# BEFORE THE FIRST EQUATION: SELECTIVE SCENE CONTROL IN LARGE LANGUAGE MODEL PHYSICS REASONING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We study how a solver's description of a physics scene shapes its answer. Across 284 questions from PHYBench, SciBench, and UGPhysics, four fixed model interfaces, and three repeated calls per condition, a question-matched six-field contract reaches 94.22% judged accuracy. This is +8.27 points above a no-contract scene baseline collected in a separate run (85.94%; descriptive). Within a shared response run, the matched contract exceeds a mismatched contract from another question by 74.15 points (Holm-adjusted paired sign-flip $p < 2 \times 10^{-5}$). Its benefit depends on correspondence: a one-field solution-relevant edit in the contract yields 56.19%, compared with 79.28% for an edit designated irrelevant to the registered solution in the contract, a 23.09-point paired gap (Holm-adjusted $p < 2 \times 10^{-5}$). The sensitivity is selective but not exclusive; even the designated irrelevant edit has a cost. In a same-call, the solve with designed general six-direction self-check reaches 93.57%, versus 85.71% for direct solving (+7.86 points across separate runs, descriptive). Judgments use condition-specific mixtures of programmatic checks, model-assisted judging, and human QC. A well-matched physical description, and potentially an inline check of one's own solve, can support more accurate physics answers.

## 1 INTRODUCTION

A physics solution starts before its first equation. The solver chooses a system boundary, decides which interactions and constraints matter, and identifies the quantity and stage the question asks about. If that scene is wrong, a fluent derivation can still solve the wrong problem.

Consider a two-body mechanics problem that asks for a speed after the bodies interact. The solver must decide which objects form the system, which interaction acts during the interval, and what physical state the requested speed refers to. Treating the bodies as one system or as separate systems changes which forces count as external and which constraints couple their motion. The algebra can be correct under either model; only one may answer the question that was asked. Scene construction is therefore a distinct part of physics reasoning, not a synonym for calculation.

Benchmark scores can reflect missing assumptions, faulty references, grading decisions, or model reasoning. An expert audit of physics benchmarks found that rechecking questions, references, and grading can change measured performance (Ansari et al., 2026). Accuracy therefore speaks to model reasoning only when the judgment process is sound. Holding that process fixed, we ask a separate question: how does the solve change when a model checks its work or receives a different physical scene?

In the main experiment, we vary the scene information supplied before solving. The contract is equation-free and records six commitments: target quantity, system boundary, physical state and interactions, constraints and independent components, stages and transitions, and assumptions and approximations. In the question-matched condition (G), the contract describes the question at hand. S-primary substitutes a complete contract from another question in the same benchmark. X-rel and X-irr each edit one field of the matched contract while leaving the other five unchanged; X-rel changes a choice used by the registered solution, whereas X-irr changes a field designated to leave



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.

that solution path intact. These four conditions share one response run. A separately collected condition, $D_{\text{scene}}$, uses the same scene interface without a contract.

We also ask whether the solver can improve its own answer before submission. AU-inline prompts a check of the current solve within the same call; $D_{\text{clean}}$ solves directly. Their observed accuracy difference is positive across the four interfaces and three calls, although the runs are separate and the comparison is descriptive. The older cross-call AU procedure, which reviews a frozen earlier answer, is documented in the appendix.

## 2 Related Work

**Physics reasoning and evaluation.** We use 84 PHYBench, 100 SciBench, and 100 UGPhysics questions, three benchmarks that assess quantitative scientific or undergraduate physics problem solving (Qiu et al., 2025; Wang et al., 2024; Xu et al., 2025). Ansari et al. audit questions, references, model responses, and evaluators, separating benchmark defects from grader errors and model errors (Ansari et al., 2026). Their expert-led audit repairs or excludes cases before recomputing scores. Our study changes the information supplied to the solver and measures final-answer judgments under the selected anchors.

**Checking and perturbing a solution.** Iterative refinement asks models to revise their own outputs, while process supervision trains verifiers on intermediate reasoning steps (Madaan et al., 2023; Lightman et al., 2024). Huang et al. find that intrinsic self-correction without external feedback can fail or reduce reasoning accuracy (Huang et al., 2024). Our same-call check targets six concrete decisions within a physics solve, while the main intervention specifies the physical scene before solving. Irrelevant context can distract mathematical reasoning (Shi et al., 2023), and GSM-Symbolic reports performance drops after adding a clause that does not contribute to the solution (Mirzadeh et al., 2025). We ask whether the analogous cost appears inside a structured physics contract: X-irr changes a field outside the registered solution path, while X-rel changes one used by that solution. Step-Back prompting elicits an abstraction from the model itself (Zheng et al., 2024); our scene contract is authored and frozen outside the solver.

