HOME-AWAY EVALUATION: A HOME-FIELD FRAMEWORK FOR PAIRWISE LLM EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Static benchmark accuracy summarizes how often a model succeeds, but obscures where its strengths differ from those of another model. We introduce HOME-AWAY EVAL, a pairwise evaluation framework that tests whether comparative strengths transfer to new questions. Each model generates questions from HOME seeds that it answers correctly and its opponent answers incorrectly. Independent referees validate the questions and establish their answers; both players then answer the accepted questions on both fields. Only an exclusive correct answer earns a point. In our controlled benchmark-exposure experiments, fine-tuning improves static accuracy, but the fine-tuned checkpoints lose to their clean bases on the generated fields we evaluate. Training on the questions a comparison decides, by contrast, wins the match while leaving static accuracy slightly improved. An authorship study shows that each model answers its own questions more accurately than other authors' questions, which motivates reciprocal fields. The fields expose domain-dependent strengths, distinguish question posing from answering, and focus comparative evidence on questions the two models answer differently. HOME-AWAY EVAL complements static benchmarks with an explicit account of who challenges whom, on which questions, and with what evidence.

1 INTRODUCTION

Modern LLM evaluation relies heavily on fixed benchmarks, often summarized by aggregate accuracy over a fixed collection of questions (Hendrycks et al., 2021; Wang et al., 2024). Such scores can conceal differences in the questions on which models succeed: two models with similar scores may perform well on different items, and a question easy for one model may be adversarial for another. Fixed test sets are also vulnerable to memorization, benchmark-specific optimization, and source-test leakage. Dynamic or refreshed evaluation (Kiela et al., 2021; Wu et al., 2025; White et al., 2025) does not guarantee independence, since new benchmarks may still inherit source-test templates, paraphrases, or memorized structures. This motivates testing whether performance on exposed benchmark items transfers to newly generated questions.

Moreover, evaluation reflects the choices of its authors, and their habits, such as topics, phrasing, difficulty, and format, can favor evaluated models whose strengths align with them. When a single computational model writes the questions for every answerer, the bias can remain implicit and unequally distributed, so part of a measured performance gap between models may reflect the generator's affinity as well as the answerers' ability. This motivates giving both evaluated models a role in constructing the questions, with each drawing on its own comparative strengths. Conventional benchmarks also do not ask whether a model can generate valid, answerable questions from a region of demonstrated competence. We treat question generation as a probe of knowledge use, following educational research on question generation (King, 1992; Rosenshine et al., 1996; Yang et al., 2025).

We propose HOME-AWAY EVAL, a pairwise evaluation framework built on home fields: Given two models, we identify seed examples that model A answers correctly and model B incorrectly, and vice versa; these examples form each model's HOME seed pool, representing its observed comparative strengths. The HOME model generates new questions from few-shot examples drawn from its own pool. The AWAY model influences pool selection, but its identity, behavior, and failure rationale are not provided to the generator. Both models answer every generated candidate accepted by independent referee models that agree on its validity and answer. An exclusive correct answer earns

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.

one point for that model, including an Away point when the comparator alone solves a Home-field question; both correct and both incorrect earn none. Scores across reciprocal fields (i.e., Home-field and Away-field) decide the match.

In this paper, we make the following contributions:

- We introduce Home-Away Eval, a pairwise framework in which both evaluated models generate questions from their observed comparative strengths, with independent referees validating the questions and establishing their answers.
- We evaluate this framework through contamination and authorship experiments. Contaminated checkpoints improve static accuracy on their training benchmarks yet lose to their clean bases on generated fields. Models answer their own questions more accurately than other authors' questions, motivating reciprocal participation in constructing the evaluation.
- We examine how field construction, question posing, referee agreement, and generation budgets shape model comparisons, distinguishing posing from answering and characterizing the evidence behind domain-specific rankings. We also demonstrate that decisive questions can provide useful training data: fine-tuning improves performance against the clean base on newly generated fields.

2 Related Work

Static, Dynamic, and Contamination-Aware Benchmarks. General-purpose LLM evaluation is organized around fixed datasets, whose aggregate scores can obscure model-specific competence regions and are vulnerable to saturation, benchmark-specific optimization, and contamination. Dynamic or contamination-aware evaluation, including adversarial or model-in-the-loop collection, frequently refreshed benchmarks, and benchmarks built from updated real-world knowledge (Kiela et al., 2021; White et al., 2025; Wu et al., 2025), reduces dependence on stale test sets, but newly collected items can still inherit source-test templates, paraphrases, or memorized structures. Home-Away Eval addresses this risk to some extent by having the evaluated models generate candidates from demonstrated-competence seeds and by auditing the validity, novelty, and non-paraphrase status of every accepted question.

Pairwise and Examiner-Based LLM Evaluation. Chatbot Arena (Chiang et al., 2024) ranks models through human-judged head-to-head comparisons, MT-Bench and related LLM-as-judge work use strong models as automated evaluators (Zheng et al., 2023), examiner models generate prompts or adaptive evaluation sessions (Zhao et al., 2025; Bai et al., 2023; Xu et al., 2025). A fixed benchmark or a single examiner supplies a common question distribution whose authorship may favor some evaluated models. Home-Away Eval makes authorship bias explicit and reciprocal: each player model generates questions from its demonstrated comparative strengths, giving both models an opportunity to construct challenges that favor them. Both models answer both fields, and the match combines their exclusive correct answers across the two authors' fields. The distinction is reciprocal participation in authoring the evaluation, rather than the elimination of authorship bias.

Problem Posing and Duel-Style Evaluation. Dr.Academy (Chen et al., 2024) evaluates the quality of LLM-generated questions in teaching-oriented settings, and AutoBencher (LI et al., 2025) constructs datasets declaratively by optimizing salience, novelty, and difficulty objectives. The closest prior work is MathDuels (Xu et al., 2026), in which models generate mathematics problems from a domain specification, other models solve them under a conditional verifier, and a Rasch model rates solving and posing separately. Home-Away Eval differs in two respects: generation is conditioned on the pair's own disagreement items rather than on a domain and a difficulty objective, and each accepted question is posed for one comparison and answered by both models. Home-Away Eval covers both multiple-choice question answering (MCQA) and mathematics.

