Some thoughts on the unreasonable aspects of FrontierPhysics

Author: Benchmarks' Last Exam Lab 2026-11-10

FrontierPhysics, called FP below, uses 56 physics research tasks to test whether AI can carry out research on its own. A model has to search the literature, write code and process data, then submit its computed results and a paper.

Grading has two parts. A checking program compares the model's computed results with the task author's reference values. A second AI, the reviewer, reads the model's code, action log and paper and scores them item by item against the rubric the task author wrote. Some rubric items are veto items: if any one of them fails, the whole attempt counts as a failure.

Each model attempts each task three times, 168 attempts in all. The best performer, Claude Opus 5.5, passes 46 of them, a pass rate of 27.4%.

This article draws on the FP paper, the official site examples and the public development code. It examines five aspects of the evaluation that are not entirely reasonable and how they affect what the score means: the rubric for some tasks restricts the method, the reference conclusions lack public independent verification, veto items amplify misjudgments, grading error may not be uniform across models, and models may have seen the answers beforehand. The FP team's own review already found examples of such problems. When it reviewed GPT-6 Astra's attempts, 9 failures should have been changed to passes, because of faults in the tasks, the checking programs and the grading. The development version is the set of tasks and grading code that the task authors made public while writing and debugging them. Whether it matches the version used for the official leaderboard has not been confirmed.

The rubric restricts the solution method

The FP paper lists "do not prescribe a solution method in advance" as a design principle. The laser noise task in the public development version departs from it.

When lasers control atoms in quantum computing, small fluctuations in the laser frequency cause errors in quantum operations. The task provides four sets of measured noise data. It asks the model to calculate the error these fluctuations cause and to plot how the error changes with the laser drive strength. The task statement does not specify a method.

The error can be computed in at least two ways. The first calculates how sensitive the quantum operation is to fluctuations at each frequency, multiplies that sensitivity by the measured noise frequency by frequency, and sums the products. The second repeatedly generates random fluctuations on a computer, simulates their effect and averages the results. Both approaches are valid physics.

The rubric accepts only the first and makes it a veto item. A model that uses random simulation without explaining how it relates to the first method fails the whole attempt, even when the numeric check passes.

The numeric check, by contrast, allows reasonable disagreement. The noise data extend only to 8 MHz. The task author assumed that noise above that frequency is either zero or unchanged, which gives two error curves. A result passes if it falls between the two curves, with some margin. The numeric check accepts different assumptions, yet the method grading accepts only one path. If the rubric stated what conditions another method must meet to pass, it would fit FP's own design principle better.

The reference conclusions await independent verification

FP's reference answers come from the researchers who wrote the tasks. Of the 56 tasks, 18 come from published research, and 4 of those were published in Nature. A publication record does not mean the conclusions have been independently verified.

FP ran two kinds of checks. The first is domain peer review. Researchers in closely related fields read each task to judge whether it gives away too much of the solution and whether the rubric tests the physics knowledge that is actually needed. They also read the model trial records to find unclear task statements and misjudgments. The second kind gave the original research paper or manuscript attached to 33 tasks to the AI reviewer, scoring only the items the manuscript itself could answer. The manuscripts averaged 96.2% of the full score. This shows that the rubric agrees with the original authors' materials.

Both checks examine the tasks and the rubric. Neither tests whether the original research conclusions hold. Testing a conclusion takes other researchers who recalculate, reanalyze the data or repeat the experiment. The FP paper does not report such verification across all tasks.

For tasks that compute a value from given equations and parameters, the reference value can be checked directly, so the exposure is limited. Tasks that ask whether the data support a finding are different. Suppose a model sees an effect in one segment of the data, and the effect disappears when another equally reasonable range is used. Pointing out that the conclusion depends on the data range is itself a research result. If the rubric accepts only the expected finding, this kind of careful objection can be marked wrong. FP's grading basis would be clearer if it stated what checks each task's reference conclusion went through, and how it grades a model that finds no expected effect or finds counter-evidence.

Veto items amplify misjudgments

A veto item turns a single error into the failure of a whole attempt. An ambiguous task statement, a faulty checking program or evidence the AI reviewer missed can each cause a correct attempt to be marked as a failure.

The FP team and the task authors went through all 168 of GPT-6 Astra's attempts at xhigh reasoning effort, one by one. The higher the reasoning effort, the more computation a model spends on thinking while it works. The review found 9 failures that should be changed to passes. Eight came from the task, the checking program or the rubric, and one from a mistake by the AI reviewer.

The paper gives two examples. In the lattice geometry task, one attempt had computed the correct value, but the task prescribed a computation too large to run and the program crashed. In the neutron star task, one attempt lost points because the task did not say how to set the prior before the analysis, meaning the possible values of the parameters and their probabilities.

After the changes, passes in the same batch rose from 25 to 34 and the pass rate from 14.9% to 20.2% (Figure 1a). The FP team then repaired the tasks and had models re-run or be regraded, and the final table gives 19.6% for this setting. The 20.2% comes from changing verdicts on the original attempts. The 19.6% comes from fresh results after the tasks were repaired. Analyzing which research abilities a model lacks first requires removing misjudgments of this kind.

