HONKAKUBENCH: CAN AI SOLVE FAIR-PLAY MURDER MYSTERIES?

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating whether a large language model can *reason*, rather than rely solely on retrieval or pattern matching, requires assessing not only its final answer but also the validity of the reasoning steps supporting it. Honkaku mysteries offer a natural setting for this evaluation as fair-play puzzles challenge readers to explain seemingly impossible events using clues in the story. We introduce HonkakuBench, a benchmark of 70 mystery puzzles for evaluating evidence-grounded reconstruction, paired with unique official solutions written by human experts. Given a complete case file and any accompanying diagrams, models must produce an open-ended solution explaining what happened, how it was possible, and which evidence supports the reconstruction. Across eight reasoning models evaluated through a common harness, GPT-5.6-SOL and FABLE 5 recover 61.3% and 59.3% of rubric credit on average but receive full credit on only 6 and 8 of 70 cases. All eight models score 18.1–26.8 points lower on the cases with visual evidence than on the text-only cases. Our case studies reveal how partial progress fails to build a coherent reconstruction: a model may recover the mechanism while naming the wrong culprit, or propagate one misread clue through consistent arguments. HONKAKUBENCH thus tests whether models can turn scattered textual and visual clues into one explanation consistent with the entire case.

1 INTRODUCTION

Since the advent of reasoning capabilities in large language models, a wide variety of benchmarks have been proposed to evaluate the reasoning of these models. These benchmarks now span broad knowledge and multitask understanding, grade-school mathematics, science question answering, implicit multistep inference, and graduate-level scientific reasoning (Hendrycks et al., 2021; Cobbe et al., 2021; Clark et al., 2018; Geva et al., 2021; Rein et al., 2023). Despite this breadth and increasing difficulty, many of these benchmarks ultimately reduce a response to whether its final answer is correct. For many applications, such outcome-focused evaluation is appropriate since the final result determines utility.

Nevertheless, when it comes to assessing the reasoning capabilities of large language models, we need to go beyond the final answer since a correct answer alone does not establish that a model reasoned correctly or followed a valid chain of inference. The model may have guessed, matched a familiar pattern, or produced a plausible explanation that does not faithfully reflect how it reached the answer (Turpin et al., 2023). Conversely, a wrong final answer can conceal substantial progress on the mechanism that produced it. Assessing reasoning therefore requires more than checking the model's final answer; it requires examining how the evidence in the problem is turned into an explanation.

On the other hand, a typical mystery story brings this reasoning problem into focus as a model must perform three key steps in tandem to solve it. First, it must notice which details matter as any decisive clue may be only one sentence among many, and become meaningful only after a later observation changes how it should be read. This resembles needle-in-a-haystack evaluation, where a target fact must be retrieved from long distractor text (Hsieh et al., 2024), with a crucial difference that the model is not told what needle to look for. Instead, it must decide which details might be clues, keep plausible interpretations, and revisit them as new evidence appears.

Additionally, solving a mystery can require interpreting visual evidence alongside the text. Detective mysteries often use floor plans, maps, and other diagrams as evidence. Such a diagram is part of

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.

the mystery as its meaning depends on statements spread throughout the different narrative, and any proposed solution must be consistent with the visual evidence it shows. A solver must therefore combine textual and visual evidence in the same reconstruction.

The clues also must fit one coherent solution. Times, locations, causes, what each character knows, and often real-world constraints, all place limits on what could have happened. A clue may support several interpretations when it first appears, so the model must test each one against the rest of the case and revise its solution when the pieces do not fit. Unlike a targeted multi-hop question with a predefined answer and annotated supporting facts (Yang et al., 2018), a mystery does not specify which facts will matter or how they should be assembled. The model must work out which pieces to connect and explain why the resulting solution explains the whole story.

Fair-play detective mysteries therefore bring these three challenges together while also providing a clear evaluation framework. We introduce HONKAKUBENCH, a benchmark of 70 fair-play mysteries prepared with a team of mystery-writing experts, with official solutions and scoring rubrics supplied by the original authors. Under the *honkaku*[1] tradition, each model receives the complete mystery case upfront including the narrative and any accompanying visual evidence, and produces a whole-case free-form reconstruction. The model must decide which details are clues, resolve their possible meanings across text and diagrams, and assemble them into a complete reconstruction of the culprit, mechanism, and supporting evidence. The evaluation target is criterion-level credit against the author-provided rubric, covering the culprit, mechanism, supporting evidence, and the intermediate deductions required by the case.

Across eight reasoning models, GPT-5.6-SOL and FABLE 5 form the leading pair under the primary judge, recovering 61.3% and 59.3% of the rubric on average. Yet they receive full credit on only 6 and 8 of 70 cases. This gap is the central finding of our study: even the strongest reasoning models often recover many of the right pieces but fail to fit them into one complete reconstruction. The pattern remains under alternative judges and further analysis shows that all eight models perform worse when a case includes visual evidence. In individual cases analysis, we discovered that a model often explain the mechanism correctly yet accuse the wrong culprit, or build an otherwise coherent argument on a clue it misread or misinterpreted.

We summarize our contributions as follows:

- We introduce HONKAKUBENCH, 70 fair-play mysteries prepared with a team of mystery-writing experts, including complete case packages, author-provided solutions and rubrics, and visual evidence in 46 cases.[2]
- We formulate mystery solving as whole-case free-form reconstruction and evaluate it with criterion-level credit for conclusions, mechanisms, evidence, and connecting deductions.
- Across eight reasoning models, we find a substantial gap between recovering correct components and completing the reconstruction: the leading models recover roughly 60% of the rubric on average but receive full credit on only 6/70 and 8/70 cases. The broad finding persists across alternative judges and a matched harness ablation.
- We diagnose this gap across evidence modality and individual rubric criteria: every model scores lower on cases with both textual and visual evidence, and the case studies show how a wrong culprit or one misread clue can invalidate an otherwise substantial reconstruction.

2 RELATED WORK

Answer correctness does not expose the reasoning process. Benchmarks such as MMLU and GPQA score final answers (Hendrycks et al., 2021; Rein et al., 2023). Process reward models evaluate intermediate mathematical steps (Lightman et al., 2024; Wang et al., 2024b), but usually assume an ordered derivation or a verifiable answer. Mystery solutions allow many valid inference orders and equivalent phrasings, so HONKAKUBENCH instead scores semantic components including the culprit, mechanism, evidence, and connecting deductions, against an author-provided rubric.

[1] *Honkaku*, often translated as "orthodox," refers to a Japanese tradition of puzzle-centered detective fiction built around fair play: readers receive the clues needed to reach the intended solution before the reveal.

[2] Artifact available at `https://anonymous.4open.science/r/honkaku-bench-A491`

Detective benchmarks isolate parts of mystery solving. True Detective presents human-authored mysteries but uses a multiple-choice answer as its required output (Del & Fishel, 2023), while WhoDunIt asks for the culprit in public-domain mystery fiction (Gupta, 2025). Because WhoDunIt draws from public-domain texts, prior exposure is a particular concern. MuSR generates complete narratives from latent reasoning trees, providing controlled multistep structure while retaining answer accuracy as the evaluation target (Sprague et al., 2024). DetectBench tests implicit-evidence discovery through multiple-choice questions (Gu et al., 2024), while DetectiveQA pairs four-way questions over long novel contexts with human-annotated reference steps (Xu et al., 2025). TurnaboutLLM uses testimony and evidence descriptions from detective games and asks for the contradicting pair (Yuan et al., 2025). Together, these benchmarks score an answer, clue set, or local contradiction rather than a complete free-form reconstruction.