3 Method

Figure 1 illustrates how Home-Away Eval compares two models through reciprocal question fields. Benchmark disagreements determine the seeds, each player generates questions from its own seeds,

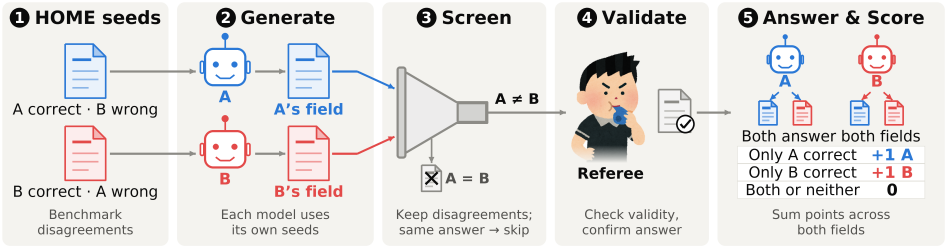

Figure 1: An illustration of Home-Away Eval. Each model generates from its Home seeds. After screening and referee validation, both players answer both fields; only an exclusive correct answer earns a point; both-correct and both-incorrect outcomes earn none.

and independent referees assess questions for validity and answerability to establish which can be scored. Both players then answer the accepted questions to determine the match outcome.

3.1 Problem Setup

Let M_A and M_B be the player models and $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ a seed benchmark or compatible benchmark pool. Evaluating both players under a common correctness procedure gives

$$c_A(i) = \mathbf{1}\{M_A(x_i) \equiv y_i\}, \qquad c_B(i) = \mathbf{1}\{M_B(x_i) \equiv y_i\},$$

where model outputs denote extracted answers and $\equiv$ denotes task-specific answer equivalence. In our experiments, seed selection additionally requires stable correctness across repeated evaluations (Section 4). The source items condition generation and never contribute to the final match score.

3.2 Constructing Home Seed Pools

For each player model, Home seeds are the items the player solves and its opponent misses:

$$\mathcal{H}_A = \{i : c_A(i) = 1,\ c_B(i) = 0\}, \qquad \mathcal{H}_B = \{i : c_B(i) = 1,\ c_A(i) = 0\}.$$

These pools identify observed comparative strengths within the pair. Questions generated by M_A form its Home field and M_B's Away field; the roles reverse for questions generated by M_B. Each field is specific to its pair, so adding a model (e.g., to extend the comparison to three models) requires new comparisons involving that model while leaving existing fields fixed.

3.3 Question Generation

Each player generates candidates from few-shot examples drawn from its own Home pool so that its field reflects its comparative strengths, using a shared prompt format. The prompt provides the model with the examples and a task cue, without the opponent's identity, responses, or failure explanations. The opponent thus only affects which seeds are selected. Answers produced during question generation are discarded because they could be incorrect; the answer is determined during validation.

3.4 Screening and Validation

Both players first answer each question candidate in a screening pass. If both answers can be extracted and agree after normalization, the candidate is masked to save referee calls[1]. The framework supports a single referee, two referees, or more. Each referee receives the source seed and generated question, checks that the question is well-formed, answerable, novel relative to the seed, and not a superficial paraphrase, and proposes an answer.

In our dual-referee variation, a candidate is accepted only when both referees approve it and their answers agree: the same option label for multiple choice or equivalent final answers for mathematics. Consequently, reported match outcomes are conditional on disagreement at screening and on acceptance by both referees with agreeing answers.

[1]We acknowledge that screening before validation makes the workflow less intuitive. However, this ordering significantly reduces cost because referee calls are generally more expensive than player calls.

3.5 Answering and Scoring

After validation, both players answer each accepted question afresh, and their answers are scored against the answer agreed upon by the referees. Finally, only an exclusive correct answer earns a point, and we call a question with an exclusive correct answer *decisive*. Both-correct and both-wrong outcomes earn none. A player's points on its own field are Home points; those on its opponent's field are Away points.

It is important to note that accepted field sizes are not equalized and the points are not normalized. Producing more accepted questions creates more scoring opportunities for both players, so a model benefits from posing only when it alone answers correctly, which implicitly incorporates question-generation ability into the evaluation. The player with more total points wins.

4 Experiment Setup

Models and Benchmarks. We evaluate models spanning families, versions, and sizes: Qwen3.6-35B-A3B (Qwen Team, 2026b), Qwen3.5-35B-A3B and Qwen3.5-9B, 4B, 2B (Qwen Team, 2026a), Qwen3-8B (Qwen Team, 2025), Gemma-4-26B-A4B-it, Gemma-4-E4B-it (Google DeepMind, 2026), and GPT-OSS-20B (OpenAI, 2025). The main comparison uses 6 models: Qwen3.6-35B-A3B, Qwen3.5-35B-A3B, Qwen3.5-9B, Qwen3-8B, Gemma-4-26B-A4B-it, and GPT-OSS-20B. The contamination experiments fine-tune Qwen3.5-9B and Gemma-4-E4B-it, and the scaling experiment adds Qwen3.5-4B and Qwen3.5-2B. We construct fields from 5 benchmarks: MMLU (Hendrycks et al., 2021), MMLU-Pro (Wang et al., 2024), and GPQA-Main (Rein et al., 2023) for multiple-choice QA, and GSM8K (Cobbe et al., 2021) and MATH-500 (Lightman et al., 2024) for free-response mathematics.

Two pooled tasks test how comparisons change when generation draws on broader seed pools. Global-MC draws seeds from MMLU and GPQA-Main; Global-Math draws from GSM8K and MATH-500. Each pool caps every source at 200 Home seeds for each model in each pair, limiting each source's contribution while allowing a scarce source to contribute all of its seeds. MMLU-Pro is excluded from the pools because its option count differs. Other benchmarks, such as ARC-Challenge (Clark et al., 2018), are excluded since their disagreements are too sparse either to form separate fields or to contribute to a pool. Dataset splits and disjoint demonstration pools are detailed in Appendix A.

Seed Selection and Generation. We select Home seeds using answering with thinking mode enabled and three-response (K=3) self-consistency (Wang et al., 2022) to reduce reliance on a single sampled answer. A seed enters a model's Home pool only when all three of that model's responses are correct and the opponent's majority answer is incorrect. This strict rule selects seeds with stable pairwise correctness on the source benchmark. It does not establish that the advantage transfers to generated questions. Generation uses 8-shot prompts without thinking mode.

Validation and Answering. A question is accepted only when both referees approve it and agree on its answer. We use GPT-5.6 Terra and Gemini-3.6-Flash[2] as independent referees, both at low reasoning effort. We choose low effort because medium effort improves referee agreement but yields nearly the same accepted volume (Table 7 in the Appendix). We enable thinking mode for answering, allowing up to 24,576 reasoning tokens unless otherwise specified. Generated-field mathematics answers match when their numeric values agree within 10^{-6} and otherwise by normalized string comparison; when scoring, trailing unit words in the reference answer are ignored (e.g., "45 cups" matches 45).