These lines of work change different parts of a solve. Refinement changes what the model does after an initial attempt; irrelevant-context tests alter surrounding text; a scene contract specifies the physical commitments that precede the derivation. Our design holds the question and answer format fixed while changing whether that specification matches the item, then edits one field at a time. The comparison is therefore about alignment and stability, not simply about whether a longer prompt helps.

The benchmark-audit question is complementary. Audits ask whether the problem, reference, or judgment supports the score; our experiment reuses selected benchmark items and varies the information supplied to the solver. An end-to-end score cannot by itself tell whether a miss began in scene selection or in the subsequent derivation. The matched, mismatched, and single-field conditions make that distinction experimentally visible at the level of final answers, without treating the benchmark items as a new task.

## 3 Experiments

### 3.1 Questions, models, and interfaces

The fixed manifest contains 84 PHYBench, 100 SciBench, and 100 UGPhysics questions (Qiu et al., 2025; Wang et al., 2024; Xu et al., 2025). We evaluate DeepSeek V4 Flash, GPT-5.6 Sol, GPT-5.6 Terra, and Kimi K3 as four fixed model interfaces rather than a random sample of models. Calls are stateless; tools, retrieval, and web access are disabled. Each condition contains three responses per item–model cell. Outputs request JSON strings `reasoning` and `final_answer`; judgment concerns the latter. Exact prompts and recoverable call settings are in the Supplement; run provenance is summarized in Appendix A.

**Prompt operations.** Direct solving gives the question alone. AU-inline adds an instruction to recheck the interpretation, physical model, constraints, derivation, units, and requested quantity

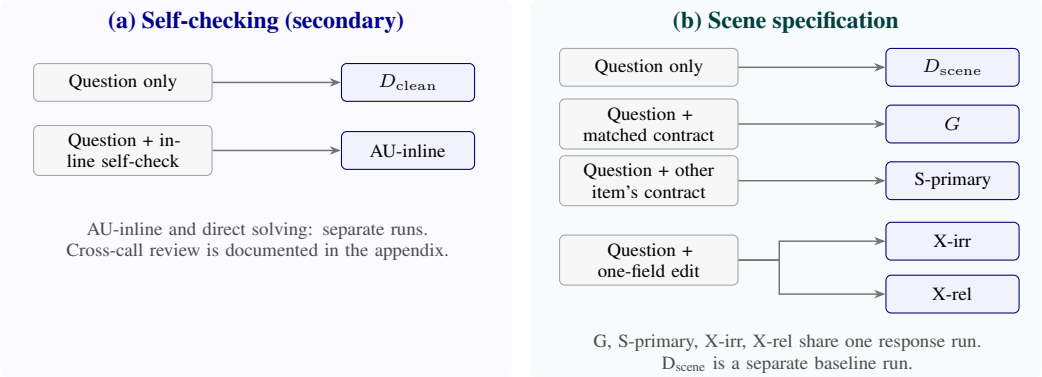


Figure 1: Experimental map. G, S-primary, X-irr, and X-rel share one response run; $D_{\text{scene}}$ is a separate no-contract baseline.

Table 1: The six commitments in a scene contract and the modeling question each makes explicit.

| Contract field | Modeling question |
|---|---|
| Target quantity | What quantity does the problem ask the solver to determine? |
| System boundary | Which object or collection of objects is the system? |
| State and interactions | Which physical states, forces, contacts, or interactions matter? |
| Constraints and independent components | Which relations constrain the system, and which components can be treated independently? |
| Stages and transitions | Which stages or events change the physical description to be used? |
| Assumptions and approximations | Which idealizations or approximations define the model? |

before submitting; it provides neither an earlier answer nor correctness feedback. In cross-call AU, a second call receives a frozen earlier answer for review. The scene interface appends a binding six-field contract to the question: G uses the question's own contract, S-primary uses another item's contract, and X-irr/X-rel replace one field of the matched contract. Across conditions the response schema is unchanged, so the scored answer is always the same field.

**Inline self-checking.** $D_{\text{clean}}$ receives the question alone. Before finalizing, AU-inline is asked to verify six aspects of its current solve: (1) interpretation of the question, (2) choice of physical model, (3) constraints, (4) derivation, (5) units, and (6) the requested quantity. It may correct its work within that same call but receives neither a previous answer nor correctness feedback. The two conditions were collected in separate runs, so their contrast describes an observed difference rather than an isolated self-correction effect. Cross-call AU and its absolute result are in Appendix C.

**Scene contracts.** A contract makes explicit the physical choices that sit between reading a word problem and writing its first equation. The solver is instructed to treat it as binding. $D_{\text{scene}}$ asks the same scene-interface solver to answer without a contract. G supplies the question-matched contract. S-primary substitutes a complete contract from another question in the same benchmark. X-rel and X-irr replace one field of G while keeping the other five fields unchanged; the former targets a choice used by the registered solution, whereas the latter is designed to leave that solution path intact. The relevant/irrelevant designation describes the authored edit relative to the registered solution, not a guarantee about what a model will attend to.

