THE WRONG KIND OF BENCHMARK DIFFICULTY: DIAGNOSING BENCHMARK ERRORS USING MODEL DISAGREEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Current AI benchmarking practice has aimed for harder and harder questions. Yet increasingly, evidence suggests that hard questions may just be fallacious ones. We offer and test the reliability of a remarkably simple diagnostic approach for surfacing such questions: expanding answer choices to allow abstention or multiple correct answers and assessing model disagreement (from keyed answer and convergence on the same non-keyed answer) for M frontier models. First, we show that a panel as small as $M = 3$, with full disagreement and any amount of convergence, recovers the per-subject quality profile established by previously published, independent expert audits of two popular benchmarks, Humanity's Last Exam (HLE) and Massive Multitask Language Understanding (MMLU), without per-benchmark tuning. Second, we demonstrate that this simple three-model disagreement signal outperforms prior disagreement-based screening and an LLM-as-judge panel, recovering independently verified errors with 72–84% precision and 61–65% recall. Third, we assess key assumptions, including independence, and provide evidence that model diversity enhances performance more so than panel size. Fourth, we highlight the recency of these diagnostic capabilities by running the same procedure with earlier model generations and tracing improvements, particularly in precision, across successive model generations. Fifth, we show that the current trend to design harder benchmarks may have produced rising rates of fallacious items in newer benchmarks. Our findings have considerable implications for current benchmarking practice, and offer a scalable method for detecting fallacious items.

1 INTRODUCTION

A model's standing on recent benchmarks such as Humanity's Last Exam (HLE) (Phan et al., 2025), OSWorld (Xie et al., 2024), DeepSWE (Huang et al., 2026), and FrontierMath (Glazer et al., 2025), as well as on earlier benchmarks such as GPQA (Rein et al., 2023) and Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2021), drives release narratives, research direction, and adoption. As older benchmarks saturate, the field builds harder ones at an accelerating cadence to keep pace with model improvement. The premise is that harder questions measure more.

However, recent manual verification has increasingly surfaced items that are ambiguous, mis-keyed, malformed, or admit multiple or no correct answers in MMLU (Gema et al., 2025), HLE (Zhai et al., 2026), and many other "saturated" benchmarks (Vendrow et al., 2025). In other words, questions that are increasingly pressured to admit hard answers may risk trafficking in the wrong kind of benchmark difficulty: they may simply be unanswerable.

We draw on jury theory (Grofman et al., 1983) to propose a remarkably straightforward approach to diagnose, design, and correct such questions. First, we ensure that the option set includes two often-missing choices: that more than one of the listed options is correct, and that none are. Then, we query a small panel M of frontier models from different developers independently, flagging questions for which a threshold number D of models disagree with the keyed answer and, as an optional additive condition, a threshold C of those disagreements converge on a single non-keyed answer. We formalize three assumptions (§3.1)—model competence, model independence, and

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.

answer expressibility—that undergird theoretical probabilities of full model disagreement and full convergence of model disagreement (§3.2). Our study empirically tests both the degree to which those assumptions might be violated and, nonetheless, the practical utility of our diagnostic approach for surfacing fallaciousness in benchmarks.

We make five contributions. *(1)* We show that our primary diagnostic signal, specified as simply the rate at which three frontier models all disagree with the keyed answer across a set of benchmark questions, closely recovers the per-subject quality profile of arduous expert verification on MMLU and HLE (§4.1). *(2)* We measure our item-level diagnostic performance against verified MMLU and HLE errors and achieve precision of 72.4% and 84.4% and recall of 61.7% and 65.4%, respectively, outperforming prior disagreement-based screening and an LLM-as-judge panel through a simple three-model disagreement signal (§4.3). *(3)* We show that diversity across developers, not panel size, drives strong diagnostic signal: three models from three developers outperform three temperature-varied samples from a single model, consistent with the role of independence in label aggregation (§4.2). *(4)* We show that this diagnostic capability is recent: rerunning the same pipeline with frontier models from a year ago, recall on HLE is high but precision is only 64.1% (Early-2025) compared to 84.4% (Sep-2026), because earlier models disagree with enough valid items that full disagreement provides a poor diagnostic signal (§4.3). *(5)* Applied across benchmarks released over the past six years, the rate of diagnostic flags rises with release date: newer, harder benchmarks may contain more likely-fallacious items, not fewer (§4.5). We conclude with discussion of limitations, future work, and substantial implications for the benchmarking community.

2 Related Work

Concerns about the validity of AI benchmarks have grown alongside their influence on model release decisions and research direction. Raji et al. (2021) argue that benchmarks framed as general-capability tests do not actually measure general capability, raising construct-validity concerns about inferences drawn from leaderboards. Bowman & Dahl (2021) argue that many benchmarks fail on the criteria of validity, reliable annotation, statistical power, and bias. Sainz et al. (2023) and Kapoor & Narayanan (2023) document widespread leakage of test data. Singh et al. (2025) argue that selective disclosure and unequal access confer measurable Chatbot Arena advantages to some providers.

Recent efforts have focused on verifying benchmark question integrity. A re-annotation of 5,700 MMLU questions found an overall error rate of 6.5% across 57 subjects, with error rates exceeding 50% in some subjects (Gema et al., 2025). HLE-Verified, which found that a substantial fraction of HLE items require revision or cannot be conclusively validated (Zhai et al., 2026), utilized model outputs (*e.g.*, answer replication, equivalence checking, multi-model collaborative repair) as diagnostic signals and relied on domain experts for final adjudication. PlatinumBench found similar results based on expert review of frontier-model outputs across fifteen benchmarks (Vendrow et al., 2025). Earlier research has established that the problem of benchmark errors extends far beyond any specific dataset or evaluation domain (Northcutt et al., 2021). Such existing work yields high-quality corrections but is costly, must be designed and implemented on a case-by-case basis for each new evaluation, and is hence difficult to scale.

Several automated approaches use model response patterns to prioritize items for review. Truong et al. (2026) run benchmark questions through up to 91 LLMs, compute psychometric statistics from their responses, and manually verify only the highest-ranked candidates. Similarly, Land & Bikel (2026) fit responses from 114 models using an Item Response Theory-based indicator and report 95% precision for the top 200 flags. Notably, both studies focus their precision-based validation on a subset of high-confidence benchmark items rather than running a diagnostic across the full benchmark in order to capture both precision and recall. Vendrow et al. (2025) route any item for which at least one model disagrees with the keyed answer to human review, which, while more selective than full re-verification, still expands considerably relative to benchmark difficulty and the error rate of the weakest constituent model: applied to HLE, the method would route over 80% of the dataset to human review. Nahum et al. (2025) use smaller ensembles of four LLMs, average class-token probabilities across four prompt variants, and flag high-confidence disagreements on binary or ordinal labels, while Wittlinger et al. (2026) count agreements among four locally hosted LLMs and human labels in a structured extraction task for clinical free-text in order to triage risk-stratified review.

A different line of work on LLM-as-a-Judge methods uses language models as evaluators (Zheng et al., 2023; Liu et al., 2023). Calderon et al. (2025) develop a statistical test for determining whether an LLM can replace a human annotator by comparing it with a small panel of human annotators. These works ask whether models can evaluate generated outputs or substitute for human annotators, rather than whether model disagreement with an answer key reveals defective benchmark questions.

Following prior work on label aggregation problems (Dawid & Skene, 1979; Snow et al., 2008; Ratner et al., 2017), we contribute to these efforts by drawing on jury voting. In the labeling context, under Condorcet assumptions (Grofman et al., 1983), majority aggregation of a panel of noisy LLM annotators may recover the correct label as the panel grows. Refinements (Ben-Yashar & Nitzan, 2017; Nitzan & Paroush, 1982; Karotkin & Paroush, 2003) show that low-competence annotators can reduce group accuracy, and thus small panels of comparably strong models may be preferable. Our work asks whether three hard answer selections from cross-provider frontier models are sufficient to produce a scalable and precise item-level diagnostic for prompts, options, and answer keys, without confidence calibration or large response matrices. We apply one fixed rule uniformly to every item in each evaluated set, rather than validating only a selected set of top-ranked candidates, and report both precision and recall against previously published, independent verification labels.

3 METHODOLOGY

3.1 PROBLEM SETUP AND ASSUMPTIONS

Setup. Benchmark item j, a question with a finite option set $\mathcal{O}_j$ and a keyed answer $y_j \in \mathcal{O}_j$, is *sound* if y_j is in fact the sole correct response; otherwise it is *fallacious*. Fallaciousness comprises *key errors* (wrong y_j), *construction* failures (no unique correct response), *grounding* failures (false presupposition or contested source), *staleness* (once-correct y_j no longer holds), and *multiplicity* (more than one defensible answer). We query M frontier language models as annotators, $i \in \{1, \ldots, M\}$, on each benchmark item, such that x_{ij} denotes annotator i's response for item j. Our diagnostic approach rests on three assumptions, stated formally as follows:

Assumption 1 (Competence): The probability p that annotator i selects y_j in a sound item j is sufficiently high. Namely, our diagnostic assumes that $p := \Pr(x_{ij} = y_j \mid j \text{ sound}) \geq 0.5$, per Condorcet assumptions. This is grounded in the basic results that empirical per-model accuracies on independently verified sound items are well above 50% (see §4.2; 91.3–96.3% on MMLU and 63.0–80.5% on HLE).

Assumption 2 (Independence): Annotator responses x_{ij} are independent. This is the standard independence condition in label aggregation (Dawid & Skene, 1979; Grofman et al., 1983) and the assumption most likely to be violated for language models. We use it to derive analytical baselines (§3.2), not as a claim about real-world model behavior, and provide measures of the degree to which independence appears to be violated for our study in §4.2 and Appendix A.1. What matters is not whether independence violations are completely avoidable, but whether their potential contributions to diagnostic errors (*i.e.*, false positives on sound items) remain small enough for our diagnostic to still provide an effective signal for fallaciousness. We show empirically that this condition is met for cross-developer panels at $M = 3$, with the cross-developer construction likely reducing—though not eliminating—independence violations relative to within-model alternatives.

