BEYOND ACCURACY: RE-EVALUATING FAILURES IN LLM PHYSICS REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating physics reasoning in large language models (LLMs) requires more than identifying where models fail: it requires determining whether those failures are evidence of a recurring reasoning weakness. An observed error may disappear under repetition, a persistent error may not survive controlled reformulation, and a new failure may arise from a different reasoning breakdown. This distinction is especially consequential in physics, where errors in physical modeling, formula applicability, and symbolic or numerical execution can produce similar incorrect outcomes. We introduce PACE (**P**robe-based **A**ssessment and **C**haracterization of **E**rrors), a physics-grounded diagnostic framework that distinguishes *observed*, *stable*, and *attributable* failure evidence. PACE combines selective repeated measurement, mechanism-abstracted Difficulty Cards, controlled probe synthesis, scientific validation, and source-relative attribution to test not only whether a model fails again, but whether the same reasoning breakdown recurs. Across frontier reasoning models and multiple physics subfields, we find that repeated failure, failure on a source-guided probe, and broad error-category agreement each provide useful but incomplete diagnostic evidence. In particular, source-guided probe failures do not necessarily reproduce the source reasoning breakdown, and category-level similarity can conceal differences in the underlying reasoning breakdown. These results expose a gap between reproducing an error and diagnosing a recurring weakness, motivating a reliability-aware view of failure evaluation, in which evidence strength is measured rather than assumed.

1 INTRODUCTION

Large language models (LLMs) have become increasingly capable on scientific reasoning tasks, particularly in physics, a domain that demands both deep physical knowledge and multi-step inference (Rein et al., 2024; Xu et al., 2025; He et al., 2024; Qiu et al., 2025). Evaluation has correspondingly moved beyond static problem sets, from human-in-the-loop protocols to automated and dynamically generated tests (Nie et al., 2020; Kiela et al., 2021; Zhu et al., 2024; Zhang et al., 2024; Mirzadeh et al., 2025). More recently, failure-driven methods have begun to use observed model errors to guide subsequent test generation (Chen et al., 2026; Wang et al., 2025; Brown et al., 2025; Huang et al., 2026). Yet discovering another failure does not by itself establish that the same reasoning weakness has been reproduced: a model may answer a source-guided probe correctly, or fail on it for an entirely different reason. The central question is therefore not what failures can be discovered, but what diagnostic claims those failures can actually support.

The difficulty is that increasingly strong forms of behavioral evidence do not necessarily support equally strong diagnostic conclusions. A single incorrect response may disappear under repetition, as prior work has documented for repeated querying (Chen et al., 2024; He et al., 2025). Repeated failure strengthens the evidence for behavioral persistence, but does not show that the same weakness survives controlled reformulation. A source-guided probe provides a stronger test, yet a new failure may still arise from a different decisive reasoning step. Broad error categories offer another plausible shortcut, but two failures assigned to the same category can reflect different concrete breakdowns. **What evidence, then, is needed to move from an observed error to a defensible diagnosis of a recurring reasoning weakness?**

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.

This evidential ambiguity is not unique to physics, but physics makes it particularly consequential because correctness depends jointly on physical modeling, regime and formula selection, and symbolic or numerical execution. Similar incorrect outcomes can therefore arise from qualitatively different breakdowns, while the same underlying weakness may manifest differently across problem formulations. Figure 1 illustrates two representative limits of failure-based diagnosis. In one case, an initially failed source problem yields mixed pass–failure outcomes under repeated evaluation, showing that a one-shot error need not be behaviorally stable. In the other, a source-guided probe fails repeatedly, yet the decisive errors differ from the source breakdown, showing that behavioral recurrence need not imply mechanism recurrence.

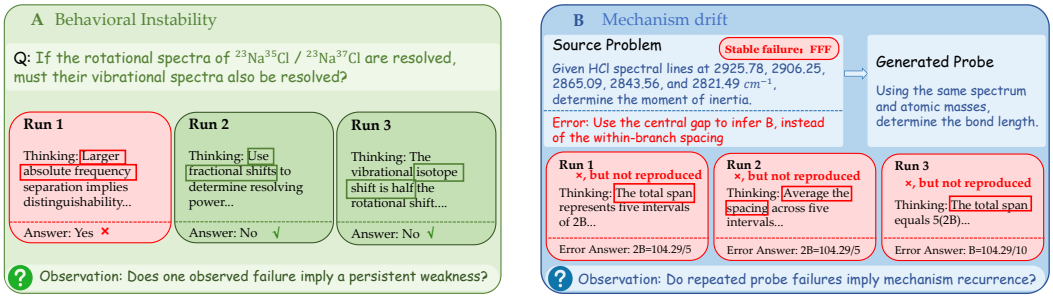

Figure 1: Two limits of failure-based diagnosis in physics. (A) Repeated evaluation reveals behavioral instability in an initially failed source problem. (B) A source-guided probe fails repeatedly, but the failed runs do not reproduce the source reasoning breakdown. Red highlights mark the decisive reasoning errors.

To make these distinctions explicit, we introduce PACE (**P**robe-based **A**ssessment and **C**haracterization of **E**rrors), a physics-grounded diagnostic evaluation framework. PACE separates three complementary forms of failure evidence: an *observed failure*, a *stable failure* supported by repeated measurement, and an *attributable failure* whose probe-induced failure supports recurrence of the source reasoning breakdown. These are complementary evidence dimensions rather than interchangeable labels. PACE operationalizes them through selective reliability auditing, mechanism abstraction, controlled and scientifically validated probe synthesis, and source-relative attribution, separating whether a model fails from whether it fails for the same reason.

Across frontier reasoning models and multiple physics subfields, our experiments expose three successive limits of common diagnostic shortcuts. Behavioral persistence does not by itself establish that a source failure is diagnostically usable. Failure on a targeted and scientifically validated probe does not guarantee recurrence of the source reasoning breakdown: in the atomic cross-model evaluation, only 56.4% of failed probe-route observations reproduce the source mechanism under strict attribution. Broad error-category agreement does not recover this distinction: in the same atomic setting, even among probes whose failed runs all align with the source error category, 23.7% are not attributed as full source-mechanism recurrence. Together, these results show that common proxies for diagnostic evidence each capture only part of the picture, **and that reproducing an error is not the same as diagnosing a recurring weakness.**

Our main contributions are:

1. We formalize *observed*, *stable*, and *attributable* failures as distinct forms of diagnostic evidence, clarifying the gap between observing an error and diagnosing a recurring reasoning weakness.
2. We introduce PACE, a physics-grounded framework that operationalizes this distinction through selective auditing, mechanism abstraction, validated probe synthesis, and source-relative attribution.
3. Across frontier reasoning models and physics domains, we show that behavioral persistence, source-guided probe failure, and broad category agreement are informative but insufficient proxies for source-relative mechanism recurrence.

2 Related Work

Scientific Reasoning and Process-Level Evaluation. Scientific reasoning benchmarks have evolved from broad-coverage suites such as MMLU (Hendrycks et al., 2021a), MMLU-Pro (Wang et al., 2024), ScienceQA (Lu et al., 2022), and GPQA (Rein et al., 2024) to more specialized evaluations of mathematical and physics reasoning, including MATH (Hendrycks et al., 2021b), OlympiadBench (He et al., 2024), UGPhysics (Xu et al., 2025), and PHYBench (Qiu et al., 2025). These benchmarks provide valuable measurements of reasoning performance, while process-level evaluation and step-level verifiers identify intermediate reasoning errors (Lightman et al., 2024). Multi-sample evaluation provides another view of reasoning reliability: self-consistency aggregates sampled completions, while pass@k characterizes success across multiple attempts (Wang et al., 2023; Chen et al., 2021). More recently, SymPyBench (Imani et al., 2026) introduces parameterized physics problems with executable solutions and consistency-oriented metrics across variants. These approaches provide complementary views of reasoning performance and reliability. PACE focuses on a distinct diagnostic question: whether an observed, source-specific reasoning breakdown persists behaviorally and whether it recurs under scientifically validated structural reformulation.