**Contract construction and assignment.** Contracts were drafted in a separate structured stage and reviewed against each question and its registered solution before the intervention manifest was frozen. Review checked for omitted commitments, internal physical inconsistency, and equation leakage. For S-primary, a seeded derangement assigns complete contracts within each benchmark, with no self-matches and a screen against near-duplicate matches. Each one-field edit replaces the full field

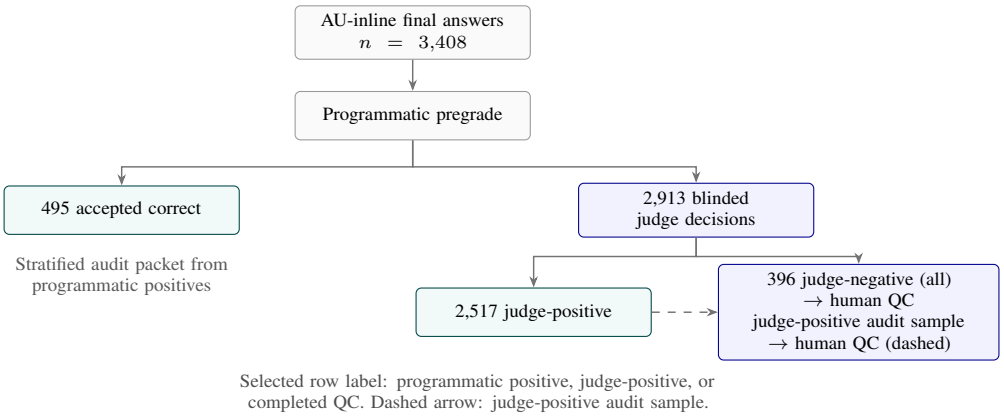


Figure 2: AU-inline adjudication path. The blinded judge sees the question, answer type, reference answer and unit, and candidate final answer, but not model or condition identity. All judge-negative cases and a sample of judge-positive cases enter human QC; the dashed arrow marks the positive-case audit. Counts report selected row-label sources, not the audit-sample size.

value while preserving the other five fields byte-for-byte. Models do not receive the condition name, source item, edited-field identity, or expected direction.

The contrasts vary the scope of the scene change. G–S-primary changes which problem the complete six-field description belongs to while retaining its format and solve instruction. G–X-irr measures the effect of altering one commitment, and X-irr–X-rel compares edits designated outside or inside the registered solution path. $D_{\text{scene}}$ provides an absolute reference for solving through the scene interface without a contract, but comes from a separate response run. Read together, the conditions distinguish scene correspondence from sensitivity to a local edit; no single score is asked to stand in for both.

### 3.2 Judgment and analysis

Correctness is a binary judgment of `final_answer`; the generated reasoning string is not scored. For example, Figure 2 traces the AU-inline path. A programmatic pregrade accepts established answer matches as correct and sends all other cases, including unresolved formats, to a blinded binary judge. The judge receives the question, answer type, reference answer and unit, and candidate final answer, but not the model or condition identity. A judge-negative outcome is routed to human QC rather than accepted as a final error; the completed packet's `human_answer_correct` value supplies that row's selected label. Programmatic positives also generate a stratified audit packet, and a sample of judge-positive decisions receives human QC as an audit. Other conditions have different selected-source compositions and human-review coverage (Appendix C, Table 7). The analysis retains one selected judgment per condition, model, item, and call; it does not vote across calls.

For a condition $c$, we report $\hat{p}_c = 3408^{-1} \sum_{i,m,r} Y_{imrc}$, where $Y_{imrc}$ indicates a correct final answer. Each item contributes 12 rows, one for each of four interfaces and three calls, so items receive equal weight and benchmarks contribute in proportion to their size in the 284-item manifest. Rows are matched by item, model, and repeat. We call comparisons across separate response runs cross-run descriptive comparisons: item matching does not isolate a prompt condition from run-level differences. AU-inline–$D_{\text{clean}}$ and G–$D_{\text{scene}}$ are cross-run contrasts. Their item-bootstrap intervals resample items within the recorded runs and do not estimate run-to-run drift. Ninety-five-percent intervals for differences are percentile intervals from 20,000 benchmark-stratified item-bootstrap samples; each sampled item carries all four models, three repeats, and paired conditions together, retaining the dependence among its 12 judgments.

For G–S-primary and X-irr–X-rel, the paired sign-flip test operates on the 284 item-level mean differences (100,000 two-sided sign flips), with Holm correction over these two protocol-designated endpoints; G–X-irr is secondary. Model and repeat observations are averaged within each item, making the question the inferential unit. The test assumes sign symmetry under the paired null and is not a randomization test. Intervals and tests describe item variation for these four fixed interfaces.

