Eliminating Discriminative Shortcuts in Multiple Choice Evaluations with Answer Matching

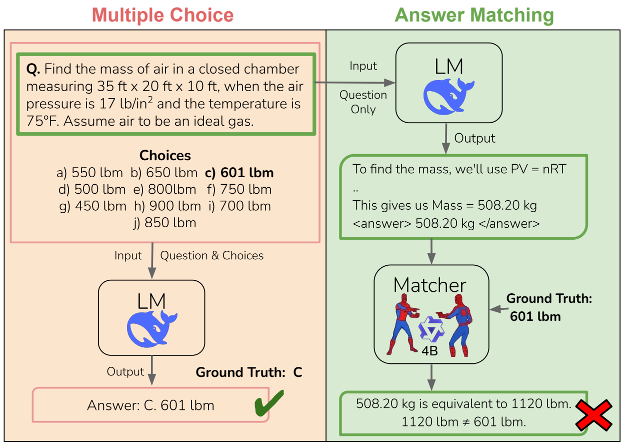
Nikhil Chandak [* 1] **Shashwat Goel** [* 1 2] **Ameya Prabhu** [3 4] **Moritz Hardt**[† 1 3] **Jonas Geiping**[† 1 2 3]

Abstract

Multiple choice benchmarks have long been the workhorse of language model evaluation because grading multiple choice is objective and easy to automate. However, we show that popular multiple-choice benchmarks admit superficial shortcuts that yield high accuracy without even looking at the questions, reflecting a fundamental limitation of discriminative evaluation not shared by evaluations of the model's free-form, generative answers. To circumvent this issue, we consider a scalable method for generative evaluation, which we call *answer matching*: Give the candidate model the question without the options, have it generate a free-form response, then use a modern language model with the reference answer to determine if the answer matches the reference. Comparing multiple-choice, "LLM-as-judge" without references, and answer-matching evaluations against human grading, we find that multiple-choice aligns poorly with humans, while answer matching using recent models—even small ones—achieves near-perfect alignment within inter-grader agreement. In light of this, we discuss how to move the evaluation format from multiple choice to answer matching.

1. Introduction

Evaluating generative models is challenging, as there is no straightforward way to score the unconstrained text output of a language model. Benchmarks try to avoid the hard problem of evaluating free-form, generative responses altogether by moving to multiple choice questionnaires. Grading multiple choice responses is fast, objective, and easy to automate. But multiple choice does not directly evaluate generative capabilities; picking one out of multiple choices is rather a discriminative problem. A recent scalable alternative to multiple choice is LLM-as-judge, where a strong judge model directly scores a candidate model's answer, or, more commonly, compares the answers provided by two models (Zheng, 2023). Although compelling as a direct means of generative evaluation, LLM-as-judge runs into numerous problems and pitfalls (Tan et al., 2024a; Wang et al., 2024a).


Multiple Choice
Answer Matching
Q. Find the mass of air in a closed chamber measuring 35 ft x 20 ft x 10 ft, when the air pressure is 17 lb/in² and the temperature is 75°F. Assume air to be an ideal gas.
Choices
a) 550 lbm b) 650 lbm c) 601 lbm
d) 500 lbm e) 800lbm f) 750 lbm
g) 450 lbm h) 900 lbm i) 700 lbm
j) 850 lbm
Input
Question & Choices
LM
Output
Ground Truth: C
Answer: C. 601 lbm
Input
Question Only
LM
Output
To find the mass, we'll use PV = nRT
..
This gives us Mass = 508.20 kg
<answer> 508.20 kg </answer>
Matcher
4B
Ground Truth: 601 lbm
508.20 kg is equivalent to 1120 lbm.
1120 lbm ≠ 601 lbm.

Observation: Model can't solve the question but can guess the correct choice

Figure 1: In this work, we show how multiple choice evaluations measure a discriminative task, rather than the generative capabilities language models (LMs) are used for. The above example from MMLU-Pro illustrates this, where DeepSeek v3 picks the correct answer, perhaps due to choice-only shortcuts like "odd one out", while giving the wrong response when prompted to give a free-form response with **just the question** (without choices).