Assumption 3 (Expressibility): A sole correct response is actually contained in $\mathcal{O}_j$. This is trivially satisfied in sound items, but *construction* failures and *multiplicity* violate this assumption. Faced with only incorrect options in $\mathcal{O}_j$, annotators may be forced to select an indefensible x_{ij}. Therefore, we augment each item j with the *expanded* option set $\mathcal{O}_j^+ = \mathcal{O}_j \cup \{\text{NONE}, \text{MULTIPLE}\}$ (de-duplicating as appropriate), where NONE asserts that no element of $\mathcal{O}_j$ is correct and MULTIPLE asserts more than one is. By allowing $x_{ij} \in \mathcal{O}_j^+$, we de-noise the signal annotators might usefully provide for these particular forms of fallaciousness. We clarify how this approach extends to open-form question formats with an HLE example in Appendix A.2.

3.2 DIAGNOSTIC SPECIFICATIONS

Notation. Two counts summarize a panel's response on item j: *disagreement* $D = |\{i : x_{ij} \neq y_j\}|$ and *convergence* $C = \max_{a \neq y_j} |\{i : x_{ij} = a\}|$, with $M \geq D \geq C$. We denote a particular

diagnostic specification as $M_m/D_d/C_c$ for panel size $M = m$, disagreement count $D = d$, and convergence count $C = c$. The last term is dropped when C is unconstrained (*i.e.*, M_m/D_d aggregates over all C consistent with $D = d$).

Theoretical probabilities. While our diagnostic approach allows many possible specifications, we focus on *full disagreement* (M_3/D_3), which flags items with $D = M$ (*i.e.*, every annotator rejects y_j) and has a theoretical probability of triggering on a sound item under Assumptions 1–3 of $(1-p)^M$. *Full convergence* ($M_3/D_3/C_3$) additionally requires $C_j = M$ (*i.e.*, every annotator picks the same alternative to y_j), has a theoretical probability of $\sum_{a \neq y_j} q_a^M$ where q_a is the probability of selecting non-keyed response a on a sound item, with $\sum_{a \neq y_j} q_a = 1 - p$, and is the highest-precision measure reported in Appendix B.1 alongside other $M_m/D_d/C_c$ specifications.

To illustrate, for $p = 0.8$ and $M = 3$, full disagreement occurs on a *sound* item with probability $(1-p)^M = 8 \times 10^{-3}$. In other words, even for a benchmark question with no errors, there is always some chance, albeit vanishingly low, for three competent models to all disagree with an actually correctly-keyed answer, which would be a false positive flag by our diagnostic. However, if we observe an occurrence rate of full disagreement well above 8-in-1000 when running our M_3/D_3 diagnostic across a large number of questions, then one or more of the following must be true: Assumptions 1, 2, or 3 are violated, or many benchmark questions are in fact *fallacious*. As we empirically show in Section 4 and Appendix A, our diagnostic approach appears to rest on robust enough forms of these assumptions so as to reliably surface the remaining culprit: benchmark fallaciousness.

3.3 Experiments

Benchmarks. We evaluate diagnostic performance against independent ground truth on two large verified benchmarks, MMLU and HLE, and separately apply our diagnostic to five additional benchmarks with partial or no post-hoc verification. For MMLU (Hendrycks et al., 2021), we use 2,489 verified questions from MMLU-Redux (Gema et al., 2025), stratifying the 57 subjects by verified error rate and selecting 25 text-only subjects (the 10 highest, the 10 lowest, and 5 drawn from the middle). For HLE (Phan et al., 2025), we use 1,147 text-only items from HLE-Verified (Zhai et al., 2026), specifically the 575 *Gold* (verified correct) and 572 *Revision* (verified as requiring revision) questions. We exclude the *Uncertain* subset that the verifiers did not adjudicate.

The remaining benchmarks differ in their verification status. Vendrow et al. (2025) verified a 300-question subset of GSM8K (Cobbe et al., 2021) as part of the PlatinumBench effort, identifying 27 bad questions (26 poorly written and 1 mislabeled). MedXpertQA (Zuo et al., 2025) has not been externally verified; the authors performed multi-round expert review during construction, but no independent post-hoc audit has been published. MedQA (Jin et al., 2021) was re-annotated by Saab et al. (2024) using at least three primary care physicians per question. The released annotations record per-rater answers but were not adjudicated into a corrected ground truth: physicians frequently disagreed on judgment-heavy items, and the reported 7.4% unfit rate is a threshold applied over rater agreement rather than a settled relabeling. The underlying ambiguity is therefore documented but unresolved. OlympicArena (Huang et al., 2024) releases ground-truth answers only for its small validation split, with the larger test split held out for a public leaderboard; we therefore evaluate on the 227 items that are simultaneously text-only, English, and accompanied by a public answer key, drawn from the validation splits of all seven subject configurations on Hugging Face. PHYBench (Qiu et al., 2025) similarly withholds answers for 400 of its 500 physics problems; we evaluate on all 100 unique questions released with public reference answers. We apply our diagnostic to the items released with ground-truth answers in each benchmark and use the results to flag the items, subjects, and benchmarks that should be prioritized for review. All experiments used API-based model inference followed by local CPU post-processing; no GPU training or large-scale local compute was used.

Annotator panel and detection thresholds. The panel comprises three of the highest-performing general-purpose models available from distinct providers at the time of evaluation: Claude Opus 5.5, Gemini 3.8 Flash, and GPT-6 Sol, all released in September 2026. Each model receives the prompt and the expanded option set $\mathcal{O}_j^+$ and returns a single selection. We report results using M_3/D_3 as the primary measure, and alternative specifications in selected comparisons.

Evaluation metrics. For benchmarks with verified labels (MMLU and HLE), we report *precision* (the fraction of diagnostic-flagged items the verifier labeled as problematic) and *recall* (the fraction of verifier-labeled items our diagnostic flags). MMLU-Redux positives are items with `error_type` other than `ok`; HLE-Verified positives are `Verified_Classes` = `Revision`. For unverified or partially verified benchmarks (GSM8K, MedQA, MedXpertQA, OlympicArena, PHYBench), we report benchmark- and subject-level flag rates to identify benchmarks and subjects warranting further review and to examine how these rates vary with benchmark release date. We do not interpret these flag rates as estimates of defect prevalence.

4 Results

4.1 Full Disagreement Recovers the Verified Quality Profile

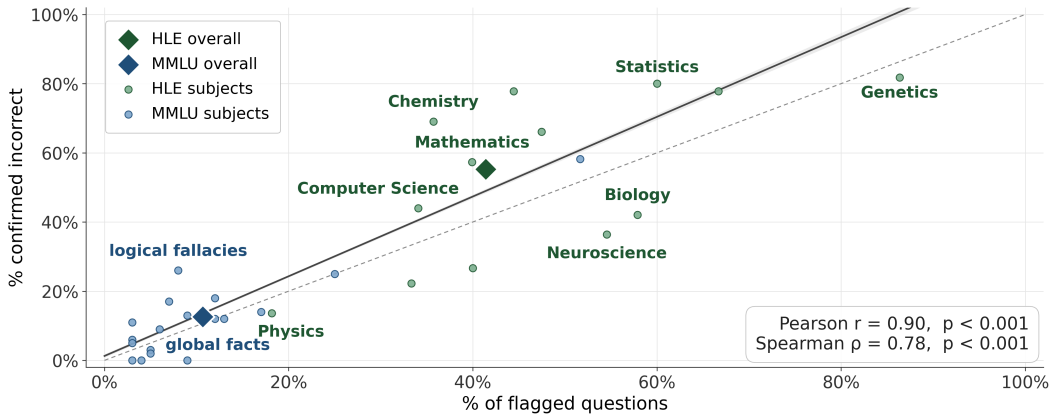


Figure 1: Per-subject M_3/D_3 flag rate against verified defect rate for MMLU and HLE using the September 2026 panel. Circles show subjects with at least three flagged questions; diamonds mark aggregates over the displayed subjects for each benchmark. The dashed line is $y = x$, and the solid line is the N-weighted regression fit with a 95% confidence band. The regression and correlations include all 134 subjects, including those omitted from display. Flag rates closely track independently verified defect rates (weighted Pearson $r = 0.90$, weighted Spearman $\rho = 0.78$, both $p < 0.001$).

M_3/D_3 produces a per-subject signal that tracks the per-subject distribution of errors found by independent expert verification (Figure 1). Outlier subjects in fallaciousness, *e.g.*, statistics and genetics per HLE-Verified (Zhai et al., 2026), are similarly identified as outliers by M_3/D_3. At the aggregate benchmark level, among questions for which all three panel models returned an answer, the recovered signals of fallaciousness closely mirror independent expert verification: M_3/D_3 flags 7.9% of MMLU compared with an independently verified defect rate of 9.1%, and 39.7% of HLE compared with 49.9%. This suggests our approach could serve as an automatic, cost-effective guide to prioritize benchmark- or subject-level human review.

Benchmark	Questions	Defects	Flags	Precision	Recall
MMLU	2,489	230	196	72.4%	61.7%
HLE	1,147	572	443	84.4%	65.4%

Table 1: Item-level M_3/D_3 performance against the independently published MMLU-Redux and HLE-Verified labels. Missing or unusable responses cannot produce a flag; recall is computed against the full set of independently verified defects.

At the item level, most questions flagged by unanimous disagreement are independently verified as defective, while the diagnostic recovers approximately three-fifths of verified MMLU defects and two-thirds of verified HLE defects. Thus, the subject-level relationship also translates into a selective item-level screening signal.

4.2 Robustness of Assumptions

Assumption 1 (Competence). To test whether our diagnostic operates in a high-competence regime, we measure p directly as the fraction of independently verified sound items answered correctly by each model, restricting attention to items for which all three returned a scoreable answer. On MMLU, p is 96.3% for Claude Opus 5.5, 95.3% for Gemini 3.8 Flash, and 91.3% for GPT-6 Sol; on HLE, it is 80.5% for Claude, 66.7% for Gemini, and 63.0% for GPT.

Assumption 2 (Independence). We test how the MMLU and HLE results change as the expected independence among panel members varies. First, we vary M and compute all single- and two-model combinations on the common set of questions answered by all three models (full results in Table 2 in Appendix A.1). On MMLU, single-model precision ranges from 68–73.1%, while two-provider panels achieve 75.4–78.6%. On HLE, the corresponding ranges are 69.2–77.7% and 76.5–80.4%. The full M_3/D_3 panel achieves 84.0% precision on MMLU and 84.1% precision on HLE, exceeding every single-model and two-provider specification on both benchmarks in the tested sample.