Dynamic and Failure-Driven Evaluation. Fixed benchmark collections face contamination, saturation, and limited responsiveness to evolving models (Srivastava et al., 2023). Performance estimates on fixed benchmarks can also be unstable (Schaeffer et al., 2023; Zhang et al., 2026), and repeated evaluations can produce substantial score variation (Gonzalez et al., 2025; Potamitis et al., 2025). Dynamic evaluation addresses these limitations through human-in-the-loop adversarial collection (Nie et al., 2020; Kiela et al., 2021) and automated or structured test generation (Zhu et al., 2024; Zhang et al., 2024; Mirzadeh et al., 2025). More closely related to our work, failure-driven methods exploit observed errors to guide subsequent evaluation. StressEval extracts difficulty factors from failed reasoning steps and synthesizes grounded instances that preserve these factors (Chen et al., 2026); self-evolving evaluation frameworks iteratively adapt test instances (Wang et al., 2025); task elicitation identifies systematic weaknesses through automatically constructed evaluations (Brown et al., 2025); and ProbeLLM searches verifiable failure regions and consolidates discovered cases into interpretable failure modes (Huang et al., 2026). These studies advance failure discovery and characterization. Complementing these objectives, PACE asks what diagnostic conclusions are supported by a source error and its generated probes. It separately assesses behavioral persistence, failure occurrence on validated probes, and source-relative mechanism recurrence, rather than treating repeated failure or categorical similarity as sufficient evidence of a recurring reasoning breakdown.

3 Method

The diagnostic questions in Section 1 cannot be resolved from a single evaluation. PACE operationalizes them through three phases (Figure 2): Phase I assesses the evidence supporting an observed source failure; Phase II constructs scientifically valid probes designed to test its diagnostic target under controlled variation; and Phase III measures target behavior and determines whether the observed failures support recurrence of the source-specific breakdown. The resulting records support subsequent characterization across models and physics domains. Throughout, let v denote an evaluated item (a source problem x or a synthesized probe q), and let $z_t(v) \in \{\mathrm{P}, \mathrm{F}, \mathrm{U}\}$ denote the outcome of its t-th evaluation. Unresolved outcomes are not treated as failures (Appendix A.2).

3.1 Phase I: Auditing and Abstraction

A single incorrect response establishes neither behavioral persistence nor the reliability of a source-specific diagnosis. Phase I therefore concentrates repeated measurement on initially failed source items until K trials are complete; unresolved items undergo bounded recovery, and a sparse pass-side sample serves as a control. Each source item is characterized by the sequence $(z_1(v), \ldots, z_K(v))$ and classified as stable failure, mixed pass–failure, stable pass, or unresolved. Any source with at least one completed failure remains eligible for mechanism analysis: behavioral persistence is measured rather than imposed as a prerequisite for attribution.

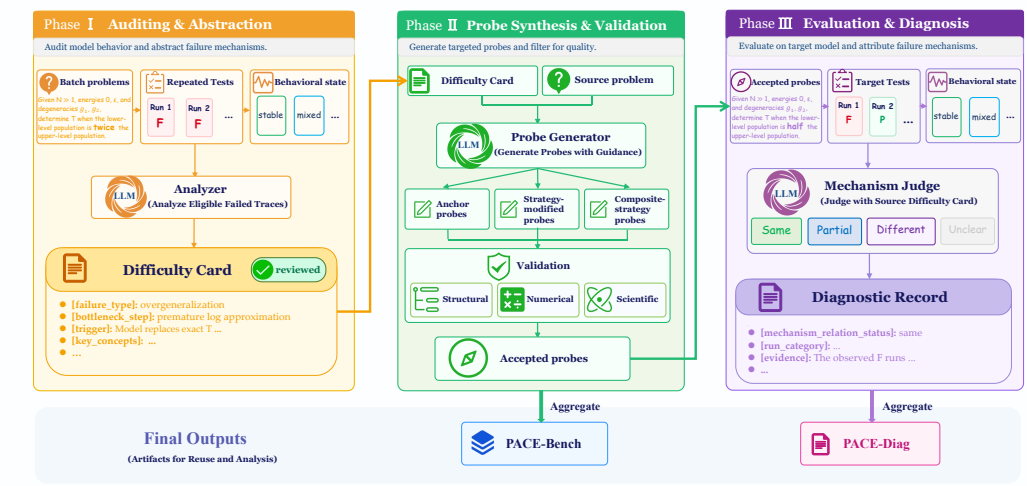

Figure 2: Overview of PACE. Phase I audits source-model behavior and abstracts supported failure mechanisms into reviewed Difficulty Cards. Phase II synthesizes targeted probe variants under a preservation–variation contract and filters them through structural, numerical, and scientific validation. Phase III evaluates accepted probes on target models and separately characterizes behavioral outcomes and source-mechanism recurrence. Validated probes are aggregated into PACE-Bench, while mechanism-aware diagnostic records form PACE-Diag.

For each eligible source item, all failed traces are jointly examined, and up to two diagnostically meaningful breakdowns may be abstracted into *Difficulty Cards*; items without a sufficiently supported breakdown yield no card. Each card specifies a concrete reasoning bottleneck, its triggering conditions, and the characteristic error signature supported by the source traces. The Card representation and a representative example are provided in Appendix B.2. Cards are reviewed by a separate reviewer model for diagnostic clarity and scientific validity. Source items with incorrect references, missing information, or equivalent answers are discarded, and admitted items establish a scientifically defensible evaluation context applied symmetrically to reference and model responses.

3.2 Phase II: Probe Synthesis and Validation

Phase II treats each Difficulty Card as an interface between source-side diagnosis and probe construction. Under a preservation–variation contract, probe generation is conditioned on the source Card to target the intended diagnostic weakness while varying selected problem attributes. The contract constrains generation toward the source diagnostic target, but does not determine the target model's eventual error: whether the intended breakdown actually recurs is determined only after target evaluation and attribution in Phase III. Probe synthesis follows three routes of increasing intended challenge: *anchor* probes apply only minimal parameter or surface changes, *strategy-modified* probes apply a single strategy from a predefined inventory, and *composite-strategy* probes combine two or more strategies to introduce stronger structural variation. Historical strategy usage is used only as a soft tie-breaker to encourage coverage and diversity (Appendix C.1).

Because a probe that is not scientifically valid cannot support a diagnostic claim, each candidate must pass a multi-stage validation gate. Structural validation checks schema completeness, answer form, and leakage; numerical validation checks equation, unit, arithmetic, and answer consistency through executable checks; and scientific validation combines an ordinary scientific validator with an independent physics-consistency audit. Candidates with locally repairable issues at any gate may undergo one bounded repair cycle; repaired candidates are revalidated in full, and candidates that remain invalid are rejected (Appendix C.2).

3.3 Phase III: Evaluation and Diagnosis

A probe that passes validation is well-formed, but this does not determine how the target will fail on it. The diagnostic task is to measure both whether the target fails and, whenever failures oc-

cur, whether they support recurrence of the source mechanism. Phase III therefore evaluates each accepted probe K times and treats behavioral state and mechanism attribution as separate branches.

The resulting evaluation trajectory is retained for behavioral stability analysis. Any probe with at least one observed failure remains eligible for mechanism attribution, regardless of its behavioral-state label.

In parallel, the mechanism judge compares all failed responses $\mathcal{F}(q)$ against the source Difficulty Card and assigns a probe-level relation $m(q) \in \{Same, Partial, Different, Unclear\}$. Same denotes recurrence of the source mechanism without a substantive competing breakdown; Partial denotes incomplete or non-uniform support, including co-occurring breakdowns; Different indicates an alternative breakdown; Unclear preserves cases with insufficient evidence. Same provides full recurrence support, whereas Partial provides qualified or non-uniform support; both are retained as attributable evidence but reported separately where diagnostic strength matters (Figure 1; Appendices C.3–C.4).

Let $\mathcal{Q}$ denote the set of probes admitted after validation for which target-evaluation records are available. To exclude probes with no resolved target outcome, we define $\mathcal{Q}_R = \{q \in \mathcal{Q} : \exists t,\ z_t(q) \in \{\mathrm{P}, \mathrm{F}\}\}$ as the subset with at least one resolved outcome, and $\mathcal{Q}_F = \{q \in \mathcal{Q}_R : \exists t,\ z_t(q) = \mathrm{F}\}$ as the subset inducing at least one failure. We report two complementary probe-level measurements. *Failure occurrence* is the fraction of behaviorally evaluable probes that induce at least one failure, $|\mathcal{Q}_F|/|\mathcal{Q}_R|$. *Mechanism recurrence*, conditional on an observed probe failure, is reported under two thresholds:

$$R_{\text{strict}} = |\{q \in \mathcal{Q}_F : m(q) = \text{Same}\}|/|\mathcal{Q}_F|, \tag{1}$$

$$R_{\text{broad}} = |\{q \in \mathcal{Q}_F : m(q) \in \{\text{Same}, \text{Partial}\}\}|/|\mathcal{Q}_F|. \tag{2}$$

We report R_{broad} when comparing overall recurrence support and R_{strict} when distinguishing full from qualified recurrence. Source admission and probe-validation yields are reported as pipeline statistics rather than diagnostic metrics.

