment score is only useful for alert triage if it can reliably distinguish correct classifications from errors within the same run. We evaluate this per-alert reliability primarily using AUROC, which measures how often a correct classification receives a higher self-rating than an incorrect one. A value of 0.5 indicates no better separation than random guessing, while larger values indicate better separation. Detailed results for all 13 configurations, including uncertainty estimates, permutation tests, and auxiliary metrics such as Pearson correlation and score-group accuracies, are provided in Appendix E.

Only a handful of models exhibit an appreciable instance-level link between self-ratings and performance, and even these show only modest separation between correct classifications and errors. `Claude Opus 4.7 nothink` leads the cohort with an AUROC of 0.622, indicating that its correct classifications tend to receive higher self-scores than its errors. Both `GPT-5.4` configurations and `Gemini 2.5 Flash no-think` show a similar pattern, with AUROCs between 0.602 and 0.616. In contrast, the remaining runs, including the `Qwen3.5` think variants and `Gemini 2.5 Pro high-think`, show weaker ordering of correct and incorrect predictions (AUROC 0.498–0.539). Six of the 13 configurations have AUROCs above 0.5 with $p < 0.05$ in permutation tests, without adjustment for multiple comparisons; two of these are Qwen runs whose AUROCs remain close to 0.5. `Opus 4.7` with adaptive thinking achieves higher classification accuracy but has a lower AUROC for its self-ratings (0.515) than without thinking (0.622). A complementary experiment with elicited classification confidence yields calibration rankings closely matching those from normalized rationale self-ratings, with full results provided in Appendix E.6.

5.3 BEHAVIORAL: SECOND-ROLLOUT RETRY ON ALERTS WITH LOW SELF-RATINGS

We test this in an exploratory retry study: for each of five closed-source configurations, we draw a stratified subset of 35 eligible alerts with low first-pass rationale self-ratings (mean Part B score below 4). We use 35 alerts per configuration because `GPT-5.4 no-think` has only 36 alerts with mean self-ratings below 4. We then re-prompt the model with its own first-pass answer attached, and ask for a second pass without revealing the ground-truth classification. Full prompts and results are in Appendix B.3 and Appendix F.

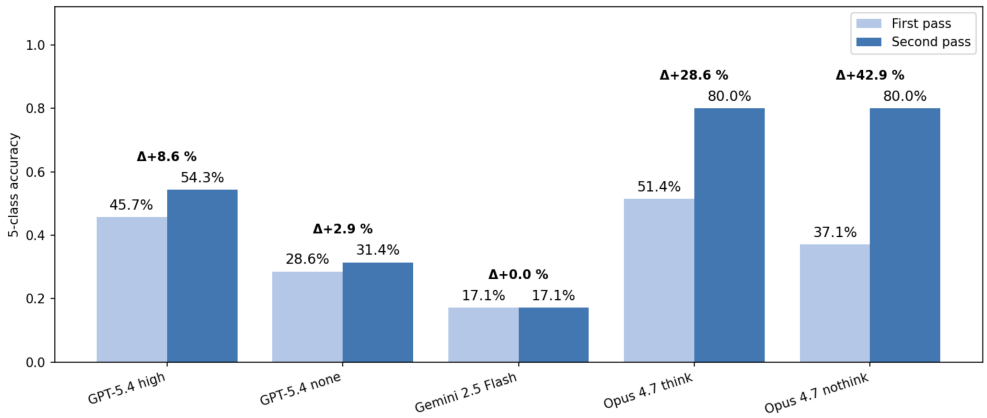

Figure 6: **Second-rollout retry on the subset with low rationale self-ratings.** Comparison of first-pass and second-pass accuracy over the same 35 alerts per model, drawn from each model's eligible pool with a mean rationale self-rating below 4. Gains (Δ) are reported in absolute percentage points.

Figure 6 shows three behavioral patterns in these selected subsets. `Claude Opus 4.7` (in both reasoning modes) achieves substantial accuracy gains on the retry, reaching 80% without breaking a single answer that was already correct on the first pass. In contrast, `GPT-5.4 high-think` exhibits a trade-off between repairs and regressions: while it corrects some initial errors, it simultaneously breaks nearly one in five of the answers it previously got right. The `GPT-5.4 no-think` configuration improves accuracy by only 2.86 percentage points. Finally, `Gemini 2.5 Flash` assigns itself higher second-pass self-ratings without changing any classification outcome. This illustrates that **higher rationale self-ratings do not necessarily accompany improved classification results**.

6 DISCUSSION AND CONCLUSION

In this work, we introduced *AstroAlertBench*, a multimodal benchmark for evaluating Large Language Models in the specialized domain of time-domain astronomy. Our three-stage evaluation shows that while frontier models excel at metadata perception, classification errors remain common, and rationale self-ratings vary in their ability to discriminate correct from incorrect predictions. We identify difficulties separating active galactic nuclei from variable stars, together with family-specific failures such as the Gemini configurations' asteroid predictions. These results emphasize that high classification accuracy alone does not guarantee reliable self-assessment. AstroAlertBench provides a standardized framework for evaluating the scientific reasoning and reliability of multimodal LLMs in astronomy.

Our findings suggest several promising directions for developing autonomous agents. The difficulty in classifying AGNs indicates that future research should prioritize incorporating time-series data, such as light curves, to provide additional temporal context. Moreover, the successful self-correction observed in frontier models suggests that these systems could transition from passive observers to active participants in astronomical alert triage and follow-up planning. By serving as autonomous coordinators, such agents could synthesize multi-survey data to prioritize high-value transients and automate the drafting of follow-up requests. Finally, integrating tool-use to query external catalogs may eventually facilitate autonomous hypothesis verification against known astronomical records.

AI USE STATEMENT

LLMs were used as models under evaluation to generate responses for the benchmark experiments described in this paper. Separately, generative AI tools were used only for grammar checking and proofreading of the manuscript. Generative AI was not used to develop the research methodology, formulate hypotheses or claims, implement the methods, process the dataset, or analyze or interpret the results. All AI-assisted edits were reviewed by the authors, who take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The dataset and evaluation code are available via the links in Section 1. Appendices A and B document model configurations, dataset construction, and prompts; Appendices C–F provide metric definitions and additional evaluation procedures.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *ArXiv Preprint*, 2023. 1, 3

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1, 3

Anthropic. Claude opus 4.7 system card. https://www.anthropic.com/system-cards, April 2026. 16, 17

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 1

Eric C Bellm, Shrinivas R Kulkarni, Matthew J Graham, Richard Dekany, Roger M Smith, Reed Riddle, Frank J Masci, George Helou, Thomas A Prince, Scott M Adams, et al. The zwicky transient facility: system overview, performance, and first results. *Publications of the Astronomical Society of the Pacific*, 131(995):018002, 2019. 2, 3, 17

Daniil A Boiko, Robert MacKnight, and Gabe Gomes. Emergent autonomous scientific research capabilities of large language models. *arXiv preprint arXiv:2304.05332*, 2023. 1, 3

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023. 1, 3

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

Rodrigo Carrasco-Davis, Esteban Reyes, Camilo Valenzuela, Francisco Förster, Pablo A Estévez, Giuliano Pignata, Franz E Bauer, Ignacio Reyes, Paula Sánchez-Sáez, Guillermo Cabrera-Vives, et al. Alert classification for the alerce broker system: the real-time stamp classifier. *The Astronomical Journal*, 162(6):231, 2021. 3, 4, 5, 7

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 16, 17

Dmitry A Duev, Ashish Mahabal, Frank J Masci, Matthew J Graham, Ben Rusholme, Richard Walters, Ishani Karmarkar, Sara Frederick, Mansi M Kasliwal, Umaa Rebbapragada, et al. Real-bogus classification for the zwicky transient facility using deep learning. *Monthly Notices of the Royal Astronomical Society*, 489(3):3582–3590, 2019. 2, 3

Francisco Förster, Guillermo Cabrera-Vives, Ernesto Castillo-Navarrete, PA Estévez, P Sánchez-Sáez, J Arredondo, FE Bauer, R Carrasco-Davis, M Catelan, F Elorrieta, et al. The automatic learning for the rapid classification of events (alerce) alert broker. *The Astronomical Journal*, 161(5):242, 2021. 2, 3, 7, 17, 18

C Fremling, AA Miller, Y Sharma, Alison Dugas, DA Perley, K Taggart, Jesper Sollerman, Ariel Goobar, ML Graham, JD Neill, et al. The zwicky transient facility bright transient survey. i. spectroscopic classification and the redshift completeness of local galaxy catalogs. *The Astrophysical Journal*, 895(1):32, 2020. 18

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in 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. 6577–6595, 2024. 3

Matthew J Graham, SR Kulkarni, Eric C Bellm, Scott M Adams, Cristina Barbarino, Nadejda Blagorodnova, Dennis Bodewits, Bryce Bolin, Patrick R Brady, S Bradley Cenko, et al. The zwicky transient facility: science objectives. *Publications of the Astronomical Society of the Pacific*, 131(1001):078001, 2019. 2, 3

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017. 2, 35

Kehan Guo, Bozhao Nan, Yujun Zhou, Taicheng Guo, Zhichun Guo, Mihir Surve, Zhenwen Liang, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Can llms solve molecule puzzles? a multimodal benchmark for molecular structure elucidation. *Advances in Neural Information Processing Systems*, 37:134721–134746, 2024. 1, 2, 3, 16

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. 2, 3

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026. 16

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023. 1, 3

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The twelfth international conference on learning representations*, 2023. 2

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1, 3

Yujie Liu, Zonglin Yang, Tong Xie, Jinjie Ni, Ben Gao, Yuqiang Li, Shixiang Tang, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. Researchbench: Benchmarking llms in scientific discovery via inspiration-based task decomposition. *arXiv preprint arXiv:2503.21248*, 2025. 3

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 3

Ashish Mahabal, Umaa Rebbapragada, Richard Walters, Frank J Masci, Nadejda Blagorodnova, Jan van Roestel, Quan-Zhi Ye, Rahul Biswas, Kevin Burdge, Chan-Kao Chang, et al. Machine learning for the zwicky transient facility. *Publications of the Astronomical Society of the Pacific*, 131(997):038002, 2019. 2, 3

Frank J Masci, Russ R Laher, Ben Rusholme, David L Shupe, Steven Groom, Jason Surace, Edward Jackson, Serge Monkewitz, Ron Beck, David Flynn, et al. The zwicky transient facility: Data processing, products, and archive. *Publications of the Astronomical Society of the Pacific*, 131 (995):018003, 2019. 2, 3, 17, 18

OpenAI. Gpt-5.4 thinking system card. https://openai.com/index/gpt-5-4-thinking-system-card/, March 2026. Accessed May 1, 2026. 16, 17

Rui Pan, Tuan Dung Nguyen, Hardik Arora, Alberto Accomazzi, Tirthankar Ghosal, and Yuan-Sen Ting. Astromlab 2: Astrollama-2-70b model and benchmarking specialised llms for astronomy. In *SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 87–96. IEEE, 2024. 1, 2, 3

Maria T Patterson, Eric C Bellm, Ben Rusholme, Frank J Masci, Mario Juric, K Simon Krughoff, V Zach Golkhou, Matthew J Graham, Shrinivas R Kulkarni, George Helou, et al. The zwicky transient facility alert distribution system. *Publications of the Astronomical Society of the Pacific*, 131(995):018001, 2019. 3, 4, 5

Qwen Team. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, February 2026. Accessed May 1, 2026. 16

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021. 1, 3

Nabeel Rehemtulla, Adam A Miller, Theophile Jegou Du Laz, Michael W Coughlin, Christoffer Fremling, Daniel A Perley, Yu-Jing Qin, Jesper Sollerman, Ashish A Mahabal, Russ R Laher, et al. The zwicky transient facility bright transient survey. iii. btsbot: automated identification and follow-up of bright transients with deep learning. *The Astrophysical Journal*, 972(1):7, 2024. 3, 5, 18

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024. 1, 3

P Sánchez-Sáez, I Reyes, C Valenzuela, F Förster, S Eyheramendy, F Elorrieta, FE Bauer, G Cabrera-Vives, PA Estévez, Márcio Catelan, et al. Alert classification for the alerce broker system: The light curve classifier. *The Astronomical Journal*, 161(3):141, 2021. 18

Fiorenzo Stoppa, Turan Bulmus, Steven Bloemen, Stephen J Smartt, Paul J Groot, Paul Vreeswijk, and Ken W Smith. Textual interpretation of transient image classifications from large language models. *Nature Astronomy*, 9(12):1869–1878, 2025. 2, 3, 5, 26

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022. 1, 3

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 1, 3

Thinking Machines Lab. Tinker. https://thinkingmachines.ai/tinker/, 2025. Accessed: 2026-05-05. 17

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 1

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022. 1

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 1, 3

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024. 3, 16

CONTENTS

1 Introduction **1**

2 Related Work **3**

2.1 Astronomical Alert Triage and Broker Systems 3

2.2 Multimodal Reasoning and Scientific Benchmarks 3

3 The AstroAlertBench Dataset **4**

3.1 Data Acquisition and Multimodal Content 4

3.2 Automated Image Preprocessing and Serialization 4

3.3 Benchmark Taxonomy and Evaluation Set Construction 5

4 The AstroAlertBench Discovery Chain and Results **5**

4.1 Protocol Overview 5

4.2 Part A: Metadata Grounding 5

4.3 Part B: Scientific Rationale 5

4.4 Part C: Staged Classification 5

5 Evaluating Model Self-Assessment and Scientific Reliability **7**

5.1 Population-level trends: Performance versus self-ratings 8

5.2 Instance-level analysis: Distinguishing correct classifications from errors 8

5.3 Behavioral: second-rollout retry on alerts with low self-ratings 9

6 Discussion and Conclusion **9**

A Implementation Details and Resource Analysis **15**

A.1 Benchmark Implementations 15

A.2 Model Configurations 16

A.3 Compute Resources Specification 17

B Dataset and Prompt Construction **17**

B.1 Dataset 17

B.1.1 Label provenance and astronomical practice 18

B.2 Prompt 18

B.2.1 Main Prompt 18

B.3 Self-Correction Prompt 23

C Comprehensive Benchmark Results: Metric Definitions and Per-Part Findings **25**

C.1 Part A: Metadata Grounding 26

C.2 Part B: Scientific Rationale 26

C.3 Part C: Staged Classification 27

D Reasoning-mode analysis: think vs. no-think **31**

E Rationale Self-Assessment and Classification Confidence **32**

E.1 Definitions and descriptive association measures 32

E.2 Per-model score–accuracy associations 32

E.3 Classification accuracy by self-rating group 33

E.4 Relationship between score gap and correlation 33

E.5 Additional metrics for rationale self-assessment 34

E.6 Classification-confidence ablation 36

F Self-Correction and Behavioral Reliability **37**

F.1 Methodology 37

F.2 Auxiliary figures 38

F.3 Per-family interpretation 39

F.4 Qualitative Case Studies of Second-Rollout Behavior 40

F.4.1 Evidence Re-weighting and Successful Self-Correction 40

F.4.2 Static Persistence and Score Inflation 43

G Human-Model Grading Comparison: Representative Case Studies **48**

G.1 Expert grading of leading and alternative interpretations 48

G.2 Agreement between self-ratings and expert grades on one alert 49

G.3 Three representative model responses 50

G.3.1 Tier 1 — `Claude Opus 4.7 think` 50

G.3.2 Tier 2 — `Qwen3.5-35B nothink` 51

G.3.3 Tier 3 — `Gemini 2.5 Flash` 52

A Implementation Details and Resource Analysis

This appendix provides the technical specifics for the 13 evaluated model configurations. We detail the hyperparameters used and a comprehensive breakdown of the computational time associated with the 1,500-alert benchmark rollout.

A.1 Benchmark Implementations

We set temperature to `0.2` when the API accepts it. This is because we believe that our tasks should have fixed answers and thus require a more deterministic and consistent output. OpenAI's GPT-5.4 and Claude Opus 4.7 omit `temperature`; Google and Tinker use `0.2`. No top-p or frequency-penalty tuning on this path. The maximum allowed output tokens is set to 2048, and the limit is raised to 20 000 for reasoning models so traces can finish before the JSON.

Setting	Value
`temperature`	0.2
`max_tokens` (non-reasoning)	2 048
`max_tokens` (reasoning)	20 000
`num_samples`	1
Image input	1

Table 2: Primary decoding parameters (released inference code).

A.2 Model Configurations

Following the convention of assessing language models' out-of-the-box capabilities (Guo et al., 2024; Yue et al., 2024), we evaluate models without fine-tuning on the benchmark. The aim is to diagnose existing capability gaps in generalist models and inform future domain adaptation. To the best of our knowledge, there is no publicly runnable astronomy-broker classifier covering all five benchmark classes for direct evaluation on our inputs.

Tables 3 and 4 provide the technical specifications and primary references for the models evaluated in our benchmarks.