Interactive benchmarks mix reasoning with evidence acquisition. WhodunitBench evaluates multimodal agents in murder-mystery games through perception questions, decisions, and game outcomes (Xie et al., 2024). In the closest open-ended setting, Watson & Holmes lets a model choose locations, receive evidence incrementally, and answer unconstrained questions scored by an automated rubric (Leelawat & Griffin, 2026). Both settings evaluate evidence acquisition as part of the interaction. HONKAKUBENCH instead provides the complete case upfront and scores one final reconstruction, isolating how a model interprets and connects a fixed body of evidence (Table 1).

Open-ended reconstruction requires rubric-based evaluation. At scale, rubric scoring of free-form answers often relies on an LLM judge and therefore inherits concerns about evaluator bias and reliability (Zheng et al., 2023; Wang et al., 2024a; Tan et al., 2025). We use one primary judge and report a judge-sensitivity analysis on identical submissions.

3 HONKAKUBENCH

3.1 TASK FORMULATION

HONKAKUBENCH is a end-to-end mystery-solving benchmark built on the fair-play premise: all evidence needed for the solution appears before the final reveal. A benchmark instance, *called a case*, contains one or more narrative documents describing the mystery and any accompanying visual evidence, such as floor plans, maps, grids, or schematics. The model receives the complete case package and is required to output a free-form *whole-case reconstruction* that addresses all case-specific questions in a single solution document. The required components are case-specific and may include the culprit, mechanism, chronology, supporting evidence, or explanation of any context presented in the case. As a result, the model's final reconstruction must account for the complete solution and connect its conclusions to the evidence rather than answer one isolated question.

3.2 CASE SOURCES AND COMPOSITION

The benchmark contains 70 submissions to a national mystery-writing competition, selected and prepared with a team of mystery-writing experts. With the contributing experts' consent, we package each submission as one case containing the complete puzzle materials, official solution, and scoring rubric. Before inclusion, the expert team reviewed each case and verified that, under its stated rules and narrative assumptions, the available evidence supports a single complete solution. For each of the scoring rubrics, we further annotate cases with reasoning demands categorized into five categories. **Spatial** reasoning covers routes, visibility, geometry, and architectural constraints. **Temporal** reasoning covers event order, timing windows, alibis, and delayed actions. **Physical** reasoning covers material, mechanical, biological, chemical, and forensic mechanisms. **Identity** reasoning covers disguise, perspective, role, representation, and mistaken identity. **Rule-based** reasoning covers explicit games, numerical relations, logical rules, and constraint systems.

These reasoning coverage tags are multilabel because a single case can combine several reasoning demands. Figure 1 summarizes their composition together with the solver task and evaluation path. Furthermore, HONKAKUBENCH contains 46 cases with visual evidence such as floor plans, maps, grids, or schematics that form an integral part of the mystery. For instance, a floor plan may determine visibility or movement, while a grid or schematic may rule out a plausible mechanism. Solving these cases therefore requires reconciling the depicted structure presented in the visual evidences with claims distributed throughout the narrative.

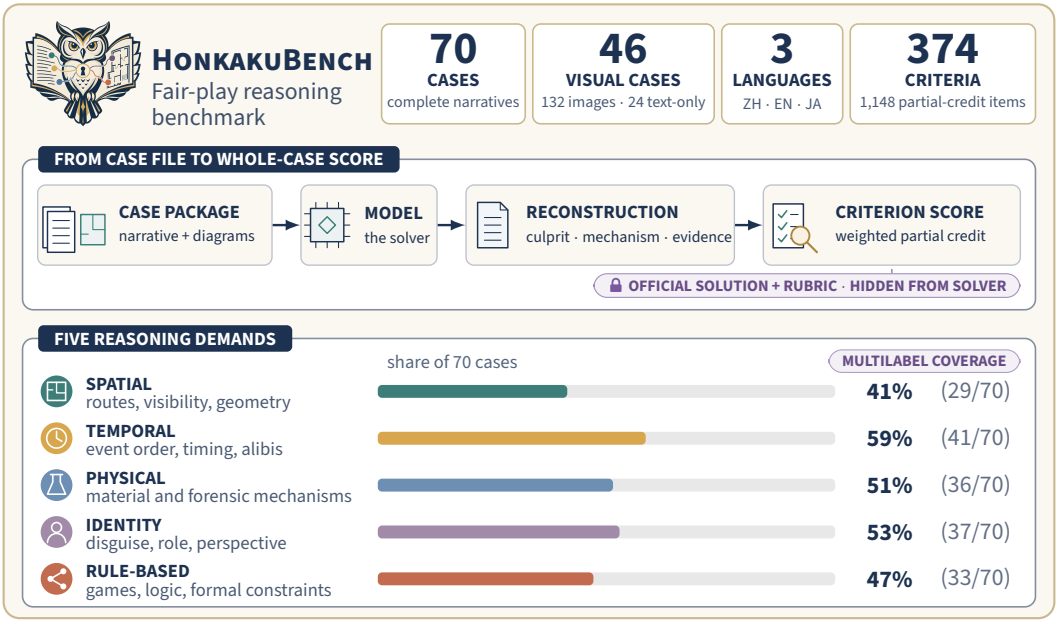

Figure 1: Overview of HONKAKUBENCH composition and evaluation. "Visual cases" denotes 46 cases containing 132 diagrams or other images; the remaining 24 cases are text-only. The official solution and author-verified rubric are hidden from the solver and used for criterion-level grading.

3.3 RUBRICS AND CRITERION-LEVEL SCORING

For each case, the original author supplies the official solution and scoring rubric. We restructure each rubric once into independently scorable criteria, after which the original author verifies every final criterion and weight. Each verified rubric is then stored and reused, remaining unchanged across all model evaluations. Specifically, let criterion i have maximum weight w_i and receive $a_i \in [0, w_i]$ points. Because rubric totals vary across cases, we report the normalized case score as

$$S = 100 \cdot \frac{\sum_{i=1}^{K} a_i}{\sum_{i=1}^{K} w_i}. \tag{1}$$

This decomposition distinguishes partial reconstruction from complete solution as a response can identify the culprit while losing credit for an incompatible mechanism or unsupported evidential chain. We further combine these criterion-level scores with LLM-assigned failure labels to distinguish missed evidence, incomplete reasoning, flawed deduction, and wrong conclusions in Section 6.1.

3.4 POSITIONING AMONG MYSTERY BENCHMARKS

We note that existing mystery benchmarks differ in what evidence the model receives, what output it must produce, and which parts of that output are scored. Table 1 compares five properties that directly determine the capability being evaluated. We summarize the properties as follows: *Human source* denotes game cases written by actual experts; *fixed evidence* requires the model to receive the complete evidence before responding; *visual clues* carry visual information needed for the solution; *whole case* requires reconstruction of the complete mystery; and *criterion credit* scores components of that reconstruction rather than only its final answer or outcome.