Beyond these aggregate metrics, PACE-Diag retains target-side failure classifications and probe-level source-relative attribution as complementary outputs. Target-side classifications are assigned at the failed-run level: each observed failed run receives one failure category, and repeated failures of the same probe contribute separate observations, including runs from probes with relation $Different$. Probe-level attribution is determined jointly over all failed responses and yields a single relation $m(q)$. These outputs serve different diagnostic purposes: run-level categories characterize the observed failure composition across models and domains, whereas probe-level attribution assesses recurrence relative to a specified source Difficulty Card. Their joint analysis exposes differences that would be obscured by either behavioral outcomes or broad failure categories alone.

4 EXPERIMENTS

We evaluate PACE around five questions: (1) when does an observed benchmark failure become a usable diagnostic seed; (2) how does failure on a source-guided probe relate to recurrence of the source reasoning breakdown; (3) how much mechanism-level information is preserved by conventional error categories; (4) how reliable are the resulting diagnostic judgments; and (5) whether the protocol remains operational on an independent physics corpus.

4.1 EXPERIMENTAL SETUP

Datasets & Models. We draw source problems from UGPhysics (Xu et al., 2025), focusing on electromagnetism, statistical mechanics, and atomic physics—1,865 problems in total—which provide sufficient source problems after evidence attrition and span distinct knowledge and tool sets. We explicitly distinguish the evaluated target models from the LLMs powering the diagnostic pipeline. The cross-domain experiment evaluates Gemini 3.7 Flash (Google DeepMind, 2026) across all three subfields. The cross-model experiment fixes atomic physics. For each of GPT-5.6 Sol (OpenAI, 2026), Kimi K3 (Kimi Team, 2026), and Gemini 3.7 Flash, all incorrect responses within the fixed atomic-physics subset are passed through the PACE pipeline; the resulting probes are then evaluated on all three models, yielding a 3×3 transfer matrix that includes self-evaluation on the diagonal.

The PACE pipeline itself uses MiniMax (MiniMax, 2026) for failure analysis and probe generation, and DeepSeek (DeepSeek-AI, 2026) for validation and mechanism attribution (Appendix A.1).

Evaluation Protocol. Initial failures are not determined solely by the original benchmark labels: we adapt and extend the released grading scripts and add an LLM-based final-answer verification step, so that the failure pool entering source auditing reflects both answer-level and script-level consistency. Repeated measurement uses $K = 3$ throughout, with at most two bounded recovery attempts for unresolved items. We apply a 20% pass-side control sample in the cross-domain setting; the cross-model experiment reuses existing failure seeds without an additional pass-side control. External validation on PHYSICS (Feng et al., 2025) and human audit procedures are detailed in Section 4.6 and Section 4.5, respectively.

4.2 When Does a Benchmark Failure Become a Diagnostic Seed?

A natural first safeguard is repetition: if the same item fails across repeated evaluations, the error seems more likely to reflect a persistent weakness. We therefore first ask whether behavioral persistence is sufficient to make an observed failure diagnostically usable. Figure 3 contrasts behavioral persistence with source-side qualification and summarizes the audit outcomes that separate operational failures from admitted diagnostic seeds.

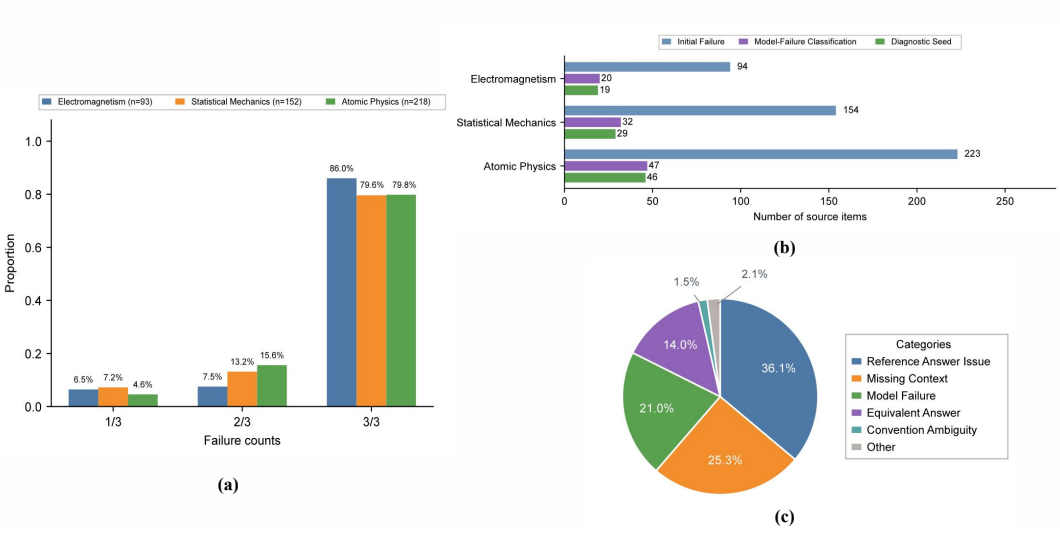

Figure 3: (a) Failure persistence across three repeated evaluations, computed over complete P/F trajectories and grouped by subfield. (b) Source-side qualification from operational initial failures to admitted source-level diagnostic seeds. (c) Overall source-audit outcomes across all initial failures, based on analyzer classifications.

Most observed failures in our setting satisfy this behavioral criterion: about 81% of complete P/F trajectories fail in all three evaluations, while 19% are mixed. **Behavioral persistence, however, does not imply diagnostic usability.** Auditing the 471 operational initial failures admits only 94 as diagnostic seeds.

The reasons for exclusion are heterogeneous. Among the 471 initial failures, 236 involve reference or scoring issues or equivalent symbolic answers, 126 involve under-specified context or convention-dependent interpretations, and 10 remain unattributable or lack sufficient evidence for analysis. The remaining 99 are classified as model failures; 94 yield accepted Difficulty Cards. **An operational failure label alone is insufficient for diagnostic attribution.**

These outcomes are conditional on operational failure and should not be interpreted as a benchmark-wide invalidity rate. They instead show that a source item must support a sufficiently specific and scientifically defensible reasoning breakdown before it can seed a probe. One representative case involves mathematically equivalent inverse-trigonometric forms judged differently by the operational grading pipeline; additional cases are provided in Appendix B.1.

A pass-side control suggests a complementary asymmetry: re-evaluating 20% of initially correct responses ($n = 268$) yields 91.8% stable-pass, indicating that correct responses may be more stable than failed ones under repeated evaluation. Because the control sample is smaller and not matched to the failure-side population, we treat it as a control signal rather than a symmetric stability estimate.

Source qualification yields a set of diagnostically usable failures, but whether those failures recur under controlled variation remains to be tested.

4.3 How Does Probe Failure Relate to Mechanism Recurrence?

Source auditing establishes usable diagnostic seeds, but not whether the diagnosed weakness survives controlled reformulation. These probes are not arbitrary new problems: each is generated from an admitted Difficulty Card under a preservation–variation contract, with controlled structural or numerical variation, and passes structural, numerical, and scientific validation before evaluation. If source-guided construction were sufficient to make probe failure a reliable indicator of recurrence, failures on these probes should reproduce the source breakdown. But only 56.4% do.

Across 649 observed-F probe–target instances, 366 are $Same$, 99 $Partial$, and 184 $Different$. Strict recurrence is 56.4%, rising to 71.6% when $Partial$ cases are included. Among Gemini self-route FFF probes across the three domains, 23.9% are still non-$Same$ (Appendix D.3). **Generation-time diagnostic targeting does not guarantee evaluation-time mechanism recurrence**: even under controlled construction and validation, a substantial fraction of failed probes only partially reproduce the source breakdown or fail through a different one.

Table 1: Cross-model failure occurrence and strict mechanism recurrence. Occurrence is computed over probes with at least one resolved outcome; strict recurrence counts only $Same$ attributions over failed probes. Within each source-model pool, all targets receive the same frozen probe set. All values are percentages.

(a) Failure occurrence

Source model	Target model		
	GPT	Gemini	Kimi
GPT	63.8	49.2	63.9
Gemini	65.3	74.2	63.8
Kimi	48.3	39.7	62.3

(b) Strict recurrence

Source model	Target model		
	GPT	Gemini	Kimi
GPT	67.6	58.3	48.7
Gemini	48.1	68.4	44.4
Kimi	57.1	64.6	52.6

GPT-derived probes induce failures on Kimi at 63.9%, yet only 48.7% of those failures reproduce the source mechanism; conversely, Kimi-derived probes induce failures on Gemini at 39.7%, while 64.6% are strictly $Same$. Failure susceptibility and source-relative recurrence therefore need not move together, and neither follows a uniform diagonal pattern across models. The broad recurrence matrix is reported in Appendix D.1.