The stable-correctness rule restricts eligible source seeds, while dual-referee agreement determines which generated questions are retained. Prompt templates and details, including sampling parameters, are provided in Appendices A and C.

[2] `https://api.openai.com/`, `https://generativelanguage.googleapis.com`

Table 1: Static accuracy and contamination matches. Points list the clean base first and contaminated checkpoint second. Scoreable counts referee-accepted questions.

	Static Accuracy					vs. Contaminated	
	Multiple Choice			Mathematics			
Model	**MMLU**	**M-Pro**	**GPQA**	**GSM8K**	**MATH**	**Points**	**Scoreable**
Qwen3.5-9B	87.9	80.7	63.5	93.8	83.0	–	–
+ MCQA contam.	92.3	85.9	64.7	–	–	**66**–40	320
+ Math contam.	–	–	–	96.3	87.1	**21**–2	257
+ All contam.	87.9	82.1	66.8	96.1	90.9	**91**–64	399
Gemma-4-E4B-it	77.1	67.4	46.3	88.7	85.1	–	–
+ MCQA contam.	96.9	90.6	52.3	–	–	**204**–79	379
+ Math contam.	–	–	–	94.7	91.7	**45**–11	252
+ All contam.	95.2	86.2	50.7	93.0	91.7	**163**–72	267

Table 2: Qwen3.5 scaling under contamination. Static accuracy is in percent; Static ranks the equal-weight mean across all benchmarks. H-A ranks aggregate MCQA matches among the 5 models. Losses of 35B are in GPQA.

	Multiple Choice			Mathematics		Joint Rank		
Model	**MMLU**	**M-Pro**	**GPQA**	**GSM8K**	**MATH**	**Static (Avg.)**	**H-A**	**H-A W/T/L**
Qwen3.5-9B (contam-all)	87.9	82.1	66.8	96.1	90.9	1	4	5/1/6
Qwen3.5-35B-A3B	90.7	83.8	67.3	95.6	84.5	2	1	10/0/2
Qwen3.5-9B	87.9	80.7	63.5	93.8	83.0	3	2	8/1/3
Qwen3.5-4B	85.2	79.1	57.5	93.2	77.8	4	3	6/0/6
Qwen3.5-2B	72.4	59.9	39.3	81.6	55.4	5	5	0/0/12

5 Results

5.1 Contamination Stress Test

Design. We fine-tune Qwen3.5-9B and Gemma-4-E4B-it on benchmark test items under three conditions: MCQA (MMLU, MMLU-Pro, and GPQA-Main), mathematics (GSM8K and MATH-500), and all 5 benchmarks. We compare each contaminated checkpoint with its clean base on static accuracy and in a Home-Away match (Table 1). Fine-tuning uses answer-conditioned reasoning traces to retain the reasoning format during training; implementation details appear in Appendix A.

The scaling experiment jointly ranks clean Qwen3.5 bases from 2B to 35B and the all-benchmark-contaminated 9B checkpoint (Table 2). It tests whether contamination lets a smaller model outrank larger clean bases on static benchmarks, and whether Home-Away gives the same ordering.

Results. In Table 1, all 6 contaminated checkpoints improve average static accuracy on their training benchmarks, yet lose to their clean bases on Home-Away fields. For example, mathematics contamination raises Qwen3.5-9B's GSM8K accuracy from 93.8 to 96.3 and MATH accuracy from 83.0 to 87.1, but the checkpoint loses 21–2 on 257 accepted questions. These results show that higher accuracy on exposed benchmark items need not transfer to newly generated questions.

Contamination also changes the apparent ordering across model sizes, as presented in Table 2. The contaminated Qwen3.5-9B ranks 1st in average static accuracy, ahead of the clean 35B model, but 4th under Home-Away, below the clean 35B, 9B, and 4B models. Thus, its benchmark improvement does not translate into a higher position in the generated-field comparison.

5.2 Authorship and Reciprocal Evaluation

Design. We compare Qwen3.6-35B-A3B, Gemma-4-26B-A4B-it, and GPT-OSS-20B as question authors and answerers. In the shared-seed control, each model generates three candidates per MMLU test item (42,126 per model from 14,042 items). All three models answer each question candidate for screening; only candidates on which their answers are not unanimous are eligible for referee validation. After validation, all three models answer every retained field afresh for scoring. We compare each model's accuracy on its own questions with its mean accuracy on the other two authors' fields. Δ is the difference between own-question accuracy and the other-question accuracy. Under

Table 3: Accuracy on own questions and the other authors' questions, averaged equally across their two fields. HOME-AWAY reports observed wins–losses and points scored–conceded.

	Shared Question			HOME-AWAY	
Model	**Own**	**Others'**	**Δ**	**W–L**	**Total Points**
Qwen3.6-35B-A3B	72.8	66.5	+6.3	1–1	46–16
Gemma-4-26B-A4B-it	81.6	66.6	**+14.9**	**2–0**	31–16
GPT-OSS-20B	51.7	50.3	+1.4	0–2	14–59

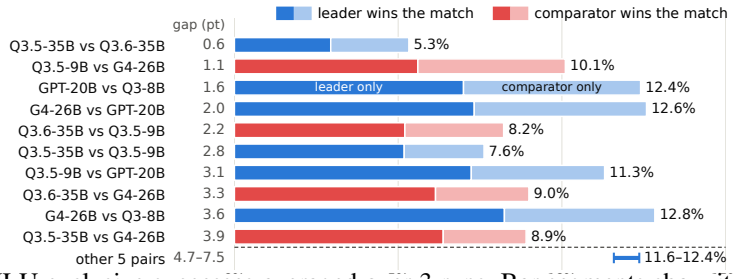

Figure 2: MMLU exclusive successes averaged over 3 runs. Bar segments show items solved only by the static leader or its comparator; end labels give their sum. Gap is the static accuracy difference in points. Red marks observed HOME-AWAY reversals. The other 5 pairs are summarized.

the HOME-AWAY condition, each model generates from its own pair-disagreement seeds. The final columns report each model's win–loss record and points scored–conceded over its two matches.