We note that with benchmark with attributes such as human source, fixed-evidence are well established, but prior work only ask for an answer, clue set, or local contradiction rather than a complete reconstruction (Del & Fishel, 2023; Gu et al., 2024; Xu et al., 2025; Gupta, 2025; Yuan et al., 2025). MuSR pairs narratives with answer accuracy, while WhodunitBench evaluates visual evidence through interactive play (Sprague et al., 2024; Xie et al., 2024). Other benchmarks like DetectiveQA and Watson & Holmes provide step-level or rubric-level credit but do not require a fixed-evidence

whole-case reconstruction (Xu et al., 2025; Leelawat & Griffin, 2026). HONKAKUBENCH alone combines human-authored cases, fixed evidence, visual clues, whole-case output, and criterion-level credit.

Table 1: Comparison of mystery-reasoning benchmarks. A teal check indicates that a property is present; a coral cross indicates that it is absent.

	Case and evidence			Evaluation	
Benchmark	**Human source**	**Fixed evidence**	**Visual clues**	**Whole case**	**Criterion credit**
True Detective (2023)	✓	✓	×	×	×
MuSR (2024)	×	✓	×	×	×
DetectBench (2024)	✓	✓	×	×	×
DetectiveQA (2025)	✓	✓	×	×	✓
WhoDunIt (2025)	✓	✓	×	×	×
TurnaboutLLM (2025)	✓	✓	×	×	×
WhodunitBench (2024)	✓	×	✓	×	×
Watson & Holmes (2026)	✓	×	×	×	✓
HONKAKUBENCH (ours)	✓	✓	✓	✓	✓

4 EXPERIMENTAL SETUP

Model execution. For each case on different models, we start a fresh run with the case package and task instruction while withholding the official solution and rubric from the model being evaluated. Models can read files, inspect images, and execute code in an isolated environment before producing one solution. The instruction prompt requests the culprit, method, trick mechanism, motive, and supporting evidence and reasoning chain as presented in Appendix A.2. We further ablate the default evaluation harness and experiment with direct inference in Section 5.3. In Section 5.1, we compare eight models on all 70 cases under the same resource constraint using Harbor (Harbor Framework Team, 2026) as the evaluation harness. We refer the model coverage and runner details to Appendix B.

Grading. For grading on model's final reconstruction, we use OPUS 4.8 as the sole primary judge. Given the case, official solution, author-verified rubric, and model submission, it assigns criterion-level credit with supporting evidence. For judge sensitivity, GPT-5.6-SOL and GEMINI 3.7 FLASH regrade the same submissions with the same structured rubrics, grading criteria, and output schema. We refer the judge details to Appendix D.1.

Metrics and failure handling. Our primary results are the mean and median normalized rubric score defined in Equation 1. We additionally report the number of reconstructions receiving at least 90% of the available credit and the number receiving full credit, i.e., a fully reconstructed case. We retry grading attempts that fail to return a structured score and retain every score, including zero.

5 RESULTS AND ROBUSTNESS

We first ask how much of each complete case current frontier models can reconstruct in Section 5.1. We then use the case-level reasoning-demand labels to examine performance across evidence and reasoning conditions in Section 5.2. Finally, we test robustness to the evaluation harness, language, and grading model in Section 5.3. Unless stated otherwise, results use the primary harness and judge.

5.1 MODELS MAKE PARTIAL PROGRESS BUT RARELY SOLVE THE WHOLE CASE

As one can see from the primary results presented in Table 2, GPT-5.6-SOL and FABLE 5 form the leading pair with means of 61.3 and 59.3 and medians of 62.9 and 58.5. Their two-point mean difference is small relative to the variation across cases, so we do not treat it as a stable ordering and

Table 2: Whole-case reconstruction on all 70 cases. The final columns count cases meeting each completion threshold; the best value in each column is bold.

Model	Rubric credit (%)		Case completion	
	Mean	Median	Nearly solved ($S \geq 90\%$)	Fully solved ($S = 100\%$)
GPT-5.6-SOL	**61.3**	**62.9**	11/70	6/70
FABLE 5	59.3	58.5	**14/70**	**8/70**
KIMI K3	50.1	48.5	7/70	5/70
GEMINI 3.7 FLASH	49.5	45.5	8/70	2/70
OPUS 4.8	45.5	42.8	5/70	2/70
GPT-5.5	43.9	46.0	5/70	0/70
OPUS 4.6	41.7	35.0	4/70	2/70
SONNET 5	34.5	33.2	1/70	0/70

instead treat them both as the leading reasoning models in this task. KIMI K3 and GEMINI 3.7 FLASH follow at 50.1 and 49.5, while the remaining four models range from 45.5 to 34.5. For both leaders, we see that the median remains close to the mean, meaning that their results are not explained by exceptional success cases and instead perform in a uniformly stable manner.

Moreover, FABLE 5 reaches 90% most often (14 out of 70 cases), but it is still difficult for both reasoning models to achieve full credit, i.e., perfectly reconstruct the whole case: GPT-5.6-SOL and FABLE 5 solve only 6/70 and 8/70 cases. Their roughly 60% average coverage therefore separates models' capability in recovering correct components from assembling one reconstruction that satisfies the entire rubric. Across all eight models, only 55 of 560 reconstructions reach 90% credit, and only 25 receive full credit. Higher mean performance thus reflects recovering more solution components across cases, without yet making complete reconstruction reliable.

5.2 PERFORMANCE ACROSS EVIDENCE AND REASONING CONDITIONS

We further examine how performance varies across evidence and reasoning conditions as aggregate scores can conceal which reconstruction remain difficult. By using the reasoning-demand annotations introduced in Section 3.2, Figure 2(a) reports the mean over all cases carrying each tag. Since a case can enter multiple columns, these slices describe performance on cases involving a given demand rather than isolate that demand's causal effect. On the other hand, Figure 2(b) reports the mean score over a modality split to inspect whether cases with visual clues are as hard as text-only cases or not.

(a) Reasoning-demand profile

Model	Spatial	Temporal	Physical	Identity	Rule-based
GPT-5.6-SOL	59.7	61.3	65.7	59.2	58.7
FABLE 5	56.3	56.8	65.4	53.7	57.2
KIMI K3	50.9	51.6	55.6	45.7	44.1
GEMINI 3.7 FLASH	49.1	44.7	52.5	47.1	47.7
OPUS 4.8	40.9	43.5	47.5	46.2	46.8
GPT-5.5	42.9	39.9	44.9	43.4	42.6
OPUS 4.6	39.5	39.9	41.7	42.9	40.9
SONNET 5	28.1	34.4	36.5	33.3	33.9

Cell shade encodes mean score; darker is higher.

(b) Evidence modality

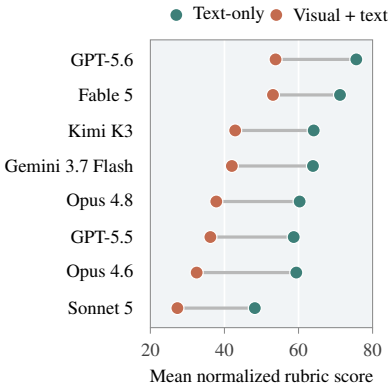

Figure 2: Performance profiles under the primary results. (a) Mean normalized rubric score for cases carrying each multilabel reasoning-demand tag. (b) Mean score on text-only and visual + text cases; each line connects the two group means for one model.