Exact mechanism recurrence is therefore not determined by either failure frequency or source–target identity alone. We next ask whether conventional error categories capture this distinction.

4.4 Do Error Categories Capture Mechanism Recurrence?

Section 4.3 showed that failure on a probe does not ensure mechanism recurrence. Once failure occurrence proves insufficient, a natural alternative is to examine run-level error categories: if a source failure and a target failure belong to the same broad category, does that establish that the same mechanism has recurred? We test this category-level proxy against probe-level source-relative attribution. The two outputs are distinct: run-level categories describe how a target failure manifests, while probe-level attribution determines whether it reproduces the source reasoning breakdown.

Error categories are informative for describing failures. As shown in Figure 4(a), failure-mode composition varies substantially across physics domains: hidden-assumption failures dominate electromagnetism (47.2%), formula-applicability failures concentrate in atomic physics (32.9%), and boundary-condition or approximation failures are more prominent in statistical mechanics (26.7%). The taxonomy therefore provides a useful descriptive view of domain-level failure structure.

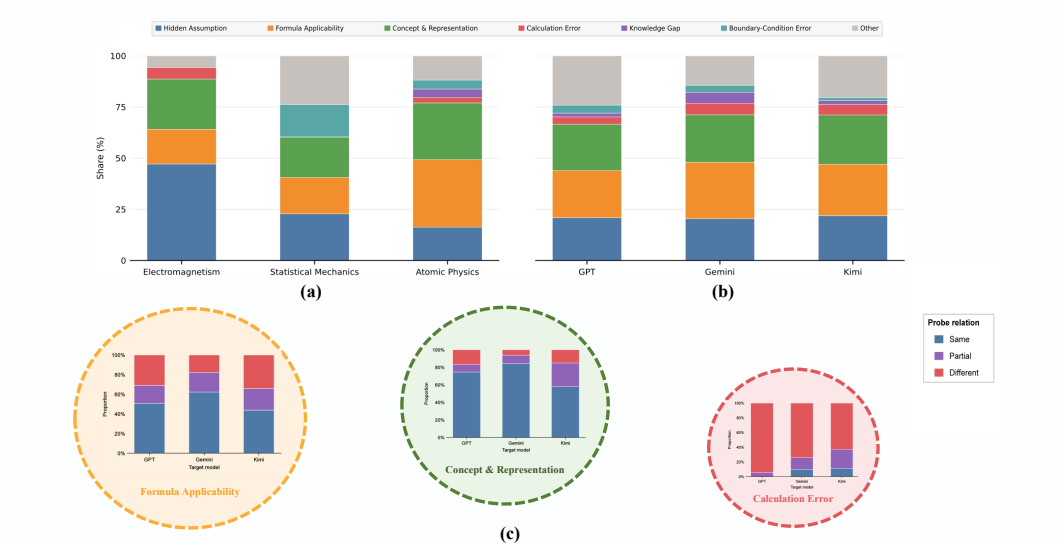

Figure 4: Attribution-aware failure characterization. (a) Run-level failure-category composition for Gemini across the three physics domains. (b) Corresponding composition for GPT, Gemini, and Kimi on the shared atomic probe pools. (c) Source-relative attribution within selected target-side failure categories, showing the proportions of $Same$, $Partial$, and $Different$ probe relations across target models. Panels (a) and (b) are normalized over failed target runs; panel (c) operates at the probe level. A probe is counted in every run-level category appearing in its failed runs. Display labels in the legend use a compact form; the canonical taxonomy is in Appendix B.3.

Figure 4(b) shows a different pattern across models. Across the atomic cross-model evaluation, the three target models share a comparatively stable dominant profile: formula applicability, concept/representation, and hidden assumption together account for 66.6–71.2% of failed runs in each model, with the clearest model-level differences appearing in smaller tail categories. **Broad error categories capture how failures vary across domains and models, but this does not yet establish whether the source reasoning breakdown recurs.**

After normalizing source and target labels into a shared taxonomy, 451 probes have all failed runs aligned with the source category. Yet 107 (23.7%) are still not attributed as $Same$, comprising 46 $Partial$ and 61 $Different$ (Appendix D.2). Figure 4(c) shows this at the category level: within formula-applicability probes, only 48.0% are $Same$, while 24.6% are $Partial$ and 27.4% are $Different$. A single target-side category can therefore contain substantially different mechanism relations. **Categorical similarity does not establish mechanism recurrence.**

The converse also occurs: 22 probes attributed as $Same$ have no category-aligned failed run at all. **Category alignment is neither sufficient nor necessary for strict mechanism recurrence.** Error categories describe the form of a failure; source-relative attribution tests whether the source reasoning breakdown recurs. Because these fine-grained distinctions rely on trace-level judgments, we next evaluate their reliability through human audit.

4.5 Human Audit

We assess the reliability of automated judgments through a human audit with six independently sampled groups of 50 cases each (300 cases in total): source auditing, card abstraction, probe synthesis, run-level diagnosis, mechanism attribution, and end-to-end chains. Four reviewers with physics backgrounds independently evaluated each case, using three rating levels: fully supported, minor caveats (a correctable limitation that does not overturn the judgment), or invalidating issue. We report both fully supported rates and a broader supported rate (full + caveats); the latter indicates the absence of invalidating issues rather than full correctness. Sampling and aggregation procedures are detailed in Appendix E.

Table 2: Human audit of diagnostic judgments. Supported (Full + Caveats) indicates that no invalidating issue was found. Percentages are over reviewer ratings (200 per stage).

Stage	Fully Supported	Minor Caveats	Supported	Invalidating Issue
Source Auditing	93.5%	6.0%	99.5%	0.5%
Card Abstraction	93.0%	5.5%	98.5%	1.5%
Probe Synthesis	92.5%	7.0%	99.5%	0.5%
Run-level Diagnosis	81.5%	11.5%	93.0%	7.0%
Mechanism Attribution	89.5%	5.5%	95.0%	5.0%
End-to-End Chain	84.0%	11.0%	95.0%	5.0%

Table 2 reports stage-level and end-to-end results. Supported rates remain high across all groups, but the two downstream stages—run-level diagnosis and probe-level attribution—show lower fully supported rates than source auditing, card abstraction, and probe synthesis. This pattern suggests that validating a problem and its scientific context is more reliable than explaining why a target trace fails or whether two failures share a source mechanism. Case-level review highlights three recurring patterns among disputed cases: context-dependent physical assumptions that are not fully explicit in the source problem, traces in which multiple reasoning errors co-occur and the earliest critical step is ambiguous, and probes whose failures partially reproduce the source mechanism while introducing a new deviation. These are case-observed patterns rather than statistically inferred causes; the audit sample is stratified and does not support frequency estimates. The end-to-end audit directly evaluates complete diagnostic chains rather than multiplying stage-level rates; supported chains remain the majority, and residual issues concentrate in the attribution stages.

4.6 External Validation

We further evaluate PACE on PHYSICS (Feng et al., 2025), an independent multimodal benchmark spanning six physics domains. Its official text-only subset, which covers all six domains and contains 999 problems (approximately 77% of the corpus), serves as our evaluation pool. On Gemini 3.7 Flash, the accepted probes achieve 73.8% failure occurrence and 79.2% broad mechanism recurrence (Table 3) under the same definitions as Section 3.3. This indicates that PACE remains operational beyond its primary source pool, though it does not by itself establish cross-domain generalization.

Table 3: External validation on the PHYSICS benchmark (official text-only subset).

Stage	Value
Retained text-only problems	999
Admitted diagnostic seeds	24
Candidate probes	66
Passed validation	65
Failure occurrence	73.8%
Broad mechanism recurrence	79.2%

5 Discussion

Summary. We introduced PACE, a diagnostic framework for calibrating what an observed model failure can legitimately support. Across physics domains and models, our results show that commonly used proxies for diagnostic evidence—persistence, source-guided probing, and category agreement—capture different aspects of failure behavior but are not interchangeable with source-mechanism recurrence. This motivates treating failure diagnosis as an evidence-grading problem rather than a binary extension of benchmark accuracy.