We examine response dependence more directly by collecting three fresh responses from each current panel model on a stratified 1,000-item sample comprising 500 MMLU and 500 HLE questions. For Gemini 3.8 Flash, we additionally collect three responses at each of three temperatures: 0.0, 0.7, and 1.0. On MMLU, repeated calls to the same model return exactly the same answer on 94.8–95.6% of complete items, while combining Gemini responses across temperatures leaves exact agreement similarly high at 95.0–96.4%. Cross-provider agreement remains high at 87.0–87.4%, but is approximately eight percentage points lower than repeated-model agreement. HLE is predominantly open-form, so literal matching understates similarity; after independently judging each response against the reference, all three responses receive the same correct-versus-incorrect judgment on 85.6–89.7% of complete items within repeated-model panels, compared with 66.0–68.7% across providers. Further results in Appendix A.1 report precision and recall across all repeated-model, temperature-varied, and cross-provider specifications; no repeated-model or temperature-varied specification provides a consistently better precision–recall trade-off than the primary panel across both benchmarks. Together, these results provide empirical support for provider diversity as a source of less-correlated responses that repeated sampling or temperature variation alone does not reproduce.

Assumption 3 (Expressibility). Our expansion of $\mathcal{O}_j$ to $\mathcal{O}_j^+$, which ensures that MULTIPLE and NONE are available answers, is intended to reduce the likelihood that expressibility is violated. We isolate its effect by restricting this comparison to questions with fixed-answer choices and evaluating all 2,489 MMLU questions and all 293 HLE multiple-choice questions with and without these additional responses. For MMLU, precision increases by 4.7 pp from $\mathcal{O}_j$ to $\mathcal{O}_j^+$, from 67.8% to 72.4%, while recall increases by 26.1 pp, from 35.7% to 61.7%. For HLE, precision increases by 3.3 pp, from 75.0% to 78.3%, while recall increases by 24.2 pp, from 45.0% to 69.2%. For open-form questions, models can express such cases directly in their free-text responses. Across the 833 HLE open-form questions with complete responses under both formulations, explicitly instructing models to return MULTIPLE or NONE when appropriate increases recall by 5.2 pp while decreasing precision by 0.4 pp, indicating that the standard formulation already captures most of the available benefit; we report the full comparison In Appendix A.2. By contrast, explicitly accommodating multiple-correct and no-correct-answer cases materially strengthens the diagnostic for the multiple-choice questions in both benchmarks, as expanding the answer space simultaneously increases precision and substantially improves recall for these question formats.

4.3 Capability Comparisons

Alternative D and C thresholds. While M_3/D_3 is the configuration used throughout the paper, our diagnostic framework also supports alternative operating points for users prioritizing precision or recall. $M_3/D_3/C_3$, which additionally requires all three models to converge on the same alternative answer, reaches 70.7% precision and 46.1% recall on MMLU and 90.7% precision and 27.3% recall on HLE. Convergence therefore improves precision on HLE, from 84.4% under M_3/D_3 to 90.7%, but not on MMLU, where precision instead decreases slightly from 72.4% to 70.7%. Its effect is thus benchmark-dependent, and the recall cost is especially pronounced for HLE open-form questions, where models may reject the reference while producing substantively different alternatives. At the other end of the trade-off, M_3/D_1, which flags any item for which at least one of the three models

disagrees with the keyed answer, reaches 67% and 87% recall on MMLU and HLE, at the cost of 41% and 64% precision. Full results across D and C thresholds are reported in Appendix B.1.

Capability over time. We expect diagnostic reliability to scale with frontier capability: stronger models err less often on sound items, lowering the empirical false-positive rate. We test this by rerunning M_3/D_3 on the same MMLU and HLE pools at four panel vintages, substituting in the strongest contemporaneous frontier model from each provider. The Early-2025 panel uses Claude Sonnet 4, OpenAI o1, and Gemini 2.5 Pro; Mid-2025 uses Claude Opus 4, OpenAI o3, and Gemini 2.5 Pro; Early-2026 uses Claude Opus 4.5, GPT-5.2, and Gemini 3 Pro; and September-2026, our primary specification, uses Claude Opus 5.5, GPT-6 Sol, and Gemini 3.8 Flash. Pipeline, prompt, expanded option set, and independently verified reference labels are held fixed.

Figure 2 shows that diagnostic precision improves across successive frontier panels on both benchmarks, with the largest gains occurring among the most recent model generations. Recall improves on MMLU but declines on HLE, where the high recall of earlier panels accompanies much lower precision and substantially more over-flagging of sound items. As frontier capability improves, models become less likely to unanimously reject valid keys, producing a more precise diagnostic signal without any methodological change. This improvement depends on model capability rather than panel size: adding six previously evaluated weaker models produces modest and non-monotonic changes in precision while consistently reducing recall. These results are reported in Figure 6 in Appendix B.2 and show that a small cross-provider frontier panel provides a better precision–recall balance than a larger panel containing weaker models.

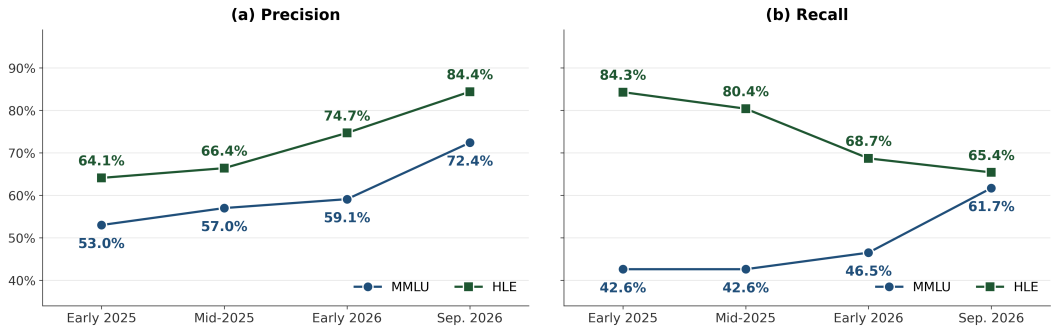


Figure 2: M_3/D_3 performance against independently verified labels across four successive frontier panels, shown separately for (a) precision and (b) recall. Lines denote MMLU and HLE. Missing or unusable responses cannot produce an M_3/D_3 flag; recall is computed against the full fixed set of independently verified defects.

LLM-as-a-Judge. A natural alternative to our diagnostic's disagreement-based signal is to ask each frontier model to directly judge whether a benchmark item is defective. To do this, we show the question including its original options and the labeled correct answer, and request a single GOOD/BAD/UNSURE verdict. When all three judges return BAD, a question is deemed panel-flagged. Appendix B.3 contains the full prompt and per-model verdict distributions.

Single-model judges over-flag on both benchmarks, while unanimous verdicts tighten the flagged set and improve precision (Table 4 in Appendix B.3), benefiting from the same aggregation dynamics as our disagreement-based approach. However, the unanimous judge panel trails our primary M_3/D_3 specification on both precision and recall for MMLU, achieving 54.2% precision and 45.2% recall compared with 72.4% and 61.7%, respectively. On HLE, the judge panel achieves similar precision, 83.2% compared with 84.4%, but lower recall, 53.8% compared with 65.4%. Directly asking models to identify defective items therefore does not improve upon the simpler disagreement signal at the same number of model calls. Moreover, answer selections expose both disagreement with the key and convergence on a common alternative, supporting operating points based on D and C that are unavailable from GOOD/BAD/UNSURE verdicts alone.

4.4 Supplementary Review of Apparent False Positives

Our primary evaluation uses the independently published MMLU-Redux and HLE-Verified labels as fixed ground truth. We additionally examine apparent false positives, for which all three models reject the keyed answer but the independent verification does not classify the item as defective. To avoid reviewing only disagreements favorable to our method, we construct a broader pool containing disagreements in both directions across multiple frontier-panel vintages. Because this review is conducted by the authors and informed by model outputs, we treat it as supplementary evidence rather than a replacement for the independent verification.

Among the current panel's apparent false positives, our review finds that the panel correctly identifies a problem with the question, options, or reference answer in 37 of 54 MMLU cases and 14 of 69 HLE cases. These often involve ambiguity, multiple defensible answers, or contested reference answers. Figure 9 in Appendix D presents a representative example in which two answer choices name sovereign states, leading all three models to reject the single-answer key.

As a sensitivity analysis, incorporating the author-review decisions increases both precision and recall on MMLU and HLE. We report these results only as supplementary evidence; all headline findings and comparisons use the independently published labels. The review protocol and additional examples are provided in Appendix D.

4.5 Unverified Benchmarks

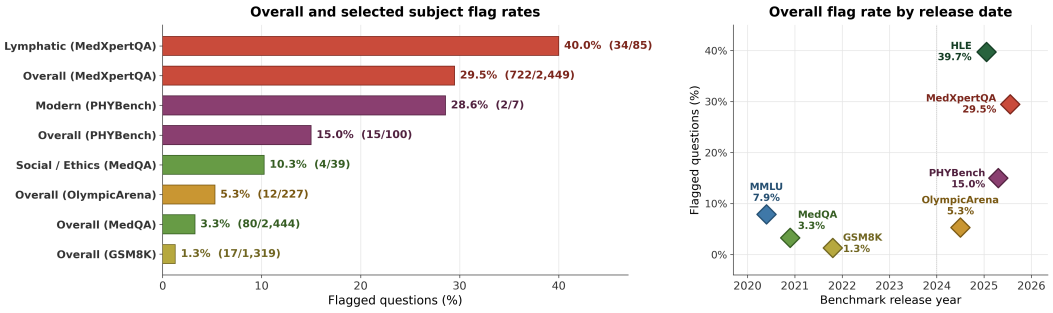


Figure 3: Diagnostic flag rates across benchmarks. Left: selected subject-level and overall rates for PHYBench (plum), MedXpertQA (red), MedQA (green), OlympicArena (gold), and GSM8K (olive). Right: overall flag rate by benchmark release date, with independently verified MMLU and HLE included as reference points. For benchmarks without complete independent verification, flag rates indicate priorities for review, not estimated defect rates.

Figure 3 applies the same diagnostic to benchmarks without complete independent verification. These flag rates should not be interpreted as estimates of defect prevalence: higher rates may reflect genuine item defects, greater difficulty, specialized subject matter, answer-format differences, or lower panel competence in a particular domain. Nevertheless, the signal has demonstrated diagnostic value on independently verified data. On MMLU and HLE, M_3/D_3 achieves 72.4% and 84.4% precision, respectively, and subject-level flag rates closely track the distribution of independently verified defects (Pearson $r = 0.90$; Spearman $\rho = 0.78$; both $p < 0.001$).