As a result, even recent benchmarking efforts fall back to multiple choice (Wang et al., 2024b; Zhang et al., 2025b), and frontier model releases continue to evaluate on multiple choice benchmarks (Yang et al., 2025; Liu et al., 2024; Google, 2025; Team Gemma et al., 2024). Recent work even attempts to automatically generate multiple choice questions using language models, either from scratch (Yu et al., 2024), or by converting open-ended questions (Zhang et al., 2025b). It almost appears as though there is no viable, scalable alternative to multiple choice, except in a few, specialized domains like code or math which support

[1] Max Planck Institute for Intelligent Systems [2] ELLIS Institute Tübingen [3] Tübingen AI Center [4] University of Tübingen. Correspondence to: Nikhil Chandak <nikhil.chandak@tuebingen.mpg.de>, Shashwat Goel <shashwat.goel@tuebingen.mpg.de>.

ICML 2025 Workshop on Assessing World Models. Copyright 2025 by the author(s).

automatic ground truth verification.

Our contributions. In this work, we take a step back to comprehensively revisit the problem of evaluating generative models. We start from a lightweight formal discussion that makes this problem of *generative evaluation* more precise and delineates it from discriminative evaluation. Against this backdrop, we show why multiple choice fails to evaluate generative models. The reason is that *discriminative shortcuts* arising from the multiple choice format can sidestep generative evaluation. We show simple experiments with striking outcomes that reveal the propensity for shortcut learning in multiple choice problems.

Our primary contribution, however, is to motivate a compelling, scalable means of generative evaluation, we call *answer matching*: Let the model generate a candidate answer given only the question. Then provide a second model with the correct answer and let this model decide whether the candidate answer *matches* the correct answer. At the outset, answer matching, also refered to as reference-guided grading/scoring, is a lesser known cousin of LLM-as-judge and it would seem to run into similar issues (Zheng, 2023; Zhu et al., 2024a) if not graded by human raters. On the contrary, we find that answer matching, if done right, strongly outperforms both multiple choice and LLM-as-judge for generative evaluation.

In order to rigorously compare one evaluation method to another, we examine how well the evaluation method aligns with ground truth evaluations in three popular benchmarks: MATH (Hendrycks et al., 2021), MMLU-Pro (Wang et al., 2024b), and GPQA-Diamond (Rein et al., 2024). While answers to questions in MATH are automatically verifiable, answers to questions in MMLU-Pro and GPQA-Diamond are not, so we carefully annotate model responses using manual grading. Due to space limitations, we only report results for MMLU-Pro and GPQA-Diamond in the main paper and defer results on MATH to the Appendix.

Based on our new annotations, we find that answer matching achieves alignment with ground truth evaluations that is vastly superior to the alternatives. Even relatively small judge models—when used for matching and not direct evaluation—achieve agreement rates comparable to the agreement between two humans. What's important here is that the matching model is *recent*. If you had evaluated answer matching even just two years ago, it would have fared a lot less convincingly. The superiority of answer matching is a recent phenomenon.

Summary and outlook. Answer matching isn't new, but its superiority is. We argue that this qualitative change should inform future benchmark design. In principle, we can use answer matching on any questions from multiple choice benchmarks, as long as they are specific enough that the correct choice, or an equivalent paraphrase, can be uniquely arrived at. It might also be helpful to specifically design benchmarks for answer matching (Wei et al., 2024; 2025). What would help, for example, is to provide a reference list of multiple correct solutions for each question. Overall, the success of language models has recently been met with efforts to make harder multiple choice benchmarks. We show that multiple choice, by allowing discriminative shortcuts, is fundamentally easier than the generating correct solutions. Rather than making multiple choice harder, the path forward may be to leverage the newfound capabilities of language models to better align our evaluations with the generative capabilities we care about.

2. Discriminative Shortcuts to Multiple Choice Evaluations

Whether answering a question, or solving a task, generation can be formalized as presenting a model $\mathcal{F}$ with a question Q, for which it generates a response $R = \mathcal{F}(Q)$, where $R \in \mathcal{S}$, the universe of all possible finite length outputs. Let $\mathcal{A}_Q \subseteq \mathcal{S}$ be the set of correct answers for the question Q. Evaluating a generative model can therefore be formalized as the decision problem—Is the generated response R a member of the set of correct outputs $\mathcal{A}_Q$?

Unfortunately, directly checking whether $R \in \mathcal{A}_Q$ can be difficult. In the special case of a single correct response, $|\mathcal{A}_Q| = 1$, we can evaluate using using string matching, as done in classical NLP benchmarks, such as SQuaD (Rajpurkar et al., 2016). Circumventing this challenge is what popular question answering formats like multiple choice attempt to solve. Multiple choice evaluations give the model a question Q, and a list of choices consisting of a correct answer $a \in \mathcal{A}$ and n incorrect choices $\{w_i\}_{i=1}^{n} \subset \mathcal{S} \backslash \mathcal{A}_Q$ called *distractors*. The model's response $\hat{R} = \mathcal{F}(Q, \{a\} \cup \{w_i\}_Q)$[1] is marked correct only if $\hat{R} = a$. In this way, the set of correct answers is now reduced to singleton, only a, enabling automatic grading. At first, it seems that multiple choice solves the problem of $|\mathcal{A}_Q| > 1$ we outlined above.