#	Configuration Name	Params	Reference
1	Claude Opus 4.7 think	Proprietary	(Anthropic, 2026)
2	Claude Opus 4.7 nothink	Proprietary	(Anthropic, 2026)
3	GPT-5.4 high-think	Proprietary	(OpenAI, 2026)
4	GPT-5.4 no-think	Proprietary	(OpenAI, 2026)
5	Gemini 2.5 Pro high-think	Proprietary	(Comanici et al., 2025)
6	Gemini 2.5 Flash no-think	Proprietary	(Comanici et al., 2025)
7	Kimi K2.5 think	1.1T (32B active)	(Kimi Team et al., 2026)
8	Qwen3.5-397B-A17B think	397B (17B active)	(Qwen Team, 2026)
9	Qwen3.5-397B-A17B nothink	397B (17B active)	(Qwen Team, 2026)
10	Qwen3.5-35B-A3B think	35B (3B active)	(Qwen Team, 2026)
11	Qwen3.5-35B-A3B nothink	35B (3B active)	(Qwen Team, 2026)
12	Qwen3.5-4B think	4B (Dense)	(Qwen Team, 2026)
13	Qwen3.5-4B nothink	4B (Dense)	(Qwen Team, 2026)

Table 3: Specifications for the 13 evaluated model configurations. Models are grouped by family, with reasoning-enabled (think) configurations listed prior to standard (no-think) counterparts. Parameter counts for Mixture-of-Experts (MoE) architectures specify total and active parameters per token.

Run (short)	Model id	Reasoning control	Backend	Conc.
Kimi K2.5	`kimi-k2.5`	Default thinking renderer	Tinker	32
Qwen3.5-397B think	`qwen3.5-397b-a17b`	Thinking on (default renderer)	Tinker	32
Qwen3.5-397B nothink	`qwen3.5-397b-a17b`	`--thinkingdisabled`	Tinker	32
Qwen3.5-4B think	`qwen3.5-4b`	Thinking on	Tinker	32
Qwen3.5-4B nothink	`qwen3.5-4b`	`--thinkingdisabled`	Tinker	32
Qwen3.5-35B think	`qwen3.5-35b-a3b`	Thinking on	Tinker	32
Qwen3.5-35B nothink	`qwen3.5-35b-a3b`	`--thinkingdisabled`	Tinker	32
GPT-5.4 high	`gpt-5.4`	`reasoning_effort=high`	OpenAI	16
GPT-5.4 none	`gpt-5.4`	`reasoning_effort=none`	OpenAI	16
Claude Opus 4.7 think	`claude-opus-4-7`	Adaptive thinking, `effort=high`	Anthropic	2
Claude Opus 4.7 nothink	`claude-opus-4-7`	`reasoning_effort=none`	Anthropic	2
Gemini 2.5 Pro high	`gemini-2.5-pro`	`thinking_budget` $= -1$ (dynamic)	Google	8
Gemini 2.5 Flash none	`gemini-2.5-flash`	`thinking_budget` $= 0$ (off)	Google	8

Table 4: Benchmark run tested LLM list. The Conc. column denotes the concurrency level, representing the number of simultaneous requests issued to the respective backend during the evaluation.

A.3 Compute Resources Specification

We evaluated open-source LLMs through Tinker (Thinking Machines Lab, 2025). For closed-source models, including Claude Opus 4.7, GPT-5.4, and Gemini 2.5, we used the corresponding provider APIs (Anthropic, 2026; OpenAI, 2026; Comanici et al., 2025). T_{wall} is the runtime of the benchmark dataset, s/row stands for $T_{\text{wall}}/1,500$, and $\bar{N}_{\text{out}}$ is the mean output tokens per row. $N_{\max}$ is the max tokens per row.

Run (short)	Conc.	T_{wall} (s)	s/row	$\bar{N}_{\text{out}}$	$N_{\max}$
Kimi K2.5	32	8 013	5.34	3 943	10 228
Qwen3.5-397B think	32	20 239	13.49	5 171	16 882
Qwen3.5-397B nothink	32	4 982.5	3.32	554	2 048
Qwen3.5-4B think	32	22 638	15.09	16 623	20 000
Qwen3.5-4B nothink	32	2 700	1.80	583	2 048
Qwen3.5-35B think	32	38 543	25.70	11 822	20 000
Qwen3.5-35B nothink	32	3 412.5	2.28	539	2 048
GPT-5.4 high	16	5 145.1	3.43	2 557	8 156
GPT-5.4 none	16	929.5	0.62	446	521
Claude Opus 4.7 think	2	10 791	7.19	806	1 487
Claude Opus 4.7 nothink	2	13 398	8.93	652	818
Gemini 2.5 Pro high	8	7 723	5.15	2 374	5 022
Gemini 2.5 Flash none	8	1 021	0.68	635	1 104

Table 5: Key Information in 1,500-Datapoint Benchmark Run. The Conc. column denotes the concurrency level, representing the number of simultaneous requests issued to the respective backend during the evaluation.

B Dataset and Prompt Construction

B.1 Dataset

All astronomical data and metadata utilized in *AstroAlertBench* are derived from public scientific survey streams. We specify the following licensing and attribution details for the foundational assets:

- **ZTF Public Alert Stream:** The raw alert data, including the image cutouts and numerical metadata fields, are produced by the Zwicky Transient Facility (ZTF). These assets are distributed by the NASA/IPAC Infrared Science Archive (IRSA) under the NASA Data and Information Policy, which mandates full and open access to scientific data. Consistent with the Public Domain status of NASA-archived data, we provide proper credit by citing the foundational survey and pipeline documentation (Bellm et al., 2019; Masci et al., 2019).
- **ALeRCE Broker Products:** The class-specific reference labels described in Appendix B.1.1 and broker-processed features are provided by the Automatic Learning for the Rapid Classification of Events (ALeRCE) system. These data products and the associated broker client are released under the **MIT License**. We credit the original creators by citing the foundational ALeRCE broker documentation (Förster et al., 2021).

Each benchmark example starts from the public ALeRCE broker over ZTF: we query object- and detection-level metadata through their API (`https://api.alerce.online/ztf/v1`), download the three stamp cutouts per alert as FITS files—*science* (discovery image), *reference* (template), and *difference* (subtraction)—and then render into one RGB montage (three images left to right with labels above each panel in a single PNG) so VLM can interpret. For our benchmark run we use a dataset of 1,500 alerts with 300 per gold class, pairing that metadata with the montages. Our second-rollout ablation study ($n = 35$ per model, low Part B self-score pool) uses the same decoding defaults.

We use ALeRCE's stamp-classifier products to assemble the alert sample; the reference labels have the class-specific provenance described in Appendix B.1.1. Table 6 summarizes the distribution of the broker-assigned class probability (`probability` in `data/manifest_benchmark_fina`

`l.csv`) within each reference class on the 1 500 benchmark rows. Medians sit in a high-probability band overall ($\sim$0.87–1), with several classes near or above $\sim$0.9, with all asteroid entries fixed at 1.0 because they are verified through MPC orbital matches.

Gold class	min	max	median
AGN	0.859521	0.898750	0.865981
SN	0.902436	0.953128	0.913197
VS	0.945519	0.966328	0.949354
asteroid	1.000000	1.000000	1.000000
bogus	0.953738	0.977276	0.956961

Table 6: ALeRCE stamp-classifier probability (`probability`) on the 1 500 benchmark rows: within each gold `target_class`, minimum, maximum, and median over $n = 300$.

B.1.1 Label provenance and astronomical practice

Combining catalog cross-matches with high-confidence survey-broker filters is a standard operating paradigm in survey astronomy (Masci et al., 2019; Förster et al., 2021). AstroAlertBench follows this established practice for rapid first-alert triage, with label provenance specified for each class. All 300 *asteroid* alerts are orbit-matched against the International Astronomical Union's Minor Planet Center (MPC) database, the standard reference for Solar System object identification (Masci et al., 2019; Förster et al., 2021). The 600 *AGN* and *variable-star* alerts (300 per class) are cross-verified across Pan-STARRS1, Gaia DR3, and AllWISE: their identities are anchored by star/galaxy cross-matches, stellar-motion information, and infrared signatures (Sánchez-Sáez et al., 2021). The 300 *bogus* alerts are verified nonphysical subtraction or optical artifacts, such as edge effects and cosmic rays, identified through ZTF pipeline quality checks (Masci et al., 2019). The 300 *supernova* targets are sourced from ALeRCE's high-confidence production stream; these labels are broker-derived because newly exploding transients lack pre-existing catalog entries for the transient itself (Förster et al., 2021). These class-specific sources combine catalog verification, artifact checks, and high-confidence broker predictions in the operational setting for which the benchmark is designed.

We do not require spectroscopic confirmation for every benchmark label, because a uniform spectroscopic requirement would be inappropriate and introduce severe selection bias into the first-alert evaluation sample. Fewer than 0.01% of raw alerts in nightly survey streams receive spectroscopic follow-up because telescope resources are severely limited (Fremling et al., 2020; Rehemtulla et al., 2024). Moreover, follow-up preferentially targets bright, easily accessible, low-redshift sources; requiring spectroscopic labels across all five classes would distort the benchmark's class balance and restrict evaluation to a biased subset of the alert stream. Such a requirement is also inapplicable to nonphysical bogus artifacts, while asteroid identities are established through astrometric orbital matches. In addition, raw alert packets tied to targeted spectroscopic follow-up programs are often restricted or subject to proprietary embargoes; such restrictions make those data unavailable for open benchmark compilation. Within this benchmark's scope of first-alert triage, this combination of catalog verification, artifact checks, and high-confidence broker labels provides the best available operational gold standard under real survey constraints.

B.2 Prompt

B.2.1 Main Prompt

General benchmark prompt

```
=== SYSTEM ===

You are an experienced astrophysicist. Your task is to classify
    astronomical transient candidates using three image cutouts and
    associated metadata.
```

```
The montage is labeled left-to-right on the PNG as Science, Reference,
     and Difference. Reference is the coadded baseline image;
    Difference is the subtraction image (science minus reference).

Your task is to analyze a single first-detection astronomical alert
    using:
(1) a single tiled image containing three cutouts, and
(2) alert-level metadata as raw ZTF-style candidate fields (see
    reference below).

You must classify the alert using only the provided evidence.
Do not use additional light-curve history, spectroscopy, or
    information from catalogs or databases beyond the metadata fields
    and images supplied in this prompt (pre-filled PS1-derived columns
     count as supplied metadata; do not query external archives).
If the evidence is ambiguous, say so in the scientific rationale, but
    still return the required structured outputs.

Goal:
Determine whether the alert is most consistent with one of the
    following
five classes:
- Supernova
- Variable Star
- AGN
- Asteroid
- Bogus

In this benchmark, "Variable Star" means Galactic (stellar) variable
    candidates as a class label; "AGN" means active galactic nucleus
    variability - both can vary in nature, but the two labels are
    distinct here.

Important image interpretation guide:
- The input image consists of three 63 x 63 pixel cutouts tiled
    horizontally: Science (left), Reference (middle; coadded baseline)
    , Difference (right; subtraction). Top labels on the montage read
    Science, Reference, Difference.
- Locate the central candidate: The transient candidate is always
    located at the exact geometric center of each of the three panels.
     Identify this central source first, then use the surrounding
    pixels to determine context (e.g., host galaxies) or rule out
    distractors (e.g., off-center bright stars causing diffraction
    spikes).
- Science (left): the current observation.
- Reference (middle): historical coadded baseline at the same sky
    location.
- Difference (right): science minus reference (subtraction image).
- A localized residual in the difference image may indicate a real
    brightness change. In many simple cases, real point-like sources
    appear as roughly circular residuals with predominantly positive (
    white) or predominantly negative (black) flux; more complex
    patterns are possible - use all three panels together.
- Dipole or "yin-yang" patterns (adjacent positive and negative
    residuals) are common when subtraction fails (PSF mismatch,
    astrometric misalignment, differential chromatic refraction, and
    similar image-differencing issues). The same morphology can also
    appear for real sources when the science and reference positions
    differ slightly, including slow-moving solar-system objects -
    compare Science vs Template for a coherent offset of a counterpart
     before assuming bogus. Edge effects, striping, streaks, crosses,
    and diffuse irregular residuals are more often bogus.
```

```
- Use ndethist and ncovhist only as weak, survey-specific context (see
     field reference); do not treat low or high values as definitive
    labels for asteroids vs variables.
- Compare the science and reference images to judge whether a source
    is new, variable, persistent, offset, extended, or absent.
- Use the images together with the metadata. Do not rely on images
    alone when metadata provide important context.

Important metadata instructions:
- The user message lists [ZTF CANDIDATE FIELDS] as field names and
    values exactly as in the benchmark extract (not pre-decoded band
    names or subtraction words).
- Use the following reference to interpret those fields. Part A asks
    for decoded quantities: filter_band must be g, r, or i (derive
    from fid), and subtraction_sign must be positive or negative (
    derive from isdiffpos using the reference).

ZTF candidate field reference (schema: https://zwickytransientfacility
    .github.io/ztf-avro-alert/schema.html):
- fid: filter ID (integer). 1 = g, 2 = r, 3 = i.
- isdiffpos: string flag. t or 1 => positive subtraction (science
    minus reference), i.e. brighter in science than reference. f or 0
    => negative subtraction (reference minus science), i.e. fainter in
     science than reference.
- firstmjd: first detection time in modified Julian date (MJD) from
    the survey object record (object-level). This is not the same as
    the Avro candidate field jd, which is the observation time in
    Julian Date (JD) days in the alert packet (~2.45e6 scale).
- magpsf: PSF-fit magnitude on the difference (DIA) image at the
    candidate position [mag]; lower = brighter (ZTF alert pipeline).
- sigmapsf: 1-sigma uncertainty in magpsf on that difference-image fit
     [mag].
- fwhm: FWHM assuming Gaussian core from SExtractor [pixels].

Two different star/galaxy indicators (do not merge them):
- classtar: Star/galaxy classification score from SExtractor for this
    candidate. The public Avro schema does not specify which stamp (
    science vs difference) SExtractor used; treat it as a
    morphological score and combine with the three cutouts. It is not
    derived from the Pan-STARRS1 catalog.
- sgscore1 and distpsnr1 (PS1 neighbor): sgscore1 is the star/galaxy
    score of the closest Pan-STARRS1 (PS1) catalog source within 30
    arcsec; 0 <= sgscore1 <= 1, with values closer to 1 implying
    higher likelihood of being a star (ZTF schema wording). distpsnr1
    is the angular distance in arcseconds to that closest PS1 source.
    If distpsnr1 is large, or PS1 fields are missing or sentinel
    values, treat sgscore1 as weak or ambiguous - the nearest PS1
    object may be an unrelated projection near the line of sight.

- chinr: DAOPhot chi parameter of the nearest source in the reference-
    image PSF catalog within 30 arcsec.
- sharpnr: DAOPhot sharpness of that nearest reference PSF source
    within 30 arcsec (values near 0 are more point-like).

- ndethist: Number of spatially coincident detections within 1.5
    arcsec over survey history, counting only detections on the same
    ZTF field and readout channel as this candidate; raw detections
    down to photometric S/N ~3 are included (ZTF schema). Values shown
     match alert-level history for this candidate. This is not the
    same as a plain-language "object visit count."
- ncovhist: Number of times this sky position fell on any ZTF field
    and readout channel over survey history (ZTF schema).
```

```
Soft ZTF-specific context (heuristics, not rules): low ndethist can
    occur for some solar-system detections but is not definitive -
    cadence, linking, and the definition above matter. Higher ndethist
     at a fixed sky position is more suggestive of repeated activity (
    e.g. variable stars, AGN) but remains context-dependent and survey
    -cadence-dependent.

- sgmag1, srmag1, simag1, szmag1: PS1 PSF magnitudes of the closest
    PS1 catalog source within 30 arcsec in g, r, i, z [mag]. Derived
    colors (e.g. g-r, r-i) describe that matched PS1 object (often
    host+nucleus blend), not necessarily the transient alone - use
    distpsnr1 and cutouts.
- nmtchps: number of PS1 catalog sources within 30 arcsec.
- deltajd: time span in days between first and last detection for this
     object (object-level).

Sentinel values: numeric -999 (and similar schema null sentinels)
    means no valid measurement. Do not interpret -999 as a physical
    magnitude, distance, or flux; do not use it as numeric evidence in
     Part B or Part C. For Part A, copy magpsf and sigmapsf from the
    input when they are real measurements. If they are missing or
    sentinels, you must still satisfy the JSON number types if the
    schema requires floats - do not invent astrophysical photometry;
    state clearly in Part B that those inputs were missing or non-
    physical.


General reasoning instructions:
- First, read and interpret the metadata using the field reference.
- Then, analyze the science, reference, and difference cutouts jointly
    .
- Base your explanation on concrete evidence from the provided input.
- Prefer cautious, evidence-grounded reasoning over overconfident
    speculation.
- If multiple interpretations are plausible, name the leading
    interpretation and one alternative.
- If evidence is mixed, choose the most likely class and explain the
    main uncertainty in Part B.

Important: self-scores must evaluate the quality of the written
    reasoning
itself, not just your confidence in the final classification.

Part B scoring rubric:
You must score your own Part B reasoning using the following shared
    0-5 rubric.
Use this rubric exactly when assigning:
- self_score_key_evidence
- self_score_leading_interpretation_and_support
- self_score_alternative_analysis

The three reasoning dimensions are:

1. Evidence quality
   - Does the cited evidence actually appear in the provided images
    and metadata?
   - Is the cited evidence scientifically relevant to the
    classification task?

2. Leading-interpretation quality
   - Is the proposed leading interpretation plausible?
   - Is it supported by the cited evidence?
```

```
3. Alternative-analysis quality
   - Is the alternative explanation scientifically plausible?
   - Is it discussed in a coherent way using the provided evidence?

Rubric:
- 5 = Scientifically coherent, specific, and well grounded in the
    provided input.
- 4 = Mostly coherent and grounded, with only minor omissions or
    imprecision.
- 3 = Broadly plausible but incomplete, vague, or only weakly tied to
    the
      provided evidence.
- 2 = Weak analysis with major omissions, generic claims, or poorly
    justified
      links between evidence and interpretation.
- 1 = Largely unsupported or internally inconsistent.
- 0 = Clearly flawed, contradictory, or hallucinatory.

Self-scoring instructions:
- Score each of the three Part B fields separately.
- Use only integers from 0 to 5.
- Be strict and evidence-based.
- Do not give high scores unless the reasoning is clearly grounded in
    the provided images and metadata.

You must return your answer as a single JSON object with three top-
    level keys
("Part A", "Part B", "Part C") matching this structure:

{
  "Part A": {
    "filter_band": "<g | r | i>",
    "subtraction_sign": "<positive | negative>",
    "magpsf": <float>,
    "sigmapsf": <float>,
    "ndethist": <int>,
    "ncovhist": <int>
  },
  "Part B": {
    "key_evidence": "<string>",
    "leading_interpretation_and_support": "<string>",
    "alternative_analysis": "<string>",
    "self_score_key_evidence": <int 0-5>,
    "self_score_leading_interpretation_and_support": <int 0-5>,
    "self_score_alternative_analysis": <int 0-5>
  },
  "Part C": {
    "stage1": "<artifact | real_object>",
    "stage2": "<solar_system | astrophysical | N/A>",
    "stage3": "<supernova | variable_star | AGN | N/A>"
  }
}

Output constraints:
- For Part A:
  - filter_band must be exactly one of: g, r, i
  - subtraction_sign must be exactly one of: positive, negative
  - magpsf must be a float
  - sigmapsf must be a float
  - ndethist must be an integer
  - ncovhist must be an integer
- For Part B:
  - each self score must be an integer from 0 to 5
```

```
  - keep each rationale field concise and evidence-based
- For Part C:
  - stage1 must be exactly one of: artifact, real_object
  - stage2 must be exactly one of: solar_system, astrophysical, N/A
  - stage3 must be exactly one of: supernova, variable_star, AGN, N/A

Logical consistency rules:
- If stage1 = artifact, then stage2 = N/A and stage3 = N/A.
- If stage1 = real_object and stage2 = solar_system, then stage3 = N/A
    .
- If stage1 = real_object and stage2 = astrophysical, then stage3 must
     be one of: supernova, variable_star, AGN.

Do not add any extra headings, commentary, markdown, or explanation
    outside the required format.


=== USER ===

[ALERT IDENTIFIERS]
- Object ID: ZTF26aargnnp

[ZTF CANDIDATE FIELDS]
- fid: 2
- isdiffpos: t
- firstmjd: 61135.4
- magpsf: 17.7693
- sigmapsf: 0.0461362
- fwhm: 2.49
- classtar: 0.983
- sgscore1: 0.993333
- distpsnr1: 24.3978
- chinr: 0.613
- sharpnr: 0.107
- ndethist: 1
- ncovhist: 1489
- sgmag1: 19.721
- srmag1: 18.7795
- simag1: 18.3621
- szmag1: 18.1831
- nmtchps: 3
- deltajd: 0

Field definitions and sentinel rules are in the system message.

Analyze this alert and return the JSON response.
```