Limitations and future work. Our study has three main boundaries. First, the cross-model analysis covers three frontier models at fixed snapshots; extending the protocol to more models and model versions is a natural next step. Second, source qualification is evaluated on two physics benchmarks, and extending the protocol to other scientific domains may require domain-specific validation criteria. Third, mechanism attribution is based on observable reasoning traces and source-relative comparison against a Difficulty Card, so our claims concern operational recurrence rather than direct access to internal model mechanisms.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks with required disclosure: generating synthetic datasets (i.e., the diagnostic probes produced by our pipeline). We also used generative AI tools for the following tasks with recommended disclosure: aiding and polishing writing, retrieval and discovery (e.g., finding related work), research ideation or execution, and drafting sections of the paper. We have not used generative AI tools for the following tasks with required disclosure: formulating mathematical claims or proofs, and finalizing result interpretations that require human judgment. All AI-assisted work has been reviewed by the authors. For instance, LLM-generated code was tested and verified for correctness, and all numerical results and claims were checked against the saved experimental artifacts. We take full responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work evaluates reasoning failures of large language models on existing physics benchmarks and generated diagnostic probes. All source problems are drawn from publicly available benchmarks, and the generated probes contain no personal or sensitive information. The study does not involve sensitive personal data or deployment on human users. Human reviewers participated only in the retrospective audit of diagnostic judgments and were not subjects of behavioral or psychological experimentation. We have reviewed and adhered to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

All target-model evaluations in this work were conducted through public model APIs. We report model roles, inference settings, outcome adjudication rules, and evaluation protocols in Appendix A. The appendix further documents construction, validation, attribution criteria, representative diagnostic cases, extended analyses, and human-audit procedures. PACE-Bench and the full executed prompt templates for the canonical pipeline components are included in the supplementary material. Additional structured diagnostic materials will be released publicly later. Raw model transcripts, intermediate API logs are not included.

REFERENCES

Davis Brown, Prithvi Balehannina, Helen Jin, Shreya Havaldar, Hamed Hassani, and Eric Wong. Adaptively profiling models with task elicitation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24985–25020, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1270. URL https://aclanthology.org/2025.emnlp-main.1270/.

Angelica Chen, Jason Phang, Alicia Parrish, Vishakh Padmakumar, Chen Zhao, Samuel R. Bowman, and Kyunghyun Cho. Two failures of self-consistency in the multi-step reasoning of LLMs. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=5nBqY1y96B.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, and et al Henrique Ponde de Oliveira Pinto. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

Yongrui Chen, Yangyang Ma, Xiaoying Huang, Shenyu Zhang, Huajun Chen, Haofen Wang, and Guilin Qi. Stresseval: Failure-driven dynamic benchmarking for knowledge-intensive reasoning in large language models, 2026. URL https://arxiv.org/abs/2605.01939.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026. URL https://arxiv.org/abs/2606.19348.

Kaiyue Feng, Yilun Zhao, Yixin Liu, Tianyu Yang, Chen Zhao, John Sous, and Arman Cohan. Physics: Benchmarking foundation models on university-level physics problem solving. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11717–11743, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.610. URL https://aclanthology.org/2025.findings-acl.610/.

Miguel Angel Alvarado Gonzalez, Michelle Bruno Hernandez, Miguel Angel Peñaloza Perez, Bruno Lopez Orozco, Jesus Tadeo Cruz Soto, and Sandra Malagon. Do repetitions matter? strengthening reliability in llm evaluations, 2025. URL https://arxiv.org/abs/2509.24086.

Google DeepMind. Gemini 3.7 Flash model card. Technical report, Google DeepMind, August 2026. URL https://deepmind.google/models/model-cards/gemini-3-7-flash/.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.211. URL https://aclanthology.org/2024.acl-long.211/.

Jiahang He, Rishi Ramachandran, Neel Ramachandran, Aryan Katakam, Kevin Zhu, Sunishchal Dev, Ashwinee Panda, and Aryan Shrivastava. Modeling and predicting multi-turn answer instability in large language models, 2025. URL https://arxiv.org/abs/2511.10688.

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*, 2021a. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021b. URL https://openreview.net/forum?id=7Bywt2mQsCe.

Yue Huang, Zhengzhe Jiang, Yuchen Ma, Yu Jiang, Xiangqi Wang, Yujun Zhou, Yuexing Hao, Kehan Guo, Pin-Yu Chen, Marzyeh Ghassemi, Stefan Feuerriegel, and Xiangliang Zhang. ProbeLLM: Automating principled diagnosis of LLM failures. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=2rHDObXYYH.

Shima Imani, Seungwhan Moon, Adel Ahmadyan, Lu Zhang, Ahmed Kirmani, and Babak Damavandi. SymPyBench: A dynamic benchmark for scientific reasoning with executable python code. In Yevgen Matusevych, Gülşen Eryiğit, and Nikolaos Aletras (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track)*, pp. 105–118, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-384-5. doi: 10.18653/v1/2026.eacl-industry.8. URL https://aclanthology.org/2026.eacl-industry.8/.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4110–4124, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.324. URL https://aclanthology.org/2021.naacl-main.324/.

Kimi Team. Kimi K3: Open frontier intelligence, July 2026. URL https://www.kimi.com/en/blog/kimi-k3.

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*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=HjwK-Tc_Bc.

MiniMax. MiniMax M3: Frontier coding, 1m context, native multimodality—all in one model, May 2026. URL https://www.minimax.io/blog/minimax-m3. Accessed: 2026-09-24.

Seyed Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=AjXkRZIvjB.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A new benchmark for natural language understanding. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4885–4901, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.441. URL https://aclanthology.org/2020.acl-main.441/.

OpenAI. GPT-5.6 system card. Technical report, OpenAI, 2026. URL https://deploymentsafety.openai.com/gpt-5-6.

Nearchos Potamitis, Vansh Ramani, Har Ashish Arora, Dhairya Kuchhal, Lars Klein, and Akhil Arora. Reasonbench: Benchmarking the (in)stability of llm reasoning, 2025. URL https://arxiv.org/abs/2512.07795.

Shi Qiu, Shaoyang Guo, Zhuo-Yang Song, Yunbo Sun, Zeyu Cai, and et al Jiashen Wei. PHYBench: Holistic evaluation of physical perception and reasoning in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=brG8FPq1cf.

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. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=ITw9edRDlD.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, and et al Adam Fisch. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=uyTL5Bvosj. Featured Certification.

Siyuan Wang, Zhuohan Long, Zhihao Fan, Xuanjing Huang, and Zhongyu Wei. Benchmark self-evolving: A multi-agent framework for dynamic LLM evaluation. In Owen Rambow,

Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 3310–3328, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.223/.

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

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, and et al Xuan He. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=y10DM6R2r3.

Xin Xu, Qiyun Xu, Tong Xiao, Tianhao Chen, Yuchen Yan, Jiaxin ZHANG, Shizhe Diao, Can Yang, and Yang Wang. UGPhysics: A comprehensive benchmark for undergraduate physics reasoning with large language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=EmLiyZGvrR.

Wenbo Zhang, Hengrui Cai, and Wenyu Chen. Beyond the singular: Revealing the value of multiple generations in benchmark evaluation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 10033–10043, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.488. URL https://aclanthology.org/2026.findings-acl.488/.

Zhehao Zhang, Jiaao Chen, and Diyi Yang. DARG: Dynamic evaluation of large language models via adaptive reasoning graph. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=5IFeCNA7zR.

Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. Dyval: Dynamic evaluation of large language models for reasoning tasks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gjfOL9z5Xr.

APPENDIX

A EXPERIMENTAL SETUP AND EVALUATION PROTOCOL

A.1 MODEL AND INFERENCE PROFILES

Table 4: Model roles across the diagnostic pipeline.

Role	Model	Reasoning setting
Target evaluation	Gemini 3.7 Flash	high reasoning
Target evaluation	GPT-5.6 Sol	max reasoning effort
Target evaluation	Kimi K3	max reasoning effort
Failure analysis and probe generation	MiniMax-M3	provider default
Validation and mechanism attribution	DeepSeek V4 Flash	provider default

All calls use a 900-second timeout, and sampling parameters follow the provider defaults unless otherwise specified.

A.2 OUTCOME ADJUDICATION AND REPEATED MEASUREMENT

Each model response is adjudicated as Pass (P), Fail (F), or Unresolved (U). Deterministic answer checks are used where applicable, and LLM adjudication is reserved for cases that require semantic or scientific equivalence judgments. Unresolved outcomes, such as timeout or unparseable responses, are never recoded as failures. Repeated measurement uses three evaluation slots; request-level retries may recover transport or parsing failures within a slot but do not create new observations, and an unresolved slot may be re-evaluated within the three-slot budget.

The resulting trajectory is classified as stable pass (all P), stable failure (all F), mixed pass/fail (P and F without U), or unresolved (any U). A U-and-F trajectory remains behaviorally unresolved, but its observed failures still enter mechanism attribution. Any trajectory with at least one observed failure remains eligible for mechanism attribution, regardless of its behavioral-state label.