However, on a closer look, changing the input from just Q to $Q, a, \{w_i\}$ fundamentally shifts the task from *generating* a correct response to *separating* the correct answer from the incorrect choices. The latter is traditionally considered a *discriminative* problem. To demonstrate the extent to which multiple choice datasets can be solved discriminately in practice, we perform a simple experiment. We finetune a language model (Qwen3-8B) to predict the correct answer a given only the choices $\{a\} \cup \{w_i\}$ **without the question** Q. Any accuracy obtained beyond chance in this way raises uncertainty about the extent to which accuracy on the dataset

[1] For notational convenience we assume choices are unordered, but evaluations can be sensitive to order (Zheng et al., 2024)

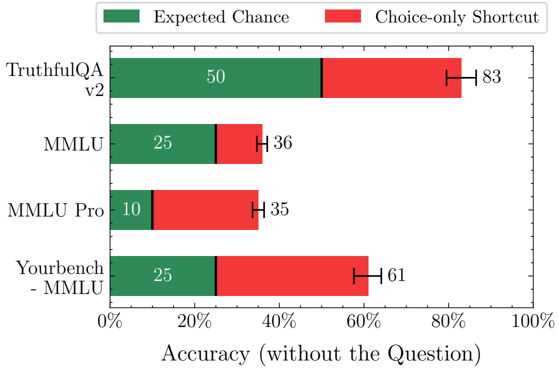


Figure 2: Shortcut accuracy achieved by finetuning a discriminative classifier that sees only the answer choices, **without any access to the question**. Strikingly, discrimination can provide high accuracies throughout, attaining 83% on TruthfulQA-v2 and 35% on MMLU-Pro, showing the *statistical separability* of correct choices from incorrect ones.

reflects generative question answering, as the model does not even know what question it is answering. Unfortunately, Figure 2 shows we can achieve strikingly high accuracies across popular datasets with *choice-only shortcuts*.

We are not the first to point out this problem. Recently, Turner & Kurzeja (2025) showed that identifying the "odd one out" can lead to large accuracies without looking at the question in the TruthfulQA dataset. This prompted the creators of TruthfulQA to release an updated version, TruthfulQA v2 (Evans et al., 2025), shifting the task from three to only one incorrect choice. However, there are myriad ways of exploiting inherent statistical separation between correct and incorrect choices, of which "odd one out" is only an instance. Indeed, we obtain an accuracy of 83% on Truthful QA v2, without even showing the question to the model. One could argue that reducing the choices from four to two only makes the discriminative task easier!

TruthfulQA is not special in being affected by this problem. Even on widely used hard, cross-domain benchmarks like MMLU, a non-trivial shortcut accuracy of 36% is seen, which might seem low considering chance accuracy is 25%, but is still interesting as it consists of questions from human examinations like GRE and USMLE.

It seems like the rising trend of using language model generated choices (Shashidhar et al., 2025) exacerbates the presence of choice-only shortcuts. For example, MMLU-Pro uses GPT4-Turbo to generate additional incorrect choices shifting the number of choices from 4 to 10. However, this also increases our classifier's shortcut accuracy significantly, from chance 10% to 35%, when compared to MMLU. YourBench (Shashidhar et al., 2025) released by Huggingface entirely generates the question and all choices from a document using an LLM, and on their "replication" of MMLU, we obtain a much higher shortcut accuracy (61%).

3. Answer Matching for Generative Evaluation

A simple way to prevent discriminative shortcuts is by not providing the model with choices in the input. In this section, we compare many evaluation methods of this form. What stands out as a compelling alternative is what we term *Answer matching*—where the model is simply tasked with providing a free-form response R, and then, another model checks whether the response R matches with a provided reference answer a. Empirically we find that answer matching achieves alignment with ground truth evaluations that is vastly superior to all available alternatives. Even relatively small (but recent) grading models—when used for answer matching, not directly correctness assessment—achieve agreement rates comparable to the agreement between two human graders.