Grading error may not be uniform across models

After the numeric check, the AI reviewer's verdict has a large effect on the result. Converting the main table's pass rates into counts, 77 of Opus 5.5's attempts passed the numeric check, and 31 of those failed a veto item, about 40%. For GPT-6 Astra at max reasoning effort the figure is 27 of 66, a similar share (Figure 1b). These rejections include real research flaws and may include misjudgments. Telling them apart takes a check of each attempt.

Every FP result comes from the same AI reviewer, GPT-5.6 Sol. The paper also regraded the first-round attempts with Claude Opus 5.5 as the reviewer. For GPT-6 Astra's attempts, the two reviewers agreed on 48 of the 53 attempts that passed the numeric check, and the pass counts were 25 and 26. For Opus 5.5's own attempts, they agreed on 58 of 67, and the pass count rose from 37 under the default reviewer to 44. The Opus reviewer alone passed 8 attempts, while the default reviewer alone passed 1. The paper notes that the Opus 5.5 reviewer is more lenient toward attempts from models of its own family. Only these two models were regraded, and no other configuration was checked with a different reviewer.

The scope of the misjudgment review is also limited. The paper explains that other model and tool combinations may fail in ways outside its classification, so it could accurately label and correct only GPT-6 Astra's misjudgments. The paper does not estimate how many misjudgments the other configurations' pass rates contain. Scores of different models on the leaderboard may therefore carry different amounts of evaluation error.

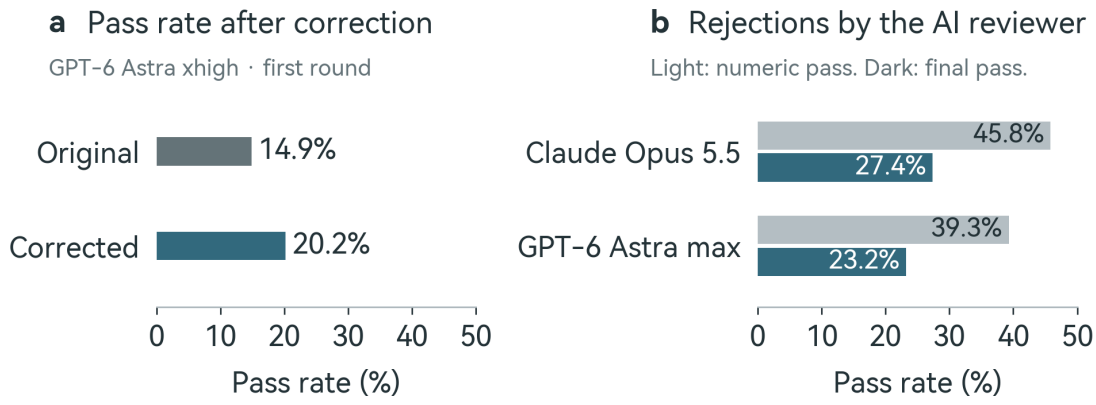


Figure 1. Panel a shows the same 168 attempts: 25 passes under the original grading and 34 after manual correction. The 19.6% in the text comes from re-running or regrading after the tasks were fixed. Panel b shows the two configurations with the highest pass rate in the paper's main table. Light bars give the share that passed the numeric check. Dark bars give the final pass rate, which also requires every veto item to pass. The gap between them is what the AI reviewer rejected.

Models may have seen the answers during training

The rubric of the laser noise task states that the task author's doctoral thesis is publicly available and contains the complete solution. The task authors' public project code also holds four development tasks whose names match tasks in the official task list, together with their reference answers and rubrics.

FP's safeguards target the period when a model is working. The official site example states that a set of web addresses is blocked. The development version of the laser noise task allows web access, but grading checks whether the model found the original author's paper and copied results from it. A model that read these solutions during training could complete the task from memory without searching. So the absence of any trace of answer lookup in the records does not yet show that the model derived the result on its own.

Conclusion

These five aspects complicate what 27.4% means. It reflects how models performed on these 56 tasks, under this rubric and this AI reviewer, and it remains some distance from independent research ability. The specific reasons for failure deserve more attention. The FP paper labels each of the 143 failures in GPT-6 Astra xhigh's original grading with its causes, and lists misjudgments as a separate category. One failure can carry several labels. The most frequent are insufficiently rigorous reasoning at 48%, lack of expert intuition at 34%, and literature errors at 27%. These labels say more about where models are weak in research than the total score or the ranking does.

References

[1] FrontierPhysics paper v1, November 4, 2026. Tasks and rubrics are in Section 2, the grading of the authors' reference materials in Section 3.3, the review, repairs and failure classification in Section 3 and Figure 6, the main-table pass rates in Table 2, and the regrading with a second reviewer in appendix Table 6.

[2] Public development tasks and grading code, pinned at commit 2f3ec886. The code evidence for the method restriction and the answer exposure comes from this version.

[3] FrontierPhysics website, example task, network settings and leaderboard, accessed November 10, 2026.

Contact Us

benchmarks-last-exam.com Benchmarks-Last-Exam@outlook.com

HarmonyOS Sans SC · Copyright 2021 Huawei Device Co., Ltd.