Results. In Table 3, all three models achieve higher accuracy on their own questions than on the other authors' questions, with differences ranging from 1.4 to 14.9 percentage points. Gemma-4 shows the largest difference. These comparisons reflect performance across authored question distributions, whose difficulty also varies. Under HOME-AWAY, Gemma-4 wins both matches, Qwen3.6 splits its matches, and GPT-OSS loses both, although Gemma-4's win over Qwen3.6 has a margin of only two points on 26 accepted questions. Reciprocal evaluation gives both models an opportunity to construct favorable questions and combines their performance across both fields, making authorship an explicit part of the comparison.

5.3 PAIRWISE DIFFERENCES AND MATCH RESOLUTION

Design. We compare item-level MMLU correctness across three static runs with HOME-AWAY results, then examine whether additional generation resolves comparisons with few decisive questions. The budget study uses Qwen3.5-35B-A3B vs. Qwen3.6-35B-A3B, Gemma-4-26B-A4B-it vs. GPT-OSS-20B, and Qwen3.5-35B-A3B vs. Qwen3-8B, which represent small (0.6), intermediate (2.0), and large (7.5) gaps under MMLU scores, respectively. The models generate up to 10k, 8k, and 4k questions from their MMLU HOME seeds, respectively.

Results. Figure 2 shows why similar accuracy need not imply similar strengths. The closest MMLU pair differs by only 0.6pt, yet the models disagree in correctness on 5.3% of items. Across all 15 pairs of the 6 main models (Section 4), this disagreement ranges from 5.3% to 12.8%. Each bar separates the models' exclusive successes, whereas the accuracy gap retains only their difference.

Figure 3 shows how additional generation resolves these comparisons: in the three pairs, the sampled match trajectories become decisive as more candidates are generated. With few generated questions, only about 10 are accepted. With more generation, the static leader pulls away in every pair. Qwen3.5-35B leads Qwen3-8B 44–2 after 4k generated questions. Gemma-4 trails GPT-OSS-20B briefly at the start but leads 99–55 after 8k. Even the closest pair, 0.6 accuracy points apart, separates to 17–5 after 10k generated questions. The larger the static gap, the fewer generated questions a match needs to become decisive.

5.4 DOMAIN RANKINGS AND EVIDENCE PER QUESTION

Design. We compare mean static accuracy, aggregated benchmark fields, and a newly generated pooled field in each domain. Aggregated rankings combine matches conducted separately on each

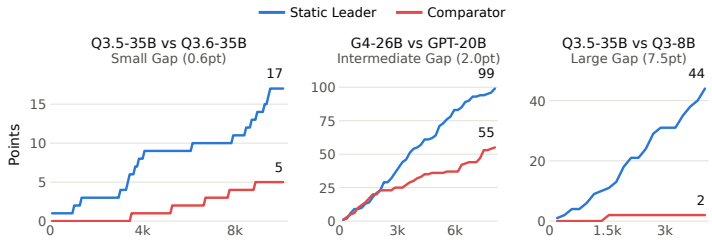

Figure 3: MMLU match resolution by the number of generated questions for the three pairs in Figure 2. Lines show the total points of the static leader and its comparator on the accepted questions among the generated questions, counted over both fields.

Table 4: Domain ranks and mean static accuracy; Aggr. combines benchmark matches; Global uses a pooled field. We also report Kendall's τ_b. Rank correlations are descriptive across 6 models.

	Multiple Choice						Mathematics				
	Static				Home-Away		**Static**			Home-Away	
Model	MMLU	M-Pro	GPQA	Avg.	Aggr.	Global	GSM8K	MATH	Avg.	Aggr.	Global
Qwen3.5-35B-A3B	**1**	2	**1**	**1** (80.6)	**1**	**1**	2	**1**	**1** (90.1)	**1**	3
Qwen3.6-35B-A3B	2	**1**	2	2 (79.8)	3	2	**1**	2	2 (89.9)	2	2
Qwen3.5-9B	3	4	3	3 (77.4)	4	3	4	4	4 (88.4)	3	4
Gemma-4-26B-A4B-it	4	3	4	4 (75.8)	2	4	2	3	3 (89.6)	4	**1**
GPT-OSS-20B	5	5	5	5 (72.0)	5	5	5	5	5 (81.5)	6	5
Qwen3-8B	6	6	6	6 (68.8)	6	6	6	6	6 (68.4)	5	6
τ vs Aggr.	.733	.733	.733	.733	–	–	.552	.733	.733	–	–
τ vs Global	1.000	.733	1.000	1.000	–	–	.828	.600	.600	–	–

benchmark; pooled rankings use new questions generated from combined seed pools with at most 200 seeds per source. Static accuracy weights each benchmark equally within MCQA or mathematics. Match-based rankings give 1 point per win and 0.5 per tie, with total question points breaking remaining ties. Per-benchmark static accuracies are reported in Table 10, pairwise win/tie/loss outcomes in Figure 4, and per-task decisive and saturated shares in Table 11, all in the Appendix.

Results. Table 4 reports that generated MCQA fields preserve the highest and lowest positions, while aggregated fields change the middle of the ranking. Qwen3.5-35B ranks 1st under all three measures, while GPT-OSS-20B and Qwen3-8B remain 5th and 6th. Gemma-4 rises from 4th under static accuracy to 2nd on aggregated fields but stays 4th on the pooled field. Applying match scoring to the static items reproduces their accuracy ranking, so the observed changes are not explained by the scoring rule alone. Agreement with static accuracy is τ_b=.733 for aggregated fields and 1.000 for the pooled field; the lower agreement of aggregated fields indicates that Home-Away is a complementary evaluation framework.

Mathematics rankings depend more strongly on field construction. Qwen3.5-35B ranks 1st on aggregated fields, whereas Gemma-4 rises from 4th on aggregated fields to 1st on the pooled field. Aggregated fields have τ_b=.733 with static accuracy and the pooled field .600. Combining seed sources can therefore produce a different ordering from combining separate benchmark matches.

5.5 Posing and Answering

Design. We compare accepted questions by author, Home-Away points, and own-field and opposing-field accuracy on the 5 direct tasks. Posing is measured by the acceptance rate, the percentage of generated candidates retained after disagreement screening and referee validation, on the same basis in both domains. This rate reflects the full generation and selection procedure. Pearson correlations describe associations across the 6 evaluated models.

Results. In Table 13 in the Appendix, acceptance rate correlates positively with total points on MCQA (r=+.69), but shows little association on mathematics (r=+.13). Own-field and opposing-field accuracy correlate positively in both domains (r=+.98 and +.89), indicating that models answering more accurately on their own fields also tend to answer more accurately on opposing fields. Per-task Home and Away points are in Table 12 in the Appendix.