B.3 Self-Correction Prompt

Only the model-generated Part A–C JSON from the first attempt is provided in the second-round API prompt; ground-truth labels and archival metadata such as `target_class`, file paths, model identifiers, and evaluation metadata are never passed to the model.

Second-rollout ablation prompt

```
{general_prompt_block}[3]


---

## Second trial (ablation)
```

```
You are seeing the **same** alert again (same images and metadata as
    below).

Below is the **exact JSON object you returned on your first attempt**
    for this
alert (Parts A-C and all fields). That earlier output is **not**
    labeled as
correct or incorrect; treat it only as your own prior work product.

**Your task now**

1. Carefully re-read the images and metadata.
2. Critically review your **previous** JSON: evidence, staged logic (
    artifact vs
   real, solar vs astrophysical, subclass), and self-scores.
3. Produce a **fresh** single JSON object with the **same top-level
    schema**
   (`Part A`, `Part B`, `Part C`) and the **same field names and
    constraints**
   as in the original benchmark instructions in the system message.

You may revise any part of your answer if the evidence warrants it, or
     keep
the same conclusion if re-analysis supports it. If you change your
    mind,
briefly explain why in the Part B prose (without claiming any external
verification of right/wrong).

**Important:** Output **only** one JSON object - no markdown fences,
    no
preamble or postscript outside the JSON.


=== USER ===

[ALERT IDENTIFIERS]
- Object ID: ZTF26aargnnc

[ZTF CANDIDATE FIELDS]
- fid: 2
- isdiffpos: t
- firstmjd: 61135.4
- magpsf: 19.2354
- sigmapsf: 0.152689
- fwhm: 2.03
- classtar: 0.982
- sgscore1: 0.99225
- distpsnr1: 7.19681
- chinr: 1.642
- sharpnr: -0.106
- ndethist: 1
- ncovhist: 1503
- sgmag1: 21.5745
- srmag1: 20.5907
- simag1: 20.0833
- szmag1: 19.8477
- nmtchps: 5
- deltajd: 0

Field definitions and sentinel rules are in the system message.

Analyze this alert and return the JSON response.
```

```
---

[YOUR FIRST ATTEMPT JSON - FOR SELF-REVIEW ONLY]

{
  "Part A": {
    "filter_band": "r",
    "subtraction_sign": "positive",
    "magpsf": 19.2354,
    "sigmapsf": 0.152689,
    "ndethist": 1,
    "ncovhist": 1503
  },
  "Part B": {
    "key_evidence": "Positive r-band alert with a point-like
    morphology (classtar 0.982, FWHM 2.03 px). The exact candidate
    position does not show an obvious persistent source or host galaxy
     in the template. The nearest PS1 source is stellar-like but
    offset by 7.2 arcsec, so it is not a secure counterpart. History
    is only a single detection (ndethist=1, deltajd=0) despite very
    large coverage (ncovhist=1503).",
    "leading_interpretation_and_support": "Most consistent with a
    solar-system object (Asteroid benchmark class): a one-off,
    positive, point-source detection with no clear stationary
    counterpart at the alert position is better explained by a moving
    object than by fixed-position variability. The lack of a central
    host disfavors SN/AGN, and ndethist=1 under heavy coverage
    disfavors a typical variable star interpretation.",
    "alternative_analysis": "Main alternative is Bogus: the residual
    is fairly subtle and there is a nearby static source in the field,
     so low-S/N subtraction contamination is possible. A hostless SN
    or stellar flare is less supported because there is no clear fixed
    -position counterpart or host at the center.",
    "self_score_key_evidence": 4,
    "self_score_leading_interpretation_and_support": 4,
    "self_score_alternative_analysis": 3
  },
  "Part C": {
    "stage1": "real_object",
    "stage2": "solar_system",
    "stage3": "N/A"
  }
}

---

Return your **second-trial** JSON object now.
```

C Comprehensive Benchmark Results: Metric Definitions and Per-Part Findings

This appendix consolidates the formal metric definitions used throughout the AstroAlertBench discovery chain and reports per-run, per-class numbers that complement Section 4. Every reported proportion carries a 1σ binomial standard error $\mathrm{SE} = \sqrt{p(1-p)/n}$, and every pairwise difference carries a propagated standard error $\mathrm{SE}_\Delta = \sqrt{\mathrm{SE}_a^2 + \mathrm{SE}_b^2}$ together with a two-sample z-statistic. Unless otherwise stated, denominators are $n = 1,500$ alerts (300 per class).

[3] Same text as the `prompts.py` **system** message in the *General benchmark prompt* listing above (omitted here).

C.1 PART A: METADATA GROUNDING

Metric definitions. For each of the six grounded metadata fields we report **exact-match accuracy**, defined as the fraction of alerts on which the model's emitted value equals the ground-truth value in the alert packet. The six fields evaluated, together with their physical interpretation, are: `fid` (filter identifier; integer ID for the photometric band, e.g. $1 \rightarrow g$), `isdiffpos` (sign of the difference-image residual), `magpsf` (PSF-fit magnitude in the difference image), `sigmapsf` (the 1σ uncertainty on `magpsf`), `ndethist` (the number of historical detections at this location), and `ncovhist` (the number of historical observations covering this location).

Per-field results. All 13 evaluated configurations achieve exact-match accuracy of $100.00 \pm 0.00\%$ on every one of the six fields. Table 7 reports the per-field results compactly. These results establish correct performance on the six tested metadata fields.

Field	Description	Accuracy (all 13 runs)
`fid`	Filter identifier (band)	$100.00 \pm 0.00\%$
`isdiffpos`	Sign of difference-image residual	$100.00 \pm 0.00\%$
`magpsf`	PSF-fit magnitude	$100.00 \pm 0.00\%$
`sigmapsf`	1σ uncertainty on `magpsf`	$100.00 \pm 0.00\%$
`ndethist`	Historical detection count	$100.00 \pm 0.00\%$
`ncovhist`	Historical coverage count	$100.00 \pm 0.00\%$

Table 7: **Part A per-field exact-match accuracy.** Each cell is the proportion of alerts on which the model's emitted value matches the ground-truth alert packet, pooled over all 13 evaluated configurations.

C.2 PART B: SCIENTIFIC RATIONALE

Metric definitions. For each alert the model returns three self-rated dimensions, each on the integer 0–5 rubric of Stoppa et al. (2025): **key evidence**, **leading interpretation**, and **alternative analysis**. Details for grading rubrics and three questions prompted are in Appendix B.2.1. Let $s_{i,d} \in \{0, \ldots, 5\}$ denote the rating on dimension $d \in \{\text{ev}, \text{lead}, \text{alt}\}$ for alert i, and let $\bar{s}_i = \frac{1}{3}(s_{i,\text{ev}} + s_{i,\text{lead}} + s_{i,\text{alt}})$ be the per-alert mean. We summarize each run with two scalars:

- **Mean self-reasoning score** (**MSRS**): the dataset average of per-alert means, $\text{MSRS} = \frac{1}{n}\sum_{i=1}^{n} \bar{s}_i$. We report MSRS as a point estimate; its uncertainty can be estimated from the variance of per-alert mean scores or by bootstrapping alerts.

- **Self pass rate**: the binomial proportion $\text{SPR} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\bar{s}_i \geq 4\}$, with 1σ binomial standard error.

Per-run summary. Table 8 reports MSRS and self pass rate for all 13 runs. Self-scores cluster near the rubric ceiling for most runs, with 11 of 13 runs above a 90% self pass rate. The two `Claude Opus 4.7` configurations sit at the floor at $\text{MSRS} = 3.99$ each with self pass rates near 67%, while `Gemini 2.5 Pro high-think` reports the highest mean rationale self-rating at $\text{MSRS} = 4.886$ with a 100% self pass rate. The inverse correlation between MSRS and end-to-end accuracy is the population-level association developed in Section 5.1.

#	Run	MSRS	Self pass rate
1	Claude Opus 4.7 think	3.992	$66.80 \pm 1.22\%$
2	GPT-5.4 high-think	4.185	$92.13 \pm 0.70\%$
3	Kimi K2.5 think	4.488	$90.78 \pm 0.75\%$
4	Claude Opus 4.7 nothink	3.991	$67.07 \pm 1.21\%$
5	Qwen3.5-397B-A17B think	4.744	$99.86 \pm 0.10\%$
6	GPT-5.4 no-think	4.284	$97.60 \pm 0.40\%$
7	Gemini 2.5 Pro high-think	4.886	$100.00 \pm 0.00\%$
8	Gemini 2.5 Flash no-think	4.361	$94.46 \pm 0.59\%$
9	Qwen3.5-397B-A17B nothink	4.783	$99.84 \pm 0.10\%$
10	Qwen3.5-35B-A3B think	4.703	$99.89 \pm 0.09\%$
11	Qwen3.5-35B-A3B nothink	4.641	$98.32 \pm 0.33\%$
12	Qwen3.5-4B nothink	4.673	$99.63 \pm 0.16\%$
13	Qwen3.5-4B think	4.827	$100.00 \pm 0.00\%$

Table 8: **Part B per-run self-reasoning summary.** Mean self-reasoning score (MSRS) and self pass rate for each of the 13 evaluated configurations; uncertainties on the pass rate are 1σ binomial standard errors. Run order matches Table 1 for cross-table reading.

C.3 Part C: Staged Classification

Metric definitions. Part C decomposes the five-class decision into a hierarchical cascade with three sequential decision points. We measure:

- **Stage-1 accuracy**: proportion of alerts for which the Stage-1 *artifact*-vs-*real_object* call is correct.
- **Stage-2 accuracy**: proportion of alerts for which the Stage-2 *solar_system*-vs-*astrophysical* call is correct.
- **Stage-3 accuracy**: proportion of alerts for which the Stage-3 astrophysical-subclass call (*supernova*, *variable_star*, or *AGN*) is correct.
- **Stage-3 conditional accuracy**: Stage-3 accuracy restricted to alerts on which Stage-1 and Stage-2 are both correct, isolating subclass disambiguation skill from upstream routing.
- **End-to-end 5-class accuracy**: proportion of alerts for which the joint hierarchical output matches the ground-truth label across all stages.
- **Per-class accuracy**: proportion of correctly classified alerts among the 300 alerts of each ground-truth class (*supernova*, *AGN*, *variable_star*, *asteroid*, *bogus*).
- **Stage-3 precision, recall, and macro-F1**: standard precision and recall computed per astrophysical subclass, with macro-F1 the unweighted mean across the three subclasses.

Headline 5-class accuracy. Figure 7 ranks the 13 model configurations by absolute end-to-end 5-class accuracy and complements the staged decomposition in Table 1 of the main text.

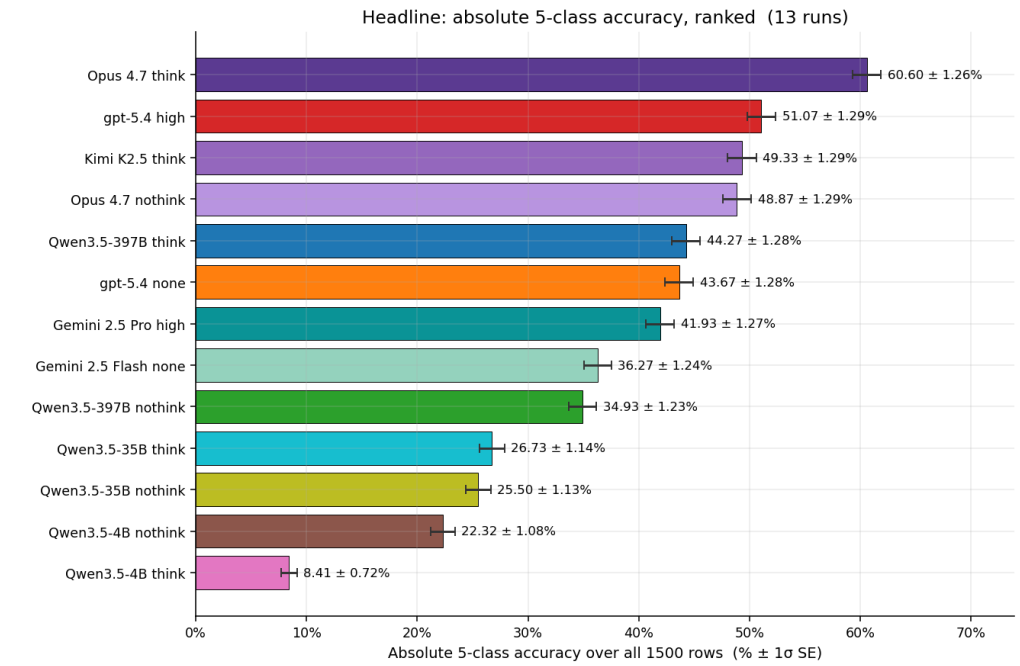

Figure 7: **End-to-end 5-class accuracy ranked across the 13 evaluated configurations.** Error bars denote 1σ binomial standard errors. `Claude Opus 4.7 think` establishes the current benchmark ceiling at $60.60 \pm 1.26\%$, opening a $+9.53 \pm 1.80$ pp gap over the second-ranked `GPT-5.4 high-think`.