The results therefore identify where expert review is likely to be most valuable rather than establishing how many items are defective. The higher rates observed for several recent, demanding benchmarks suggest that familiar quality-control problems may be recurring as benchmark difficulty increases, while subject-level breakdowns like in Figure 1 may provide a more targeted basis for allocating review effort. Representative auditing of both flagged and unflagged questions would still be required to estimate defect prevalence and recall. Examples of flagged questions, additional per-benchmark details, and a discussion of model memorization are provided in Appendix C.

5 DISCUSSION

As benchmarks have grown harder, the gap between what scores measure and what they are taken to measure has widened. If a substantial fraction of items are fallacious, a benchmark score mixes accuracy on well-formed items with performance on ambiguous, mis-keyed, or otherwise defective items. M_3/D_3 provides a cheap and sufficiently accurate way to diagnose the mixture's composition.

Our diagnostic signal has utility at two granularities. At the aggregate level, over a benchmark or a subject within one, the flag rate recovers the quality profile established by independent expert verification without per-benchmark tuning (§4.1). On benchmarks without complete verification, these rates indicate where review should be prioritized rather than directly estimating defect prevalence (§4.5). At the item level, individual flags serve as candidates for human review. The flexibility of diagnostic specifications (Appendix B.1) lets practitioners select the most practical operating point.

The diagnostic also bears on benchmark contamination. If models inflate scores by reproducing memorized keyed answers, those are precisely the items on which the panel would converge with the key, not unanimously reject it. Our diagnostic results therefore appear largely robust to contamination of this form. The extent of full disagreement on contested items suggests that concern about memorization should be accompanied by equal attention to benchmark validity.

Closed-form benchmarks remain an efficient form of evaluation for the broader population of models the field uses, including open-source releases, smaller models, and fine-tuned variants for which large-pool psychometric audits are impractical. Our approach offers a scalable way to make these benchmarks more reliable. At the same time, recent benchmarks increasingly release a small public subset while holding the remaining items behind a hosted leaderboard to prevent memorization (Qiu et al., 2025; Huang et al., 2024). This limits external review of the items driving reported results. The elevated diagnostic flag rates on several public subsets do not establish their defect prevalence, but they identify benchmarks and subjects that warrant closer examination. Greater access for community auditing would help ensure the quality and reliability of benchmark results.

Limitations. Our approach applies to benchmarks with answer-keyed reference targets, including multiple-choice and open-form items whose references can be compared with model responses. Agentic and behavioral evaluations based on trajectories, tool use, interaction, or other non-static outcomes require different auditing methods.

Our primary evaluation relies on MMLU-Redux and HLE-Verified, which provide high-quality, independently produced ground truth but, like any large verification effort, retain a limited number of difficult or contestable cases. The supplementary review in §4.4 and Appendix D illustrates several such examples without replacing the published labels used for our headline results.

The method also depends on competence, independence, and expressibility. We test each assumption empirically in §4.2: the panel operates above chance on verified sound items, cross-provider models provide substantially more diverse evidence than repeated samples from one model, and expanding the answer space improves diagnostic performance. These results support the assumptions in our setting without establishing that they will hold equally in low-resource languages or specialist domains where current frontier models are less capable. Finally, the diagnostic identifies items for review but does not itself repair them; scalable correction is a natural direction for future work.

6 CONCLUSION

"Agree to disagree" is normally where a conversation ends; for benchmark design, it may be where more rigorous validation begins. We have provided empirical evidence of our diagnostic approach's practical and emerging utility, particularly when tuned to the full disagreement of three frontier, cross-developer models. Our results suggest that the high-pressure pursuit of benchmark difficulty has increasingly settled for the wrong kind—fallaciousness—which raises urgent questions about whether benchmarking may be leading us astray as a measure of progress in the field. However, M_3/D_3 and its variants leverage recent and growing model capabilities to empower more rigorous benchmark design and review, which have previously proved too resource-intensive, and offer a promising path forward.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology and experiments, assist with method implementation, support qualitative and thematic analysis, and help interpret results. We did not use generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, contribute to or write mathematical proofs, propose or refine hypotheses, assist with translation, or clean and reformat datasets. No other required-disclosure tasks were applicable. Additionally, we used generative AI tools and coding agents to write and debug analysis code, prepare figures, and revise and polish portions of the manuscript. The authors reviewed all AI-assisted work, tested the resulting code, and verified the reported results against the underlying experimental outputs. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study analyzes existing benchmark datasets and outputs from language models. It does not involve human subjects, collect personal or sensitive information, deploy systems that affect individuals, or introduce a new dataset or model. Accordingly, no institutional review, human-subject protections, or additional privacy and security safeguards were required. We identify no other material ethical concerns specific to this work.

REPRODUCIBILITY STATEMENT

The main contribution of this work can be reproduced without a newly collected dataset or trained model. The benchmark datasets are publicly available, and the models used in each panel are accessible through public APIs. Section 3.3 documents the benchmark subsets, model configurations, prompts, diagnostic thresholds, and evaluation metrics, while the appendices provide the complete prompts, additional specifications, and item-level analyses.

REFERENCES

Ruth Ben-Yashar and Shmuel Nitzan. Are two better than one? A note. *Public Choice*, 171 (3):323–329, June 2017. ISSN 1573-7101. doi: 10.1007/s11127-017-0439-7. URL https://doi.org/10.1007/s11127-017-0439-7.

Samuel Bowman and George Dahl. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4843–4855, 2021. URL https://aclanthology.org/2021.naacl-main.385/.

Nitay Calderon, Roi Reichart, and Rotem Dror. The alternative annotator test for llm-as-a-judge: How to statistically justify replacing human annotators with llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16051–16081, 2025. URL https://aclanthology.org/2025.acl-long.782/.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems, November 2021. URL http://arxiv.org/abs/2110.14168. arXiv:2110.14168 [cs].

A. P. Dawid and A. M. Skene. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28, 1979. ISSN 0035-9254. doi: 10.2307/2346806. URL https://www.jstor.org/stable/2346806.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile Van Krieken, and Pasquale Minervini. Are We Done with MMLU? In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for*

Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 5069–5096, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.262. URL https://aclanthology.org/2025.naacl-long.262/.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, Olli Järviniemi, Matthew Barnett, Robert Sandler, Matej Vrzala, Jaime Sevilla, Qiuyu Ren, Elizabeth Pratt, Lionel Levine, Grant Barkley, Natalie Stewart, Bogdan Grechuk, Tetiana Grechuk, Shreepranav Varma Enugandla, and Mark Wildon. FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI, December 2025. URL http://arxiv.org/abs/2411.04872. arXiv:2411.04872 [cs].

Bernard Grofman, Guillermo Owen, and Scott L. Feld. Thirteen theorems in search of the truth. *Theory and Decision*, 15(3):261–278, September 1983. ISSN 1573-7187. doi: 10.1007/BF00125672. URL https://doi.org/10.1007/BF00125672.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding, January 2021. URL http://arxiv.org/abs/2009.03300. arXiv:2009.03300 [cs].

Wenqi Huang, Charley Lee, Leonard Tng, and Serena Ge. Deepswe: Measuring frontier coding agents on original, long-horizon engineering tasks. *arXiv preprint arXiv:2607.07946*, 2026. URL https://arxiv.org/abs/2607.07946.

Zhen Huang, Zengzhi Wang, Shijie Xia, Xuefeng Li, Haoyang Zou, Ruijie Xu, Run-Ze Fan, Lyumanshan Ye, Ethan Chern, Yixin Ye, Yikai Zhang, Yuqing Yang, Ting Wu, Binjie Wang, Shichao Sun, Yang Xiao, Yiyuan Li, Fan Zhou, Steffi Chern, Yiwei Qin, Yan Ma, Jiadi Su, Yixiu Liu, Yuxiang Zheng, Shaoting Zhang, Dahua Lin, Yu Qiao, and Pengfei Liu. OlympicArena: Benchmarking Multi-discipline Cognitive Reasoning for Superintelligent AI, June 2024. URL http://arxiv.org/abs/2406.12753. arXiv:2406.12753 [cs].

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. *Applied Sciences*, 11(14):6421, January 2021. ISSN 2076-3417. doi: 10.3390/app11146421. URL https://www.mdpi.com/2076-3417/11/14/6421.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, September 2023. ISSN 26663899. doi: 10.1016/j.patter.2023.100804. URL https://linkinghub.elsevier.com/retrieve/pii/S2666389923001599.

Drora Karotkin and Jacob Paroush. Optimum committee size: Quality-versus-quantity dilemma. *Social Choice and Welfare*, 20(3):429–441, June 2003. ISSN 1432-217X. doi: 10.1007/s003550200190. URL https://doi.org/10.1007/s003550200190.

Sander Land and Daniel M Bikel. Auditing llm benchmarks with item response theory. *arXiv preprint arXiv:2605.30504*, 2026. URL https://arxiv.org/abs/2605.30504.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment, May 2023. URL http://arxiv.org/abs/2303.16634. arXiv:2303.16634 [cs].

Omer Nahum, Nitay Calderon, Orgad Keller, Idan Szpektor, and Roi Reichart. Are llms better than reported? detecting label errors and mitigating their effect on model performance. In *Proceedings of the 2025 conference on empirical methods in natural language processing*, 2025. URL https://aclanthology.org/2025.emnlp-main.1360/.

Shmuel Nitzan and Jacob Paroush. Optimal Decision Rules in Uncertain Dichotomous Choice Situations. *International Economic Review*, 23(2):289–297, 1982. ISSN 0020-6598. doi: 10.2307/2526438. URL https://www.jstor.org/stable/2526438.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks, November 2021. URL http://arxiv.org/abs/2103.14749. arXiv:2103.14749 [stat].

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025. URL https://arxiv.org/abs/2501.14249.