Table 5: Decisive-question examples. Category is the subject label both classifiers agree on, with the seed's main category in parentheses. Gold is the answer both referees agree on. Answers lists both players' final answers, the field's author first. More examples are presented in Table 9.

Category	Question	Gold	Answers (HOME first)
Philosophy (moral scenarios)	Which of the following best characterizes Susan Wolf's specific objection to the concept of a moral saint? (A) … lack the capacity to experience genuine joy or happiness. (B) … would likely be unattractive because the pursuit of moral perfection would crowd out other valuable human pursuits and personality traits. (C) … impossible because human nature is inherently selfish … (D) … dangerous because they would inevitably try to impose their moral views on others by force.	(B)	**Qwen3.5-35B**: (B) Qwen3-8B: (A)
Geometry (algebra)	In triangle $\triangle ABC$, we are given that $AB = AC = 25$ and $BC = 14$. Find the length of the median $\overline{AM}$, where M is the midpoint of $\overline{BC}$.	24	Qwen3-8B: $\sqrt{551}$ **Gemma-4**: 24
Grade-school word problem	A bakery sells cupcakes for $3.00 each and muffins for $2.50 each. If a customer buys 4 cupcakes and 6 muffins, and pays with a $50 bill, how much change should they receive?	$23.00	Qwen3.6-35B: 27.00 **Gemma-4**: $23.00

High acceptance and answering accuracy do not necessarily produce many points. Qwen3.5-35B has the highest mathematics acceptance rate, yet earns 74 HOME points, compared with 133 for Qwen3.6-35B, whose rate is about half: Qwen3.5-35B generates 414 candidates and has 104 accepted questions, whereas Qwen3.6-35B generates 1,662 and has 230. AWAY points likewise depend on both the available questions and exclusive correctness. Qwen3-8B earns 105 MCQA AWAY points across 487 opposing questions, the most after Gemma-4 (133), despite having the lowest opposing-field accuracy. Field size, accuracy, and exclusive-correct points should be considered together.

6 FURTHER ANALYSIS

6.1 REFEREE ABLATION

Design. We ablate the referee requirements, which accept a question only when GPT-5.6 Terra and Gemini-3.6-Flash both approve it and agree on its answer. Using the same screened candidates, we compare this rule with acceptance by either referee alone and with joint acceptance without requiring answer agreement. These comparisons examine how referee choice and answer agreement affect question retention.

Results. Relaxing the main acceptance rule increases question retention across all tasks (see Table 8 in the Appendix). Gemini alone accepts more generated MCQA candidates than GPT-5.6, showing that a single referee's judgments substantially affect the retained question distribution. Requiring both referees to approve a question reduces retention, but does not ensure that they agree on its answer. Answer disagreement is more frequent in mathematics: among jointly accepted questions, the referees disagree on 64.5% of GSM8K answers, 38.3% of MATH answers, and 52.0% of Global-Math answers, compared with 14.3–18.6% across MCQA tasks. On GSM8K, removing the answer-agreement requirement increases retention from 2.8% to 7.9% of generated candidates.

Three coauthors each audited 455 accepted questions, about 100 per benchmark and 75 for GPQA-Main. Human–referee agreement on answer correctness was 90.0% overall (1,229/1,365), 100% on MCQA (825/825), and 74.8% on mathematics (404/540). Most mathematics mismatches were due to answer formatting.

6.2 WHAT THE DECISIVE QUESTIONS ASK

Design. We classify the subjects of the finalized decisive Global-MC and Global-Math questions with GPT-5.6 Terra and Gemini-3.6-Flash, both at low reasoning effort (the same setting as the referees) and constrained to the category list. We retain only questions on which the classifiers agree exactly. For MCQA, agreement is checked on MMLU's 57 subject labels before grouping them into humanities, STEM, social sciences, and other subjects; mathematics uses MATH-500 topics and grade-school word problems. The classification prompt is presented in Appendix C.3.

Results. Among decisive questions with agreed labels, humanities accounts for 59.6% of MCQA questions, followed by STEM (22.2%) and other subjects (15.2%). In mathematics, geometry is the largest category (36.8%), followed by grade-school word problems (28.9%), algebra (13.2%),

Table 6: Static accuracy (%) and the HOME-AWAY match for decisive-question training. Points list the clean base first and the fine-tuned model second; Scoreable counts referee-accepted questions.

	Multiple Choice			Mathematics			
Model	**MMLU**	**M-Pro**	**GPQA**	**GSM8K**	**MATH**	**Points**	**Scoreable**
Gemma-4-E4B-it	77.1	67.4	46.3	88.7	85.1	78–**134**	342
+ decisive	79.1	69.3	46.0	88.6	85.1	–	–

and number theory (7.9%). These percentages describe the subject distributions observed in our experiments and may vary with the sampled seed pools. Table 5 and Appendix D show examples.

Generated questions also differ in category from their source seeds. Among agreed questions with comparable seed labels, 45.1% of MCQA questions and 60.5% of mathematics questions change category. Conditioning generation on a model's comparative-strength seeds therefore does not necessarily preserve their subject categories.

6.3 TRAINING ON DECISIVE QUESTIONS

Design. We fine-tune Gemma-4-E4B-it on decisive questions derived from comparisons among the 6 main models, using the same training recipe as the contamination experiments. After the rationale generation, the training set contains 2,512 questions: 1,981 multiple-choice and 531 mathematics from Global-MC and Global-Math. We compare the fine-tuned model with its clean base on static benchmarks and in HOME-AWAY. The evaluation field uses newly generated questions from seed pools, with no exact duplicates of training questions. Also, the source seeds are not reused.

Results. The decisive-question fine-tune wins over its clean base 134–78 on 342 accepted questions, as presented in Table 6. MMLU and MMLU-Pro improve, GPQA and GSM8K remain almost the same. This supports using HOME-AWAY comparative outcomes to construct training data.

7 DISCUSSION

HOME-AWAY EVAL examines whether comparative strengths on benchmark items persist on questions generated by both models. All contaminated checkpoints improve average static accuracy on their training benchmarks yet lose to their clean bases on generated fields, revealing that benchmark gains alone do not establish comparative improvement on new questions. Matches between models with similar accuracy need more generated questions to become decisive, but even a 0.6-point pair separates with enough generation.

Reciprocal fields make authorship explicit and show how models perform on their own and their opponents' questions. The posing analysis and referee ablation show that acceptance, answer agreement, and exclusive-correct points capture distinct aspects of a match. Generated questions also often change subject from their seeds, so a field is not confined to its seeds' subjects. Decisive questions can improves performance against the clean base on newly generated fields, with small gains.