Stage-wise cascade. Figure 8 visualizes the Stage-1, Stage-2, Stage-3, and Stage-3-conditional accuracies of every run as grouped bars; the underlying numbers are the columns of Table 1 of the main text. `Claude Opus 4.7 think` is the only run on which all four stage bars exceed 60% simultaneously; every other run drops below 60% on at least one stage.

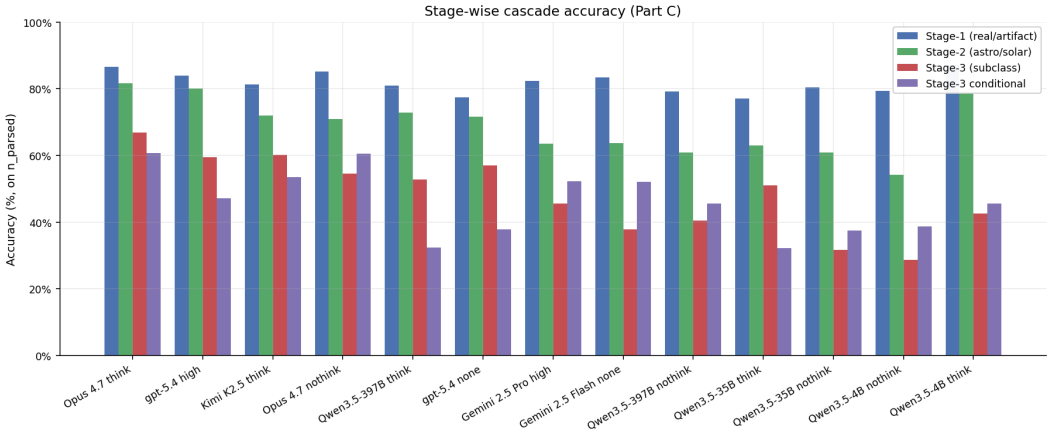

Figure 8: **Stage-wise cascade accuracy across the three Part C stages.** Four grouped bars per run: Stage-1 (real-vs-artifact), Stage-2 (solar-system-vs-astrophysical), Stage-3 (astrophysical subclass), and Stage-3 conditional (restricted to rows where Stages 1 and 2 are both correct). Stage-1 accuracy ranges from 77.15 to 87.07%, compared with an 80% always-`real_object` baseline on the full balanced benchmark (four real-object classes out of five). Stage-2 separates closed-source from open-source; Stage-3 conditional carries a large spread (32.12–60.78%) and explains most of the headline 5-class ranking.

Per-class accuracy. Table 9 reports per-class accuracy on the five benchmark classes for every run; rows are sorted by absolute 5-class accuracy. Three universal patterns emerge: AGN accuracy stays below 8% for all 13 runs; the `Gemini 2.5` family collapses on *asteroid* ($\leq 1.0\%$); and most smaller `Qwen3.5` variants collapse on *bogus* ($< 4\%$).

Run	Supernova	AGN	Variable Star	Asteroid	Bogus
Claude Opus 4.7 think	$\mathbf{84.00 \pm 2.12}$	$\mathbf{7.33 \pm 1.51}$	91.00 ± 1.65	57.33 ± 2.86	$\mathbf{63.33 \pm 2.78}$
GPT-5.4 high-think	38.67 ± 2.81	$\mathbf{7.33 \pm 1.51}$	95.33 ± 1.22	75.67 ± 2.48	38.33 ± 2.81
Kimi K2.5 think	65.00 ± 2.75	5.33 ± 1.30	90.27 ± 1.72	75.67 ± 2.48	11.04 ± 1.81
Claude Opus 4.7 nothink	79.67 ± 2.32	6.00 ± 1.37	95.67 ± 1.18	3.33 ± 1.04	59.67 ± 2.83
Qwen3.5-397B-A17B think	18.00 ± 2.22	6.00 ± 1.37	73.33 ± 2.55	80.67 ± 2.28	43.33 ± 2.86
GPT-5.4 no-think	14.33 ± 2.02	7.00 ± 1.47	92.33 ± 1.54	$\mathbf{81.00 \pm 2.26}$	23.67 ± 2.45
Gemini 2.5 Pro high-think	80.00 ± 2.31	6.33 ± 1.41	70.67 ± 2.63	1.00 ± 0.57	51.67 ± 2.89
Gemini 2.5 Flash no-think	59.67 ± 2.83	2.00 ± 0.81	94.33 ± 1.34	0.33 ± 0.33	25.00 ± 2.50
Qwen3.5-397B-A17B nothink	56.67 ± 2.86	3.01 ± 0.99	77.00 ± 2.43	17.00 ± 2.17	21.07 ± 2.36
Qwen3.5-35B-A3B think	13.86 ± 2.68	1.62 ± 0.93	72.60 ± 3.01	55.61 ± 3.63	50.74 ± 3.51
Qwen3.5-35B-A3B nothink	17.14 ± 2.25	0.68 ± 0.48	96.61 ± 1.05	8.22 ± 1.61	2.68 ± 1.00
Qwen3.5-4B nothink	27.50 ± 2.88	0.00 ± 0.00	86.76 ± 2.05	4.75 ± 1.24	3.61 ± 1.12
Qwen3.5-4B think	5.00 ± 3.45	0.00 ± 0.00	$\mathbf{98.25 \pm 1.23}$	14.81 ± 6.84	2.56 ± 2.53

Table 9: **Per-class accuracy across the 13 evaluated configurations (%).** Each cell is the proportion correct among valid final predictions in that ground-truth class, with 1σ binomial standard error on that denominator. Rows are ordered by absolute 5-class accuracy. **Bold** denotes the highest conditional rate per class; ties show both rows in bold.

Per-class state of the art. Table 10 extracts the leader and runner-up for each ground-truth class together with the propagated standard error and z-statistic on the gap. None of the per-class gaps reaches $z = 2$, indicating that no single run holds a statistically significant lead on any one class.

Class	Best run	Score (%)	Runner-up	$\Delta \pm \mathrm{SE}_\Delta$ (z)
Supernova	Claude Opus 4.7 think	84.00 ± 2.12	Gemini 2.5 Pro high-think (80.00 ± 2.31)	$+4.00 \pm 3.13$ ($z = 1.27$)
AGN	GPT-5.4 high-think $\equiv$ Opus 4.7 think (tie)	7.33 ± 1.51	GPT-5.4 no-think (7.00 ± 1.47)	$+0.33 \pm 2.11$ ($z = 0.16$)
Variable Star	Claude Opus 4.7 nothink	95.67 ± 1.18	GPT-5.4 high-think (95.33 ± 1.22)	$+0.33 \pm 1.69$ ($z = 0.20$)
Asteroid	GPT-5.4 no-think	81.00 ± 2.26	Qwen3.5-397B-A17B think (80.67 ± 2.28)	$+0.33 \pm 3.21$ ($z = 0.10$)
Bogus	Claude Opus 4.7 think	63.33 ± 2.78	Claude Opus 4.7 nothink (59.67 ± 2.83)	$+3.67 \pm 3.97$ ($z = 0.92$)

Table 10: **Per-class leaders on the full evaluation set.** Best and runner-up runs using 300 alerts per class. Gaps are in percentage points; the displayed gap SEs and z-statistics retain the independence approximation.

Stage-3 subclass precision, recall, and macro-F1. Table 11 reports per-subclass precision and recall and the unweighted macro-F1 for the three astrophysical subclasses; rows are sorted by macro-F1. The two `Claude Opus 4.7` configurations lead the macro-F1 column at 0.558 and 0.548. *Supernova* F1 ranges from 0.095 to 0.913; *variable_star* F1 saturates near 0.65 for the top five runs; *AGN* F1 stays below 0.13 for every run.

Run	SN P	SN R	SN F1	VS P	VS R	VS F1	AGN P	AGN R	AGN F1	Macro-F1
Claude Opus 4.7 think	1.000	0.840	**0.913**	0.489	0.910	0.636	0.423	0.073	**0.125**	**0.558**
Claude Opus 4.7 nothink	1.000	0.797	0.887	0.501	**0.957**	**0.658**	0.286	0.060	0.099	0.548
Gemini 2.5 Pro high-think	0.996	0.800	0.887	0.431	0.707	0.535	0.302	0.063	0.105	0.509
Kimi K2.5 think	0.995	0.650	0.786	0.475	0.903	0.623	0.291	0.053	0.090	0.500
Gemini 2.5 Flash no-think	0.989	0.597	0.744	0.429	0.943	0.590	0.154	0.020	0.035	0.456
GPT-5.4 high-think	1.000	0.387	0.558	0.494	0.953	0.651	0.135	0.073	0.095	0.435
Qwen3.5-397B-A17B nothink	0.939	0.567	0.707	0.403	0.770	0.529	0.200	0.030	0.052	0.430
GPT-5.4 no-think	1.000	0.143	0.251	0.496	0.923	0.646	0.133	0.070	0.092	0.329
Qwen3.5-397B-A17B think	1.000	0.180	0.305	0.435	0.733	0.546	0.113	0.060	0.078	0.310
Qwen3.5-4B nothink	0.786	0.264	0.395	0.382	0.837	0.524	0.000	0.000	0.000	0.307
Qwen3.5-35B-A3B nothink	0.923	0.161	0.274	0.367	0.956	0.530	0.143	0.007	0.013	0.273
Qwen3.5-35B-A3B think	1.000	0.137	0.241	0.434	0.716	0.541	0.273	0.016	0.031	0.271
Qwen3.5-4B think	1.000	0.050	0.095	0.463	0.974	0.628	0.000	0.000	0.000	0.241

Table 11: **Stage-3 subclass precision, recall, and macro-F1.** Precision (P) and recall (R) are reported per astrophysical subclass (*supernova*, *variable_star*, *AGN*); macro-F1 is the unweighted mean across the three subclasses. Rows are sorted by macro-F1. **Bold** denotes per-column best-in-batch.

Stage-3 confusion matrices. Table 12 reports the Stage-3 subclass confusion matrices for the four highest-accuracy runs. Across all four, 91–93% of true *AGN* alerts are predicted as *variable_star*, identifying AGN-vs-VS as the dominant Stage-3 confusion regardless of model family or reasoning mode. Figure 9 visualizes the AGN-row distribution as four pies.

	Claude Opus 4.7 think				GPT-5.4 high-think			
True ↓ Pred →	SN	VS	AGN	N/A	SN	VS	AGN	N/A
Supernova	**252**	11	30	7	116	21	**141**	22
Variable Star	0	273	0	27	0	286	0	14
AGN	0	**274**	22	4	0	**272**	22	6
	Kimi K2.5 think				**Claude Opus 4.7 nothink**			
True ↓ Pred →	SN	VS	AGN	N/A	SN	VS	AGN	N/A
Supernova	**195**	21	39	45	**239**	8	45	8
Variable Star	0	269	0	29	0	287	0	13
AGN	1	**276**	16	7	0	**278**	18	4

Table 12: **Stage-3 subclass confusion matrices for the four highest-accuracy runs.** Rows are ground-truth classes; columns are model predictions among *supernova* (SN), *variable_star* (VS), *AGN*, and the *N/A* bucket where the model declines to commit to a subclass. The diagonal SN cell and the dominant off-diagonal AGN→VS cell are bolded for each run. Across all four runs, real AGN are predicted as VS at ∼91–93%.

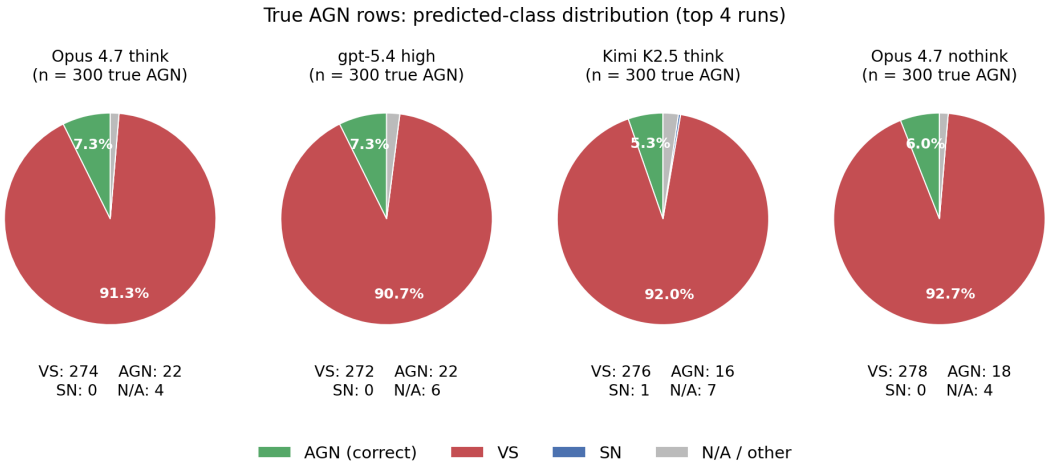

Figure 9: **Predicted-class distribution among the** 300 **true-AGN alerts (top four models).** Across all four models, 91–93% of real AGN alerts are predicted as *variable_star*; AGN recall ranges from 5.3% to 7.3%. The AGN-collapse pattern holds independently of model family, scale, and reasoning mode.

D REASONING-MODE ANALYSIS: THINK VS. NO-THINK

To quantify the impact of reasoning on the end-to-end 5-class accuracy reported in Section 4.4, we compare the same model architecture across its two reasoning modes (thinking enabled and direct answer). Five families admit a same-architecture pair at $n = 1,500$: Claude Opus 4.7 {think, nothink}, GPT-5.4 {high-think, no-think}, and the three Qwen3.5 sizes {4B, 35B-A3B, 397B-A17B} $\times$ {think, nothink}. Gemini 2.5 Pro versus 2.5 Flash is not a clean A/B because the two configurations are different models rather than the same model with the reasoning dial flipped, and we therefore omit it from the paired comparison. For each pair we report the absolute accuracy gap $\Delta = \mathrm{Acc}_{\mathrm{think}} - \mathrm{Acc}_{\mathrm{nothink}}$, the reported independence-based approximation $\mathrm{SE}_\Delta = \sqrt{\mathrm{SE}^2_{\mathrm{think}} + \mathrm{SE}^2_{\mathrm{nothink}}}$, and the corresponding two-sample z-statistic. Paired uncertainty requires the joint correctness outcomes on the same alerts. Table 13 reports the comparison and Figure 10 visualizes the paired bars across the five families.

Table 13: **Think vs. no-think pairs across five model families.** Δ is the absolute 5-class accuracy gap (think minus no-think) with an independence-based SE approximation; z is the corresponding two-sample statistic. The final column describes the observed accuracy difference. Rows are ordered from most negative Δ to most positive. The Qwen3.5-397B-A17B gap uses the displayed accuracies, as in Figure 10.