B SOURCE QUALIFICATION AND DIFFICULTY CARDS

B.1 SOURCE QUALIFICATION AND REPRESENTATIVE CASES

The Source Analyzer assigns each source failure one of five mutually exclusive decisions: model failure, equivalent answer, reference issue, ambiguous problem, or unattributable failure. Only valid source failures with a sufficiently supported reasoning breakdown can produce Difficulty Cards.

Table 5: Source-audit decisions and their operational interpretation.

Decision	Interpretation
Model failure	The source and grading basis are defensible, and at least one response contains a specific and diagnostically reusable scientific error.
Equivalent answer	The apparent failure disappears after algebraic, numerical, unit, rounding, or expression-equivalence checks.
Reference issue	The reference answer or grading basis is not scientifically correct for the stated problem.
Ambiguous problem	Essential information is absent, or equally standard interpretations yield incompatible answers.
Unattributable failure	The response is wrong, but the saved evidence does not support a reusable reasoning breakdown.

Figure 3(c) groups the detailed Analyzer decisions together with the remaining operational outcomes into six mutually exclusive display categories for visualization. A benchmark-scored wrong answer

can therefore be a reference defect, an equivalent answer, an ambiguous item, or a valid but non-attributable failure, and exclusion from the Card pool does not imply that the benchmark item is intrinsically invalid.

Case presentations are abbreviated for readability and include only the evidence relevant to the illustrated decision; intermediate reasoning not material to the reported adjudication or attribution is omitted. The two cases below show the two most common exclusion reasons: a response that is actually equivalent to the reference, and a reference value that is itself incorrect.

Excluded Failure: Equivalent Answer

Source problem.
A conducting sphere of radius R carries charge Q. A point charge q is placed at distance $a > R$ in a dielectric of permittivity ε. Determine the potential at the location of q using the image-charge construction.

Model answer.
$\frac{1}{4\pi\varepsilon}\Big[\frac{Q}{a} - \frac{qR^3}{a^2(a^2-R^2)}\Big]$.

Reference.
$\frac{1}{4\pi\varepsilon}\Big[\frac{Q}{a} + \frac{qR}{a}\Big(\frac{1}{a} - \frac{1}{a - R^2/a}\Big)\Big]$.

Audit finding.
The two expressions are algebraically identical; the apparent discrepancy is rearrangement, and the permittivity notation is a convention difference. The repeated mismatch with the reference is not a model failure.

Decision.
Equivalent answer; no Difficulty Card.

Excluded Failure: Reference Issue

Source problem.
A soap bubble of radius $1\,\mathrm{cm}$ is at potential $100\,\mathrm{V}$. Determine its charge Q.

Model answer.
$Q = RV/k \approx 1.11 \times 10^{-10}\,\mathrm{C}$.

Reference.
$Q \approx 1.11 \times 10^{-11}\,\mathrm{C}$.

Audit finding.
An independent calculation confirms the model value. The apparent failure is caused by an incorrect reference value rather than an error in the model response.

Decision.
Reference issue; no Difficulty Card.

B.2 Difficulty Card Representation

A Difficulty Card records a source-side reasoning breakdown rather than a final answer mismatch. Its fields cover source context, the diagnostic target, the physics context, the supporting evidence, and review metadata. Card review is a separate decision with values accept, revise, reject, or unclear, and the Card serves as the interface from source diagnosis to probe synthesis and subsequent source–target attribution. The example below shows an accepted Card from the Gemini Atomic source pool.

Difficulty Card Example

Source problem.
The electronic shell of a diatomic molecule consists of three electrons in the σ, π, and δ orbitals. Determine the possible spectroscopic terms of the electronic state.

Bottleneck step.
Combining Λ and S into final spectroscopic term symbols.

Model action.
Treats Clebsch–Gordan state multiplicity as separate spectroscopic terms and omits the Ω sublevels.

Correct action.
List one term per (Λ, S) pair—one $^2\Pi$, one $^4\Pi$, one $^2\Phi$, and one $^4\Phi$—each with its allowed $\Omega = |\Lambda+\Sigma|$ sublevels.

Error mechanism.
Confusion between Clebsch–Gordan state multiplicity and spectroscopic term multiplicity.

Trigger.
Molecular term-symbol problems where a single (Λ, S) pair admits multiple spin-coupling states.

Core equations.
$\Lambda = \left|\sum_i \lambda_i\right|$; $\frac{1}{2} \otimes \frac{1}{2} \otimes \frac{1}{2} = \frac{1}{2} \oplus \frac{1}{2} \oplus \frac{3}{2}$; $\Omega = |\Lambda + \Sigma|$.

Misused equations.
Counts $S = 1/2$ twice within one (Λ, S) block as two independent doublet terms.

Supporting runs.
Three confirmed source failures, all reproducing the same multiplicity confusion.

Critical reasoning requirements.
Combine Λ and S, identify one term per pair, and enumerate the allowed Ω sublevels without duplicating spin-coupling states.

B.3 Failure Taxonomy

The failure taxonomy comprises boundary misjudgment, hidden assumption, knowledge gap, concept representation, formula applicability, convention sign, approximation simplification, multi-step reasoning, calculation algebra, unit dimension, instruction interpretation, uncertainty incomplete, and other. Failure category, strategy alignment, and source–target mechanism relation are reported as separate variables. Three boundaries are worth making explicit: formula applicability concerns whether a relation is valid in the stated regime, whereas a purely algebraic slip is calculation algebra; approximation simplification requires an unjustified approximation or truncation, whereas a regime-selection error may instead be boundary misjudgment; and uncertainty incomplete is a target-side failure category distinct from the behavioral U state.

C Probe Construction, Validation, and Attribution

C.1 Generation Routes and Strategy Inventory

Each admitted Difficulty Card can produce up to one candidate probe per synthesis route—an anchor probe with minimal parameter or surface changes, a strategy-modified probe applying a single strategy, and a composite-strategy probe combining two or more strategies—and all candidates pass through the same validation gate, so the number of retained probes per Card is at most three. The preservation–variation contract constrains generation toward the source diagnostic target, but does not establish that the target model will reproduce the source breakdown; recurrence is assessed only after target evaluation and mechanism attribution. The nine strategies used for strategy-modified and composite-strategy probes are listed in Table 6. Historical strategy usage serves only as lightweight context to encourage coverage and diversity, and is not optimized against downstream failure or attribution outcomes.

Table 6: Strategy inventory.

Strategy	Intended transformation
Parameter extrapolation	Push a controlling parameter into a demanding but valid regime.
Formula confusion	Create a temptation between similar-looking relations with different validity domains.
Boundary stress	Place the decision near a physical regime or applicability boundary.
Approximation trap	Make a simple approximation locally tempting although an exact or higher-order relation is required.
Slope/concept swap	Distinguish directional, slope, rate, or representation quantities with similar numerical roles.
Hidden-condition removal	Weaken an idealizing condition and require the solver to recognize its failure.
Multi-step decoupling	Preserve a coupled dependency that appears separable across multiple reasoning steps.
Empirical-formula disambiguation	Distinguish competing empirical relations by configuration and validity conditions.
Data-conflict rejection	Present inconsistent evidence and test whether the solver rejects a fabricated unique answer.

Card-to-Probe Example

The Difficulty Card in Appendix B.2 is used to generate a probe along the strategy-modified route with the **Formula confusion** strategy.

Generated probe.
The orbital set σ, π, δ is replaced by π, δ, ϕ, increasing the number of Λ values. The probe asks for the complete list of spectroscopic terms in ${}^{2S+1}\Lambda_{\Omega}$ notation, including all allowed Ω sublevels.

Validation.
Passes the structural, numerical, and scientific gates, including the physics-consistency audit, and is admitted to the benchmark pool.

C.2 Validation Protocol

Each candidate passes a multi-stage validation gate before admission to target evaluation.

Table 7: Validation gates.

Gate	Main checks	Action
Structural	Schema, required fields, standalone completeness, answer form, leakage.	Pass, bounded repair, or reject.
Numerical	Equations, units, arithmetic, rounding, and answer consistency, verified through executable checks.	Pass, bounded repair, or reject.
Scientific	Physical validity, unique solvability, semantic consistency, and strategy manifestation, checked by two complementary review passes.	Accept, bounded repair, or reject.

The numerical gate is complemented by a deterministic consistency check that inspects candidate semantics, symbol mapping, and answer conditions before scientific review, and a deterministic rejection cannot be overridden by an LLM acceptance. A candidate may receive at most one bounded repair; a repaired candidate must pass all gates again before admission, and a candidate that still requires repair or is rejected after repair is excluded from the benchmark pool. Operational failures such as API or parsing errors are recorded as `pending` rather than as scientific failures.