This kind of reference-guided scoring has been occasionally considered in LLM-as-a-Judge literature (Thakur et al., 2024), but we show the distinction is crucial: LLM-as-a-Judge tasks a judge model J with *verification*—given the question Q and response R, it has to decide whether R is correct ($R \in \mathcal{A}_Q$). Traditionally (Zheng et al., 2023), the judge does not have access to a reference answer and has to assess the goodness or correctness of a response, which leads to a host of issues documented in prior work (Tan et al., 2024b; Goel et al., 2025). On the other hand, using an LLM for Answer Matching only requires it to check if the model response is semantically equivalent to the reference answer in the context of the question, $R \equiv a$ given Q.

Intuitively, matching seems easier than verifying the correctness of an arbitrary response. For example, we have highly reliable systems for matching mathematical expressions to a reference numeric answer such as MATH-Verify (Kydlicek et al., 2025), while such systems are harder to design for verifying whether an arbitrary mathematical expression R is the correct answer for a numerical question Q. On the contrary, consider tasks which require providing a graph with a certain property, say having a cycle. Here, matching to a reference answer becomes a graph-isomorphism problem, known to be NP-Hard, whereas direct verification of the graph being a correct solution can be simpler. Thus, whether using LLMs for matching works better than verification becomes a property of the task being tested.

So then how do we decide what works best for generative evaluations? We collect ground-truth evaluations of free-form model responses on popular benchmarks, and measure sample-wise outcome alignment. Evaluations that are both scalable and lead to outcomes more aligned with ground-truth evaluations can be considered better. We measure alignment using Scott's π, an inter-annotator agreement metric recommended in recent literature (Thakur et al., 2024).

Alignment with Human Grading. We use MMLU-Pro and GPQA-Diamond as our main benchmarks (but also report re-

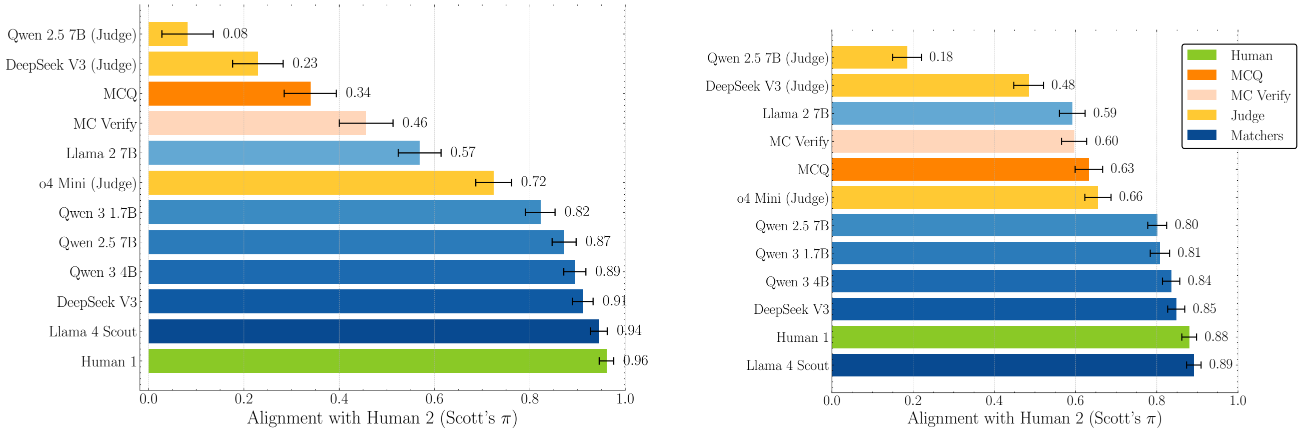


Figure 3: **Human-agreement comparison on GPQA-Diamond (left) and MMLU-Pro (right).** Each panel plots Scott's π between Human 2 and various automatic graders. Solid green bars show human–human consistency; matchers (blue) follow closely; LLM-as-Judge (yellow) and MCQs (orange) trail behind. Even Qwen3-4B approaches "human-level" grading.

sults on MATH in Appendix A)[2]. As there is no ground truth for free-form verification, we *manually evaluate* 800 model responses for correctness (with respect to the reference answer provided), across four frontier models (200 responses each from DeepSeek v3, GPT-4o, Llama-4-Maverick and Qwen3-32B). We then study how well different automatic evaluations align with human judgment. Note that questions from these datasets often rely on the choices to convey the style and specificity of the intended answer. Thus, they are not unambiguously answerable in generative style, just given the question. Further, they often have multiple possible answers. Due to the cross-domain, knowledge intensive nature of these questions, a human can only grade by comparing responses to the reference answer. This is only a ground-truth evaluation where sans semantic or functional equivalence, only one (set of) concept(s) is the correct answer. Thus, in our human study, we also prompt humans to rate whether the provided question and reference answer are specific enough, and can be arrived at uniquely. We report results on a subset of 422 questions on MMLU-Pro and 92 questions on GPQA-Diamond where both annotators think these properties are satisfied.