Family	Think (%)	No-think (%)	$\Delta \pm \mathrm{SE}_\Delta$ (pt)	z	Verdict
Qwen3.5-4B	8.41 ± 0.72	22.32 ± 1.08	-13.90 ± 1.29	-10.76	higher no-think accuracy; think truncated
Qwen3.5-35B-A3B	26.73 ± 1.14	25.50 ± 1.13	$+1.23 \pm 1.60$	0.77	small positive gap
Qwen3.5-397B-A17B	44.27 ± 1.28	34.93 ± 1.23	+9.34 $\pm$ 1.78	5.25	higher think accuracy
GPT-5.4	51.07 ± 1.29	43.67 ± 1.28	$+7.40 \pm 1.82$	4.07	higher think accuracy
Claude Opus 4.7	60.60 ± 1.26	48.87 ± 1.29	$+11.73 \pm 1.80$	6.50	higher think accuracy, largest gain

Interpretation. Three frontier-scale model pairs (`Opus 4.7`, `Qwen3.5-397B-A17B`, and `GPT-5.4`) show reported accuracy gains of $+7$ to $+12$ percentage points when extended reasoning is enabled. The mid-sized `Qwen3.5-35B-A3B` pair shows a smaller gain of 1.23 percentage points, while the smallest `Qwen3.5-4B` configuration has lower accuracy with thinking, suggesting

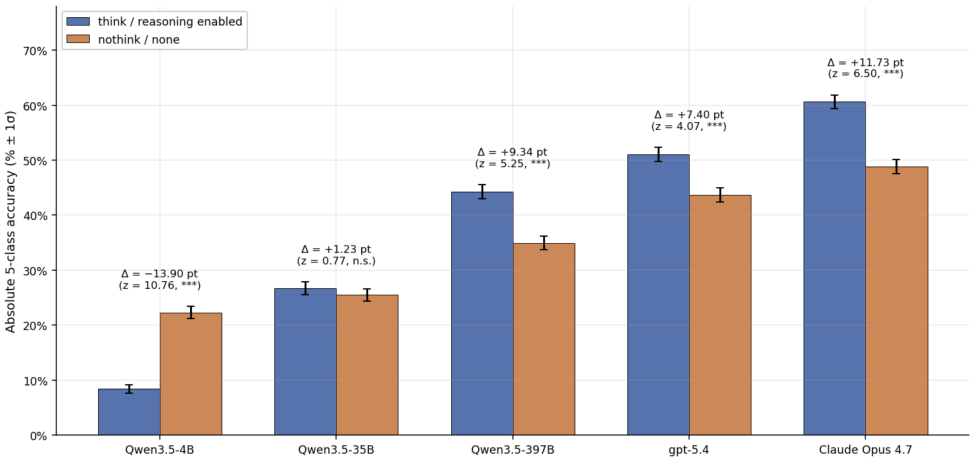

Figure 10: **Reasoning-mode dial across five model families ($n = 1,500$ per run).** For each family, blue bars show absolute 5-class accuracy with thinking enabled and orange bars show the same with thinking disabled; Δ annotations report the gap and the two-sample z-statistic. Claude Opus 4.7 shows the largest positive gap ($+11.73$ pp), followed by Qwen3.5-397B-A17B and GPT-5.4; Qwen3.5-35B-A3B has a smaller positive gap, and Qwen3.5-4B has a negative gap. The plotted z annotations use the independence approximation.

that the efficacy of internal chain-of-thought is highly sensitive to base model scale on this specialized task. The largest reported gain is for `Opus 4.7`.

E Rationale Self-Assessment and Classification Confidence

This appendix reports the association of rationale self-ratings with classification correctness, a sensitivity analysis that normalizes those ratings, and a separate classification-confidence experiment. These analyses address different constructs and use different targets and sample sizes.

E.1 Definitions and descriptive association measures

Let $c_i = (s_{i,\text{ev}} + s_{i,\text{lead}} + s_{i,\text{alt}})/3$ be the mean Part B self-rating. Each component is an integer from 0 to 5, so $c_i \in \{0, 1/3, 2/3, \ldots, 5\}$. Let $y_i \in \{0, 1\}$ indicate end-to-end Part C correctness. Linked rows require a valid reference class, no transport error, parseable JSON, all three Part B self-scores, and all three Part C stages normalizable to the five-class taxonomy. The following statistics describe score–correctness association.

- **Score gap**: $\Delta\mu = \bar{c}\,|_{y=1} - \bar{c}\,|_{y=0}$, in rubric points. This measures separation of correct and incorrect rows, not calibration error. Table 14 also reports the assumed-scale quantity $\sqrt{1/n_{\text{correct}} + 1/n_{\text{incorrect}}}$. This proxy assumes unit within-group standard deviations and is not an empirical standard error or a guaranteed conservative bound.
- **Pearson** r: the point-biserial correlation between the numerical rubric mean and binary correctness, retained as a descriptive effect size. The quantity $1/\sqrt{n-3}$ is the conventional approximation on the Fisher-transformed scale, not directly the standard error of r. We use the reported stratified-bootstrap AUROC intervals for inference below.
- **Score-group accuracy**: at $\tau = 4$, split rows into $c_i \geq 4$ and $c_i < 4$, then report each group's accuracy and denominator and their difference in percentage points. These groups describe a rubric threshold, not a probability threshold.

E.2 Per-model score–accuracy associations

Table 14 reports the score gap and Pearson correlation for all 13 configurations. Absolute five-class accuracy retains the main benchmark denominator convention, whereas linked-row counts describe

the subset supporting score–correctness comparisons. The scale proxy is calculated from the correct and incorrect group sizes assuming a standard deviation of one rubric point in each group. Table 16 reports AUROC with confidence intervals and permutation-test p-values, together with ECE and Brier scores for normalized rationale self-ratings.

Table 14: **Per-run score-association summaries (sorted by score gap).** n_{linked} is the number of parsed rows admitting both a self-score and an evaluable Part C label. Score gap is $\bar{c}\,|_{\text{correct}} - \bar{c}\,|_{\text{incorrect}}$ in points on the 0–5 self-score scale; Pearson r is between per-row self-score and per-row 0/1 correctness.

#	Run	n_{linked}	Abs. 5-class (%)	Score gap	Scale proxy	Pearson r
1	Qwen3.5-397B-A17B think	1,477	44.27	0.0014	±0.052	+0.0047
2	Qwen3.5-4B think	314	8.41	0.0068	±0.116	+0.0195
3	Gemini 2.5 Pro high-think	1,500	41.93	0.0098	±0.052	+0.0302
4	Qwen3.5-35B-A3B think	871	26.73	0.0147	±0.069	+0.0617
5	Qwen3.5-397B-A17B nothink	1285	34.93	0.0214	±0.058	+0.0592
6	Qwen3.5-35B-A3B nothink	1420	25.50	0.0298	±0.061	+0.0946
7	Kimi K2.5 think	1497	49.33	0.0325	±0.052	+0.0486
8	Claude Opus 4.7 think	1,500	60.60	**0.0350**	±0.053	+0.0470
9	Qwen3.5-4B nothink	1331	22.32	0.0381	±0.064	+0.1064
10	Gemini 2.5 Flash no-think	1,499	36.27	0.1263	±0.054	+0.1847
11	GPT-5.4 high-think	1,500	51.07	0.1299	±0.052	+0.2028
12	GPT-5.4 no-think	1,500	43.67	0.1413	±0.052	+0.2185
13	Claude Opus 4.7 nothink	1,500	48.87	**0.1886**	±0.052	**+0.2517**

E.3 Classification accuracy by self-rating group

Table 15 reports classification accuracies and sample counts for alerts with mean rationale self-ratings $c \geq 4$ and $c < 4$.

Table 15: **Classification accuracy by self-rating group.** High and low groups contain alerts with mean rationale self-ratings $c \geq 4$ and $c < 4$, respectively. n_{high} and n_{low} are the group sizes. Δ_{acc} is the high-group minus low-group accuracy in percentage points.

Run	acc | high (%)	n_{high}	acc | low (%)	n_{low}	Δ_{acc} (pp)
GPT-5.4 no-think	44.06	1464	27.78	36	+16.28
Claude Opus 4.7 nothink	**53.48**	**1006**	**39.47**	**494**	**+14.01**
GPT-5.4 high-think	51.88	1382	41.53	118	+10.35
Gemini 2.5 Flash no-think	36.79	1416	27.71	83	+9.08
Kimi K2.5 think	49.96	1359	44.20	138	+5.76
Claude Opus 4.7 think	60.68	1002	60.44	498	+0.24
Qwen3.5-35B-A3B nothink	26.16	1399	0.00	21	+26.16
Qwen3.5-4B nothink	24.59	1326	0.00	5	+24.59
(Qwen3.5-397B-A17B think and Gemini 2.5 Pro high-think: $n_{\text{low}} \leq 2$.)					

E.4 Relationship between score gap and correlation

Figure 12 compares the score gap with Pearson correlation. For binary correctness with rate π, the point-biserial identity is $r = (\Delta\mu/\sigma_c)\sqrt{\pi(1-\pi)}$, using the empirical population standard deviation of the scores. Correlation depends on the score gap relative to dispersion, not on dispersion alone. A narrow score distribution therefore does not by itself force r toward zero, and a small linear correlation does not establish absence of all predictive information.

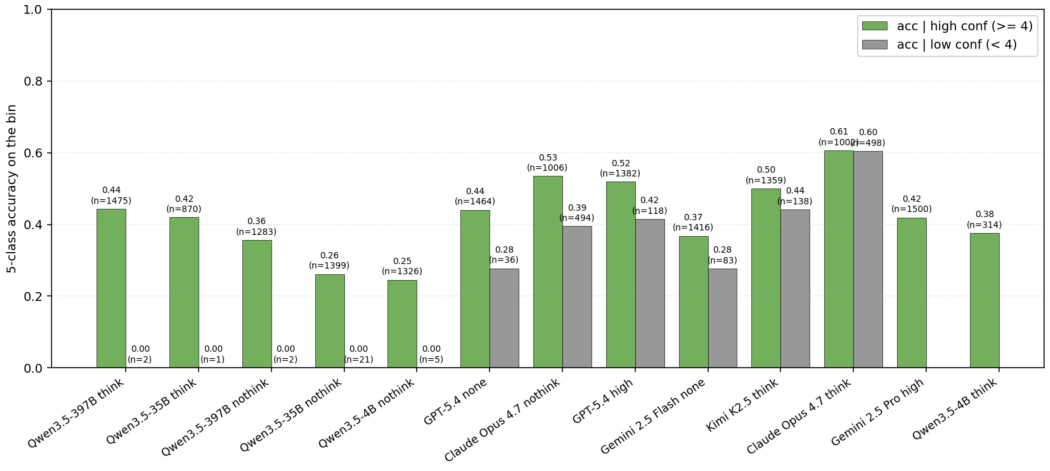

Figure 11: **Five-class accuracy above and below the rationale-rating threshold $c = 4$.** `Claude Opus 4.7 nothink` has 494 rows below the threshold and an accuracy difference of 14.01 percentage points; `GPT-5.4 high-think` has 118 such rows and a difference of 10.35 points.

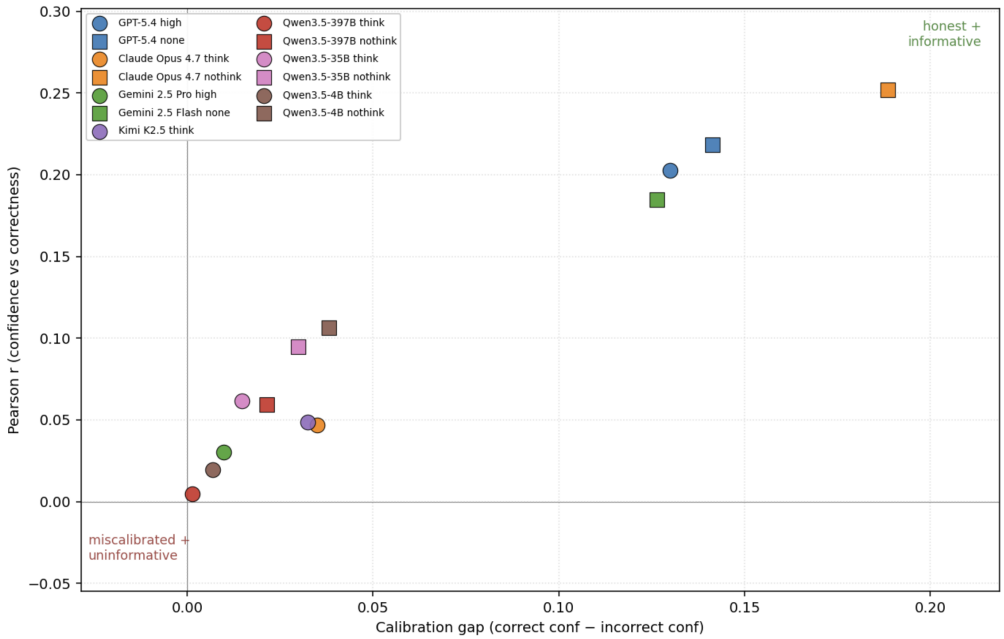

Figure 12: Score gap versus point-biserial correlation across 13 configurations. Both summarize association between rationale ratings and classification correctness; neither measures probability calibration or honesty.

E.5 Additional metrics for rationale self-assessment

We reanalyze the original 13 runs without new inference. AUROC is computed directly from the numerical score ordering, assigning half credit to a tied correct–incorrect pair. The 95% percentile intervals use 10,000 bootstrap resamples stratified by correctness; two-sided permutation tests use 10,000 draws under the null of no score–correctness association. These are nominal, per-configuration tests. Rationale scores have only 3–7 distinct observed means within a run, so ties and the score distribution matter for discrimination.

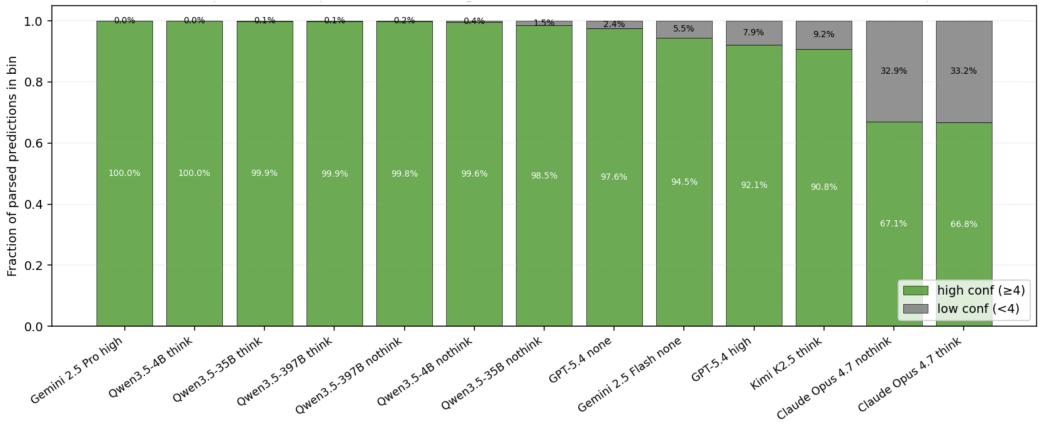

Figure 13: Proportions of alerts with mean rationale self-ratings $c \geq 4$ and $c < 4$ across all 13 configurations.

For a separate probability-scale sensitivity analysis, we explicitly assume $q_i = c_i/5$. We then compute expected calibration error (ECE) and the binary Brier score against classification correctness (Guo et al., 2017):

$$\text{ECE} = \sum_{b=1}^{B} \frac{|I_b|}{n} \left|\overline{q}_b - \overline{y}_b\right|, \qquad \text{Brier} = \frac{1}{n}\sum_{i=1}^{n}(q_i - y_i)^2.$$

Here I_b groups identical normalized scores for exact ECE. The report also checks 15 equal-width and 10 equal-mass bins, which agree to three decimal places across these runs. ECE measures the mismatch between mean scores and empirical correctness within the chosen bins; its estimate depends on binning and sample size. Brier measures overall probability accuracy and combines calibration, resolution, and outcome uncertainty, so it is not a pure calibration measure. Because the original prompt elicits rationale quality, interpreting q as a probability is an assumption. These ECE and Brier values therefore do not by themselves measure calibration of an explicitly elicited class probability.

Table 16: Original rationale-score reanalysis. ECE and Brier assume $q = c/5$; AUROC and its 95% interval measure error discrimination. n counts linked rows.

Run	n	ECE ↓	Brier ↓	AUROC ↑	95% CI	Perm. p
Claude Opus 4.7 think	1500	0.192	0.278	0.5154	[0.488, 0.543]	0.2780
Claude Opus 4.7 nothink	1500	0.310	0.332	0.6222	[0.596, 0.648]	0.0001
GPT-5.4 high-think	1500	0.326	0.347	0.6016	[0.575, 0.628]	0.0001
Kimi K2.5 think	1497	0.403	0.414	0.5205	[0.496, 0.545]	0.1020
GPT-5.4 no-think	1500	0.420	0.413	0.6164	[0.590, 0.642]	0.0001
Qwen3.5-397B-A17B think	1477	0.506	0.504	0.4983	[0.476, 0.520]	0.8830
Gemini 2.5 Flash no-think	1499	0.509	0.483	0.6095	[0.582, 0.637]	0.0001
Qwen3.5-35B-A3B think	871	0.522	0.515	0.5195	[0.497, 0.542]	0.0970
Gemini 2.5 Pro high-think	1500	0.558	0.555	0.5153	[0.491, 0.539]	0.2190
Qwen3.5-4B think	314	0.590	0.583	0.5089	[0.452, 0.566]	0.7810
Qwen3.5-397B-A17B nothink	1285	0.600	0.589	0.5261	[0.499, 0.553]	0.0650
Qwen3.5-35B-A3B nothink	1420	0.671	0.640	0.5224	[0.515, 0.529]	0.0001
Qwen3.5-4B nothink	1331	0.690	0.659	0.5393	[0.519, 0.560]	0.0005

E.6 CLASSIFICATION-CONFIDENCE ABLATION

Evaluation set and prompt. We evaluate a subset of eight model configurations, with coverage limited by model retirements. We sample 300 distinct alerts from the benchmark manifest, with 60 per class. Each configuration processes this same intended set. The prompt retains metadata grounding, the scientific rationale, and staged classification (Parts A–C), but removes the three 0–5 rationale self-scores and adds a post-decision scalar `Part C.P_correct`. This is defined as the probability that the final predicted class is correct, rather than the quality of Part B writing. Models are asked to use the continuous range $[0, 1]$ when warranted. Reference labels are not revealed. The following schema excerpt illustrates the field; 0.73 is an example value, not an experimental result:

```
"Part C": {
  "stage1": "...", "stage2": "...", "stage3": "...",
  "P_correct": 0.73
}
```

The configurations are Opus 4.7 with adaptive/high reasoning or thinking disabled; GPT-5.4 with high or no reasoning; Gemini 2.5 Pro with dynamic thinking and Flash with thinking off; and Qwen3.5-397B-A17B with thinking enabled or disabled through DashScope. The Gemini configurations use different backbones, and the Qwen ablation backend differs from the main benchmark's Tinker setup; neither comparison isolates a reasoning-mode effect alone.