These findings are limited to the evaluated models, fields, and one decisive-question fine-tuning experiment. Mathematics results also depend more on answer checking. Also, the match tables do not report variation across independent field-generation runs because of computational cost.

8 CONCLUSION

We introduced HOME-AWAY EVAL, which tests whether comparative strengths on benchmark items transfer to questions constructed by both evaluated models. Reciprocal fields make authorship explicit, distinguish posing from answering, and reveal domain-dependent strengths beyond an aggregate score. Our experiments show that benchmark gains need not transfer to generated fields, while training on decisive questions can improve performance on newly generated fields, with small static gains. HOME-AWAY EVAL complements static evaluation with model-dependent comparisons that show who challenges whom and on which questions.

AI use statement

In this work, we used generative AI tools for English-language editing, draft-level plot-generation code, and consistency checks of this paper. Also, the proposed framework relies on newly generated questions from AI models. We have not used generative AI tools for other tasks. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Reproducibility statement

We describe all the experiment details, including input prompts and sampling hyperparameters, in the Appendix.

References

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. Benchmarking foundation models with language-model-as-an-examiner. *Advances in Neural Information Processing Systems*, 36:78142–78167, 2023.

Yuyan Chen, Chenwei Wu, Songzhou Yan, Panjun Liu, and Yanghua Xiao. Dr. academy: A benchmark for evaluating questioning capability in education for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3138–3167, 2024.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

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

Google DeepMind. Gemma 4: Our most capable open models to date. `https://deepmind.google/models/gemma/gemma-4/`, 2026. Accessed: 2026-05-21.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. Dynabench: Rethinking benchmarking in nlp. In *Proceedings of the 2021 conference of the North American chapter of the Association for Computational Linguistics: human language technologies*, pp. 4110–4124, 2021.

Alison King. Facilitating elaborative learning through guided student-generated questioning. *Educational psychologist*, 27(1):111–126, 1992.

XIANG LI, Farzaan Kaiyom, Evan Liu, Yifan Mai, Percy Liang, and Tatsunori Hashimoto. Autobencher: Towards declarative benchmark construction. In *International Conference on Learning Representations*, volume 2025, pp. 94357–94386, 2025.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025. URL `https://arxiv.org/abs/2508.10925`.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026b. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

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

Barak Rosenshine, Carla Meister, and Saul Chapman. Teaching students to generate questions: A review of the intervention studies. *Review of educational research*, 66(2):181–221, 1996.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, et al. Livebench: A challenging, contamination-limited llm benchmark. In *International Conference on Learning Representations*, volume 2025, pp. 91595–91631, 2025.

Xiaobao Wu, Liangming Pan, Yuxi Xie, Ruiwen Zhou, Shuai Zhao, Yubo Ma, Mingzhe Du, Rui Mao, Luu Anh Tuan, and William Yang Wang. Antileakbench: Preventing data contamination by automatically constructing benchmarks with updated real-world knowledge. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18403–18419, 2025.

Shengwei Xu, Yuxuan Lu, Grant Schoenebeck, and Yuqing Kong. Benchmarking llms' judgments with no gold standard. In *International Conference on Learning Representations*, volume 2025, pp. 67830–67862, 2025.

Zhiqiu Xu, Shibo Jin, Shreya Arya, and Mayur Naik. Mathduels: Evaluating llms as problem posers and solvers. *arXiv preprint arXiv:2604.21916*, 2026.

Ji Won Yang, Kiroong Choe, Jaeseo Lim, and Jooyong Park. Asking questions matters: comparing the effect of learning through dialogue with peers, chatgpt, and intelligent tutoring system. *Interactive Learning Environments*, 33(10):5870–5886, 2025.

Ruochen Zhao, Wenxuan Zhang, Yew Ken Chia, Weiwen Xu, Deli Zhao, and Lidong Bing. Autoarena: Automating llm evaluations with agent peer battles and committee discussions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4440–4463, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

A Experiment Details

Datasets and Splits. We use 14,042 MMLU test items, 12,032 MMLU-Pro items, 1,319 GSM8K items, and MATH-500. Few-shot examples come from disjoint dev, validation, or training splits. For MATH, removing fixed demonstrations by identifier leaves 495 evaluated items. GPQA-Main's 448 items are deterministically split into 20 demonstrations and 428 evaluated items; option order is shuffled using an item-specific seed.

Sampling Parameters. Seed answering uses 5 examples and K=3 responses, following the best practice parameters of each model (e.g., temperature 0.7 and top-p 0.95). In the main comparison, generation samples N=2 candidates per seed, using 8 examples; here N is the number of questions generated from each seed, not the benchmark size of Section 3.1. Demonstrations are fixed per benchmark and disjoint from evaluated items. Standard answering enables thinking with a 24,576-token reasoning budget followed by extraction.

Referee. GPT-5.6 Terra and Gemini-3.6-Flash judge candidates independently at low reasoning effort with structured outputs capped at 768 tokens.

Mathematics Answer Normalization. A player's final answer is read from the extraction pass, which returns a single **#### x** line; when that pass is truncated or malformed, the answer is taken from the reasoning trace instead (the last **\boxed{}**, **####** line, or "the answer is" statement). Answers are normalized by stripping LaTeX delimiters and wrappers (**$**, **\left**, **\right**, **\text{}**), rewriting **\frac{a}{b}** as a/b, and removing thousands separators, currency signs, and common units. Two answers are equivalent when both parse as numbers, including fractions and percentages, that agree within 10^{-6}; otherwise their normalized strings must match after case folding and whitespace removal. A reference answer carrying units or words (e.g., "45 cups") also matches a prediction with the same numeric value, whereas algebraic answers are compared without this relaxation. The referees' answers are compared with the same rule, without the unit relaxation.

Referee Cost. Accounting uses input/output rates per million tokens of \$1.00/\$6.00 for GPT-5.6 Terra and \$0.375/\$1.875 for Gemini-3.6-Flash. Reasoning comprises 83% and 90% of the referees' output tokens.

Contamination Training. The prompt supplies each benchmark test item and its reference answer, and the model generates a rationale prefixed by "Explanation:". Rationales that contradict the reference answer (e.g., conclude "B" when the answer is "A") or explicitly mention the supplied answer (e.g., "since the provided answer is A") are resampled until neither occurs. Trace generation is capped at 16,384 tokens. We use the rationale as the reasoning trace and train for one epoch with full fine-tuning, a learning rate of 10^{-5}, and a batch size of 128. Training-set size varies with the contamination condition. Both model sizes can be fine-tuned on a single A100 node.