Bounded repair may correct a locally determined issue such as notation, rounding, units, or a missing standard assumption; meaningful redesign, missing indispensable information, multiple valid answers, or failure to instantiate the intended diagnostic conditions are rejections.

C.3 Attribution Rubric and Metrics

For each probe with at least one observed failure, the mechanism judge reads the source Card, the generated probe, the reference solution, the complete P/F/U trajectory, and all observed failed traces, and emits one probe-level relation.

Table 8: Probe-level mechanism relations.

Relation	Criterion
Same	The source reasoning breakdown recurs across the observed failed traces without a substantive competing breakdown.
Partial	The source breakdown recurs incompletely or non-uniformly, or co-occurs with another breakdown.
Different	The observed failed traces are better explained by a reasoning breakdown other than the source.
Unclear	The available trace evidence is insufficient for a defensible attribution.

All observed failed traces of a probe are judged jointly to produce one probe-level relation, and pass and unresolved runs are retained only as behavioral context. Failure occurrence is the fraction of behaviorally evaluable probes that induce at least one failure, whereas mechanism recurrence is conditional on an observed failure and reported under two thresholds that share the same observed-F denominator: strict recurrence counts only Same relations, and broad recurrence additionally counts Partial relations.

The example below shows a probe whose target failures do not reproduce the source mechanism, yielding a *Different* attribution.

Attribution Example

Source mechanism.
For the HCl infrared band, the source model treats the central gap ($4B$) as the relevant $2B$ spacing when extracting the rotational constant.

Target trajectory.
FFF.

Key trace evidence.
All three runs correctly identify the central gap as $4B$, but then treat the total spectral span as $10B$ and obtain an incorrect rotational constant and moment of inertia.

Run-level category.
`calculation_algebra`.

Probe-level relation.
Different.

Attribution rationale.
The target does not reuse the central-gap-versus-within-branch confusion that defines the source mechanism. All three failures instead share the same alternative counting error, so the probe is attributed as *Different* rather than *Same* or *Partial*.

C.4 Decision Contracts and Prompt Summary

Full executed prompt templates for all canonical components are provided in the **supplementary package**. The summary below covers the components most directly affecting source admission, probe construction, probe validity, and source-relative attribution.

Table 9: Canonical LLM components and their roles.

Component	Role
Source Analyzer	Source adjudication and Difficulty Card proposal.
Card Reviewer	Card acceptance, revision, or rejection.
Probe Generator	Controlled probe synthesis under the preservation–variation contract.
Probe Repairer	Bounded local repair of a candidate probe.
Scientific Validation	Two complementary review passes over candidate scientific validity and diagnostic-target alignment.
Target Evaluation	Target-model response acquisition.
Direct Probe Judge	Source-relative probe-level attribution.

Source Analyzer

- **Role.** Audit a scored source failure and decide whether it is diagnostically interpretable and reusable.
- **Inputs.** Standalone source problem, reference solution, model runs, and scoring evidence.
- **Outputs.** Decision, mechanism hypotheses, and zero to two Difficulty Cards.
- **Constraints.** The statement is audited before the reference; missing context is not reconstructed from the reference. A model failure requires one specific and diagnostically reusable error, not recurrence. Repeated final answers alone do not establish a shared mechanism. A hard ambiguity produces no Card. Necessary assumptions and standard conventions are not treated as leakage.

Probe Generator

- **Role.** Generate anchor, strategy-modified, or composite-strategy probes from a Difficulty Card.
- **Inputs.** Source problem, Difficulty Card, strategy inventory, and route assignment.
- **Outputs.** One candidate probe with problem, solution, verification code, intended failure mode, and generation metadata.
- **Constraints.** The candidate must remain scientifically well posed and preserve the conditions needed to test the diagnostic target identified by the Difficulty Card, while varying only the dimensions permitted by the assigned route. Anchor probes introduce minimal variation; strategy-modified and composite-strategy probes instantiate the assigned strategy operations without adding solution leakage. The generator does not determine whether the target model will reproduce the source breakdown. Verification code computes the reference answer from the stated givens rather than hard-coding it.

Scientific Validation

- **Role.** Determine whether a candidate probe is scientifically valid and aligned with the intended diagnostic target. The scientific-validation stage uses two complementary review passes.
- **Inputs.** Source problem, Difficulty Card, candidate probe, and deterministic evidence.
- **Outputs.** Accept, bounded repair, or reject, with semantic status, precision status, and difficulty-preservation status.
- **Ordinary validator.** Checks well-posedness, uniqueness, physical regime, and strategy manifestation; the problem must admit a well-defined target quantity and a defensible reference answer.
- **Physics-consistency audit.** Separately checks high-risk physical contradictions, including coordinate and vector consistency, quantity semantics, regime assumptions, and answer-to-statement consistency, and does not trust the generator's self-reported validity.
- **Constraints.** Diagnostic-target alignment requires that the candidate retain the observable conditions needed to test the Card-defined weakness and avoid declared leakage risks. This check concerns probe construction, not the target model's eventual failure mechanism.

Direct Probe Judge

- **Role.** Assign a probe-level mechanism relation between the source Card and the observed failed traces of a probe.
- **Inputs.** Source Card mechanism, generated probe, reference solution, complete P/F/U trajectory, and all observed failed traces.
- **Outputs.** One probe-level relation, a rationale, and per-run notes.
- **Constraints.** Same requires consistent support for the source breakdown across the available failed traces. Partial captures incomplete, non-uniform, or co-occurring reproduction. Different is used when the observed failures are better explained by another breakdown, and Unclear when the available failed-trace evidence is insufficient for a defensible probe-level attribution. Pass and unresolved runs are retained as behavioral context but are not themselves mechanism judgments. Every supplied failed run is listed in the output so that the judge's coverage of the failure set can be checked.

The four components above determine the core diagnostic chain from source admission to mechanism attribution. The following components support the pipeline without directly determining its diagnostic outcomes, and their full executed prompt templates are provided in the **supplementary package**.

Card Reviewer

- **Role.** Independently review a proposed Difficulty Card.
- **Inputs.** Proposed Card, supporting failed runs, and any deterministic lint issues.
- **Outputs.** Accept, revise, reject, or unclear, with an optional revised Card.
- **Constraints.** A non-empty deterministic lint report forbids acceptance. Recurrence is descriptive evidence and is not required for Card eligibility. A hard ambiguity is rejected rather than repaired.

Probe Repairer

- **Role.** Perform one bounded repair after an approved validation issue.
- **Inputs.** Candidate probe, validation evidence, and the approved repair instructions.
- **Outputs.** Repaired candidate content, or a repair-impossible decision.
- **Constraints.** Repair is limited to local issues such as rounding, units, a missing standard assumption, or a local code or answer correction; redesign-level issues are returned as repair-impossible.

Target Evaluation

- **Role.** Obtain a target-model response for source problems and admitted probes.
- **Inputs.** Problem text.
- **Outputs.** A boxed final answer with a concise derivation.
- **Constraints.** Responses provide a concise derivation and a boxed final answer, without unnecessary alternative solution paths.

D Extended Evidence

D.1 Broad Cross-Model Recurrence

Table 1 in the main text reports failure occurrence and strict recurrence. Broad recurrence, which additionally counts Partial attributions, is reported in Table 10 for the same nine Atomic route cells.

The gap between broad and strict recurrence is the share of observed-F probes attributed as Partial. Figure 5 shows that this gap is non-uniform across source–target routes, ranging from 10.5 to 24.4 percentage points. The largest increase appears on GPT $\rightarrow$ Kimi (24.4 pp), followed by

Table 10: Broad cross-model mechanism recurrence. Each cell is the percentage of observed-F probes with a Same or Partial relation.

Source model	Target model		
	GPT	Gemini	Kimi
GPT	78.4	75.0	73.1
Gemini	59.3	78.9	63.0
Kimi	69.6	77.1	72.4

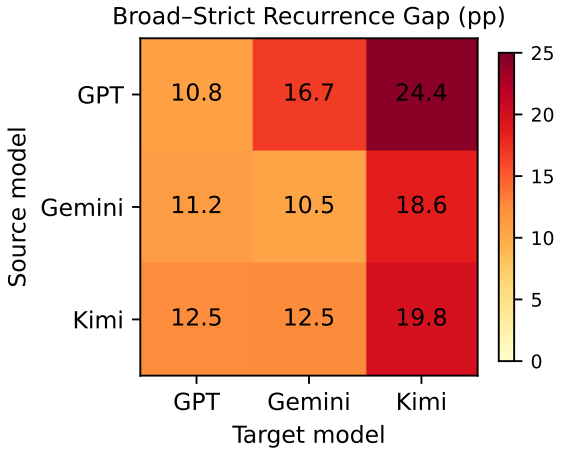

Figure 5: Increase from strict to broad mechanism recurrence (percentage points) on the Atomic cross-model probe pool. Each cell reports the contribution of Partial attributions, computed as broad recurrence minus strict recurrence for the corresponding source–target route. The non-uniform pattern shows that qualified recurrence is not a marginal phenomenon: on several routes, especially those targeting Kimi, Partial contributes a substantial share of the recurrence evidence.