# 4 RESULTS

## 4.1 MATCHED SCENES AND INLINE SELF-CHECKING

Table 2 reports individual judged answers for the full manifest. The strongest evidence comes from the four scene interventions collected in one response run; the no-contract and self-check runs add useful, separately collected points of comparison.

Table 2: Row-level final-answer accuracy. Every condition contains 3,408 judged answers; percentages are not majority-vote scores. Scene-interface conditions are shown first; the two self-check conditions are secondary. Cross-call AU is reported only in the appendix.

| Condition | Correct / 3,408 | Accuracy (%) |
|---|---|---|
| G (matched) | 3211 | 94.22 |
| $D_{\text{scene}}$ | 2929 | 85.94 |
| S-primary (mismatched) | 684 | 20.07 |
| X-irr (irrelevant edit) | 2702 | 79.28 |
| X-rel (relevant edit) | 1915 | 56.19 |
| $D_{\text{clean}}$ | 2921 | 85.71 |
| AU-inline | 3189 | 93.57 |

With a question-matched contract, G reaches 94.22% (3211 of 3,408 judged answers). The no-contract scene baseline $D_{\text{scene}}$ reaches 85.94%; the observed item-paired difference is +8.27 points (95% item-bootstrap interval [5.93, 10.80]). Because these are separate runs, this is a descriptive comparison, not an isolated contract effect. The contract supplies physical commitments rather than worked equations.

The within-run contrast tests whether those commitments belong to the question. G exceeds S-primary (20.07%) by 74.15 points (95% interval [71.80, 76.44]; paired sign-flip $p < 10^{-5}$, Holm-adjusted $p < 2 \times 10^{-5}$). Both conditions use the same six-field format and solve instruction, but S-primary receives a complete contract from another question in the same benchmark. Thus the high matched-contract accuracy is tied to question–scene correspondence, not simply to receiving a structured prompt.

The result recurs across item sources. Across PHYBench, SciBench, and UGPhysics, G exceeds S-primary by 74.50, 71.92, and 76.08 points, respectively. X-irr also exceeds X-rel in all three, though the gap varies more across sources (20.63, 31.92, and 16.33 points). Table 3 shows the condition rates behind these contrasts; contrasts in the text use unrounded row counts.

Table 3: Final-answer accuracy by benchmark for the four within-run scene conditions. Each cell reports the percentage of judged rows in that benchmark (PHYBench: 1,008; SciBench and UGPhysics: 1,200 each), computed from the released outcome ledger.

| Benchmark | G | S-primary | X-irr | X-rel |
|---|---|---|---|---|
| PHYBench | 91.67 | 17.16 | 74.50 | 53.87 |
| SciBench | 95.00 | 23.08 | 81.75 | 49.83 |
| UGPhysics | 95.58 | 19.50 | 80.83 | 64.50 |

The single-field edits test how selectively a matched scene must be preserved. X-irr reaches 79.28%, 14.94 points below G (95% interval [12.97, 16.96]). X-rel reaches 56.19%; X-irr–X-rel differs by +23.09 points (95% interval [19.75, 26.53]; paired sign-flip $p < 10^{-5}$, Holm-adjusted $p < 2 \times 10^{-5}$). The 38.03-point G–X-rel gap separates into a 14.94-point G–X-irr difference and a 23.09-point X-irr–X-rel difference. The relevant edit carries a larger cost, while the irrelevant edit shows that the selectivity is not absolute. Figure 3 summarizes the paired contrasts; Appendix D gives repeat-level counts.

The matched-versus-mismatched gap varies across the four fixed interfaces: 59.4–85.2 points. Under G, model accuracies lie between 93.3% and 95.7%; under S-primary they range from 8.8% for Sol to 36.3% for Kimi. Every interface scores lower under S-primary, but the size of that drop differs

Table 4: Correct final answers by model, each out of 852 rows (284 items $\times$ three repeats). Direct and inline conditions are followed by scene conditions; cross-call AU is retained in the appendix.

| Condition | DeepSeek | Sol | Terra | Kimi |
|---|---|---|---|---|
| $D_{\text{clean}}$ | 694/852 | 749/852 | 739/852 | 739/852 |
| AU-inline | 809/852 | 793/852 | 801/852 | 786/852 |
| $D_{\text{scene}}$ | 722/852 | 766/852 | 718/852 | 723/852 |
| G | 800/852 | 801/852 | 795/852 | 815/852 |
| S-primary | 197/852 | 75/852 | 103/852 | 309/852 |
| X-irr | 721/852 | 684/852 | 543/852 | 754/852 |
| X-rel | 529/852 | 504/852 | 371/852 | 511/852 |

Table 5: Per-model paired accuracy differences in percentage points, computed from the counts in Table 4. These model-level summaries are descriptive; confirmatory tests use the item-level paired analysis.

| Model | $G - S$-primary | $G - X$-irr | $X$-irr$-X$-rel |
|---|---|---|---|
| DeepSeek V4 Flash | 70.8 | 9.3 | 22.5 |
| GPT-5.6 Sol | 85.2 | 13.7 | 21.1 |
| GPT-5.6 Terra | 81.2 | 29.6 | 20.2 |
| Kimi K3 | 59.4 | 7.2 | 28.5 |

sharply. Kimi retains the highest accuracy under a mismatched contract, yet still loses nearly 60 points relative to its matched condition. Table 4 reports the underlying counts; Table 5 puts the three paired effects side by side. The alignment penalty is large for every interface, while the general cost of a one-field edit varies more widely. This separates a shared response to scene mismatch from interface-specific sensitivity to particular edits.