Reasoning-demand profiles. Physical-tagged cases have the highest descriptive mean for seven of the eight models, and the broad model ordering remains stable across the five columns. The leading pair illustrate why the slices are still informative: their overall two-point gap in Table 2 narrows to only 0.3 points on Physical-tagged cases, but expands to 4.5 points on Temporal-tagged cases and 5.5 points on Identity-tagged cases, both in favor of GPT-5.6-SOL. Looking only at the overall score would thus miss these differences between two otherwise similar models. A single mystery can require several kinds of reasoning, so each column describes a group of cases rather than measuring one reasoning ability in isolation.

Evidence modality. It is clear from Figure 2(b) that every model scores lower on visual + text cases: the gap is 18.1–26.8 points for the eight reasoning models. The leading pair remain strongest in both modality splits while SONNET 5 remains lowest, so the modality split shifts the score level more than the broad model ordering. Its consistent direction across models suggests that visual-evidence cases expose a shared difficulty rather than one confined to a particular model family.

5.3 ROBUSTNESS ACROSS EXECUTION, LANGUAGE, AND JUDGE

We test robustness to three evaluation choices in Figure 3: how the model is run, which language it reads, and which model grades its answer. Each panel holds the evaluated cases or submissions fixed across settings.

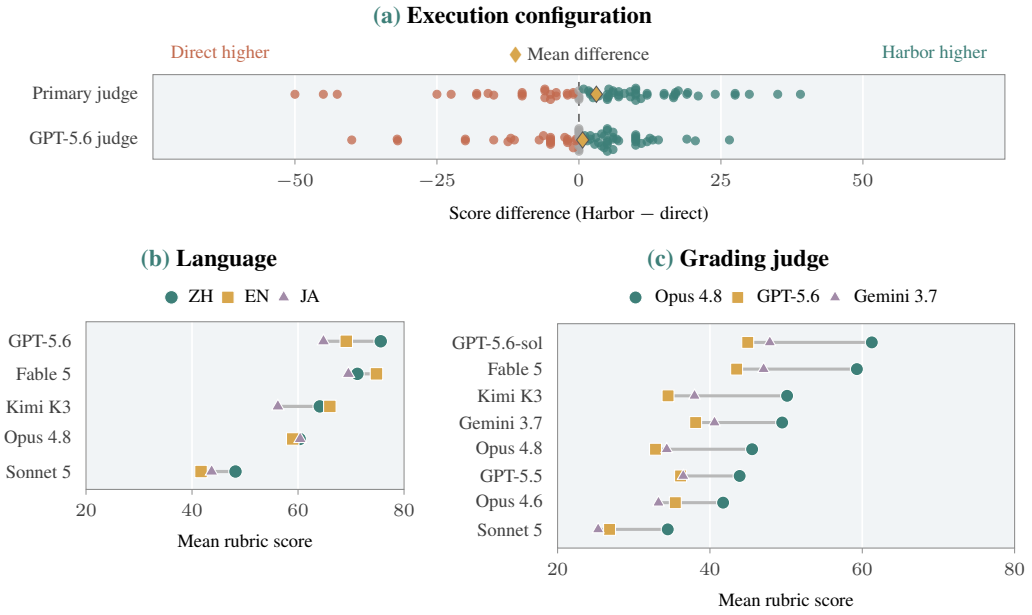

Figure 3: Robustness checks under matched comparisons. (a) Per-case score gaps between Harbor and direct inference for OPUS 4.8 under two judges. (b) Five-model means in three languages. (c) Identical submissions from all models graded by three judges. (b)–(c) share a score scale.

Execution configuration. For OPUS 4.8, the case-level differences in Figure 3(a) run in both directions. Under the primary judge, Harbor has the higher mean, while direct inference produces more 90%-plus/full-credit cases. Under the GPT-5.6 judge, the means are 32.86 versus 32.21 and the corresponding counts are 1/0 versus 1/1. Across both judges, Harbor gives higher average scores, while direct inference reaches at least as many near-complete and complete solutions.

Language. Across the five models in Figure 3(b), the range among language means is 1.4 points for OPUS 4.8 but 10.8 for GPT-5.6-SOL. The favored language also varies: FABLE 5 scores highest in English, whereas GPT-5.6-SOL, KIMI K3, and Sonnet 5 score highest in Chinese. These comparisons reflect both language and translation wording. The case version can shift a model's mean by up to 10.8 points and even change which leading model ranks first.

Grading judge. On identical submissions, the alternative judges lower means by 6.3–16.3 points relative to the primary judge. Still, GPT-5.6-SOL and FABLE 5 remain the leading pair and Sonnet 5 remains last across all three judges. The benchmark consistently identifies the same leading pair and last-place model across judges. We further conduct a human calibration experiment against the primary judge. Of all evaluated models, the primary judge closely tracks two human evaluators ($r = 0.89$ and $r = 0.90$) with themselves agreeing at $r = 0.94$, with 74.3% exact agreement across 144 criterion decisions (Appendix D.5).

6 FAILURE ANALYSIS AND CASE STUDIES

The aggregate scores in Section 5 show how often models succeed but not how their reconstructions break. We therefore pair a taxonomy over unsuccessful submissions (Section 6.1) with criterion-level case studies tracing how individual errors alter a complete solution (Section 6.2).

6.1 A TAXONOMY OF RECONSTRUCTION FAILURES

We analyze the 484 submissions scoring below 85%, excluding the borderline 85–89% band. The resulting distribution characterizes clear failed attempts rather than all reconstructions.

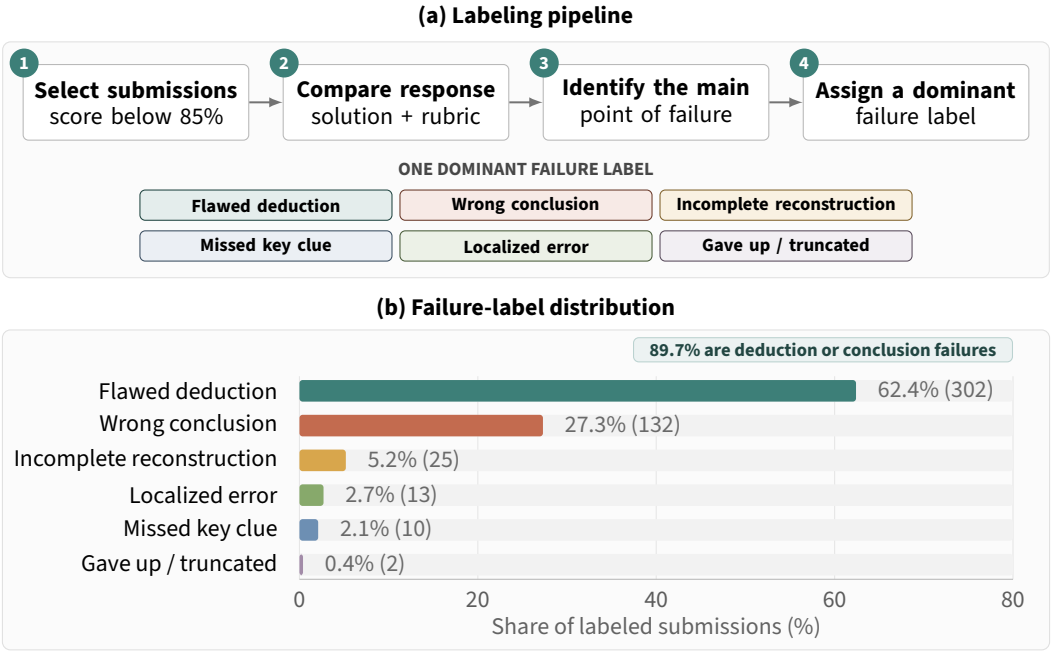

Figure 4: Failure analysis: (a) dominant-label pipeline; (b) label distribution over 484 submissions.