Modern LLMs are Human-Level Graders. Figure 3 shows alignment with human judgements of MCQs, different LLM-as-a-judge and LM matchers. We see a stark difference in alignment, with LM matchers consistently obtaining higher value of π. We also perform an error analysis for LLM-as-a-judge, finding that for the frontier models (Deepseek V3, OpenAI o4-mini), errors disproportionately (80%+) arise from false positives. That is, the judge finds responses which are marked incorrect in human annotation correct. What is however striking is that *small Qwen3 models have near-human level alignment, with the larger recent DeepSeek and Llama models having agreement within the range of inter-annotator disagreement*.

Could we fix MCQ in any other way? An alternative to standard MCQ evaluations is *multiple choice verification* (Götting et al., 2025), where the model is given each choice for a question separately, and must check independently for each choice if it is the correct answer to the question. Many recently proposed multiple choice variants like including "None of the Above" (Elhady et al., 2025) or multiple correct choices essentially boil down to this verification task (Zhu et al., 2024b). This evaluation method seems to have either similar (Fig. 3: Right) or better alignment (Fig. 3: Left) than providing all choices at once. Nonetheless, it still has significantly lower alignment than matching.

4. Discussion

In this work we show that modern LLMs excel at matching free-form responses to references answers. Via careful measurement of alignment to human grading, we find that such LLM-based *answer matching* is significantly more accurate at measuring generative capabilities than currently used alternatives, including variants of multiple choice evaluation, and LLM-as-a-Judge without a reference answer. In Appendix B, we show that this increase in validity matters for leaderboards, with several recent models doing noticeably worse when asked to provide a free-form answer to a question. Then, we also show that the cost of answer matching is at most marginally more than multiple choice evaluations making them a cheap and practical alternative. Scores on multiple choice benchmarks have long been questioned, but the lack of a scalable alternative has kept them popular in the community. The high reliability of LLM-based answer matching is a recent phenomenon, and may be a watershed moment that should inform future benchmark design.

[2] For MATH, MATH-Verify (Kydlicek et al., 2025) library supports ground truth verification so we compute alignment with that.

References

Balepur, N. and Rudinger, R. Is your large language model knowledgeable or a choices-only cheater? *arXiv preprint arXiv:2407.01992*, 2024.

Balepur, N., Ravichander, A., and Rudinger, R. Artifacts or abduction: How do LLMs answer multiple-choice questions without the question? In *ACL*, 2024. URL https://aclanthology.org/2024.acl-long.555/.

Balepur, N., Rudinger, R., and Boyd-Graber, J. L. Which of these best describes multiple choice evaluation with llms? a) forced b) flawed c) fixable d) all of the above. *arXiv preprint arXiv:2502.14127*, 2025.

Ben-Simon, A., Budescu, D. V., and Nevo, B. A comparative study of measures of partial knowledge in multiple-choice tests. *Applied Psychological Measurement*, 21(1): 65–88, 1997.

Bowman, S. R. and Dahl, G. E. What will it take to fix benchmarking in natural language understanding? *arXiv preprint arXiv:2104.02145*, 2021.

Bulian, J., Buck, C., Gajewski, W., Boerschinger, B., and Schuster, T. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. *arXiv preprint arXiv:2202.07654*, 2022.

Chen, A., Stanovsky, G., Singh, S., and Gardner, M. Evaluating question answering evaluation. In *Proceedings of the 2nd workshop on machine reading for question answering*, pp. 119–124, 2019.

Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F. P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W. H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A. N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., and Zaremba, W. Evaluating large language models trained on code. 2021.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Elhady, A., Agirre, E., and Artetxe, M. Wicked: A simple method to make multiple choice benchmarks more challenging, 2025. URL https://arxiv.org/abs/2502.18316.

Evans, O., Chua, J., and Lin, S. New, improved multiple-choice truthfulqa, 2025. URL https://www.lesswrong.com/posts/Bunfwz6JsNd44kgLT/new-improved-multiple-choice-truthfulqa.

Farr, R., Pritchard, R., and Smitten, B. A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement*, 27(3):209–226, 1990.