The edit profile is particularly different for Terra and Kimi. For Terra, the 49.8-point G–X-rel loss includes a 29.6-point drop under X-irr; for Kimi, the total 35.7-point loss includes only 7.2 points under X-irr and a 28.5-point X-irr–X-rel gap. DeepSeek and Sol lie between these profiles. A single score for "edited contracts" would hide this distinction: Terra is more affected by the designated irrelevant change, while Kimi's larger decrement appears when the changed field is solution-relevant. The table therefore characterizes not just how much accuracy changes, but how the losses divide across the two edit contrasts for each interface.

Call-level accuracy shows the same pattern. G ranges from 93.84% to 94.54% across its three calls, while S-primary ranges from 19.19% to 21.30%; the G–S-primary gap is between 72.98 and 74.82 points in each call. The X-irr–X-rel gap also persists across calls (22.27–24.38 points). Table 6 reports the full breakdown for all main-text conditions.

The direction also holds when the three calls are collapsed by majority vote: G exceeds S-primary by 79.58 points, and X-irr exceeds X-rel by 26.76 points across the 1,136 item–model cells. These sensitivity estimates use a different aggregation rule; the reported headline rates above remain individual judged answers.

**Inline self-checking.** AU-inline reaches 93.57% (3189 of 3,408), compared with 85.71% (2921 of 3,408) for $D_{\text{clean}}$: an observed +7.86-point difference (95% item-bootstrap interval [5.87, 9.92]). Its six checks ask the solver to revisit its own setup and calculation before submitting, without an earlier answer or external feedback. The direction is positive for each of the four model interfaces (Table 4) and for each of the three calls (Table 6).

The two conditions belong to separate runs and use different selected judgment-source mixes (Table 7 in the appendix); the direction of any source-related bias is unknown. We therefore treat +7.86 points as a repeatable descriptive signal in these records, not a confirmatory or causal estimate of self-correction. An interleaved direct/inline experiment would be needed to isolate that effect.

Table 6: Judged accuracy (%) by call; each cell contains 1,136 item–model rows.

| Condition | Call 1 | Call 2 | Call 3 |
|---|---|---|---|
| $D_{\text{clean}}$ | 85.56 | 85.12 | 86.44 |
| AU-inline | 93.93 | 93.75 | 93.05 |
| $D_{\text{scene}}$ | 85.21 | 86.00 | 86.62 |
| G | 94.54 | 94.28 | 93.84 |
| S-primary | 19.72 | 21.30 | 19.19 |
| X-irr | 79.75 | 79.67 | 78.43 |
| X-rel | 55.37 | 57.04 | 56.16 |

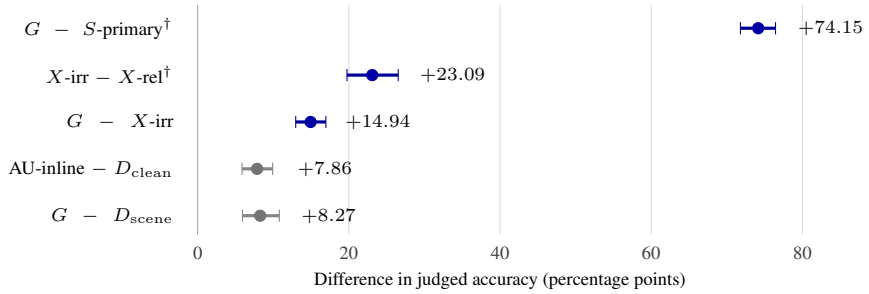


Figure 3: Paired accuracy differences and 95% benchmark-stratified item-bootstrap intervals. Gray rows compare separate runs; blue rows share the scene response run. † marks the two protocol-designated confirmatory endpoints (paired sign-flip tests with Holm correction). G–X-irr is within-run but secondary. All contrasts include 3,408 matched answer rows.

## 5 Discussion and Limitations

The question-matched contract supports high final-answer accuracy across all four interfaces and all three benchmarks. Its 74.15-point within-run advantage over a same-format contract from another question identifies correspondence as the critical feature of that structured scene. The 14.94-point G–X-irr difference shows that even an edit designed to leave the registered solution path intact can alter the answer; the further 23.09-point X-irr–X-rel difference shows that edits tied to the solution matter more. Together, these contrasts make scene alignment visible at the level of final answers, even though they cannot locate the particular reasoning step that changed.