Figure 4(a) defines the failure taxonomy and labeling pipeline. The author-defined taxonomy separates missed evidence, invalid inference, wrong outcomes, incomplete reconstructions, consequential local errors, and truncated attempts. The judge compares each response with the official solution and rubric, then assigns its main divergence one dominant label with human verification (Appendix E).

Figure 4(b) assigns 89.7% of failures to flawed deduction or wrong conclusion, versus only 2.1% to missed key clues. The dominant diagnosed breakdown therefore occurs after evidence enters the response: the model draws an invalid inference or fails to reconcile it with the conclusion. Because each response receives one label, clue omissions may coexist as secondary errors.

6.2 Local Progress Fails at Global Reconciliation

The aggregate taxonomy points to failures in deduction and conclusion rather than clue retrieval. Two cases below show the shared weakness: models recover useful clues or locally correct components without checking that the culprit, mechanism, and clue interpretations can all be true together. We call this check *global reconciliation*; Figure 5 and Table 3 show the cases and per-model credit.

Two Cases to Illustrate Global Reconciliation. Case 06, *Nisemono Maniac*, requires the hidden "Master," island motion, and a murder on an undamaged lake platform to form one account. In Case 08, *Not-Random Ball-Drawing*, a referee's gesture fixes a value linking game rules, scores, and a teammate's color blindness. It contrasts a late component mismatch with a propagated premise error.

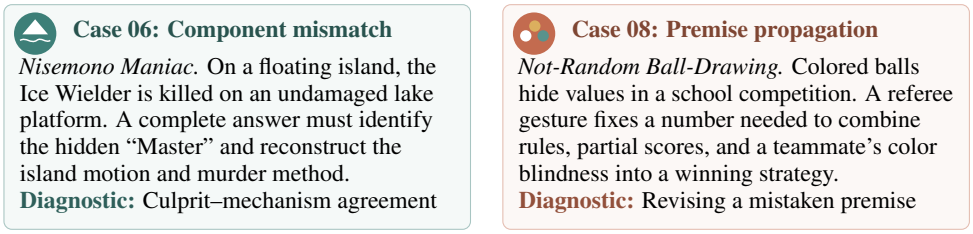

Figure 5: Cases 06 and 08 trace a component mismatch and a propagated premise error, respectively.

Table 3: Criterion credit on Cases 06 and 08. Darker cells indicate a greater share of available credit.

	Case 06: Nisemono Maniac				Case 08: Not-Random Ball-Drawing			
Model	**Culprit (identity)**	**Island motion**	**Method**	**Elimination**	**Gesture (premise)**	**Rounds**	**Value constraints**	**Final strategy**
GPT-5.6-SOL	0/25	30/30	22/22	9/23	0/5	0.5/2	2/4	1/8
FABLE 5	25/25	30/30	22/22	23/23	1/5	2/2	2/4	1/8
KIMI K3	25/25	22/30	8/22	20/23	0.5/5	0.5/2	0.5/4	1/8
GEMINI 3.7 FLASH	25/25	30/30	22/22	23/23	0/5	2/2	2/4	0/8
OPUS 4.8	0/25	22/30	22/22	9/23	1/5	0/2	0/4	1/8
GPT-5.5	25/25	30/30	22/22	21/23	0.5/5	0.5/2	1.5/4	2/8
OPUS 4.6	25/25	14/30	12/22	16/23	1/5	2/2	0/4	1/8
SONNET 5	0/25	12/30	7/22	5/23	0/5	0/2	0/4	0/8

Case 06: correct pieces remain disconnected. GPT-5.6-SOL and OPUS 4.8 recover much of the mechanism but name the wrong culprit, whereas OPUS 4.6 and KIMI K3 show the reverse pattern. Their useful pieces do not form the same causal deduction, a distinction culprit-only scoring would hide. Scoring components and their joint reconstruction exposes this gap.

Case 08: a wrong premise survives downstream reasoning. The referee's gesture encodes 16 through the year on a school crest, but all eight models misread it. Several then reason coherently from that premise and earn downstream credit, yet end with an incorrect strategy. Separate criterion scores let HONKAKUBENCH locate the first departure from the intended solution, not just the outcome.

One recurring weakness. Case 08 fails early and Case 06 at final assembly, but both produce locally coherent reasoning without global consistency, mirroring the aggregate failure profile.

7 Conclusion

HONKAKUBENCH evaluates whole-case reconstruction across 70 long, multimodal mysteries with official solutions and author-provided rubrics. The leading models recover 61.3% and 59.3% of rubric credit yet fully solve only 6 and 8 cases. Every model scores lower with visual evidence, and 89.7% of labeled failures concern deduction or conclusion. Case studies show why: models pair a correct mechanism with the wrong culprit or propagate one misread clue. HONKAKUBENCH thus measures whether partial progress becomes a complete, self-consistent explanation.

Reproducibility Statement

We will release the benchmark, prompts, selected model submissions, structured grades, run manifests, and evaluation code.[3] Section 4 and Appendix B document the evaluation protocol, and Appendix C reports expanded aggregate and distributional results.

Ethics Statement

We obtained consent from the authors and experts for all mysteries to use and release the case materials, official solutions, and rubrics. The benchmark contains fictional narratives and no personal data.

For judge calibration, two adult collaborators voluntarily provided blinded annotations of model responses. Under our institution's policy, this minimal-risk annotation activity did not require IRB review; we collect no sensitive personal attributes and report only aggregate statistics.

AI Use Statement

Generative AI systems are the objects of evaluation and also assisted with grading, analysis, and manuscript preparation. Human authors verified the benchmark rubrics and annotations, checked the reported results, and reviewed all AI-assisted material. We take responsibility for the final paper and its artifacts.

References

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

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

Maksym Del and Mark Fishel. True detective: A deep abductive reasoning benchmark undoable for GPT-3 and challenging for GPT-4. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM)*, pp. 314–322, 2023. doi: 10.18653/v1/2023.starsem-1.28.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021.

Zhouhong Gu, Lin Zhang, Xiaoxuan Zhu, Jiangjie Chen, Wenhao Huang, Yikai Zhang, Shusen Wang, Zheyu Ye, Yan Gao, Hongwei Feng, and Yanghua Xiao. DetectBench: Can large language model detect and piece together implicit evidence? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 199–222, 2024. doi: 10.18653/v1/2024.findings-emnlp.11.