Geiping, J., Stein, A., Shu, M., Saifullah, K., Wen, Y., and Goldstein, T. Coercing llms to do and reveal (almost) anything, 2024. URL https://arxiv.org/abs/2402.14020.

Goel, S., Strüber, J., Auzina, I. A., Chandra, K. K., Kumaraguru, P., Kiela, D., Prabhu, A., Bethge, M., and Geiping, J. Great models think alike and this undermines ai oversight. In *ICML*, 2025.

Google. Gemini 2.5 pro preview model card, May 2025. URL https://storage.googleapis.com/model-cards/documents/gemini-2.5-pro-preview.pdf. Accessed: 2025-05-16.

Götting, J., Medeiros, P., Sanders, J. G., Li, N., Phan, L., Elabd, K., Justen, L., Hendrycks, D., and Donoughe, S. Virology capabilities test (vct): A multimodal virology q&a benchmark, 2025. URL https://arxiv.org/abs/2504.16137.

Haladyna, T. M. and Downing, S. M. A taxonomy of multiple-choice item-writing rules. *Applied measurement in education*, 2(1):37–50, 1989.

Haladyna, T. M., Downing, S. M., and Rodriguez, M. C. A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, 15(3):309–333, 2002.

Hardt, M. The emerging science of machine learning benchmarks. Online at https://mlbenchmarks.org, 2025. Manuscript.

Hashemi, H., Eisner, J., Rosset, C., Van Durme, B., and Kedzie, C. LLM-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *ACL*, 2024. URL https://aclanthology.org/2024.acl-long.745/.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Ho, X., Huang, J., Boudin, F., and Aizawa, A. Llm-as-a-judge: Reassessing the performance of llms in extractive qa. *arXiv preprint arXiv:2504.11972*, 2025.

Hughes, J., Price, S., Lynch, A., Schaeffer, R., Barez, F., Koyejo, S., Sleight, H., Jones, E., Perez, E., and Sharma, M. Best-of-n jailbreaking. *arXiv preprint arXiv:2412.03556*, 2024.

Kelly, F. J. The kansas silent reading tests. *Journal of Educational Psychology*, 7(2):63, 1916.

Krathwohl, D. R. A revision of bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218, 2002.

Krumdick, M., Lovering, C., Reddy, V., Ebner, S., and Tanner, C. No free labels: Limitations of llm-as-a-judge without human grounding, 2025. URL https://arxiv.org/abs/2503.05061.

Kydlicek, H., Lozovskaya, A., Habib, N., and Fourrier, C. Fixing open llm leaderboard with math-verify, February 2025. URL https://huggingface.co/blog/math_verify_leaderboard. Hugging Face Blog.

Li, Z., Mondal, I., Liang, Y., Nghiem, H., and Boyd-Graber, J. L. Pedants: Cheap but effective and interpretable answer equivalence. *arXiv preprint arXiv:2402.11161*, 2024.

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Ma, Y., Cao, Y., Hong, Y., and Sun, A. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! *arXiv preprint arXiv:2303.08559*, 2023.

Myrzakhan, A., Bsharat, S. M., and Shen, Z. Open-llm-leaderboard: From multi-choice to open-style questions for llms evaluation, benchmark, and arena. *arXiv preprint arXiv:2406.07545*, 2024.

Ouyang, S., Wang, S., Liu, Y., Zhong, M., Jiao, Y., Iter, D., Pryzant, R., Zhu, C., Ji, H., and Han, J. The shifted and the overlooked: A task-oriented investigation of user-gpt interactions. *arXiv preprint arXiv:2310.12418*, 2023.

Piepho, H.-P. An algorithm for a letter-based representation of all-pairwise comparisons. *Journal of Computational and Graphical Statistics*, 13(2):456–466, 2004.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In Su, J., Duh, K., and Carreras, X. (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264/.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

Roediger III, H. L. and Marsh, E. J. The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(5):1155, 2005.

Sampson, C. and Boyer, P. G. Gre scores as predictors of minority students'success in graduate study: An argument for change. *College Student Journal*, 35(2):271–271, 2001.

Shashidhar, S., Fourrier, C., Lozovskia, A., Wolf, T., Tur, G., and Hakkani-Tür, D. Yourbench: Easy custom evaluation sets for everyone. *arXiv preprint arXiv:2504.01833*, 2025.

Simkin, M. G. and Kuechler, W. L. Multiple-choice tests and student understanding: What is the connection? *Decision Sciences Journal of Innovative Education*, 3(1):73–98, 2005.