Shi Qiu, Shaoyang Guo, Zhuo-Yang Song, Yunbo Sun, Zeyu Cai, Jiashen Wei, Tianyu Luo, Yixuan Yin, Haoxu Zhang, Yi Hu, Chenyang Wang, Chencheng Tang, Haoling Chang, Qi Liu, Ziheng Zhou, Tianyu Zhang, Jingtian Zhang, Zhangyi Liu, Minghao Li, Yuku Zhang, Boxuan Jing, Xianqi Yin, Yutong Ren, Zizhuo Fu, Jiaming Ji, Weike Wang, Xudong Tian, Anqi Lv, Laifu Man, Jianxiang Li, Feiyu Tao, Qihua Sun, Zhou Liang, Yushu Mu, Zhongxuan Li, Jing-Jun Zhang, Shutao Zhang, Xiaotian Li, Xingqi Xia, Jiawei Lin, Zheyu Shen, Jiahang Chen, Qiuhao Xiong, Binran Wang, Fengyuan Wang, Ziyang Ni, Bohan Zhang, Fan Cui, Changkun Shao, Qing-Hong Cao, Ming-xing Luo, Yaodong Yang, Muhan Zhang, and Hua Xing Zhu. PHYBench: Holistic Evaluation of Physical Perception and Reasoning in Large Language Models, April 2025. URL http://arxiv.org/abs/2504.16074. arXiv:2504.16074 [cs].

Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. AI and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*, 2021.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: rapid training data creation with weak supervision. *Proc. VLDB Endow.*, 11(3): 269–282, November 2017. ISSN 2150-8097. doi: 10.14778/3157794.3157797. URL https://dl.acm.org/doi/10.14778/3157794.3157797.

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, November 2023. URL http://arxiv.org/abs/2311.12022. arXiv:2311.12022 [cs].

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean-baptiste Alayrac, Neil Houlsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, SiWai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Ben Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathon Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale Webster, Joelle Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gottweis, Alan Karthikesalingam, and Vivek Natarajan. Capabilities of Gemini Models in Medicine, May 2024. URL http://arxiv.org/abs/2404.18416. arXiv:2404.18416 [cs].

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark, October 2023. URL http://arxiv.org/abs/2310.18018. arXiv:2310.18018 [cs].

Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D'Souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah A. Smith, Beyza Ermis, Marzieh Fadaee, and Sara Hooker. The Leaderboard Illusion, April 2025. URL https://arxiv.org/abs/2504.20879v2.

Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pp. 254–263, 2008. URL https://aclanthology.org/D08-1027/.

Sang Truong, Yuheng Tu, Michael Hardy, Anka Reuel-Lamparth, Zeyu Tang, Jirayu Burapacheep, Jonathan Perera, Chibuike Uwakwe, Benjamin Domingue, Nick Haber, et al. Fantastic bugs and where to find them in AI benchmarks. *Advances in Neural Information Processing Systems*, 38, 2026. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/5b44b99d012922011f0be1031a98c4f9-Abstract-Datasets_and_Benchmarks_Track.html.

Joshua Vendrow, Edward Vendrow, Sara Beery, and Aleksander Madry. Do Large Language Model Benchmarks Test Reliability?, February 2025. URL http://arxiv.org/abs/2502.03461. arXiv:2502.03461 [cs].

Stefan Wittlinger, Josefine Meerjansen, Fabian Wolf, Isabella C Wiest, Matthias P Ebert, Fabian Siegel, and Sebastian Belle. Multi-llm disagreement as a scalable detector of human annotation errors in structured data from clinical free-text. *medRxiv*, 2026. URL https://www.medrxiv.org/content/10.64898/2026.05.04.26352392v1.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094, 2024. URL https://neurips.cc/virtual/2024/poster/97468.

Weiqi Zhai, Zhihai Wang, Jinghang Wang, Boyu Yang, Xiaogang Li, Xander Xu, Bohan Wang, Peng Wang, Xingzhe Wu, Anfeng Li, et al. HLE-Verified: A systematic verification and structured revision of humanity's last exam. *arXiv preprint arXiv:2602.13964*, 2026. URL https://arxiv.org/abs/2602.13964.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, December 2023. URL http://arxiv.org/abs/2306.05685. arXiv:2306.05685 [cs].

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. MedXpertQA: Benchmarking Expert-Level Medical Reasoning and Understanding, 2025. URL http://arxiv.org/abs/2501.18362. arXiv:2501.18362 [cs].

APPENDIX

A FURTHER TESTS OF ASSUMPTIONS

A.1 INDEPENDENCE: CROSS-DEVELOPER VS. WITHIN-MODEL PANELS

We test the sensitivity of our MMLU and HLE results to model specifications with different levels of expected model independence. The analysis uses a stratified 1,000-item sample comprising 500 MMLU questions (100 defective and 400 sound according to MMLU-Redux) and 500 HLE questions (250 *Revision* and 250 *Gold* according to HLE-Verified). All specifications are evaluated on the same questions. Because the sample is stratified by verification outcome, these results are intended to compare panel specifications within the experiment rather than estimate precision under each benchmark's original defect prevalence.

First, we vary M to compare all single- and two-model combinations with the full three-model panel. We then hold the model fixed and combine three independently collected responses from Claude Opus 5.5, GPT-6 Sol, or Gemini 3.8 Flash. For Gemini, we additionally compare three responses collected at temperatures 0.0, 0.7, and 1.0, as well as panels containing one response from each temperature. Table 2 reports precision and recall for all specifications.

		MMLU		HLE	
Panel	**Type**	**Precision**	**Recall**	**Precision**	**Recall**
Claude	single	73.1%	49.0%	77.7%	68.4%
Gemini	single	68.0%	51.0%	69.2%	79.2%
GPT	single	68.9%	**71.0%**	71.2%	**88.8%**
Claude + Gemini	pair	78.6%	44.0%	80.4%	57.6%
Claude + GPT	pair	78.6%	44.0%	80.2%	64.8%
Gemini + GPT	pair	75.4%	49.0%	76.5%	69.2%
All 3 (M_3/D_3)	cross-developer	**84.0%**	42.0%	**84.1%**	65.6%
Claude $\times 3$	repeated	79.3%	46.0%	79.7%	56.4%
GPT $\times 3$	repeated	66.7%	56.0%	75.4%	73.6%
Gemini $\times 3$, $T = 0.0$	repeated	73.5%	50.0%	73.4%	67.2%
Gemini $\times 3$, $T = 0.7$	repeated	73.5%	50.0%	72.0%	64.8%
Gemini $\times 3$, $T = 1.0$	repeated	71.0%	49.0%	72.4%	65.2%
Gemini mixed T, run 1	mixed temperature	72.5%	50.0%	72.4%	64.0%
Gemini mixed T, run 2	mixed temperature	72.1%	49.0%	74.0%	64.8%
Gemini mixed T, run 3	mixed temperature	74.6%	50.0%	72.7%	66.0%

Table 2: Precision and recall across panel specifications on the stratified 1,000-item sample. *Single* rows flag an item when the listed model rejects the keyed answer; all other rows require unanimous rejection among the listed responses. Mixed-temperature panels contain one Gemini 3.8 Flash response from each of $T \in \{0.0, 0.7, 1.0\}$. Missing or unusable responses cannot produce a flag.

Three observations emerge. First, increasing a cross-developer panel from one to three models consistently raises precision, although stricter unanimity can reduce recall. The full cross-developer panel achieves the highest precision on both benchmarks, while the high recall of some single-model specifications is accompanied by substantially lower precision. Second, repeated calls to the same model do not reproduce the same precision–recall balance. Some repeated panels recover greater recall, but none matches the precision of the three-provider panel on either benchmark. Third, varying Gemini's temperature has little effect: both the fixed-temperature and mixed-temperature panels occupy a narrow range of precision and recall across repeated runs.

These results do not imply that responses from different providers are statistically independent. They show empirically that provider diversity contributes information that repeated sampling from one model, including sampling across temperatures, does not fully recover. The cross-developer construction is therefore preferred for its observed precision–recall balance rather than because independence holds exactly.

A.2 Expressibility: Generalization to Open-Form Items

HLE contains 854 open-form items where the answer is not selected from a fixed option set. Our diagnostic approach extends to this regime by interpreting $\mathcal{O}_j^+$ semantically rather than as a literal set of additional choices. A model can provide an answer different from y_j, indicate that no determinate answer exists, or state that multiple answers are defensible. Disagreement with the reference and agreement among annotators are therefore judged by semantic equivalence rather than literal string matching.

To test whether these possibilities must be stated explicitly, we rerun the September-2026 panel with a system prompt instructing each model to return NONE when no determinate answer exists, MULTIPLE when multiple substantively different answers are defensible, and a single answer otherwise. We compare this condition with our standard open-form prompt on the 833 questions for which all three models returned scoreable responses under both formulations, holding the questions, models, HLE-Verified labels, and scoring procedure fixed.

Open-form formulation	Flag rate	Precision	Recall
Standard semantic formulation	38.8%	84.8%	62.1%
Explicit NONE/MULTIPLE	42.3%	84.4%	67.3%

Table 3: M_3/D_3 performance on 833 HLE open-form questions with complete responses under both prompt formulations, evaluated against HLE-Verified labels.

Explicitly naming NONE and MULTIPLE increases recall by 5.2 percentage points and the flag rate by 3.5 points, while precision changes by only -0.4 points. The standard formulation therefore already captures most of the available open-form signal, although explicit instructions recover some additional verified defects. Figure 4 shows a representative case in which the models express the underlying ambiguity through their free-text answers.

HLE · Mathematics *False positive*

You have a magical rope that has 60 cm length. If a portion of the rope is measured and cut, the longer portion shrinks by 25%. It is possible to get exactly 15 cm of rope with only two cuts. What will be the length of the longer portion after the first cut?

Original benchmark answer: 36 cm
Frontier-model answers: Claude: 30 cm
Gemini: 30 cm
GPT: 30 cm (one valid answer; not unique)

Figure 4: An HLE Mathematics item where M_3/D_3 unanimously disagreed with the keyed answer. The benchmark's published rationale derives 36 cm from a specific worked example (first cut = 12 cm $\rightarrow$ 48 cm, which shrinks to 36 cm), but the problem as stated does not pin down a unique first-cut length: multiple integer cut sizes admit a valid two-cut path to the required 15 cm of rope, including the 30 cm answer on which the three frontier models converge. Manual review classified the original key as under-specified rather than uniquely correct.

B Further Comparisons of Capability

B.1 Alternative D and C Thresholds

Beyond the panel size M used in the main text, our diagnostic approach offers two further thresholds the user can tune: the disagreement threshold D and the convergence threshold C. Increasing either tightens the flag set. D controls how many panel members must reject the keyed answer, and C controls how many of those rejecting members must converge on the same alternative. Figure 5 reports per-item precision and recall for the five settings of greatest practical interest at $M = 3$: $M_3/D_3/C_3$ (full disagreement and full convergence), $M_3/D_3/C_2$ (full disagreement and at least two of the three rejecting models converging on the same alternative), M_3/D_3 (full disagreement

without a convergence requirement; the headline measure used throughout the main text), M_3/D_2 (at least two of the three models disagreeing with the keyed answer), and M_3/D_1 (at least one of the three models disagreeing with the keyed answer).