B Referee Ablation

Table 7: Low vs. medium reasoning effort of referee. We compare low and medium effort on a fixed 500-item sample spanning 34 benchmark and model-pair combinations; rates use the 479 items with valid outputs in both conditions. Rates are percentages. The sample over-represents prior referee disagreements.

Outcome	Low Effort	Medium Effort	
Accept/reject agreement	70.1	77.5	+7.3
Both accept	40.3	39.2	−1.0
Both reject	29.9	38.2	+8.4
Both accept with agreeing answers	38.4	38.8	+0.4
Answer agreement given both accept	95.3	98.9	+3.6

Table 8: Referee ablation. Acceptance rates use all generated candidates. Both & Agree is the experiment policy. Validity disagreement is measured among questions accepted by at least one referee; answer disagreement is measured among questions accepted by both.

	MMLU	M-Pro	GPQA	G-MC	GSM8K	MATH	G-Math
GPT-5.6 only	2.0	1.4	2.9	1.8	9.7	35.0	15.3
Gemini only	5.2	4.5	10.4	5.8	12.1	34.0	16.9
Both	1.8	1.1	2.7	1.6	7.9	30.4	13.0
Both & Agree (main)	1.5	0.9	2.2	1.3	2.8	18.8	6.3
Validity disagreement	67.2	**77.3**	75.0	74.1	43.1	21.1	32.1
Answer disagreement	14.3	17.3	15.4	18.6	**64.5**	38.3	52.0

C Prompt Templates

C.1 Question-Generation Prompts

Multiple-choice QA.

```
Generate one new multiple-choice question with its answer,
following the examples.

Question: <reference seed 1, including labeled options>
Answer: (<gold option label 1>)

...

Question: <current source seed, including labeled options>
Answer: (<current gold option label>)

Question:
```

The answer continuation encourages complete question formatting. Generation stops at answer or explanation delimiters; any generated answer is discarded. Referees establish the answer used for scoring.

Free-response mathematics.

```
Generate one new math problem, following the examples.

Question: <reference seed 1>

...

Question: <current source seed>

Question:
```

C.2 Referee Prompts

Multiple-choice QA.

```
Evaluate the generated question against the seed.

Accept it only if all conditions hold:
- It is answerable, unambiguous, self-contained, and well formed.
- It uses exactly these option labels: <allowed parenthesized labels>.
- It has exactly one correct option.
- It is not a superficial rewrite of the seed.
```

```
- It uses new concrete content that cannot be solved by
copying seed-specific facts.

Seed question:
<source seed the question generated>

Generated question:
<generated candidate>
```

Free-response mathematics.

```
Evaluate the generated problem against the seed.

Accept it only if all conditions hold:
- It is answerable, unambiguous, self-contained, mathematically valid,
and well formed.
- It has exactly one correct final answer.
- It is not a superficial rewrite of the seed.
- It uses new concrete content that cannot be solved
by copying seed-specific facts.

Seed problem:
<source seed>

Generated problem:
<generated candidate>
```

The structured referee output contains `accepted` and `answer`, e.g., **{ "accepted": {true/false}, "answer": {A-D}}** format. The rubrics are human-drafted and refined by an AI assistant.

C.3 Subject-Classification Prompt

```
Classify the following question into exactly one category.

Question:
<question>

Categories: <taxonomy list>
```

Each classifier's output is constrained to the category list.

D Decisive Question Examples

Table 9: Decisive questions from Global-MC and Global-Math. Category is the subject label both classifiers agree on, with the seed's category in parentheses. Gold is the answer both referees agree on. Answers lists both players' final answers, the field's author first.

Category	Question	Gold	Answers (Home first)
Global-MC			
Philosophy (moral scenarios)	Which of the following best characterizes Susan Wolf's specific objection to the concept of a moral saint? (A) … lack the capacity to experience genuine joy or happiness. (B) … would likely be unattractive because the pursuit of moral perfection would crowd out other valuable human pursuits and personality traits. (C) … impossible because human nature is inherently selfish … (D) … dangerous because they would inevitably try to impose their moral views on others by force.	(B)	**Qwen3.5-35B**: (B) Qwen3-8B: (A)
Law (professional psychology)	A defendant was charged with grand larceny for stealing a painting. The prosecution called a witness who testified that the defendant had previously been convicted of petit larceny three years ago. The defense objected to this testimony. The objection should be (A) sustained, because evidence of prior bad acts is generally inadmissible to prove character. (B) sustained, because the prior conviction was for a different degree of larceny. (C) overruled, because the prior conviction is relevant to show the defendant's intent. (D) overruled, because the statute of limitations has not expired for petit larceny.	(A)	**Qwen3.6-35B**: (A) Qwen3-8B: (C)
Chemistry (organic chemistry)	A chiral secondary alcohol, (R)-2-butanol, is treated with tosyl chloride (TsCl) in pyridine to form an intermediate, which is then reacted with sodium methoxide (NaOMe) in methanol. What is the stereochemistry and the IUPAC name of the final major organic product? (A) (S)-2-methoxybutane (B) (R)-2-methoxybutane (C) (R)- and (S)-2-methoxybutane in equal amounts (D) 1-butene	(A)	**Qwen3.6-35B**: (A) Qwen3-8B: (B)
Accounting (world history)	In a standard cost accounting system, if the actual variable overhead cost per unit is less than the standard variable overhead cost per unit, how is the variable overhead spending variance reported? (A) Favorable (B) Unfavorable (C) Zero (D) Indeterminate without knowing the standard hours	(A)	Qwen3.5-9B: (D) **GPT-OSS-20B**: (A)
Global-Math			
Geometry (algebra)	In triangle $\triangle ABC$, we are given that $AB = AC = 25$ and $BC = 14$. Find the length of the median $\overline{AM}$, where M is the midpoint of $\overline{BC}$.	24	Qwen3-8B: $\sqrt{551}$ **Gemma-4**: 24
Counting & probability (prealgebra)	Let N be the number of integers n such that $1 \leq n \leq 1000$ and the sum of the digits of n is divisible by 5. What is the value of N?	199	**Qwen3.6-35B**: 199 Qwen3-8B: 200
Number theory (algebra)	A book is said to have n leaves if it is composed of n pieces of paper; the number of pages is twice the number of leaves. If the number of pages in a book is 3 more than a multiple of 7, and the number of leaves is greater than 100, then what is the smallest possible number of leaves?	103	Qwen3-8B: 105 **Gemma-4**: 103
Grade-school word problem (grade-school word problem)	A bakery sells cupcakes for $3.00 each and muffins for $2.50 each. If a customer buys 4 cupcakes and 6 muffins, and pays with a $50 bill, how much change should they receive?	$23.00	Qwen3.6-35B: 27.00 **Gemma-4**: $23.00

E Pairwise and Domain Evidence

Table 10: Static accuracy, mean ± standard deviation across 3 runs. Mathematics uses corrected source answers. All with thinking enabled.