Sinha, S., Goel, S., Kumaraguru, P., Geiping, J., Bethge, M., and Prabhu, A. Can language models falsify? the need for inverse benchmarking. In *ICLR SSI-FM Workshop*, 2025. URL https://openreview.net/forum?id=QSQEUJfhen.

Swamy, G., Choudhury, S., Sun, W., Wu, Z. S., and Bagnell, J. A. All roads lead to likelihood: The value of reinforcement learning in fine-tuning, 2025. URL https://arxiv.org/abs/2503.01067.

Tan, S., Zhuang, S., Montgomery, K., Tang, W. Y., Cuadron, A., Wang, C., Popa, R., and Stoica, I. JudgeBench: A Benchmark for Evaluating LLM-Based Judges. In *The Thirteenth International Conference on Learning Representations*, October 2024a. URL https://openreview.net/forum?id=G0dksFayVq.

Tan, S., Zhuang, S., Montgomery, K., Tang, W. Y., Cuadron, A., Wang, C., Popa, R. A., and Stoica, I. Judgebench: A benchmark for evaluating llm-based judges, 2024b. URL https://arxiv.org/abs/2410.12784.

Taylor, W. L. “cloze procedure”: A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433, 1953.

Team Gemma, Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., Ferret, J., Liu, P., Tafti, P., Friesen, A., Casbon, M., Ramos, S., Kumar, R., Lan, C. L., Jerome, S., Tsitsulin, A., Vieillard, N., Stanczyk, P., Girgin, S., Momchev, N., Hoffman, M., Thakoor, S., Grill, J.-B., Neyshabur, B., Bachem, O., Walton, A., Severyn, A., Parrish, A., Ahmad, A., Hutchison, A., Abdagic, A., Carl, A., Shen, A., Brock, A., Coenen, A., Laforge, A., Paterson, A., Bastian, B., Piot, B., Wu, B., Royal, B., Chen, C., Kumar, C., Perry, C., Welty, C., Choquette-Choo, C. A., Sinopalnikov, D., Weinberger, D., Vijaykumar, D., Rogozińska, D., Herbison, D., Bandy, E., Wang, E., Noland, E., Moreira, E., Senter, E., Eltyshev, E., Visin, F., Rasskin, G., Wei, G., Cameron, G., Martins, G., Hashemi, H., Klimczak-Plucińska, H., Batra, H., Dhand, H., Nardini, I., Mein, J., Zhou, J., Svensson, J., Stanway, J., Chan, J., Zhou, J. P., Carrasqueira, J., Iljazi, J., Becker, J., Fernandez, J., van Amersfoort, J., Gordon, J., Lipschultz, J., Newlan, J., Ji, J.-y., Mohamed, K., Badola, K., Black, K., Millican, K., McDonell, K., Nguyen, K., Sodhia, K., Greene, K., Sjoesund, L. L., Usui, L., Sifre, L., Heuermann, L., Lago, L., McNealus, L., Soares, L. B., Kilpatrick, L., Dixon, L., Martins, L., Reid, M., Singh, M., Iverson, M., Görner, M., Velloso, M., Wirth, M., Davidow, M., Miller, M., Rahtz, M., Watson, M., Risdal, M., Kazemi, M., Moynihan, M., Zhang, M., Kahng, M., Park, M., Rahman, M., Khatwani, M., Dao, N., Bardoliwalla, N., Devanathan, N., Dumai, N., Chauhan, N., Wahltinez, O., Botarda, P., Barnes, P., Barham, P., Michel, P., Jin, P., Georgiev, P., Culliton, P., Kuppala, P., Comanescu, R., Merhej, R., Jana, R., Rokni, R. A., Agarwal, R., Mullins, R., Saadat, S., Carthy, S. M., Cogan, S., Perrin, S., Arnold, S. M. R., Krause, S., Dai, S., Garg, S., Sheth, S., Ronstrom, S., Chan, S., Jordan, T., Yu, T., Eccles, T., Hennigan, T., Kocisky, T., Doshi, T., Jain, V., Yadav, V., Meshram, V., Dharmadhikari, V., Barkley, W., Wei, W., Ye, W., Han, W., Kwon, W., Xu, X., Shen, Z., Gong, Z., Wei, Z., Cotruta, V., Kirk, P., Rao, A., Giang, M., Peran, L., Warkentin, T., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Sculley, D., Banks, J., Dragan, A., Petrov, S., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Borgeaud, S., Fiedel, N., Joulin, A., Kenealy, K., Dadashi, R., and Andreev, A. Gemma 2: Improving Open Language Models at a Practical Size. *arxiv:2408.00118[cs]*, October 2024. doi: 10.48550/arXiv.2408.00118. URL http://arxiv.org/abs/2408.00118.