Kimi → Kimi (19.8 pp) and Gemini → Kimi (18.6 pp). Partial attributions are therefore not a marginal label, but a substantial component of recurrence evidence on several routes, especially those targeting Kimi.

D.2 Category Alignment and Mechanism Recurrence

Section 4.4 compares run-level category alignment against probe-level mechanism attribution. Table 11 reports the joint distribution over the 649 Atomic observed-F probe–target instances, where each cell counts instances by whether all observed failed runs align with the source category and by the probe-level relation assigned by the mechanism judge.

Table 11: Joint distribution of category alignment and probe-level mechanism relation over 649 Atomic observed-F probe–target instances.

Relation	All aligned	Mixed	None aligned	Total
Same	344	0	22	366
Partial	46	41	12	99
Different	61	6	117	184
Total	451	47	151	649

Among the 451 probes whose failed runs all align with the source category, 107 (23.7%) are not attributed as Same, while 22 probes attributed as Same have no category-aligned failed run. No Atomic

observed-F probe in this analysis received an *Unclear* attribution. The disagreement is asymmetric: category agreement often overstates exact recurrence, while a smaller but non-negligible set of Same probes is missed by category matching altogether. The two cases below show how category alignment and mechanism attribution can diverge.

Category-Aligned but Mechanism-Different

Source mechanism.
The source Card concerns omission of the wiggler vector potential A_u from the total magnetic field when computing $\beta \times B$ in an ultra-relativistic electron beam passing through a wiggler.

Target trajectory.
FFF.

Run-level category.
All three failed runs are labeled `multi_step_reasoning`, matching the source-level category.

Key trace evidence.
All three runs correctly include A_u in the total vector potential, handling the source bottleneck. They then fail by deriving a transverse β_x from canonical momentum conservation, producing an extra z-component in $\beta \times B$.

Probe-level relation.
Different.

Attribution rationale.
Source and target fall into the same broad multi-step reasoning family while failing at different physical steps, showing that run-level category agreement does not imply mechanism recurrence.

Partial Attribution at the Category Boundary

Source mechanism.
The source model takes an algebraic shortcut by rearranging the geometric identity $V = (4\pi/3)r_0^3 A$ into $A = 3V/(4\pi r_0^3)$, bypassing the Fermi-gas phase-space calculation.

Target trajectory.
FFF.

Run-level category.
All three failed runs are labeled `formula_applicability`, matching the source-level category.

Key trace evidence.
Runs 1 and 2 reproduce the source algebraic shortcut. Run 3 instead writes the correct phase-space density-of-states relation but concludes that A is arbitrary and does not provide the requested expression in terms of V, p_F, and h.

Probe-level relation.
Partial.

Attribution rationale.
All three runs share the same broad run-level category, but only two directly reproduce the source mechanism. The third is why the aggregate probe-level relation is *Partial* rather than *Same*, showing that a broad category label does not by itself establish mechanism recurrence.

D.3 Behavioral Stability and Mechanism Recurrence

Section 4.3 notes that 23.9% of behaviorally stable Gemini self-route probes are not attributed as Same. Table 12 reports this breakdown by physics domain, with FFF probes, that is, probes that fail on all three evaluations, as the denominator.

Overall, 28 of 117 behaviorally stable probes (23.9%) are Partial or Different. The domain-wise breakdown is descriptive; the sample sizes are not used to infer a domain effect.

Table 12: Mechanism relation among FFF probes on the Gemini self-route, by domain. Percentages are computed within each row.

Domain	FFF probes	Same	Partial	Different
Electromagnetism	15	86.7	6.7	6.7
Statistical Mechanics	29	82.8	3.4	13.8
Atomic Physics	73	71.2	8.2	20.5
Overall	117	76.1	6.8	17.1

D.4 External PHYSICS Validation

The external PHYSICS validation reported in Section 4.6 uses the text-only subset. On Gemini 3.7 Flash, 48 of the 65 validated probes induce at least one observed failure. Among these observed-F probes, 34 are attributed as Same, 4 as Partial, and 10 as Different. Strict recurrence is therefore $34/48 = 70.8\%$, and broad recurrence is $(34 + 4)/48 = 79.2\%$.

E Human Audit

This appendix supplements the human-audit results reported in Section 4.5 with sampling details, inter-reviewer agreement statistics, and representative boundary cases.

E.1 Sampling and Rating Protocol

Each of the six audit stages was sampled independently with 50 units, drawn from five non-external conditions (Gemini–Electromagnetism, Gemini–Statistical Mechanics, Gemini–Atomic, GPT–Atomic, and Kimi–Atomic) at ten units per condition. Sampling used a fixed random seed and sampling without replacement from the eligible population. The audit is retrospective and unblinded: reviewers saw the system evidence, including the source problem, generated probe, target trajectory, and automated judgment. The end-to-end sample is a separate cross-stage check rather than the product of stage-level rates.

Four reviewers with physics backgrounds independently rated each unit using the three levels defined in Section 4.5. The audit comprises 300 units and 1,200 reviewer ratings; ratings are reviewer assessments of the supplied evidence, not automated ground truth.

E.2 Agreement and Disagreement

Table 13 reports Fleiss κ together with two raw agreement measures per stage: the proportion of units on which all four reviewers assigned the same rating, and the proportion on which at least three of four reviewers agreed. Stage names match the corresponding rows of the main-text human-audit table.

Table 13: Inter-reviewer agreement per audit stage.

Stage	Fleiss κ	4/4 agree	$\geq 3/4$ agree
Source Auditing	0.290	84.0%	96.0%
Card Abstraction	0.419	86.0%	96.0%
Probe Synthesis	0.235	80.0%	96.0%
Run-level Diagnosis	0.486	70.0%	92.0%
Mechanism Attribution	0.535	84.0%	94.0%
End-to-End Chain	0.309	66.0%	86.0%

Fleiss κ is sensitive to category prevalence and can be reduced when ratings are strongly concentrated in one category. Across the 1,200 individual reviewer ratings, 89% are fully supported

and only 3.25% are invalidating, so raw and chance-corrected agreement can diverge. We therefore report raw agreement alongside κ to distinguish observed reviewer concordance from chance-corrected agreement. Among all stages, mechanism attribution—the stage most directly tied to the paper's central diagnostic claim—reaches the highest κ (0.535), while stages dominated by near-unanimous ratings show lower chance-corrected agreement. Inspection of disagreement cases suggests that the remaining reviewer differences are concentrated around boundary judgments, including implicit physical assumptions, co-occurring errors, and Partial versus Different distinctions. Representative cases for two of these patterns are provided below, and the Partial versus Different boundary is illustrated in Appendix D.2.

E.3 Representative Boundary Cases

The two cases below illustrate representative disagreement patterns identified in Section 4.5. Each case reports the system judgment, the distribution of reviewer ratings, and the nature of the disagreement. Cases are selected to be representative of a disagreement pattern, not to estimate how often it occurs.

Boundary Case 1: Different Attribution under a Shared Convention

Stage. Mechanism attribution.
System judgment. *Different*, based on the target using a full relativistic momentum and a consistent mass convention instead of the source's half-momentum convention.
Key trace evidence. All three target runs use the full relativistic momentum and preserve the mass convention, avoiding the source's half-momentum error while failing at a later step.
Reviewer ratings. Three reviewers rated the judgment *fully supported*; one rated it *minor caveats*, noting that the shared physical setting leaves the attribution boundary less sharp.
Disagreement. Whether a target failure under the same physical setting but at a different downstream step should be attributed as *Different* or only *Partial* with respect to the source mechanism.
Pattern. Shared physical setting but a different downstream reasoning failure.

Boundary Case 2: Co-occurring Reasoning Errors

Stage. Run-level diagnosis.
System judgment. The earliest decisive error assigned to a single failure category.
Key trace evidence. In each failed run, multiple errors co-occur within one trajectory; two reviewers identified a different earlier step as the more decisive cause.
Reviewer ratings. Two reviewers rated the judgment *fully supported*; two rated it *minor caveats*.
Disagreement. Whether the assigned category identifies the earliest outcome-changing error when several errors co-occur within one trajectory.
Pattern. Multiple co-occurring errors with an ambiguous earliest critical step.