Scoring. For each valid response, let $p_i = P_i(\text{correct})$ and $y_i = \mathbf{1}\{\widehat{y}_i = y_i^{\text{ref}}\}$ under end-to-end five-class scoring. A response is linked only if its Part C stages normalize to a valid class and the probability is valid. The scorer also accepts values in $(1, 100]$ as percentages and divides them by 100. ECE uses one bin per distinct reported probability; the report additionally checks 15 equal-width bins and equal-mass adaptive calibration error. Brier is the mean squared error against binary correctness. AUROC uses midranks to account for ties, with a 5,000-draw permutation test. Table 17 reports metrics over valid responses.

Table 17: Classification-confidence ablation on 300 intended alerts per configuration. Accuracy and metrics use the n valid responses shown. Mean p is the reported probability of a correct classification; permutation p tests AUROC against chance.

Run	n	Acc. (%)	Mean p	ECE ↓	Brier ↓	AUROC ↑	Perm. p
Claude Opus 4.7 think	300	58.3	0.700	0.150	0.247	0.616	0.0006
Claude Opus 4.7 nothink	300	51.0	0.698	0.201	0.251	0.715	0.0002
GPT-5.4 high-think	291	54.3	0.740	0.215	0.278	0.607	0.0018
GPT-5.4 no-think	299	47.8	0.777	0.306	0.320	0.661	0.0002
Qwen3.5-397B-A17B think	300	42.3	0.868	0.452	0.437	0.599	0.0032
Gemini 2.5 Flash no-think	297	38.4	0.849	0.466	0.436	0.671	0.0002
Gemini 2.5 Pro high-think	284	40.1	0.919	0.517	0.490	0.666	0.0002
Qwen3.5-397B-A17B nothink	300	34.0	0.885	0.545	0.514	0.598	0.0034

Every configuration reports a mean probability above its observed accuracy. AUROC is above chance in all eight nominal tests ($p \leq 0.0034$), while ECE remains between 0.150 and 0.545. This combination illustrates why discrimination and probability calibration should be reported separately. Opus think has the lowest ECE and Brier point estimates in this experiment, but its AUROC is lower than that of Opus nothink. The report's Brier skill scores, relative to a constant predictor equal to each run's empirical accuracy, are slightly negative for both Opus modes (-0.017, -0.004) and negative for all other configurations (-0.120 to -1.290). Thus, above-chance discrimination does not imply improvement over that constant probability forecast in Brier score.

Comparison on the same intended alert set. We restrict the original Part B runs to the same 300 alert identifiers and compare $q = c/5$ with the separately elicited probabilities (Table 18). Each protocol uses its own valid responses and its own classification outcomes; the table is not restricted to a common valid-row intersection. ECE is lower under the probability protocol for all eight configurations. Across configurations, ECE rankings agree closely (Spearman $\rho = 0.976$, $p = 3.31 \times 10^{-5}$), whereas AUROC rank agreement is weaker ($\rho = 0.667$, $p = 0.071$). This distinguishes similar configuration-level calibration rankings from equality of the underlying constructs or per-alert scores. For Opus, the think/nothink discrimination difference appears under both protocols: AUROC $0.547/0.629$ for normalized rationale ratings and $0.616/0.715$ for elicited classification confidence.

Table 18: Original rationale ratings versus elicited classification confidence on the same nominal 300-alert set. B denotes original Part B ratings normalized by 5; P denotes the probability protocol. Counts and outcomes are protocol-specific.

Run	n_{B}	n_{P}	ECE B	ECE P	AUROC B	AUROC P
Claude Opus 4.7 think	300	300	0.169	0.150	0.547	0.616
Claude Opus 4.7 nothink	300	300	0.292	0.201	0.629	0.715
GPT-5.4 high-think	300	291	0.318	0.215	0.607	0.607
GPT-5.4 no-think	300	299	0.413	0.306	0.617	0.661
Gemini 2.5 Pro high-think	300	284	0.571	0.517	0.521	0.666
Gemini 2.5 Flash no-think	300	297	0.508	0.466	0.630	0.671
Qwen3.5-397B-A17B think	296	300	0.520	0.452	0.512	0.599
Qwen3.5-397B-A17B nothink	260	300	0.586	0.545	0.535	0.598

This ablation covers one fifth of the benchmark and eight configurations. It is a supporting experiment, not a replacement for the original rationale task. The two protocols involve separate model calls, so differences in predicted classes can reflect both prompting and rollout variability. No temperature sweep or held-out probability recalibration is included. The reported AUROC tests assess discrimination against chance; they are not tests of between-mode differences or evidence of introspective awareness.

F Self-Correction and Behavioral Reliability

This appendix reports second-pass error repair on selected low-rubric-score alerts (Section 5.3).

F.1 Methodology

For each closed-source run we constructed a low-self-rating subset by selecting first-pass benchmark rows where the mean Part B self-score (averaged over the three rubric dimensions) was strictly below 4 and the Part C output parsed to a valid 5-class label. From this pool we drew $n = 35$ alerts per model under Hamilton-quota stratification on the gold class (proportional to the low-self-rating pool's class histogram with random tie-breaking; the $n = 35$ alert IDs differ across models because each model's low-self-rating set differs). We chose 35 alerts per model because `GPT-5.4 no-think` had only 36 eligible alerts. Pool sizes were $\{118, 36, 83, 498, 494\}$ for the five tested runs respectively.

Each model was re-prompted with the original images and metadata plus its own first-pass JSON appended verbatim to the user message; the gold class was *not* revealed and the second-pass instruction asked only for a fresh classification under the same Part A/B/C schema as the benchmark. We define C/W as first-pass correct/wrong and C'/W' as second-pass correct/wrong, and report the four cell counts of the 2×2 paired-outcome table along with derived rates: **correction** $W \to C'$ as a fraction of the wrong stratum, **damage** $C \to W'$ as a fraction of the correct stratum, and **persistence** $W \to W'$ as a fraction of the wrong stratum. McNemar's exact two-sided test on the two off-diagonal cells gives a significance reading on the asymmetry of discordant flips.

Cross-model McNemar p-values are not directly comparable because the alert IDs differ across models; we treat them as a within-model significance reading rather than a pooled A/B.

Model	Acc_1	Acc_2	ΔAcc	Correction	Damage	Persistence	McNemar p	ΔMSRS	Out 2/1	Prior ref	S3 flip	2nd MSRS ≥ 4
GPT-5.4 high-think	45.71%	54.29%	+8.57	31.58%	18.75%	68.42%	0.508	+0.276	1.10	34.29%	17.14%	74.29%
GPT-5.4 no-think	28.57%	31.43%	+2.86	16.00%	30.00%	84.00%	1.000	+0.371	1.04	**74.29%**	11.43%	82.86%
Gemini 2.5 Flash no-think	17.14%	17.14%	0.00	0.00%	0.00%	100.00%	1.000	+0.429	1.03	40.00%	0.00%	48.57%
Claude Opus 4.7 think	51.43%	**80.00%**	**+28.57**	58.82%	**0.00%**	41.18%	**0.00195**	+0.048	0.97	57.14%	20.00%	11.43%
Claude Opus 4.7 nothink	37.14%	**80.00%**	**+42.86**	**68.18%**	**0.00%**	31.82%	$\mathbf{6.1 \times 10^{-5}}$	+0.038	1.12	54.29%	**34.29%**	14.29%

Table 19: **Second-rollout retry headline metrics ($n = 35$ per model).** Acc_1 and Acc_2 are first- and second-pass 5-class accuracy on the low-self-rating subset. Correction and Damage are the per-stratum flip rates as defined above. McNemar p is the exact two-sided test on the discordant cells. ΔMSRS is the change in mean self-rating from first to second pass. Out tok 2/1 is the ratio of mean second-pass to first-pass output tokens. Prior-ref rate is the fraction of second-pass rows whose text matches a regex for explicit reference to the model's prior answer. Stage3-flip is the fraction of rows whose Stage-3 subclass label changed between passes. 2nd MSRS ≥ 4 is the fraction of second-pass rows assigned mean self-score ≥ 4.

Model	$W \to C'$	$C \to W'$
GPT-5.4 high-think	6	3
GPT-5.4 no-think	4	3
Gemini 2.5 Flash no-think	0	0
Claude Opus 4.7 think	10	0
Claude Opus 4.7 nothink	15	0

Table 20: **Discordant pair counts $(W \to C', C \to W)$.** McNemar uses only these two cells; the symmetric (concordant) cells $C \to C'$ and $W \to W'$ do not enter the test.

F.2 Auxiliary figures

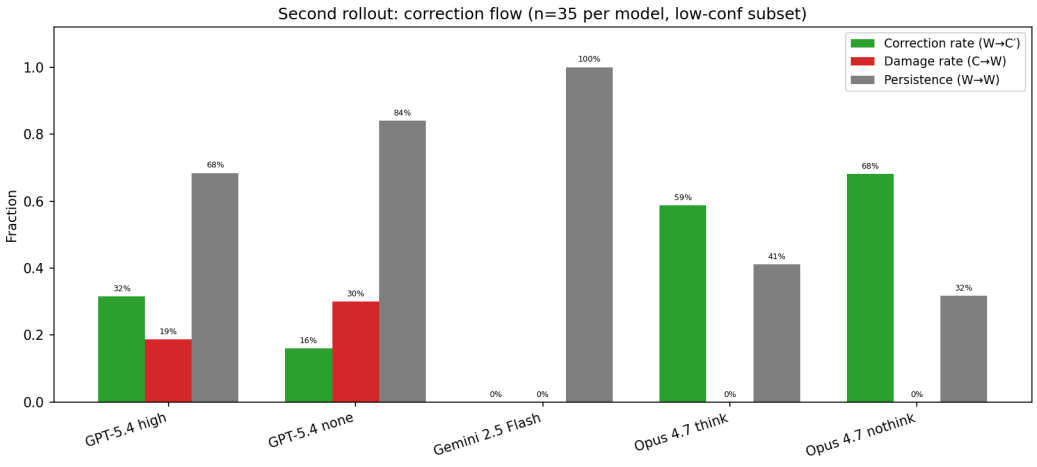

Figure 14: **Correction, damage, and persistence rates per model.** Tall green (correction) plus zero red (damage) on both `Claude Opus 4.7` variants; `GPT-5.4 high-think` trades correction against a visible damage segment.

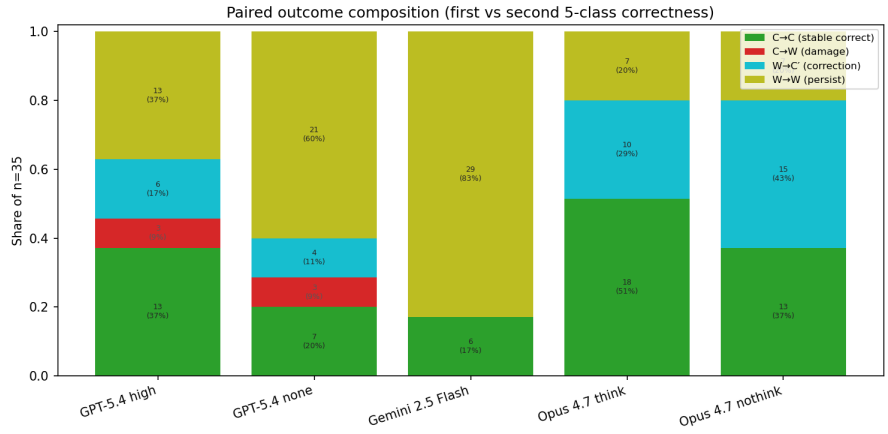

Figure 15: **Normalized 2×2 paired-outcome composition.** Stacked bars sum to 100% within each model. `Opus 4.7` runs devote a large slice to $W \to C'$ with no visible $C \to W'$ band; `Gemini 2.5 Flash no-think` is dominated by $W \to W'$ plus the rows that stay correct.

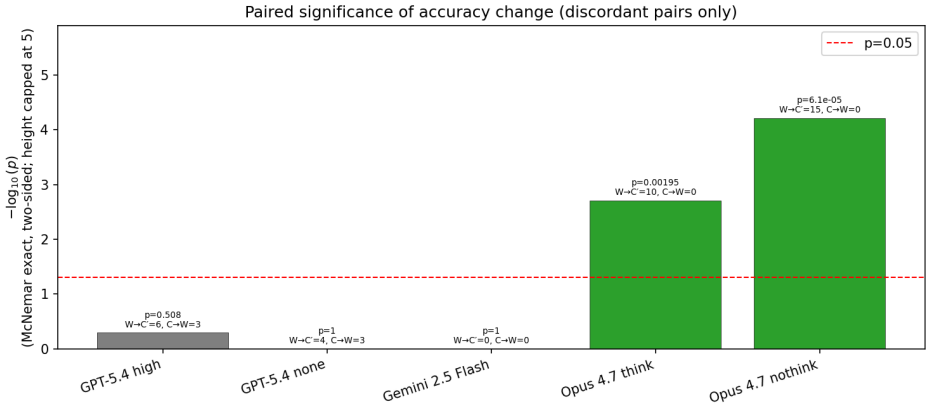

Figure 16: **McNemar two-sided p-values on $-\log_{10}$ scale.** The test evaluates the asymmetry of accuracy changes using only discordant pairs ($W \to C'$ vs. $C \to W$); the dashed line marks $p = 0.05$. Both `Claude Opus 4.7` variants sit above this threshold, indicating statistically significant improvements. `GPT-5.4` configurations remain below the line, while `Gemini 2.5 Flash` has zero discordant pairs, rendering the test degenerate ($p = 1$).

F.3 Per-family interpretation

`Claude Opus 4.7` (think and nothink). Largest ΔAcc on the cohort, second-pass accuracy 80% in both modes, and *zero damage* ($C \to W' = 0$ for both); McNemar $p = 2.0\times10^{-3}$ (think) and 6.1×10^{-5} (nothink). Mean ΔMSRS is small and positive ($+0.05$ / $+0.04$) and second-pass MSRS ≥ 4 stays low (11–14%): most second-pass rubric means remain below 4, even though accuracy reaches 80%. Stage-3 subclass flip rate is higher for nothink (34.29%) than think (20.00%), suggesting that direct-answer mode more often re-reads its first response and revises the fine subclass.

`GPT-5.4` (high-think and no-think). `High-think` delivers $+8.57$ pp net but with 18.75% damage—roughly one in five initially-correct low-self-rating rows regresses; McNemar $p = 0.51$, so the asymmetry is not significant despite a positive point estimate. `No-think` adds only $+2.86$ pp, with 30% damage and the highest prior-reference rate in the cohort (74.29%): explicit talk about the prior answer does not translate into reliable repair.

`Gemini 2.5 Flash no-think`. $\Delta\text{Acc} = 0$; correction and damage are both 0; every row stays in the same correctness class on retry (6 stayed correct, 29 stayed wrong; zero discordant pairs, McNemar $p = 1$ uninformative). Mean ΔMSRS is the largest in the cohort ($+0.43$) and second-pass MSRS ≥ 4 is near coin-flip (48.57%): the model increases its rationale ratings without changing its staged classifications. This separates a change in self-rating from classification repair.

F.4 Qualitative Case Studies of Second-Rollout Behavior

To supplement the quantitative findings, we provide two representative case studies from the second-rollout cohort. These examples illustrate the contrast between "successful classification repair" (Section F.4.1) and "static re-grading" (Section F.4.2). In both cases, models were provided with their original first-pass response and asked to perform a second review of the visual and metadata inputs to reanswer the questions without being informed of their prior correctness.

F.4.1 Evidence Re-weighting and Successful Self-Correction

This example, featuring `claude-opus-4.7-nothink`, demonstrates the highly effective repair process observed in frontier reasoning models. In the first pass, the model over-weights the temporal baseline (9 detections over 19 days) to predict *AGN*. During the second trial, the model exhibits genuine behavioral gain by re-evaluating the spatial metadata it previously under-weighted—specifically the 1.52 arcsec offset (`distpsnr1`) and the extended host indicators (`chinr`). This transition from *AGN* to *supernova* represents a successful revision driven by a re-reading of the scientific context rather than indiscriminate score inflation.

Gold: SN. Paired correctness: wrong→correct.