Kshitij Gupta. WHODUNIT: Evaluation benchmark for culprit detection in mystery stories. *arXiv preprint arXiv:2502.07747*, 2025.

Harbor Framework Team. Harbor: A framework for evaluating and optimizing agents and models in container environments. Software, 2026. URL https://doi.org/10.5281/zenodo.20953922.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

[3] Anonymized artifact: https://anonymous.4open.science/r/honkaku-bench-A491

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *Conference on Language Modeling*, 2024. arXiv:2404.06654.

Thatchawin Leelawat and Lewis D. Griffin. Watson & holmes: A naturalistic benchmark for comparing human and LLM reasoning. *arXiv preprint arXiv:2602.19914*, 2026.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. arXiv:2305.20050.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. MuSR: Testing the limits of chain-of-thought with multistep soft reasoning. In *International Conference on Learning Representations*, 2024.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. JudgeBench: A benchmark for evaluating LLM-based judges. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.12784.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, 2024a. doi: 10.18653/v1/2024.acl-long.511.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024b. doi: 10.18653/v1/2024.acl-long.510.

Junlin Xie, Ruifei Zhang, Zhihong Chen, Xiang Wan, and Guanbin Li. WhodunitBench: Evaluating large multimodal agents via murder mystery games. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024. doi: 10.52202/079017-2751.

Zhe Xu, Jiasheng Ye, Xiaoran Liu, Xiangyang Liu, Tianxiang Sun, Zhigeng Liu, Qipeng Guo, Linlin Li, Qun Liu, Xuanjing Huang, and Xipeng Qiu. DetectiveQA: Evaluating long-context reasoning on detective novels. In *ICLR 2025 Workshop on Reasoning and Planning for Large Language Models*, 2025. arXiv:2409.02465.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018. doi: 10.18653/v1/D18-1259.

Yuan Yuan, Muyu He, Muhammad Adil Shahid, Ziyang Li, Jiani Huang, and Li Zhang. TurnaboutLLM: A deductive reasoning benchmark from detective games. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1951–1965, 2025. doi: 10.18653/v1/2025.emnlp-main.101.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, 2023.

A BENCHMARK DETAILS

A.1 CORPUS SCOPE AND CASE SELECTION

HONKAKUBENCH contains 70 fair-play mystery submissions selected from a national mystery-writing competition and prepared with mystery-writing experts. Each mystery forms one instance and was screened for a complete puzzle, an identifiable intended solution, and enough evidence for criterion-level grading. The corpus includes text-only cases and cases in which a floor plan, grid, map, or other diagram constrains the reconstruction. We obtained consent from all 70 authors to use and release the narratives, diagrams, official solutions, and criterion rubrics.

A.2 CASE PACKAGE AND SOLVER TASK

Each solver receives the complete narrative, associated visuals, and a common instruction, but not the official solution or rubric. It produces one free-form whole-case reconstruction that may need to identify the culprit, recover a physical or logical mechanism, order events, eliminate alternatives, and ground those claims in evidence. After submission, the primary judge compares the saved response with independently scorable criteria and normalizes the awarded credit to 100.

Solver instruction. The shared Harbor evaluation uses the instruction below. The direct-inference ablation presents the narrative and images inline and requests the same solution components.

Solver instruction

You are an expert solver of *honkaku* mystery fiction. There is a mystery puzzle in the current working directory (`/app`). Complete the following steps:

1. Read `/app/puzzle.md`, the puzzle narrative.
2. If `/app/media/` exists, inspect every image in it; the images are part of the clues.
3. Reason carefully and provide a complete solution: identify the culprit, method, trick mechanism, motive, and the key clues and chain of reasoning supporting the conclusion.
4. Write the complete solution to `/app/solution.md` in clear Markdown, organized with headings such as "Conclusion," "Reasoning Process," "Key Clues," and "Trick Reconstruction."

Proceed directly without asking for confirmation.

A.3 DATASET COMPOSITION

Table 4: Aggregate composition of HONKAKUBENCH. Narrative lengths are measured in the canonical source files after removing task instructions and markup.

Statistic	Value	Statistic	Value
Cases	70	Text-only cases	24
Visual + text cases	46	Referenced images	132
Official criteria	374	Partial-credit items	1,148
Minimum narrative length	2,057	Median narrative length	11,112
Maximum narrative length	58,261	Available languages	zh, en, ja

HONKAKUBENCH is available in Chinese, English, and Japanese; the matched multilingual analysis uses the same 24 text-only cases in all three languages.

A.4 OFFICIAL SOLUTIONS AND CRITERION RUBRICS

Each expert-provided solution is decomposed into weighted criteria that preserve the original questions and intended deductions, with partial-credit descriptions where needed. This supports equivalent wording and reasoning order while keeping the authors' criterion content and weights authoritative.

A.5 Reasoning-Demand Annotation

Reasoning-demand annotations are nonexclusive: one mystery can carry several tags and therefore contribute to more than one analysis column.

Table 5: Reasoning-demand codebook used in Section 5.2.

Demand	Definition
Spatial	Locations, routes, visibility, geometry, or architectural relations determine the solution.
Temporal	Event order, timing windows, alibis, or delayed actions determine the solution.
Physical	A material, mechanical, biological, chemical, or forensic mechanism determines what occurred.
Identity	Disguise, perspective, role, representation, or mistaken identity determines how evidence should be read.
Rule-based	An explicit game, logical, numerical, or constraint system determines the reconstruction.

The tags were produced from official solutions and rubrics without access to model outputs or scores, independently audited for every case, and frozen before analysis.

B Experimental Configuration and Reproducibility

B.1 Evaluated Models and Shared Harness

The primary evaluation covers GPT-5.6-sol, Fable 5, Kimi K3, Gemini 3.7 Flash, Opus 4.8, GPT-5.5, Opus 4.6, and Sonnet 5. Each model produced one saved submission for all 70 cases through the common Harbor interface, yielding 560 solver outputs and 560 primary grades.

B.2 Inference and Failure Handling

Each model receives the same case-level instruction and may read files, inspect images, and execute code in the evaluation environment. A missing submission receives zero, while any returned score is retained. Grading is retried only if no structured grade is returned, reusing the saved response.

B.3 Rubric Grading

Opus 4.8 is the primary judge. Given the case, official solution, rubric, and saved submission, it returns evidence-supported criterion awards. Their sum divided by the available points gives the case percentage; we report its mean, median, number at or above 90%, and number at 100%.

Judge instruction. The primary judge uses the following instruction with a forced structured-output schema. The prompt and schema will accompany the release.

Judge instruction
You are a strict and impartial judge for a *honkaku* mystery competition. You will receive a mystery, its official solution (the sole authoritative answer), a rubric, and a submitted solution. Score the submission criterion by criterion and return the structured grade through the `emit_grade` tool.

1. Treat the official solution as authoritative; a claim that contradicts it receives no credit.
2. Award partial credit only according to the rubric items. Sum the matched items, never exceed the criterion maximum, and explain any fractional award.
3. Assign one verdict per criterion: `correct`, `partial`, `incorrect`, or `missing`.
4. Cite or closely paraphrase relevant evidence from the submission and explain the award against the official solution.
5. Credit correct reasoning and conclusions, not vague, contradictory, or unsupported guesses. A correct conclusion without the required reasoning receives only its designated conclusion credit.
6. Sum all criterion awards, divide by the available points, and round the percentage to one decimal place.