The model-level table also shows why a single "edited-contract" score can obscure the response profile. DeepSeek, Sol, and Kimi have similar total G–X-rel losses (31.8–35.7 points), but the loss is divided differently between the G–X-irr and X-irr–X-rel contrasts; Terra has a larger total loss (49.8 points), much of it already present under X-irr. Reporting the two contrasts separately preserves whether a model is affected by a field change in general or shows an additional decrement for a field tied to the registered solution. A follow-up that compares the model's stated setup with the supplied contract could help explain these outcome-level profiles.

The contrasts do not treat a contract as a generic extra hint. S-primary preserves the six-field format and solve instruction while changing which problem the contract describes; X-rel and X-irr preserve five fields and alter the sixth. This makes the study useful for evaluation design: a complete-contract substitution probes whether the scene belongs to the item, while a one-field edit probes whether the answer is stable to a local change. A final-answer score still combines scene construction and calculation, so future evaluations could record intermediate setup choices alongside correctness to locate where the trajectory changes.

For scene-aware systems, the results suggest a practical evaluation sequence: score solving with a matched contract, substitute a complete contract from another problem to test correspondence, then edit one field at a time to measure local stability. Keeping these steps separate distinguishes a scene that describes the wrong problem from a solver that reacts strongly to a small change in an otherwise matched scene. Because a contract makes its commitments explicit, it can also be checked against

the question before the derivation begins. The present experiments identify correspondence and edit stability as useful targets for such checks; they do not evaluate an automatic contract validator.

The benchmark split adds a useful distinction: the G–S-primary gap varies by 4.17 points across the three collections, and the G–X-irr loss by 3.91 points, while the X-irr–X-rel gap spans 15.58 points (Table 3). In these data, the cost of breaking question–contract correspondence and the general cost of an edit are similar across benchmarks; the additional loss associated with a solution-relevant field depends more on the benchmark slice. "Irrelevant" here refers to the edit's relation to the registered solution path, not to a claim that its content is semantically unnoticeable to a model. Future evaluations can preserve this distinction by reporting the full contrast set by benchmark and field type, rather than reducing robustness to a single average.

AU-inline supplies a second, positive observation: the same-call six-direction check has higher judged accuracy than direct solving for every model and every call. It is a practical way to ask a solver to revisit its own interpretation and derivation before the final answer. Unlike the two within-run scene endpoints, this comparison is cross-run and has a different mix of selected judgment sources, so the observed +7.86-point difference does not isolate self-correction. The scope is the 284-item manifest and four fixed interfaces, with no independent expert regrade of every item (Appendix C). An interleaved comparison could test whether the inline gain persists when run-level and adjudication differences are controlled.

## 6 Conclusion

Before a physics derivation begins, the solver has already chosen a scene. A question-matched, equation-free contract supports 94.22% judged accuracy; within the same response run it exceeds a same-format contract from another question by 74.15 points (Holm-adjusted $p < 2 \times 10^{-5}$). The advantage appears across all three benchmarks and four models. A six-direction inline check also reaches 93.57%, +7.86 points above direct solving in separate runs.Statistical analysis above indicates that these results are positive descriptive signal rather than an isolated effect. Correctly specifying the physical scene and checking a solve before submission are two promising ways to make physics answers more reliable.

## Reproducibility Statement

The supplementary package includes a row-level outcome ledger, selected judgment-file paths, per-model and per-repeat tables, bootstrap contrasts, a three-call majority-vote sensitivity analysis, source checksums, prompt files, and a standalone script that recomputes the reported statistics from the frozen ledger. The package supports calculation-level reproduction; raw model responses and adjudication packets are not part of the release. Results use the frozen 284-item manifest and selected QC anchors, with individual judged rows as the headline unit rather than majority-vote aggregation.

## AI Use Statement

Language models were the experimental subjects, were used in scene-contract drafting and answer adjudication, and assisted with manuscript drafting and analysis code. Human researchers selected the experimental anchors, reviewed the adjudication packets and contract materials, and remain responsible for the reported results and the final submission. The exact scope of human QC is specified in Appendix C; this statement does not imply human review of every automatically accepted row.

## References

Ali Ansari, Haoran Sun, Andy Zeyi Liu, Mark Jabbour, Yongshan Ding, et al. How good are frontier models at physics? Expert re-grading reveals broken evaluations and near-saturation of leading benchmarks, 2026. URL https://arxiv.org/abs/2609.13009v1.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.01798.

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. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, et al. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2303.17651.

Iman Mirzadeh, Keivan Alizadeh-Vahid, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ec2e7a896f8250986b3907f57621ce94-Abstract-Conference.html.