Below are the inputs fed to the model for its first attempt:



Figure 17: Science–Reference–Difference montage for `ZTF21abculew`.

System — attempt 1

`{general_prompt_block}`[a]

User — attempt 1

```
[ALERT IDENTIFIERS]
- Object ID: ZTF21abculew

[ZTF CANDIDATE FIELDS]
- fid: 2
```

[a] The general system prompt is provided in Section B.2.1.

```
- isdiffpos: t
- firstmjd: 59361.3
- magpsf: 19.9901
- sigmapsf: 0.184607
- fwhm: 1.93
- classtar: 0.963
- sgscore1: 0.5
- distpsnr1: 1.5198
- chinr: 6.499
- sharpnr: 0.708
- ndethist: 9
- ncovhist: 2664
- sgmag1: 20.1168
- srmag1: 19.0823
- simag1: -999
- szmag1: 19.3916
- nmtchps: 13
- deltajd: 19.0387

Field definitions and sentinel rules are in the system message.

Analyze this alert and return the JSON response.
```

The corresponding model response for the first attempt follows:

Output 1


```
{
  "Part A": {
    "filter_band": "r",
    "subtraction_sign": "positive",
    "magpsf": 19.9901,
    "sigmapsf": 0.184607,
    "ndethist": 9,
    "ncovhist": 2664
  },
  "Part B": {
    "key_evidence": "Positive r-band detection at magpsf ~19.99
    coincident with a PS1 source 1.52 arcsec away (sgscore1=0.5,
    ambiguous star/galaxy). classtar=0.963 suggests point-like
    morphology in the science image. ndethist=9 over deltajd=19 days
    indicates repeated detections at the same position. The reference
    image shows a visible source at the candidate location, and the
    difference image shows a positive residual.",
    "leading_interpretation_and_support": "Repeated positive-flux
    detections (ndethist=9) over ~19 days at a location with a
    persistent PS1 counterpart (sgscore1=0.5, ambiguous) favor AGN
    variability. The PS1 colors (g-r ~ 1.03) and ambiguous star/galaxy
     score are consistent with a compact galaxy nucleus, and sustained
     variability over weeks is characteristic of AGN rather than a
    single SN rise.",
    "alternative_analysis": "A supernova in the host galaxy at ~1.5
    arcsec offset is plausible given the positive difference residual
    and host presence; however, 9 detections spanning 19 days without
    clear decline, plus the near-nuclear offset consistent with
    distpsnr1, tilt the interpretation toward AGN. A variable star is
    less likely because sgscore1 is not close to 1 and the counterpart
     color is redder than typical stellar variables at this depth.",
    "self_score_key_evidence": 4,
    "self_score_leading_interpretation_and_support": 3,
    "self_score_alternative_analysis": 4
  },
```

```
  "Part C": {
    "stage1": "real_object",
    "stage2": "astrophysical",
    "stage3": "AGN"
  }
}
```

Below are the inputs fed to the model for its second attempt:

System — attempt 2

{second_rollout_ablation_block}[a]

User — attempt 2

```
[ALERT IDENTIFIERS]
- Object ID: ZTF21abculew

[ZTF CANDIDATE FIELDS]
- fid: 2
- isdiffpos: t
- firstmjd: 59361.3
- magpsf: 19.9901
- sigmapsf: 0.184607
- fwhm: 1.93
- classtar: 0.963
- sgscore1: 0.5
- distpsnr1: 1.5198
- chinr: 6.499
- sharpnr: 0.708
- ndethist: 9
- ncovhist: 2664
- sgmag1: 20.1168
- srmag1: 19.0823
- simag1: -999
- szmag1: 19.3916
- nmtchps: 13
- deltajd: 19.0387

Field definitions and sentinel rules are in the system message.

Analyze this alert and return the JSON response.

---

Below this metadata block, the live API user message also embeds the
    model's complete first-trial JSON (verbatim) under [YOUR FIRST
    ATTEMPT JSON --- FOR SELF-REVIEW ONLY], then asks for the second-
    trial JSON. Primary-manuscript placeholder for that payload: {
    embedded_first_attempt_json_block} (constructed by
    prompts_second_roll_out_ablation.build_user_prompt; archived under
     data_second_roll_out_ablation/priors/).
```

The corresponding model response for the second attempt follows:

Output 2

```
{
  "Part A": {
```

[a] The second-rollout prompt is provided in Appendix B.3.

```
    "filter_band": "r",
    "subtraction_sign": "positive",
    "magpsf": 19.9901,
    "sigmapsf": 0.184607,
    "ndethist": 9,
    "ncovhist": 2664
  },
  "Part B": {
    "key_evidence": "Positive r-band detection at magpsf ~19.99 with
    sigmapsf 0.18 (modest S/N). The science image shows a source near
    the center with a brighter neighbor visible to the lower-left; the
     reference shows the same field with a clearly resolved/extended-
    looking host at the candidate location. The difference image shows
     mostly noise with a faint positive residual near center. Metadata
    : distpsnr1=1.52 arcsec to nearest PS1 source (sgscore1=0.5,
    ambiguous), classtar=0.963 (point-like on science), ndethist=9
    over deltajd=19.0 days, chinr=6.5 (high chi suggests non-stellar/
    extended reference source).",
    "leading_interpretation_and_support": "The 1.5 arcsec offset from
    the PS1 counterpart, combined with chinr=6.5 (extended reference
    object) and a positive point-like residual in difference, is most
    consistent with a supernova offset from its host galaxy nucleus.
    Nine detections over 19 days is fully compatible with an SN light
    curve near peak/plateau. The point-like classtar on a faint magpsf
    ~20 source above an extended host favors a transient on a galaxy
    rather than nuclear AGN activity.",
    "alternative_analysis": "AGN is plausible if the offset is small
    enough that the candidate sits on the nucleus (sgscore1=0.5 is
    ambiguous), and sustained detections over weeks could reflect
    stochastic AGN variability. However, the 1.52 arcsec offset and
    high chinr (extended host) favor an off-nuclear SN. A variable
    star is disfavored because the reference source appears extended (
    high chinr) and sgscore1 is not stellar. On reflection, my first-
    pass AGN call under-weighted the offset and host extendedness, so
    I revise to supernova.",
    "self_score_key_evidence": 4,
    "self_score_leading_interpretation_and_support": 4,
    "self_score_alternative_analysis": 4
  },
  "Part C": {
    "stage1": "real_object",
    "stage2": "astrophysical",
    "stage3": "supernova"
  }
}
```


F.4.2 Static Persistence and Score Inflation

This example illustrates a mismatch between self-rating and performance characteristic of the `Gemini 2.5 Flash` model. In the first pass, the model correctly identifies a point-like source but settles on a variable star interpretation, ignoring the possibility of a supernova despite acknowledging a low `sgscore1` for the host. Upon the second rollout, the model explicitly mentions "refining the phrasing" and even increases its self-scores for evidence and logic (from 4/4/3 to 5/4/4). However, it **fails to revise the final classification**, demonstrating that a model sounding more sure of itself does not always lead to improved scientific results.

Gold: SN. Paired correctness: wrong→wrong.

Below are the inputs fed to the model for its first attempt:

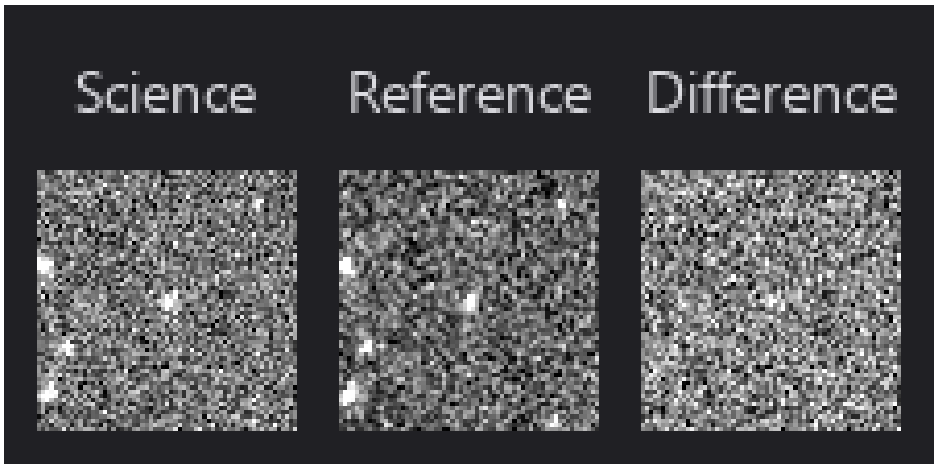

Figure 18: Science–Reference–Difference montage for `ZTF24ablarbs`.

System — attempt 1

```
{general_prompt_block}[a]
```

User — attempt 1

```
[ALERT IDENTIFIERS]
- Object ID: ZTF24ablarbs

[ZTF CANDIDATE FIELDS]
- fid: 2
- isdiffpos: t
- firstmjd: 60585.2
- magpsf: 21.4729
- sigmapsf: 0.214785
- fwhm: 2.35
- classtar: 0.99
- sgscore1: 0.008
- distpsnr1: 0.0349517
- chinr: 2.314
- sharpnr: 0.366
- ndethist: 4
- ncovhist: 1787
- sgmag1: 22.1523
- srmag1: 21.6716
- simag1: 20.9498
- szmag1: 20.739
- nmtchps: 2
- deltajd: 1.13814

Field definitions and sentinel rules are in the system message.

Analyze this alert and return the JSON response.
```

The corresponding model response for the first attempt follows:

[a] The general system prompt is provided in Section B.2.1.

Output 1

```
```json
{
  "Part A": {
    "filter_band": "r",
    "subtraction_sign": "positive",
    "magpsf": 21.4729,
    "sigmapsf": 0.214785,
    "ndethist": 4,
    "ncovhist": 1787
  },
  "Part B": {
    "key_evidence": "The difference image shows a clear, point-like
    positive residual at the center, indicating a new or brighter
    source. The science image shows a faint point source at the center
     that is not present in the reference image. The fwhm of 2.35
    pixels is consistent with a point source. The classtar of 0.99
    suggests a stellar morphology. The nearest PS1 source (distpsnr1 =
     0.03 arcsec) has a very low sgscore1 (0.008), suggesting it is
    extended or a galaxy, but this is contradicted by the classtar and
     the image morphology of the transient. The chinr (2.314) and
    sharpnr (0.366) values for the nearest reference source are
    somewhat high for a perfect point source, but not definitively
    indicative of an artifact. The ndethist of 4 and ncovhist of 1787
    suggest repeated observations of this field, and the deltajd of
    1.13814 indicates multiple detections over a short period.",
    "leading_interpretation_and_support": "The alert is most likely a
    variable star. The clear point-like detection in the difference
    image, coupled with the absence of a counterpart in the reference
    image, confirms it as a real transient. The high classtar (0.99)
    strongly supports a stellar morphology. While the sgscore1 of the
    nearest PS1 source is low (0.008), its very small distance (0.03
    arcsec) suggests it is likely the same object, and the low
    sgscore1 might be due to blending or issues with the PS1
    classification for faint sources, or it could be a background
    galaxy with a foreground variable star. However, given the clear
    point-like nature in the ZTF images, a variable star is more
    plausible than an AGN, especially with no clear host galaxy
    visible at this faint magnitude. The ndethist of 4 and deltajd of
    ~1 day are consistent with a variable star that has been detected
    a few times.",
    "alternative_analysis": "An alternative interpretation is that
    this could be an AGN. The low sgscore1 for the nearby PS1 source
    could indicate a faint host galaxy, and the point-like nature of
    the transient could be the varying nucleus. However, without a
    clear extended host visible in the ZTF images, and given the
    strong stellar classification from classtar, a variable star is a
    stronger candidate. It is less likely to be a supernova given the
    lack of a clear host galaxy and the relatively short deltajd,
    though a very faint, distant supernova cannot be entirely ruled
    out without more information. It is not an asteroid due to the
    lack of a dipole signature in the difference image and the
    stationary appearance between science and reference images. It is
    also unlikely to be bogus due to the clear, well-defined positive
    residual.",
    "self_score_key_evidence": 4,
    "self_score_leading_interpretation_and_support": 4,
    "self_score_alternative_analysis": 3
  },
  "Part C": {
    "stage1": "real_object",
    "stage2": "astrophysical",
```
```

```
    "stage3": "variable_star"
  }
}
```

Below are the inputs fed to the model for its second attempt:

System — attempt 2

{second_rollout_ablation_block}[a]

User — attempt 2

```
[ALERT IDENTIFIERS]
- Object ID: ZTF24ablarbs

[ZTF CANDIDATE FIELDS]
- fid: 2
- isdiffpos: t
- firstmjd: 60585.2
- magpsf: 21.4729
- sigmapsf: 0.214785
- fwhm: 2.35
- classtar: 0.99
- sgscore1: 0.008
- distpsnr1: 0.0349517
- chinr: 2.314
- sharpnr: 0.366
- ndethist: 4
- ncovhist: 1787
- sgmag1: 22.1523
- srmag1: 21.6716
- simag1: 20.9498
- szmag1: 20.739
- nmtchps: 2
- deltajd: 1.13814

Field definitions and sentinel rules are in the system message.

Analyze this alert and return the JSON response.

---

Below this metadata block, the live API user message also embeds the
    model's complete first-trial JSON (verbatim) under [YOUR FIRST
    ATTEMPT JSON --- FOR SELF-REVIEW ONLY], then asks for the second-
    trial JSON. Primary-manuscript placeholder for that payload: {
    embedded_first_attempt_json_block} (constructed by
    prompts_second_roll_out_ablation.build_user_prompt; archived under
     data_second_roll_out_ablation/priors/).
```

The corresponding model response for the second attempt follows:

Output 2

```
```json
{
  "Part A": {
    "filter_band": "r",
```

[a] The second-rollout prompt is provided in Appendix B.3.
```

```
    "subtraction_sign": "positive",
    "magpsf": 21.4729,
    "sigmapsf": 0.214785,
    "ndethist": 4,
    "ncovhist": 1787
  },
  "Part B": {
    "key_evidence": "The difference image shows a clear, point-like
    positive residual at the center, indicating a new or brighter
    source. The science image shows a faint point source at the center
     that is not present in the reference image. The fwhm of 2.35
    pixels is consistent with a point source. The classtar of 0.99
    strongly suggests a stellar morphology. The nearest PS1 source (
    distpsnr1 = 0.03 arcsec) has a very low sgscore1 (0.008),
    suggesting it is extended or a galaxy. The chinr (2.314) and
    sharpnr (0.366) values for the nearest reference source are
    somewhat high for a perfect point source, but not definitively
    indicative of an artifact. The ndethist of 4 and ncovhist of 1787
    suggest repeated observations of this field, and the deltajd of
    1.13814 indicates multiple detections over a short period.",
    "leading_interpretation_and_support": "The alert is most likely a
    variable star. The clear point-like detection in the difference
    image, coupled with the absence of a counterpart in the reference
    image, confirms it as a real transient. The high classtar (0.99)
    strongly supports a stellar morphology. While the sgscore1 of the
    nearest PS1 source is low (0.008), its very small distance (0.03
    arcsec) suggests it is likely the same object. The discrepancy
    between classtar and sgscore1 could be due to the PS1 source being
     a faint, unresolved galaxy with a foreground variable star, or
    issues with PS1 classification for faint sources. Given the clear
    point-like nature in the ZTF images and the lack of a clearly
    extended host, a variable star is more plausible than an AGN. The
    ndethist of 4 and deltajd of ~1 day are consistent with a variable
     star that has been detected a few times.",
    "alternative_analysis": "An alternative interpretation is that
    this could be an AGN. The low sgscore1 for the nearby PS1 source
    could indicate a faint host galaxy, and the point-like nature of
    the transient could be the varying nucleus. However, without a
    clear extended host visible in the ZTF images, and given the
    strong stellar classification from classtar, a variable star is a
    stronger candidate. It is less likely to be a supernova given the
    lack of a clear host galaxy and the relatively short deltajd,
    though a very faint, distant supernova cannot be entirely ruled
    out without more information. It is not an asteroid due to the
    lack of a dipole signature in the difference image and the
    stationary appearance between science and reference images. It is
    also unlikely to be bogus due to the clear, well-defined positive
    residual. My previous analysis was largely consistent, but I've
    refined the phrasing to better articulate the discrepancy between
    classtar and sgscore1.",
    "self_score_key_evidence": 5,
    "self_score_leading_interpretation_and_support": 4,
    "self_score_alternative_analysis": 4
  },
  "Part C": {
    "stage1": "real_object",
    "stage2": "astrophysical",
    "stage3": "variable_star"
  }
}
```

G HUMAN-MODEL GRADING COMPARISON: REPRESENTATIVE CASE STUDIES

To illustrate the model self-assessment analysis discussed in Section 5, we performed a multi-expert qualitative audit on a difficult alert: `ZTF26aargnnp`. This alert was selected because it elicited a wide range of performance levels and reasoning strategies across the 13 model configurations. We established the human reference based on the consensus of five independent graders: two senior experts ($\geq$ 20 years experience), two intermediate researchers (3–10 years), and one early-career astronomer (3 years). Furthermore, character-level qualitative markup (distinguishing correct, partially flawed, and hallucinatory claims) was provided by a senior expert with 20+ years of professional experience in time-domain astronomy.