For judge sensitivity, GPT-5.6-SOL and GEMINI 3.7 FLASH grade the same saved submissions, giving a fully crossed comparison of eight solvers and three judges. The matrix includes self-judge cells, so its judges are not independent of every solver.

B.4 RUN SELECTION AND SUMMARIES

Each experiment uses one frozen submission and grade per model–case cell, identified with its source run in the release manifest. Matched comparisons report the corresponding mean differences.

C FULL PRIMARY RESULTS

C.1 EXPANDED LEADERBOARD

Table 6: Expanded primary-judge results. Standard deviation is across 70 cases; zeros include missing submissions assigned zero by policy.

Model	Mean	Median	SD	≥ 90	100	Zeros
GPT-5.6-SOL	61.3	62.9	26.6	11	6	0
FABLE 5	59.3	58.5	28.4	14	8	1
KIMI K3	50.1	48.5	29.6	7	5	1
GEMINI 3.7 FLASH	49.5	45.5	26.7	8	2	1
OPUS 4.8	45.5	42.8	26.1	5	2	0
GPT-5.5	43.9	46.0	27.3	5	0	2
OPUS 4.6	41.7	35.0	27.4	4	2	3
SONNET 5	34.5	33.2	23.1	1	0	3

C.2 SCORE DISTRIBUTIONS

Figure 6 gives the full score distributions: right-shifted curves indicate higher scores across more cases, while their endpoints expose the scarcity of complete reconstructions.

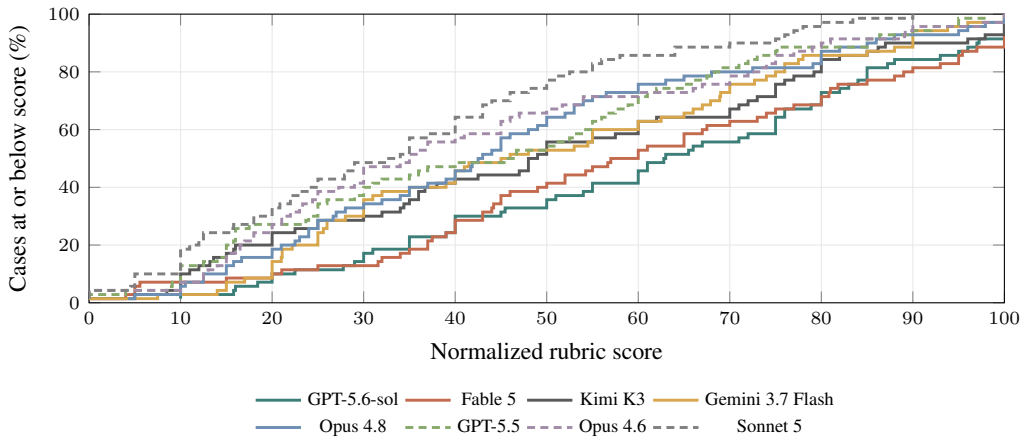

Figure 6: Empirical cumulative distributions of normalized rubric scores under the primary judge.

C.3 Reasoning-Demand Results

Table 7: Mean primary-judge score on cases carrying each reasoning-demand tag. Groups overlap because cases can receive multiple tags.

Model	**Spatial** ($n = 29$)	**Temporal** ($n = 41$)	**Physical** ($n = 36$)	**Identity** ($n = 37$)	**Rule-based** ($n = 33$)
GPT-5.6-sol	59.7	61.3	65.7	59.2	58.7
Fable 5	56.3	56.8	65.4	53.7	57.2
Kimi K3	50.9	51.6	55.6	45.7	44.1
Gemini 3.7 Flash	49.1	44.7	52.5	47.1	47.7
Opus 4.8	40.9	43.5	47.5	46.2	46.8
GPT-5.5	42.9	39.9	44.9	43.4	42.6
Opus 4.6	39.5	39.9	41.7	42.9	40.9
Sonnet 5	28.1	34.4	36.5	33.3	33.9

C.4 Evidence Modality

Table 8: Mean scores on 24 text-only and 46 visual + text cases; gap = text-only − visual + text.

Model	**Text-only**	**Visual + text**	**Gap**
GPT-5.6-sol	75.6	53.8	21.9
Fable 5	71.2	53.1	18.1
Kimi K3	64.1	42.9	21.2
Gemini 3.7 Flash	63.9	42.0	21.9
Opus 4.8	60.3	37.8	22.4
GPT-5.5	58.7	36.2	22.5
Opus 4.6	59.4	32.5	26.8
Sonnet 5	48.2	27.3	20.9

D Robustness and Secondary Experiments

D.1 Judge Sensitivity

Table 9: Mean scores when identical submissions from all eight models are graded by three judges. Diagonal cells include a judge from the same model family as the solver.

Solver	Opus 4.8 judge	GPT-5.6 judge	Gemini judge
GPT-5.6-sol	61.251	44.960	47.861
Fable 5	59.291	43.503	47.047
Kimi K3	50.129	34.500	37.976
Gemini 3.7 Flash	49.483	38.121	40.596
Opus 4.8	45.536	32.854	34.333
GPT-5.5	43.899	36.143	36.496
Opus 4.6	41.728	35.456	33.244
Sonnet 5	34.461	26.819	25.324

Across all eight models, both sensitivity judges lower absolute scores by 6.3–16.3 points. The leading pair and last-place model remain unchanged, while the middle ordering moves.

D.2 Harness Ablation

Table 10: Harbor vs. direct single-response inference on matched cases. Counts: $\geq 90\%/100\%$.

Model	Judge	Harbor	Direct	Δ	Counts H/D
Opus 4.8	Opus 4.8	45.536	42.466	+3.070	5/2; 7/4
Opus 4.8	GPT-5.6	32.856	32.210	+0.646	1/0; 1/1
GPT-5.6-sol	Opus 4.8	61.251	55.732	+5.519	11/6; 14/7

Because this ablation jointly changes the harness, prompt delivery, and execution procedure, it tests the complete Harbor setup rather than any one component.

D.3 Cross-Lingual Comparison

Table 11: Primary-judge mean scores across Chinese, English, and Japanese.

Model	Chinese	English	Japanese
GPT-5.6-sol	75.6	69.1	64.8
Fable 5	71.2	74.8	69.5
Kimi K3	64.1	66.0	56.2
Opus 4.8	60.3	59.0	60.4
Sonnet 5	48.2	41.7	43.7

The three versions contain the same mysteries, but translation can alter phrasing, clue salience, and linguistic difficulty; the comparison is matched by case, not by every linguistic property.

D.4 Run and Configuration Stability

Table 12: Headline and later runs for three models. These comparisons mix repeat variation with configuration changes and are not labeled seed variance.

Model	Headline	Later run	Δ
GPT-5.6-SOL	61.251	54.951	-6.300
FABLE 5	59.291	54.405	-4.886
OPUS 4.8	45.536	42.542	-2.993

These shifts caution against fine-grained rankings from one run but preserve the larger gap between average rubric credit and complete reconstruction.