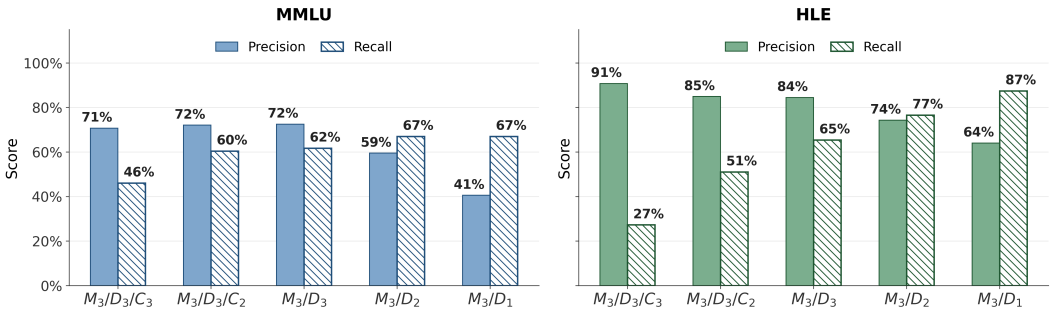


Figure 5: Precision and recall for MMLU and HLE, evaluated against independently verified labels. Each pair of bars is a configuration $M_3/D_d/C_c$, parameterized by the disagreement threshold d (the number of panel members required to disagree with the key) and the convergence threshold c (the number required to agree on the same alternative). Configurations are ordered from the strictest to the loosest flagging rule. Missing or unusable responses cannot produce a flag, and recall is computed against the full set of independently verified defects.

The effect of convergence differs across the two benchmarks. On HLE, $M_3/D_3/C_3$ delivers the highest precision, 90.7%, compared with 84.4% for M_3/D_3. The trade-off is recall, which falls from 65.4% to 27.3%, because on open-form items the models may all reject the reference while producing substantively different alternatives, thereby reducing the occurrence of full convergence. The intermediate $M_3/D_3/C_2$ specification reaches 84.9% precision and 51.0% recall. On MMLU, however, requiring convergence does not improve precision: $M_3/D_3/C_3$ reaches 70.7% precision and 46.1% recall, compared with 72.0% and 60.4% for $M_3/D_3/C_2$ and 72.4% and 61.7% for M_3/D_3. Convergence thresholds may therefore be useful when false positives are particularly costly, but their value depends on the benchmark and answer format.

Otherwise, removing the convergence requirement retains substantially more verified defects without reducing precision on MMLU and with only a moderate precision cost on HLE. We therefore recommend M_3/D_3 as the default choice.

The looser settings, M_3/D_2 and M_3/D_1, favor recall at the cost of precision. On MMLU, M_3/D_2 reaches 59.5% precision and 67.0% recall, while M_3/D_1 lowers precision to 40.5% without recovering additional verified defects. On HLE, M_3/D_2 reaches 74.2% precision and 76.6% recall, while M_3/D_1 increases recall to 87.4% at 64.0% precision. M_3/D_1 may therefore be the right choice when comprehensive review is the highest priority, although its value relative to M_3/D_2 likewise depends on the benchmark.

B.2 Adding More Models

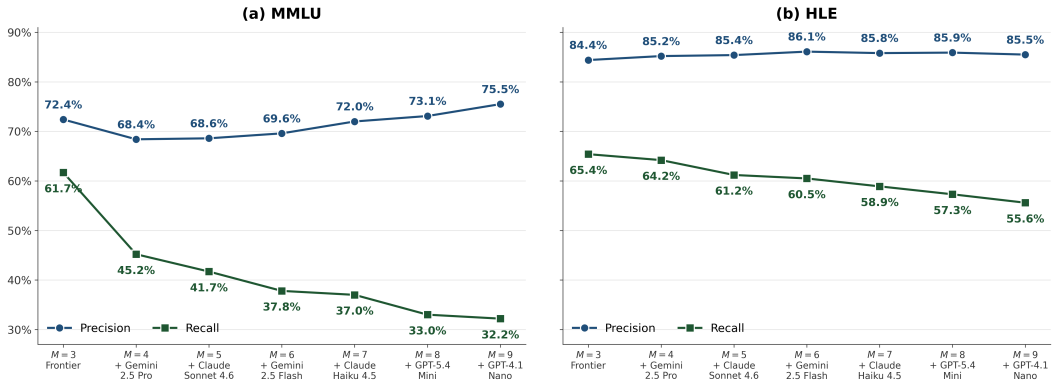


Figure 6: Precision and recall as six previously evaluated weaker models are added successively to the three-model frontier panel on (a) MMLU and (b) HLE. An item is flagged only when every model in the expanded panel disagrees with the keyed answer.

Adding models does not produce a consistent improvement. On MMLU, precision changes non-monotonically and ends only modestly higher, while recall falls from 61.7% with three models to 32.2% with all nine. On HLE, precision remains comparatively stable while recall declines from 65.4% to 55.6%. Under unanimous disagreement, each additional model effectively receives a veto: if even one weaker model accepts the keyed answer, the item is removed from the flag set. The resulting loss in recall is not offset by a reliable gain in precision. Panel quality and cross-provider diversity therefore matter more than panel size alone.

B.3 LLM-as-a-Judge

For each item we query the frontier panel used by M_3/D_3, Claude Opus 5.5, Gemini 3.8 Flash, and GPT-6 Sol, in a different framing. Each judge is shown the question, the original option set, and the labeled correct answer, and is asked whether the item is well-formed and the key is correct. The judge returns a single `VERDICT:` line and an optional one-sentence justification; the verdict is parsed by regex with no equivalence-judge step. `UNSURE` counts as not-`BAD`, so a question is panel-flagged only when all three judges commit to `BAD`.

System prompt (shared across both benchmarks and all three judges).

```
You are an expert reviewer auditing a question from an
academic benchmark used to evaluate language models.  Your
task is to judge whether the question is well-formed and has
a clear, defensible correct answer.

Respond with a single line in this exact format:
VERDICT: GOOD | BAD | UNSURE
Justification:  {one short sentence}
```

User message formats. We use separate message formats for the three item types. For MMLU, we present the question and its answer choices, label the choices with letters, and provide the keyed answer as `{letter}. {text}`. For HLE multiple-choice items, which already include lettered options in the question text, we preserve the original formatting and append `LABELED CORRECT ANSWER: {letter}`. For HLE free-response items, we present the question followed by `LABELED CORRECT ANSWER: {text}`. The corresponding multiple-choice or free-response rubric is then appended. Reference rationales are never shown to the judges.

Multiple-choice rubric.

A question is BAD if any of the following is true:

- The labeled correct answer is factually wrong.
- More than one of the listed options is defensibly correct (the labeled one is not uniquely best).
- None of the listed options is correct.
- The question is too ambiguous, underspecified, or malformed for an expert to choose a single answer.
- The question contains a factual error, broken formatting, or missing information that prevents a single correct option from being identifiable.

Otherwise it is GOOD. If you cannot decide, return UNSURE.

The free-response rubric is identical except that the second bullet is replaced with "The question admits multiple equally valid answers and the labeled answer is just one of them," and the third bullet is replaced with "The question has no determinable answer."

Per-model verdict distributions. On MMLU (n=2,489), all three judges commit to `GOOD` or `BAD` on more than 99% of items, with `BAD` rates ranging from 8.4% for Claude to 19.2% for GPT; Gemini returns `BAD` on 11.8%. HLE (n=1,147) shows greater variation, with `BAD` rates of 39.3% for Claude, 49.3% for Gemini, and 58.2% for GPT. `UNSURE` remains uncommon, accounting for 2.2% of Claude and GPT verdicts and 0.6% of Gemini verdicts on HLE. These per-model differences motivate the unanimous-`BAD` aggregation rule used in Table 4: requiring all three judges to commit to `BAD` attenuates the over-flagging visible in the less selective single-model results.

	MMLU		HLE	
Method	Precision	Recall	Precision	Recall
Claude Opus 5.5	0.522	0.474	0.796	0.628
Gemini 3.8 Flash	0.469	0.600	0.726	0.719
GPT-6 Sol	0.296	0.617	0.678	0.790
Unanimous panel	0.542	0.452	0.832	0.538

Table 4: Precision and recall of LLM-as-a-Judge baselines on MMLU and HLE, evaluated against the independently published MMLU-Redux and HLE-Verified labels. Single-model rows flag every item receiving a `BAD` verdict; the panel row flags only when all three judges return `BAD`. Missing and `UNSURE` responses do not produce flags, and recall is computed against the full set of independently verified defects.

C Further Detail on Unverified Benchmarks

We provide the per-subject breakdowns underlying Figure 3 and a short discussion of training-data contamination for GSM8K, MedQA, MedXpertQA, OlympicArena, and PHYBench. For each benchmark, we run the publicly graded portion used in our experiments through our diagnostic and report results over questions for which all three models returned scoreable answers. We record a flag, unanimous disagreement between Claude Opus 5.5, Gemini 3.8 Flash, and GPT-6 Sol with the keyed answer, as the unit of analysis. Counts are reported as "flagged / total = percentage." GSM8K (Cobbe et al., 2021) is single-domain and has no per-subject metadata; the overall result is $17/1{,}319 = 1.29\%$. For PHYBench (Qiu et al., 2025), our source snapshot contains 100 unique publicly answered questions after identical duplicate records are removed. OlympicArena (Huang et al., 2024) releases answers only for its small validation split, and we restrict the analysis to the 227 items there that are text-only and English. MedQA is classified into the 14 USMLE Step 1 organ-system categories with Gemini 2.5 Flash using the prompt below; MedXpertQA ships built-in body-system and medical-task labels.

```
You are a medical educator classifying USMLE-style questions
into the official USMLE Organ-System content outline.
Choose the SINGLE best category based on the underlying
pathology / system being tested.
Categories: Cardiovascular; Respiratory; Gastrointestinal;
Renal / Urinary; Reproductive; Endocrine; Nervous System
& Special Senses; Musculoskeletal / Skin / Connective
Tissue; Blood & Lymphoreticular / Immune; Behavioral Health;
Human Development; Multisystem Processes & Disorders;
Biostatistics & Epidemiology / Population Health; Social
Sciences (Communication / Professionalism / Ethics).
Decision rules: pick the system of the underlying
PATHOLOGY, not the system the intervention happens to act
on; use 'Multisystem Processes & Disorders' for trauma,
sepsis, polypharmacy, or items genuinely spanning 3+
systems; use 'Behavioral Health' for psychiatry, sleep,
substance use; classify by the answer choices when the stem
mentions another system.
```

Subject	Flagged / Total	Percentage
Social Sciences / Communication / Ethics	4 / 39	10.26%
Biostatistics & Epidemiology / Population Health	5 / 65	7.69%
Reproductive	11 / 222	4.95%
Blood & Lymphoreticular / Immune	11 / 252	4.37%
Cardiovascular	10 / 231	4.33%
Multisystem Processes & Disorders	4 / 126	3.17%
Behavioral Health	5 / 174	2.87%
Endocrine	5 / 178	2.81%
Musculoskeletal / Skin / Connective Tissue	6 / 239	2.51%
Nervous System & Special Senses	6 / 243	2.47%
Renal / Urinary	3 / 145	2.07%
Respiratory	3 / 155	1.94%
Gastrointestinal	5 / 263	1.90%
Human Development	2 / 112	1.79%
Overall	**80 / 2,444**	**3.27%**

Table 5: MedQA, by USMLE Step 1 organ-system category.