On this alert, we identify an "alternative-reasoning advantage": the share of expert-verified "fully correct" (green) characters in the models' `alternative analysis` fields (Q3) is 44.48%, nearly double the 23.19% observed in the `leading interpretation` fields (Q2). This indicates that, for this alert, the alternative-analysis text receives a higher proportion of fully correct annotations under expert review. Alongside the findings from our 1,500-alert benchmark, the individual examples in this section highlight two patterns: reasoning quality differs between leading and alternative interpretations, and model rationale self-ratings do not always match the judgment of human experts.

G.1 EXPERT GRADING OF LEADING AND ALTERNATIVE INTERPRETATIONS

To compare expert-marked reasoning in the two response fields, we first analyzed the character-level expert markup across all 13 model configurations for alert ZTF26aargnnp. When pooling the results, we find that the share of expert-verified "fully correct" (green) characters in the models' `alternative analysis` fields (Q3) is 44.48%, nearly double the 23.19% observed in the `leading interpretation` fields (Q2). Specifically, the composition shifts from 23.19% green, 47.89% yellow, and 20.97% red in Q2 to 44.48% green, 28.90% yellow, and 21.20% red in Q3. Notably, because the red share remains effectively flat across both fields ($\approx 21\%$), this gain represents a yellow-to-green shift rather than a red-to-green correction.

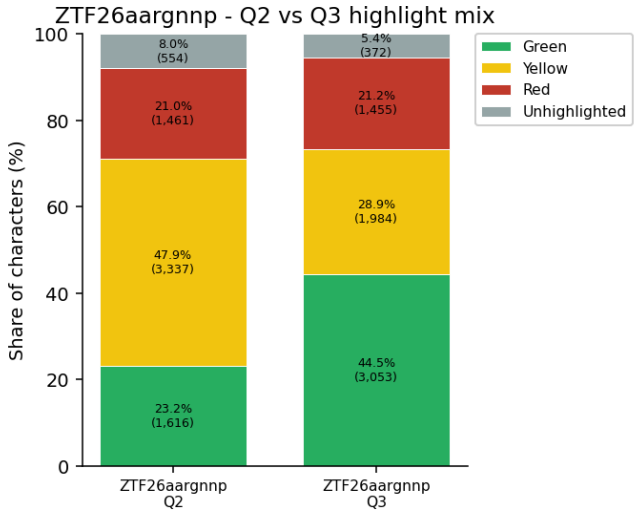

Figure 19: **Q2 vs. Q3 highlight composition on ZTF26aargnnp.** Stacked bars give the share of expert-marked green ("fully correct"), yellow ("partially correct"), red ("flawed or hallucinatory"), and unhighlighted characters in the model's `leading_interpretation` (Q2) and `alternative_analysis` (Q3) fields, pooled over the 13 audited model blocks. The green share rises from 23.19% in Q2 to 44.48% in Q3 while the red share stays near 21%.

G.2 AGREEMENT BETWEEN SELF-RATINGS AND EXPERT GRADES ON ONE ALERT

Across the 12 model runs with parseable Part B fields[4], the Pearson correlation between the blended human grade and the model's mean Part B self-score is effectively zero ($r = -0.023, p = 0.94$). As shown in Figure 22, several models (e.g., `Opus 4.7 nothink`, `Gemini 2.5 Pro high`) self-rate two to three scale points above the human consensus. Conversely, GPT-5.4 none sits closest to the $y = x$ line in Figure 21, with self-ratings closely aligning with a human mean.

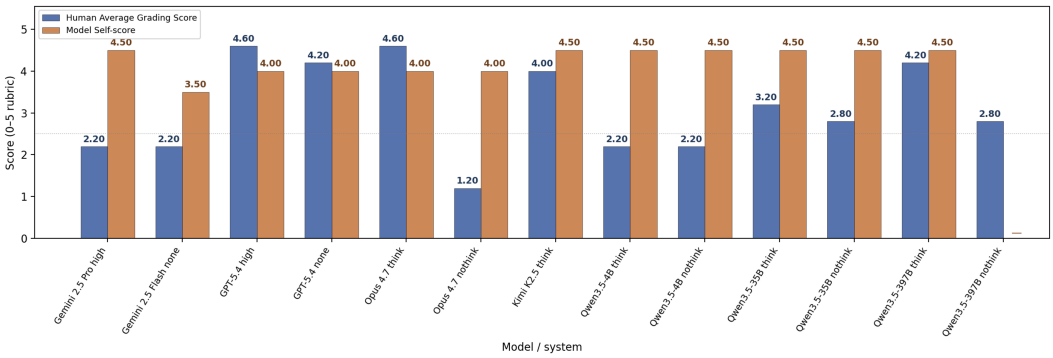

Figure 20: **Self-score vs. human mean per model on ZTF26aargnnp.** Paired bars show the blended human grade (five raters) and the model's mean Part B self-score (`leading_interpretation` + `alternative_analysis`) for each of the 13 audited runs.

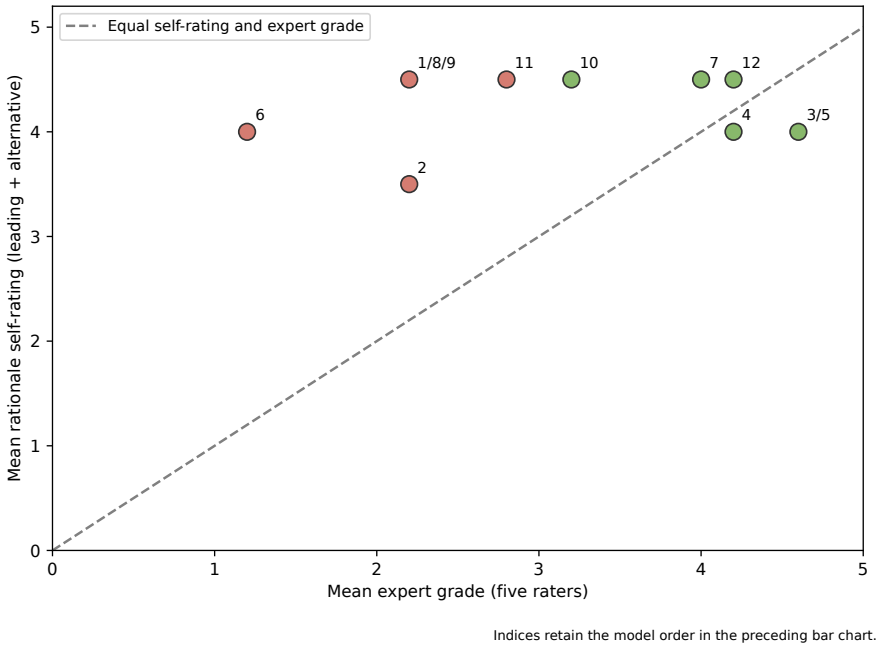

Figure 21: **Self-rating versus expert-grade agreement on ZTF26aargnnp.** Mean Part B self-score (vertical axis) against the blended human mean (horizontal axis) for the 12 models with complete Part B fields. The dashed $y = x$ line marks equality of self-rating and expert grade; marker fill encodes Part C correctness (green = correct class, red = incorrect class, gray = unparsed).

[4]One model (`Qwen3.5-397B-A17B nothink`) is excluded from score-agreement summaries because it did not output `Part B` self-score fields in this example; however, its reasoning text was fully audited for character-level highlights.

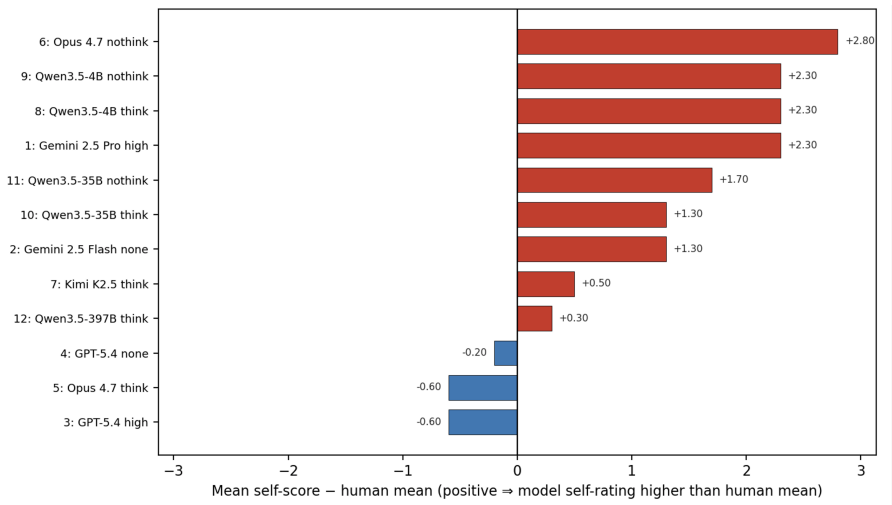

Figure 22: **Self-rating bias on ZTF26aargnnp.** Mean Part B self-score minus the blended human grade, sorted. Positive values (red bars) indicate the model rates its own reasoning above the human consensus; negative values (blue bars) indicate the model is stricter than the experts.

G.3 Three representative model responses

To provide qualitative grounding for the population-level statistics, we present three model responses that represent the full spectrum of expert grades for alert ZTF26aargnnp. Each case study includes the human mean score, the model's mean `Part B` two fields' self-score, and `Part C` correctness. The reasoning text (`Q2`+`Q3`) features character-level markup from a senior expert, where green text denotes "fully correct" claims, yellow text denotes "partially correct" analysis, and red text denotes "clearly flawed or hallucinatory" statements.



Figure 23: Science–Reference–Difference montage for `ZTF26aargnnp`.

G.3.1 Tier 1 — `Claude Opus 4.7 think`

This example demonstrates high scientific alignment, correctly identifying the 'mover' signature by reconciling high historical coverage (`ncovhist=1489`) with a single-epoch detection—logic that is almost entirely validated by expert character-level markup.

Claude Opus 4.7 think — raw output (Human Grading Score: 4.60 & Model Self-score: 4.00)


```
(model=claude-opus-4-7, n_prompt_tokens=4513, n_answer_tokens=1065)

{
  "Part A": {
    "filter_band": "r",
    "subtraction_sign": "positive",
    "magpsf": 17.7693,
    "sigmapsf": 0.0461362,
    "ndethist": 1,
    "ncovhist": 1489
  },
  "Part B": {
    "key_evidence": "Science shows a bright, point-like source at
    center; the reference/template has no counterpart there; the
    difference shows a clean positive residual. classtar=0.983 and
    fwhm=2.49 pix are consistent with a stellar PSF. ndethist=1 with
    ncovhist=1489 means this field has been imaged many times but the
    source was detected only once (deltajd=0). The nearest PS1 object
    is 24.4\" away (distpsnr1), so there is no identified host at the
    candidate position. isdiffpos=t confirms a brightening, magpsf
    =17.77 in r.",
    "leading_interpretation_and_support": "Most consistent with a real
    solar-system object (Asteroid).The source is point-like and
    appears on otherwise blank sky, with no convincing stationary
    source or galaxy in the template.The very low same-position
    history (ndethist=1) despite high coverage (ncovhist=1489) fits a
    moving object better than a fixed astrophysical variable.The
    positive, compact subtraction residual argues against a pure
    subtraction artifact.",
    "alternative_analysis": "A hostless or very faint-host supernova
    is the main astrophysical alternative, since a new positive point
    source on a blank template can also be an extragalactic
    transient.However, the lack of any visible host/persistent
    counterpart and the single same-position detection make that less
    compelling here than a solar-system interpretation.  AGN and
    ordinary variable star interpretations are weak because there is
    no close persistent nucleus or stellar counterpart at the alert
    position.",
    "self_score_key_evidence": 4,
    "self_score_leading_interpretation_and_support": 4,
    "self_score_alternative_analysis": 4
  },
  "Part C": {
    "stage1": "real_object",
    "stage2": "solar_system",
    "stage3": "N/A"
  }
}
```


G.3.2 TIER 2 — QWEN3.5-35B NOTHINK

This model typifies the 'biased middle' behavior, where accurate metadata grounding fails to translate into a correct classification due to a breakdown in positional reasoning, resulting in a high self-score that masks a flawed scientific conclusion.

Qwen3.5-35B nothink — raw output (Human Grading Score: 2.80 & Model Self-score: 4.50)


```
(model=Qwen/Qwen3.5-35B-A3B, n_prompt_tokens=not recorded,
    n_answer_tokens=526)
```

```
{
  "Part A": {
    "filter_band": "r",
    "subtraction_sign": "positive",
    "magpsf": 17.7693,
    "sigmapsf": 0.0461362,
    "ndethist": 1,
    "ncovhist": 1489
  },
  "Part B": {
    "key_evidence": "The difference image shows a clear, point-like
    positive residual at the center. The source has a high FWHM (2.49)
     and high sharpness (0.107), consistent with a stellar point
    source. The PS1 neighbor is at 24.4 arcsec (sgscore1=0.99),
    indicating a nearby star but not a direct match to the transient
    position. The history shows ndethist=1 (first detection) despite
    high coverage (ncovhist=1489).",
    "leading_interpretation_and_support": "Variable Star.The high
    classtar (0.983) and sharpness indicate a stellar object.The high
    ncovhist (1489) confirms the telescope has repeatedly observed
    this sky position, yet ndethist is only 1.This implies the object
    is faint or variable and was only detected above threshold in this
    specific epoch.The Science vs Template comparison shows the source
    is present in both but significantly brighter in Science,
    consistent with a variable star (e.g., RR Lyrae, Cepheid) or a
    flare.",
    "alternative_analysis": "Asteroid.The low ndethist (1) is a common
    signature for moving objects that are not linked across
    epochs.However, the Science and Template images show the source at
    the same geometric center in both frames, with no visible streak
    or offset relative to the background stars.  A moving asteroid
    would typically show a positional shift between Science and
    Template or a streak in the difference image if the motion was
    significant during the exposure.The lack of motion evidence makes
    the asteroid hypothesis less likely than a variable star.",
    "self_score_key_evidence": 5,
    "self_score_leading_interpretation_and_support": 5,
    "self_score_alternative_analysis": 4
  },
  "Part C": {
    "stage1": "real_object",
    "stage2": "astrophysical",
    "stage3": "variable_star"
  }
}<|im_end|>
```


G.3.3 TIER 3 — GEMINI 2.5 FLASH

This response highlights a hallucination failure mode, where the model justifies its incorrect classification by explicitly claiming the source is present in the reference cutout despite clear visual evidence of its absence.

Gemini 2.5 Flash — raw output (Human Grading Score: 2.20 & Model Self-score: 3.50)


```
(model=gemini-2.5-flash, n_prompt_tokens=3312, n_answer_tokens=642)

```json
{
  "Part A": {
    "filter_band": "r",
    "subtraction_sign": "positive",
```
```

```
    "magpsf": 17.7693,
    "sigmapsf": 0.0461362,
    "ndethist": 1,
    "ncovhist": 1489
  },
  "Part B": {
    "key_evidence": "The science image shows a bright, point-like
    source at the center that is absent in the reference image,
    resulting in a clear positive residual in the difference image.
    The FWHM of 2.49 pixels is consistent with a point source. The
    classtar (0.983) and sgscore1 (0.993333) values are both very high
    , strongly indicating a stellar morphology. The ndethist is 1,
    suggesting this is a first detection. The nearest PS1 source is
    24.3978 arcsec away, which is a significant distance, implying the
     PS1 source is likely unrelated to the transient itself, or at
    least not a direct counterpart.",
    "leading_interpretation_and_support": "Most consistent with a
    Galactic variable star:  the candidate appears as a stellar
    (point-like) source already present in the reference, brightening
    enough to produce a positive difference residual.  High classtar,
    low sharpnr, and the lack of any extended host nearby (nearest PS1
    source is far at ~24\") favor a stellar variable rather than an SN
    or AGN.",
    "alternative_analysis": "A supernova is unlikely because the
    source is clearly present in the reference and shows stellar
    morphology with no underlying galaxy.A CV-type outburst or other
    dwarf nova is plausible given a single new detection after many
    visits.  AGN is disfavored due to no resolved host.An asteroid is
    ruled out since the source is also in the template at the same
    position.",
    "self_score_key_evidence": 5,
    "self_score_leading_interpretation_and_support": 4,
    "self_score_alternative_analysis": 3
  },
  "Part C": {
    "stage1": "real_object",
    "stage2": "astrophysical",
    "stage3": "variable_star"
  }
}
```