D.5 Human Calibration of the Primary Judge

We conduct a blinded human calibration of the primary judge on 32 responses from four cases, with all eight evaluated models represented in every case. Two human evaluators independently receive the complete case, official solution, author-verified rubric, and anonymized response. Model identity, machine scores, and machine explanations remain hidden during grading. Each evaluator assigns credit independently to every rubric criterion using the same point scale as the primary judge. The two human evaluators show strong agreement at both the response and criterion levels (Table 13). Their normalized case scores have a Pearson correlation of 0.937 and a Spearman correlation of 0.920, with a mean absolute difference of 6.53 points. They assign exactly the same credit on 107 of 144 criterion decisions (74.3%), and their mean criterion-level difference is 7.9% of the corresponding criterion maximum. The primary judge also closely tracks both human evaluators: its case scores correlate at $r = 0.889$ and $r = 0.903$ with their independently assigned scores. These results establish strong alignment between the primary judge and blinded human rubric grading.

Table 13: Blinded human calibration of the primary judge. Score MAE is on the 0–100 normalized case scale. Criterion agreement compares the points assigned to each author-verified rubric criterion.

Comparison	n	Pearson r	Spearman ρ	Score MAE
Human A vs. Human B	32	0.937	0.920	6.53
Primary judge vs. Human A	32	0.889	0.908	12.17
Primary judge vs. Human B	32	0.903	0.908	11.13

E Failure Analysis and Extended Case Studies

E.1 Dominant-Failure Protocol

We apply a six-category taxonomy to submissions below 85%, excluding borderline scores of 85–89%. The primary judge compares each eligible response with the official solution and rubric, diagnoses its main divergence, and assigns one dominant label. A collaborator checked every label in the original seven-model set against the submission and grading evidence. The Kimi labels were produced by the same pipeline and remain model-generated:

Flawed deduction. The response uses relevant evidence but draws an invalid or conflicting inference.

Wrong conclusion. An incorrect culprit, mechanism, or other conclusion dominates the failure.

Incomplete reconstruction. The response omits an essential component of the reconstruction.

Missed key clue. The response does not use decisive evidence in the official solution.

Localized error. One consequential local error or omission leaves an otherwise strong reconstruction below the cutoff. Released artifacts store this as `mostly_correct`.

Gave up/truncated. The submission stops before attempting a usable reconstruction.

Table 14: Dominant-failure counts by model for submissions below 85%.

Model	n	Flawed	Wrong	Incomplete	Missed	Localized	Gave up
GPT-5.6-SOL	54	28	13	7	1	5	0
FABLE 5	53	37	13	1	1	0	1
KIMI K3	60	38	16	2	2	2	0
GEMINI 3.7 FLASH	60	41	15	1	1	2	0
OPUS 4.8	62	39	20	2	1	0	0
GPT-5.5	62	38	12	8	2	2	0
OPUS 4.6	64	41	18	1	1	2	1
SONNET 5	69	40	25	3	1	0	0
Total	**484**	**302**	**132**	**25**	**10**	**13**	**2**

E.2 CASE 06: COMPONENTS COME APART

Case 06 separately awards credit for the culprit, central moving-island trick, murder method, and elimination of alternative suspects. GPT-5.6-SOL earns full mechanism credit while naming the wrong culprit; OPUS 4.6 identifies the culprit while recovering much less of the mechanism. FABLE 5 and GEMINI 3.7 FLASH receive full credit, whereas KIMI K3 identifies the culprit but receives partial credit on all three mechanism components. Culprit accuracy thus cannot replace a whole-case score.

E.3 CASE 08: A WRONG PREMISE PROPAGATES

The official solution interprets the referee's gesture as 16 through the year printed on a school crest. Seven models instead infer 12, while OPUS 4.8 infers 14. Their downstream calculations can be internally consistent while remaining incompatible with the official premise. This case isolates a failure that a final-strategy label would obscure: the response continues to reason, but inside the wrong reconstructed world.

E.4 ADDITIONAL CASES

Cases 03 and 44 reverse the reconciliation problem: Case 03 moves from a candle trick to its user, whereas Case 44 moves from a culprit to the timeline and method. Figure 7 introduces both cases; Table 15 maps credit to their components.

Case 03: Mechanism–identity split

Blue Lantern Tale. At a mountain ghost-story gathering, a smashed jar diverts a witness as altered candles obscure her arrival. A complete answer must connect both tricks to suspect routes and the culprit.

Diagnostic: Mechanism–culprit agreement

Case 44: Identity–method split

The Little Sprite Is Playing Paid Pranks. A reclusive streamer is found poisoned after a doorbell goes unanswered. A complete answer must explain the apparent suicide, death time, repeated bell, poison route, and evidence in a van.

Diagnostic: Timeline–method agreement

Figure 7: Additional reconciliation cases: mechanism to culprit (Case 03) and culprit to full timeline and method (Case 44).

Table 15: Criterion credit in Cases 03 and 44 under the primary judge. For Case 44, Q2–Q3 and Q5–Q6 are combined; darker cells denote more credit.

	Case 03: Blue Lantern Tale				Case 44: The Little Sprite Is Playing Paid Pranks			
Model	Jardiversion	Candlemechanism	Suspectchain	Culprit +exclusion	Non-suicideevidence	Culprit +time	Doorbellmechanism	Poison +evidence
GPT-5.6-SOL	3/3	5/5	6/7	1/5	15/15	20/20	25/25	40/40
FABLE 5	3/3	5/5	6/7	2/5	13/15	16/20	25/25	39/40
KIMI K3	3/3	5/5	6/7	1/5	10/15	16/20	14/25	8/40
GEMINI 3.7 FLASH	3/3	5/5	4/7	1/5	10/15	8/20	8/25	0/40
OPUS 4.8	3/3	5/5	3/7	1/5	5/15	7/20	8/25	0/40
GPT-5.5	3/3	5/5	3/7	2/5	10/15	7/20	8/25	0/40
OPUS 4.6	3/3	3/5	4/7	0/5	5/15	7/20	8/25	0/40
SONNET 5	3/3	3/5	1/7	0/5	10/15	20/20	14/25	33/40

Opposite reconciliation gaps. In Case 03, all eight models recover the jar diversion and six recover the candle mechanism, yet no model earns more than 2/5 for culprit and exclusion; the best total is 16/20. In Case 44, all eight name the culprit, but totals range from 20 to 100: only GPT-5.6-SOL and FABLE 5 fully recover the doorbell mechanism, and four models miss the poison route and van evidence. Together with Cases 06 and 08, these reversals show why whole-case scoring must test whether correct components form one consistent account.

F LIMITATIONS, DATA GOVERNANCE, AND RELEASE

The leaderboard combines each model with its harness and provider settings. The limited harness ablation cannot isolate components, and later runs move several mean-score points, so close single-run rankings may not be durable. LLM judges produce criterion scores and failure labels; alternative judges shift the scale, self-judge cells remain, and one dominant label cannot capture every error. Author criteria clarify the target but do not remove judge error. The reasoning-demand and modality slices are descriptive: tags overlap, visual + text cases lack paired text-only versions, and translations can alter clue salience. All 70 authors consented to use and release. We will open-source narratives, diagrams, solutions, rubrics, prompts, manifests, submissions, grades, and evaluation code.