Shi Qiu et al. PHYBench: Holistic evaluation of physical perception and reasoning in large language models. *arXiv preprint arXiv:2504.16074*, 2025. URL https://arxiv.org/abs/2504.16074.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, 2023. URL https://proceedings.mlr.press/v202/shi23a.html.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, et al. SciBench: Evaluating college-level scientific problem-solving abilities of large language models. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2307.10635.

Xin Xu, Qiyun Xu, Tong Xiao, et al. UGPhysics: A comprehensive benchmark for undergraduate physics reasoning with large language models. In *International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2502.00334.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H. Chi, Quoc V. Le, and Denny Zhou. Take a step back: Evoking reasoning via abstraction in large language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.06117.

# A Experimental Interfaces and Prompts

**Collection provenance.** $D_{\text{clean}}$ and AU-inline were collected in the same study wave but in separate runs; their interleaving and order were not documented. File modification times after cross-machine transfer are not treated as collection times. $D_{\text{scene}}$ is a separate direct-input scene-interface run (run-label version v3, 2026-09-09; scoring package dated 2026-09-10). G, S-primary, X-irr, and X-rel share the scene replacement response run (source responses dated 2026-09-05; scoring package dated 2026-09-06). AU is a legacy review run fed frozen answers from an anomalous earlier direct wave; its absolute result is reported without a parent-wave contrast.

**Model and call setup.** The four model-interface labels are `deepseek-v4-flash`, `gpt-5.6-sol`, `gpt-5.6-terra`, and `kimi-k3`. Sol and Terra are operational endpoint labels rather than claims about an underlying model family. Calls were stateless, with high reasoning effort and no tools, search, or retrieval. The supplementary configuration records the call settings recoverable from the frozen artifacts and marks fields that cannot be recovered consistently; we do not infer a common temperature or token cap from the output ledgers. The benchmark manifest has 84 PHYBench, 100 SciBench, and 100 UGPhysics items, and raw repeat indices 0–2 are reported as repeats 1–3.

**Standalone direct solve.** The system instruction is: *You are solving a physics problem for a controlled research evaluation. Follow the requested output format exactly. Do not discuss the evaluation protocol.* The user instruction is *Solve the following problem.* followed by the question and *Return one JSON object with keys reasoning (string) and final_answer (string).* The exact template is `supplement/prompts/direct_clean_v23.json`.

**Inline self-check.** The system instructs the model to solve without tools, web search, external retrieval, or outside communication; to independently check its interpretation and result; to correct any errors; and to follow the output format. The user instruction asks it, *before finalizing*, to re-check interpretation, physical model, constraints, derivation, units, and requested quantity, then return only the final revised solution as the two-field JSON object. The input contains the question, not an earlier response or correctness feedback. The exact template is `supplement/prompts/au_inline_v23.json`.

**Cross-call review.** The system says to review one's earlier physics solution without external communication or tools. The user message provides the question and a frozen previous response, asks for independent verification of interpretation, scenario modeling, constraints, reasoning, and result, explicitly says no reference answer or correctness assessment is provided, and requests the two-field JSON object. The exact template is in `supplement/prompts/au_selfcheck_v1.json`. The scaffold was informed by earlier internal error analyses and fixed before these evaluations.

**Scene solve.** The system requires the solver to treat a supplied scene contract as a binding description of the intended physical model, not replace it with another interpretation, and return the two-field JSON object. The user message contains the problem followed by `Scene contract (binding modeling specification): [CONTRACT]`. The verbatim source is `supplement/prompts/gx_full_v4_v1.json`. $D_{\text{scene}}$ is the no-contract baseline in this interface.

## B Contract Schema and Intervention Construction

Each contract has six string fields: `target_quantity`, `system_boundary`, `state_and_interactions`, `constraints_and_independent_components`, `stages_and_transitions`, and `assumptions_and_approximations`. Gold contracts were drafted in a separate structured stage and reviewed against the question and registered solution before the intervention manifest was frozen. Review addressed omissions, internal physical inconsistencies, and equation leakage. The contract is an experimental annotation, not a benchmark answer.

For each item, the X-rel edit replaces one complete field value whose content participates in the registered solution. The X-irr edit changes one complete field outside that solution path and leaves the other five fields byte-identical to Gold. These labels express the edit designers' registered-solution judgment; they do not guarantee that models will ignore the irrelevant material. S-primary uses a seeded within-benchmark derangement of complete contracts, forbids self-matches, and screens near-duplicate matches. Solvers do not see the condition label, source item ID, edited-field ID, or expected direction of effect. A separate item-level audit of whether a Gold contract introduces solution-relevant information absent from the question remains incomplete, so no such audit conclusion is made here.