Training-data contamination. Cleanly measuring how much benchmark content has been ingested by a frontier model during training is genuinely hard. Training corpora and weights are not public, and partial leakage of problem text without labeled answers is difficult to bound from the outside. We do not attempt to estimate contamination directly. Instead, we note that simple recall of the

Body system	Flagged / Total	Percentage
Lymphatic	34 / 85	40.00%
Other / NA	45 / 124	36.29%
Respiratory	65 / 193	33.68%
Endocrine	58 / 175	33.14%
Nervous	127 / 387	32.82%
Urinary	40 / 124	32.26%
Cardiovascular	96 / 304	31.58%
Digestive	74 / 274	27.01%
Reproductive	54 / 201	26.87%
Skeletal	82 / 357	22.97%
Muscular	38 / 177	21.47%
Integumentary	9 / 48	18.75%
Medical task		
Basic Science	241 / 648	37.19%
Diagnosis	291 / 1,053	27.64%
Treatment	190 / 748	25.40%
Overall	**722 / 2,449**	**29.48%**

Table 6: MedXpertQA, by body system and medical task.

Tag	Flagged / Total	Percentage
MODERN	2 / 7	28.57%
ELECTRICITY	8 / 32	25.00%
MECHANICS	5 / 39	12.82%
OPTICS	0 / 8	0.00%
THERMODYNAMICS	0 / 11	0.00%
ADVANCED	0 / 3	0.00%
Overall	**15 / 100**	**15.00%**

Table 7: PHYBench, by tag, over the 100 unique publicly answered questions in our source snapshot after removing identical duplicate records.

labeled answer would push the share of flagged items *down*: a model that retrieved the labeled answer from training would score it as correct, and the unanimous disagreement required for a flag would not arise. Empirically, we still see meaningful flag rates on benchmarks that have been public long enough to plausibly be in pretraining corpora, including HLE (public since January 2025) and MMLU (public since September 2020). Across a subset of these items, the three frontier models disagreed with the labeled answer despite the answer being publicly available alongside the question, and the disagreement held up under supplementary review (Appendix D). If memorization were doing the work, the most natural target of recall is exactly that publicly available labeled answer, and the panel would converge with it rather than against it. We make this observation cautiously: we do not claim that contamination is absent, that any particular flagged item was unseen by the flagging models, or that we can cleanly separate reasoning from recall. From our results, the models disagreed with publicly available labeled answers often enough that contamination does not appear to be a serious threat to the diagnostic working as intended.

Subject	Flagged / Total	Percentage
Biology	2 / 6	33.33%
Astronomy	2 / 8	25.00%
Physics	2 / 29	6.90%
Chemistry	1 / 15	6.67%
Math	5 / 169	2.96%
Answer type		
MC (multi-correct)	2 / 5	40.00%
EX (expression)	1 / 10	10.00%
SC (single choice)	2 / 24	8.33%
NV (numeric value)	7 / 177	3.95%
EQ / TF / MA / MPV	0 / 11	0.00%
Overall	**12 / 227**	**5.29%**

Table 8: OlympicArena, by subject and answer type. Several subjects have small n and the corresponding values should be read as suggestive.

C.1 MEDXPERTQA: SIMPLE ERROR CAUGHT BY M_3/D_3

MedXpertQA · Skeletal — Treatment *Flagged item*

A 56-year-old individual initiates daily calcium supplementation to reduce their osteoporosis risk. Which preventive medicine category does this intervention represent?

(A) Quaternary prevention
(B) Disease modification prevention
(C) Secondary prevention
(D) Preventive maintenance
(E) Tertiary prevention
(F) Primary prevention
(G) Risk reduction prevention
(H) Selective prevention
(I) Targeted prevention
(J) Universal prevention
(K) More than one of the listed answers is correct
(L) None of the listed answers is correct

Original benchmark answer: E
Frontier-model answers: Claude: F Gemini: F GPT: F

Figure 7: A MedXpertQA item flagged by the September-2026 M_3/D_3 panel. The keyed answer is option (E), while all three models return (F). Taking calcium supplementation to reduce the future risk of osteoporosis is an intervention performed before the disease develops, aiming to prevent its occurrence. This is primary prevention; by definition, tertiary prevention occurs only after a disease or injury is present.

C.2 MEDXPERTQA: EXPERT-VERIFIED WORKED EXAMPLE

A medical doctor reviewed the MedXpertQA item shown in Figure 8, flagged by the M_3/D_3 panel. The item asks for the most appropriate next step in managing a patient with lower back pain and multiple risk factors; the keyed answer (physical therapy) reflects one stylized management pathway, while all three frontier models converged on a different answer (NSAIDs and observation) reflecting standard clinical practice. The expert traced the item's clinical reasoning and identified the following intersecting issues:

Incomplete workup forced into a management choice. The stem omits diagnostic steps that would precede any management decision in practice, including a CBC and urinalysis, leaving plausible differentials (renal colic, early infection, urinary tract infection) open. A clinician would typically complete those steps before committing to a management decision. The question's "most appropriate next step in management" skips past workup and forces a choice the workflow does not support.

MedXpertQA · Nervous — Diagnosis *Flagged item*

A 27-year-old incarcerated male presents to the emergency department with progressive lower back pain and reported inability to urinate. He reports nighttime chills and intermittent diarrhea, but denies trauma. History includes intravenous drug use, homelessness, antisocial personality disorder, malingering, and multiple STIs. Vital signs are stable (T 98.2°F, BP 127/77, HR 70, RR 15, SpO2 98%). Examination reveals normal heart sounds, clear lungs, intact rectal tone, no neurological deficits, and paraspinal muscle tenderness worsened by forward flexion. The patient initially refuses catheterization but later voids voluntarily with post-void residual of 50 mL. Labs: Na 140, Cl 100, K 4.0, HCO3 24, BUN 22, Glucose 116, Cr 1.2, Ca 9.9, CRP 1.0. UDS positive for benzodiazepines and opioids. Chest radiograph and ECG unremarkable. What is the most appropriate next step in management?

(A) Blood cultures and antibiotics
(B) Referral to psychiatry
(C) Epidural steroid injection
(D) NSAIDs and observation
(E) Spinal X-ray
(F) Repeat urinalysis with catheterization
(G) MRI spine
(H) Lumbar puncture
(I) Physical therapy
(J) CT spine
(K) More than one of the listed answers is correct
(L) None of the listed answers is correct

Original benchmark answer: I
Frontier-model answers: Claude: D Gemini: D GPT: D

Figure 8: A MedXpertQA Nervous–Diagnosis item discussed in this section, shown with its keyed answer and the responses from the September-2026 M_3/D_3 panel.

Sequential framing of a parallel decision. Real clinical management proceeds in parallel: symptomatic relief and referrals are initiated together, and the appropriate next step depends on patient preference and functional status. If the patient were sufficiently debilitated and willing to engage with physiotherapy, the referral would typically be initiated alongside the NSAIDs and observation rather than after them. The question's format imposes a sequential single-choice structure that does not map onto the parallel structure of the decision.

The expert's conclusion was that NSAIDs and observation (D) is what they would do in practice once the CBC and initial urinalysis are complete, with physical therapy (I) following depending on functional status and patient follow-through. The keyed answer is defensible only as the endpoint of a longer pathway the question stem skips over, and the panel's convergence on (D) reflects the standard immediate-management step rather than model error. While workup-management conflation is a known feature of board-exam-style questions, the expert judged this stem to be less complete and less well-formed than typical board items they had encountered.

D Further Detail on Supplementary Review

D.1 Procedure

As a supplementary sensitivity analysis, we manually adjudicated a broad pool of items on which M_3/D_3 disagreed with the published verifier (MMLU-Redux or HLE-Verified). Disagreements were pooled across four frontier-panel configurations spanning multiple model generations, with each item appearing exactly once in the pool: an item flagged by any one or more panels but not by the verifier counts once as an apparent false positive, and similarly for apparent false negatives. We audited 277 items on MMLU and 597 on HLE. For each item, we reviewed the full question, the reference answer

and rationale where available, the verifier's annotation, and the responses and reasoning of the panel models for the configuration that placed the item in the review pool.

Each disagreement was categorized as one of three outcomes: M_3/D_3 *right* (the review supported the panel's treatment of the item), M_3/D_3 *wrong* (the review supported the independent verifier), or *ambiguous* (the disagreement could not be cleanly resolved). Items in the ambiguous category retain the verifier's original label. Because the September-2026 panel can raise disagreements that were not raised by the configurations used to construct this pool, the review does not cover every apparent error made by the current panel. We therefore treat these decisions as supplementary evidence rather than a replacement for the independently published labels.

D.2 Examples

We provide representative examples from the supplementary review described in §4.4. Each example is a single benchmark item with the original benchmark answer, the verified-version label or correction, and the answers returned by the panel configuration that placed it in the review pool. The cases cover both apparent false positives (items M_3/D_3 flagged as defective and the verified reference retained as valid) and apparent false negatives (items M_3/D_3 did not flag and the verified reference marked as defective). In the apparent-false-positive examples below, the three models shown also exhibited full convergence on a non-keyed answer, meaning that these items would also be flagged by $M_3/D_3/C_3$.