Model	MMLU	MMLU-Pro	GPQA	GSM8K	MATH
Qwen3.6-35B	90.1 ± 0.11	84.1 ± 0.19	65.2 ± 0.83	95.9 ± 0.20	83.8 ± 0.34
Qwen3.5-35B	90.7 ± 0.20	83.8 ± 0.05	67.3 ± 2.25	95.6 ± 0.18	84.5 ± 0.38
Qwen3.5-9B	87.9 ± 0.10	80.7 ± 0.19	63.5 ± 0.61	93.8 ± 0.08	83.0 ± 0.29
Gemma-4	86.8 ± 0.01	81.4 ± 0.04	59.2 ± 0.98	95.6 ± 0.10	83.5 ± 0.10
GPT-OSS-20B	84.8 ± 0.24	75.0 ± 0.19	56.2 ± 0.77	91.9 ± 0.31	71.0 ± 1.10
Qwen3-8B	83.2 ± 0.09	73.9 ± 0.11	49.4 ± 2.04	80.5 ± 0.50	56.2 ± 0.81

Table 11: Source and generated-field outcomes by task. Source comparisons count item-pair evaluations with usable answers. Decisive means exactly one player is correct; Saturated means both are correct. Global tasks have no direct source-score counterpart.

	MMLU	**M-Pro**	**GPQA**	**GSM8K**	**MATH**	**G-MC**	**G-Math**
Source benchmarks (pairwise over the 15 pairs)							
Comparisons	210,620	179,985	6,345	19,785	7,425	–	–
Decisive (%)	10.6	14.1	27.2	9.9	23.6	–	–
Saturated (%)	82.1	72.9	47.5	86.6	59.0	–	–
HOME-AWAY field							
Decisive (%)	65.1	57.2	61.4	15.9	13.2	65.5	11.6
Saturated (%)	25.9	36.3	31.8	78.1	83.2	30.4	85.2

Table 12: HOME points / AWAY points by task. Each model plays 5 opponents per task.

Model	**MMLU**	**M-Pro**	**GPQA**	**GSM8K**	**MATH**	**G-MC**	**G-Math**
Qwen3.5-35B	139/12	65/23	3/2	8/12	66/11	33/18	56/21
Qwen3.6-35B	143/23	102/6	22/3	10/6	123/6	54/12	124/4
Qwen3.5-9B	58/29	35/28	3/3	11/4	111/3	9/15	99/6
Gemma-4	82/53	68/71	4/9	4/31	0/12	23/14	5/25
GPT-OSS-20B	50/53	19/42	2/3	26/4	15/9	26/15	33/8
Qwen3-8B	24/58	22/44	0/3	11/1	33/8	6/12	42/11

Table 13: Posing and answering on the 5 benchmarks. AccRate is acceptance after screening and validation (%); Own/Others gives accuracy on own/opposing fields (%). HOME and AWAY are exclusive-correct points on those fields.

	Multiple Choice				Mathematics			
	Posing	**Answering**			**Posing**	**Answering**		
Model	**AccRate**	HOME	AWAY	**Own / Others**	**AccRate**	HOME	AWAY	**Own / Others**
Qwen3.5-35B	**2.0**	207	37	72.9 / 57.8	**25.1**	74	23	77.9 / 61.3
Qwen3.6-35B	1.9	**267**	32	65.3 / 58.2	13.8	**133**	12	64.8 / 58.9
Qwen3.5-9B	1.5	96	60	59.3 / 46.5	15.1	122	7	59.8 / 57.9
Gemma-4	1.7	154	**133**	63.6 / 50.8	10.8	4	**43**	10.9 / 20.6
GPT-OSS-20B	1.8	71	98	44.4 / 26.3	15.6	41	13	47.2 / 28.1
Qwen3-8B	1.0	46	105	40.7 / 21.6	16.6	44	9	63.4 / 46.6
r with total points	+.69	–	–	–	+.13	–	–	–
r Own vs Others	–	–	–	+.98	–	–	–	+.89

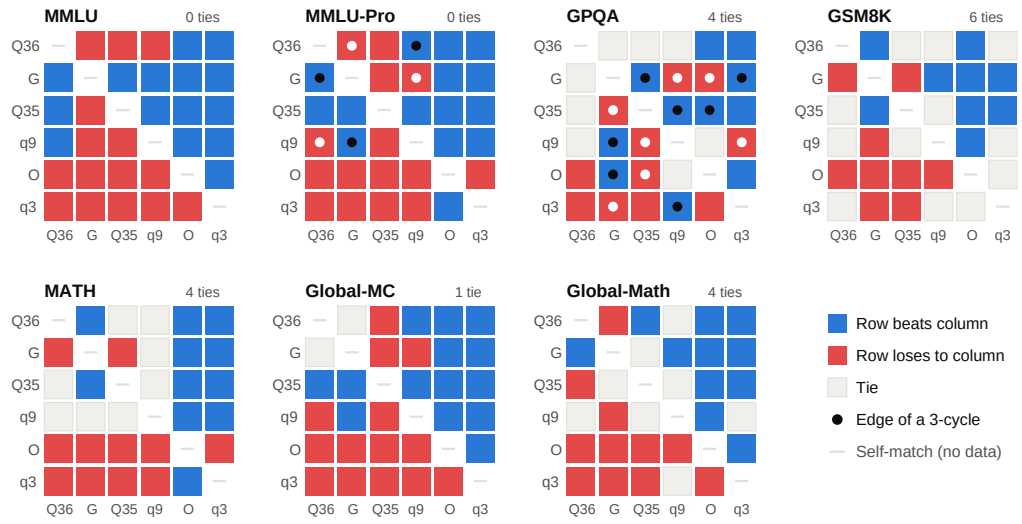

Figure 4: Pairwise outcomes: blue is a row-model win, red a loss, and gray a tie. Dots mark edges in three-cycles; panel titles report tie counts.