## C Judgment Provenance and Data Choice

The selected anchors are listed in the experiment anchor index (`ANCHOR_INDEX_20260925.md`); the supplement's `source_audit.json` records checksums of every judgment file read by the regeneration script. Where a human-QC packet supersedes a prior machine decision, the packet's `human_answer_correct` value is used. The DeepSeek $D_{\text{clean}}$ packet supplies 663 reviewed labels, joined to 189 programmatically accepted correct rows. The Sol and Terra $D_{\text{clean}}$ ledgers and the Sol $D_{\text{scene}}$ ledger use the designated human-QC fields; Kimi $D_{\text{clean}}$ uses its frozen final-label ledger. AU-inline combines 495 programmatic positives, judgments from the completed judge file for 2,517 rows, and the 396 completed human-QC resolutions, which take precedence. The resulting ledger contains one selected judgment for each of the 27,264 item–model–condition–repeat rows.

Table 7: Selected judgment-source composition for each 3,408-row condition, as recorded in the source audit. Human-review packet labels are separated where the frozen ledger distinguishes packet versions.

| Condition | Selected judgment sources (row counts) |
|---|---|
| $D_{\text{clean}}$ | 321 programmatic positives; 2,424 human-QC labels; 663 final-adopted DeepSeek QC packet labels. The DeepSeek subset is 189 programmatic positives plus 663 packet labels. |
| AU-inline | 495 programmatic positives; 2,517 model-judge outcomes; 396 human-QC resolutions replacing judge outcomes. |
| AU | 595 programmatic positives; 706 judge-confirmed correct; 147 human-QC reviewed; 1,960 human-QC reviewed-v2. |
| $D_{\text{scene}}$ | 317 programmatic positives; 852 human-QC labels; 2,239 human-QC reviewed-v3. |
| G | 470 programmatic positives; 756 fallback human-reviewed; 2,182 human-QC-v2. |
| S-primary | 125 programmatic positives; 832 fallback human-reviewed; 2,451 human-QC-v2. |
| X-irr | 420 programmatic positives; 790 fallback human-reviewed; 2,198 human-QC-v2. |
| X-rel | 222 programmatic positives; 805 fallback human-reviewed; 2,381 human-QC-v2. |

The older AU condition is retained from the frozen analysis ledger. Its row-level labels comprise 1,960 `human_qc_reviewed_v2`, 147 `human_qc_reviewed`, 706 `judge_confirmed_correct`, and 595 `programmatic_auto_correct` decisions. These labels are not copied from its parent direct-answer correctness field. However, the frozen parent responses came from an earlier wave whose direct accuracy is anomalously high relative to the rerun $D_{\text{clean}}$ result, and the cause has not been resolved. We therefore report AU accuracy as a property of its own review procedure, with no baseline subtraction or correction-transition claim. The discrepant earlier direct condition is excluded from the present analysis. The retained cross-call review condition has 3,339 correct judgments out of 3,408 (97.98%). This is an absolute result for that review procedure; no parent-wave contrast is reported.

The adjudication pipeline cannot eliminate all benchmark ambiguity or reference error. Existing validity review covers identified cases and the selected QC packets, not an independent complete validation of all 284 problems. The current paper reports the full manifest. A separately curated 100-item clean selection is reserved for further validation and is not analyzed here.

## D Complete Row-Level Results

The five reported contrasts use matching item, model, and repeat identifiers. The separate-run comparisons are G $-D_{\text{scene}}$: +8.27 points [5.93, 10.80], and AU-inline $-D_{\text{clean}}$: +7.86 [5.87, 9.92]; both are descriptive. Within the shared scene response run, G −S-primary is +74.15 [71.80, 76.44], G −X-irr is +14.94 [12.97, 16.96], and X-irr −X-rel is +23.09 [19.75, 26.53]. Paired sign-flip tests are limited to G −S-primary ($p < 10^{-5}$, Holm-adjusted $p < 2 \times 10^{-5}$) and X-irr −X-rel ($p < 10^{-5}$, Holm-adjusted $p < 2 \times 10^{-5}$). They flip the 284 item-level mean paired differences 100,000 times, averaging over four models and three repeats per item, with a two-sided statistic; they rely on sign symmetry under the paired null and are not randomization tests. These are the two protocol-designated confirmatory endpoints. G −X-irr is secondary/descriptive and is not included in the Holm family. All intervals are 95% benchmark-stratified item-bootstrap percentile intervals from 20,000 draws; each sampled item carries all models, repeats, and paired conditions.

**Reproduction files.** The bundle at `final/supplement/data/v2_3/` contains the row-level `outcomes.csv`, condition, model, repeat, and majority-vote summaries, `paired_effects.csv`, and the checksum-bearing `source_audit.json`. The script `final/paper_v2_3/analysis/build_v23.py` regenerates the ledger and numerical tables from the selected anchors. The `v23` components in these paths are frozen artifact labels, not protocol or model versions.

The cross-call AU review condition has call-level accuracies of 97.45%, 98.24%, and 98.24% across its three repeats; it is shown separately because its parent direct-answer wave is excluded.