D.2.1 Apparent False Positives: Verified-Reference Defects That M_3/D_3 Surfaces

MMLU · High School Geography *Apparent false positive*

Which of the following entities is a state?

(A) Antarctica	**(D) Marshall Islands**
(B) American Samoa	(E) More than one of the above answers is correct
(C) Egypt	(F) None of the above answers is correct

Published key: C (Egypt) **MMLU-Redux:** ok

September 2026 panel: Claude: E Gemini: E GPT: E

Figure 9: An MMLU `High School Geography` item flagged by the method but retained as `ok` by MMLU-Redux. Egypt (C) is correct, but the Marshall Islands (D) is also a sovereign UN member state; all three September-2026 panel models select (E), recognizing that multiple answers are correct.

Figure 9 and the cases below are items in which all three models shown converge on a non-keyed answer that our review finds defensible, yet the verified reference retains the original benchmark answer. Under our supplementary review, these apparent false positives reflect problems with the question, options, or reference answer rather than panel errors.

A practicing academic legal scholar reviewed the MMLU Professional Law item shown in Figure 13, flagged by the M_3/D_3 panel, and traced it back to the source shown in Figure 14. The item asks which burden of proof a defendant bears when raising insanity as an affirmative defense to murder; the keyed answer reflects one doctrinal position, while all three models shown converged on a different answer reflecting a broader consensus. The expert traced the item's authority and identified three concurrent issues:

Single-source grounding. The only online trace of the question the expert could locate was a user-generated flashcard deck of bar-exam practice questions, not an authoritative source. The keyed answer's doctrinal position is supported in that source but is not the position the larger body of case law, secondary literature, or bar-review materials would support.

Jurisdiction and time dependence. The doctrine varies across U.S. jurisdictions and has shifted over time. The prompt specifies neither, leaving the answer underdetermined: a position defensible in one jurisdiction at one period becomes indefensible in another.

MMLU · conceptual_physics *False positive*

General relativity predicts that

(A) light leaving the Sun is slowed by gravity.
(B) light passing the Sun is deflected.
(C) a clock on the Sun's surface runs faster than on Earth.
(D) All of these.
(E) None of the above answers is correct

Original benchmark answer: D
Verified-version answer: D (verified version retains the original)
Frontier-model answers: Claude: B Gemini: B GPT: B

Figure 10: MMLU · conceptual physics. The keyed answer (D, "All of these") asserts that general relativity predicts (A), (B), and (C). But (C) is wrong: general relativity predicts that a clock on the Sun's surface runs *slower* than on Earth (gravitational time dilation), not faster. (A) and (B) are correct, so (B) is the closest defensible single option. All three models shown converge on (B), surfacing the defect.

MMLU · management *False positive*

What does the acronym SMART stand for?

(A) Suitable, measurable, actionable, rewarded and timely
(B) Specific, measurable, actionable, resourced and timely
(C) Standardised, measurable, achievable, rewarded, and timely
(D) Specific, measurable, achievable, rewarded and timely
(E) More than one of the above answers is correct
(F) None of the above answers is correct

Original benchmark answer: D
Verified-version answer: D (verified version retains the original)
Frontier-model answers: Claude: F Gemini: F GPT: F

Figure 11: MMLU · management. The keyed answer is (D), "Specific, measurable, achievable, rewarded and timely." The standard SMART framework is "Specific, Measurable, Achievable, Relevant, Time-bound." None of the four content options matches a recognized version of the acronym, so the correct response is (F), none of the above. All three models shown converge on (F).

MMLU · astronomy *False positive*

Which of the following is/are true?

(A) Titan is the only outer solar system moon with a thick atmosphere
(B) Titan is the only outer solar system moon with evidence for recent geologic activity
(C) Titan's atmosphere is composed mostly of hydrocarbons
(D) A and D
(E) None of the above answers is correct

Original benchmark answer: D
Verified-version answer: D (verified version retains the original)
Frontier-model answers: Claude: A Gemini: A GPT: A

Figure 12: MMLU · astronomy. The keyed answer (D), "A and D," is malformed: option (D) references itself, leaving the choice undefined. (A) is the only correct standalone statement, since Titan is the only outer solar system moon with a thick atmosphere, and all three models shown converge on it.

MMLU · professional_law *False positive*

A husband is charged with murder in the shooting death of his brother-in-law. In the jurisdiction in which the husband is on trial for murder, the defense of insanity is an affirmative defense. The husband pleads insanity as his defense. At trial, in order to prove his defense of insanity, the husband has

(A) the burden of production.
(B) the burden of persuasion only.
(C) both the burden of production and the burden of persuasion.
(D) neither the burden of production nor the burden of persuasion.
(E) None of the above answers is correct

Original benchmark answer: A
Verified-version answer: A (verified version retains the original)
Frontier-model answers: Claude: C Gemini: C GPT: C

Figure 13: MMLU · professional law. Shown with its keyed answer and the responses from the M_3/D_3 panel. Discussed in more detail in this section.

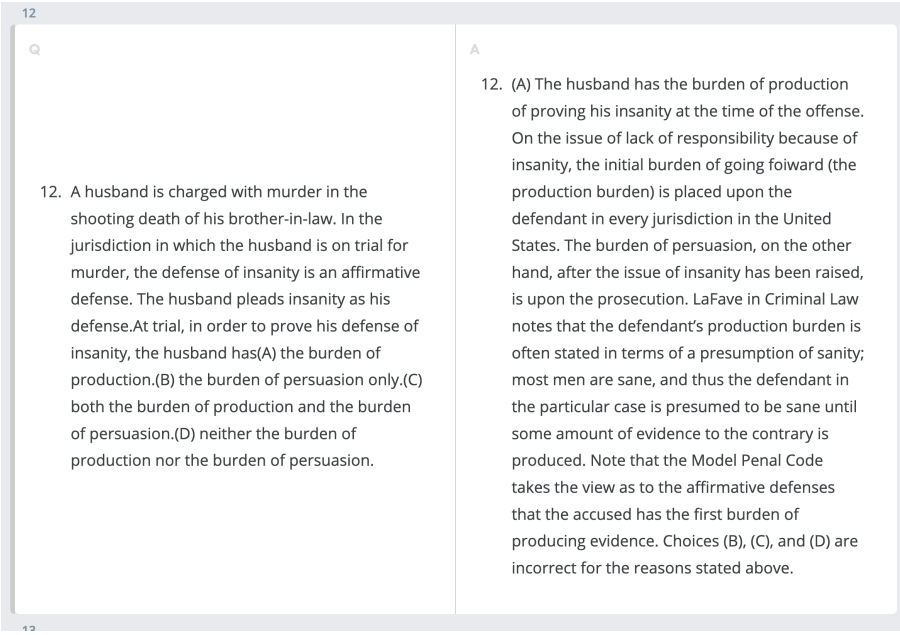


Figure 14: The source of the MMLU professional law item discussed in this section.

Unstated presuppositions. The question implicitly presupposes background facts, including procedural posture, the underlying offense, and the scope of the police action, that materially affect the legal analysis but are not specified.

The expert's conclusion was that the keyed answer is defensible under a narrow reading but that the question as posed admits a different and better-supported answer, finding that "(C) tracks the dominant doctrinal reality, even if (A) tracks the narrower Model Penal Code-style definition." The expert noted that single-source and jurisdictional patterns of this kind are characteristic of legal benchmarks built by ingesting online practice questions without verifying their provenance.

D.2.2 Apparent False Negatives: Well-Formed Items the Verified Reference Flagged

Apparent false negatives are items labeled as defective in the verification process for which M_3/D_3 returned the original benchmark answer. The cases below show several such items. Each carries the kind of minor formatting or rendering issue that motivated the verifier's flag, but in our review the question itself remains answerable and a reasonable test-taker would arrive at the original benchmark answer. We classify these items as well-formed in the supplementary analysis, while acknowledging that the verifiers' decisions to flag them for correction were not unreasonable; the issues they identified are real but minimal and do not affect a model's ability to answer the question as written.

MMLU · global_facts *False negative*

Consider the following statements: (1) In every country, people think other individuals in their country are less happy than they actually are. (2) In the United States, women□s happiness has increased both absolutely and relative to men. Which of these statements is True/False?

(A) True, True
(B) True, False
(C) False, True
(D) False, False

Original benchmark answer: B
Verified-version answer: defective — bad_question_clarity (encoding corruption: "women□s")
Frontier-model answers: Claude: B Gemini: B GPT: B

Figure 15: MMLU · global facts. The verified reference flags the item as defective due to an encoding artifact in the word "women's." The substantive content of the question is unaffected by the corruption: a reader can recover the intended text without ambiguity, and all three models return the original benchmark answer (B).

MMLU · virology *False negative*

Newborn infants in countries with limited access to safe water and low levels of education should be:

(A) Exclusively formula fed for six months
(B) Exclusively breast fed for six months
(C) Should receive both breast milk and other foods as tolerated in the first 6
(D) months
(E) More than one of the above answers is correct
(F) None of the above answers is correct

Original benchmark answer: B
Verified-version answer: defective — bad_options_clarity ("months" split off from option C)
Frontier-model answers: Claude: B Gemini: B GPT: B

Figure 16: MMLU · virology. The verified reference flags the item as defective because option (C) wraps onto two lines, with "months" rendering as a separate option (D). The flag is reasonable, but the intended option set is recoverable on inspection and the keyed answer (B) is unaffected, leaving the item meaningful as a test case under its original label. All three models return (B).

These results show that some apparent errors arise from difficult boundary cases involving ambiguity, multiple defensible answers, contested reference answers, or minor defects that do not prevent the item from being answered. Because the review was conducted by the authors and informed by model outputs, we treat it only as a sensitivity analysis. All headline results and comparisons continue to use the independently published MMLU-Redux and HLE-Verified labels.

MMLU · virology *False negative*

The prevalence of HIV among Latino-Americans compared to other ethnic groups in the U.S. is:

(A) Highest
(B) Higher than all other ethnic groups except African-Americans
(C) Higher than in all other ethnic groups except African-Americans and
(D) Native Americans
(E) More than one of the above answers is correct
(F) None of the above answers is correct

Original benchmark answer: B
Verified-version answer: defective — bad_options_clarity ("Native Americans" split off from C)
Frontier-model answers: Claude: B Gemini: B GPT: B

Figure 17: MMLU · virology. The verified reference flags the item as defective because option (C) wraps and "Native Americans" renders as option (D). As with the previous case, the intended option set is recoverable, the keyed answer (B) is unaffected, and all three models return (B).