Thakur, A. S., Choudhary, K., Ramayapally, V. S., Vaidyanathan, S., and Hupkes, D. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *arXiv preprint arXiv:2406.12624*, 2024.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Turner, A. and Kurzeja, M. Gaming truthfulqa: Simple heuristics exposed dataset weaknesses, 2025. URL https://turntrout.com/original-truthfulqa-weaknesses.

Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Kong, L., Liu, Q., Liu, T., and Sui, Z. Large Language Models are not Fair Evaluators. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024b.

Wei, J., Karina, N., Chung, H. W., Jiao, Y. J., Papay, S., Glaese, A., Schulman, J., and Fedus, W. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024.

Wei, J., Sun, Z., Papay, S., McKinney, S., Han, J., Fulford, I., Chung, H. W., Passos, A. T., Fedus, W., and Glaese, A. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025.

West, P., Lu, X., Dziri, N., Brahman, F., Li, L., Hwang, J. D., Jiang, L., Fisher, J., Ravichander, A., Chandu, K., et al. The generative ai paradox:” what it can create, it may not understand”. *arXiv preprint arXiv:2311.00059*, 2023.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. Qwen3 Technical Report. *arxiv:2505.09388[cs]*, May 2025. doi: 10.48550/arXiv.2505.09388. URL http://arxiv.org/abs/2505.09388.

Yu, H. C., Shih, Y. A., Law, K. M., Hsieh, K., Cheng, Y. C., Ho, H. C., Lin, Z. A., Hsu, W.-C., and Fan, Y.-C. Enhancing Distractor Generation for Multiple-Choice Questions with Retrieval Augmented Pretraining and Knowledge Graph Integration. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11019–11029, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.655. URL https://aclanthology.org/2024.findings-acl.655/.

Zhang, Y., Su, Y., Liu, Y., Wang, X., Burgess, J., Sui, E., Wang, C., Aklilu, J., Lozano, A., Wei, A., Schmidt, L., and Yeung-Levy, S. Automated generation of challenging multiple-choice questions for vision language model evaluation. In *CVPR*, 2025a.

Zhang, Y., Su, Y., Liu, Y., Wang, X., Burgess, J., Sui, E., Wang, C., Aklilu, J., Lozano, A., Wei, A., Schmidt, L., and Yeung-Levy, S. Automated generation of challenging multiple-choice questions for vision language model evaluation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b.

Zhang, Z., Xu, L., Jiang, Z., Hao, H., and Wang, R. Multiple-choice questions are efficient and robust llm evaluators, 2024.

Zheng, C., Zhou, H., Meng, F., Zhou, J., and Huang, M. Large language models are not robust multiple choice selectors. In *ICLR*, 2024. URL https://openreview.net/forum?id=shr9PXz7T0.

Zheng, L. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *NeurIPS Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=uccHPGDlao.

Zheng, X., Pang, T., Du, C., Liu, Q., Jiang, J., and Lin, M. Cheating automatic LLM benchmarks: Null models achieve high win rates. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=syThiTmWWm.

Zhu, L., Wang, X., and Wang, X. JudgeLM: Fine-tuned Large Language Models are Scalable Judges. In *The Thirteenth International Conference on Learning Representations*, October 2024a. URL https://openreview.net/forum?id=xsELpEPn4A.

Zhu, Z., Xu, Y., Chen, L., Yang, J., Ma, Y., Sun, Y., Wen, H., Liu, J., Cai, J., Ma, Y., Zhang, S., Zhao, Z., Sun, L., and Yu, K. Multi: Multimodal understanding leaderboard with text and images, 2024b.

A. Alignment on MATH

We start with the MATH dataset, where as discussed, MATH-Verify library (Kydlicek et al., 2025) implements rule-based ground-truth evaluations of generative responses. Further, a parallel multiple choice version is also available (Zhang et al., 2024). This allows us to compare the alignment of generative evaluations with multiple choice. We specifically focus on Level 5 questions as they are less closer to saturation. Figure 4 shows that answer matching, even with the 1.7 billion parameter Qwen3 model, achieves near-perfect alignment with ground-truth ($\pi = 0.97$). As for LLM-as-a-judge, even the much larger 671 billion parameter DeepSeek v3 model only obtains modest agreement $\pi = 0.72$, while as a matcher it achieves $\pi = 0.98$.

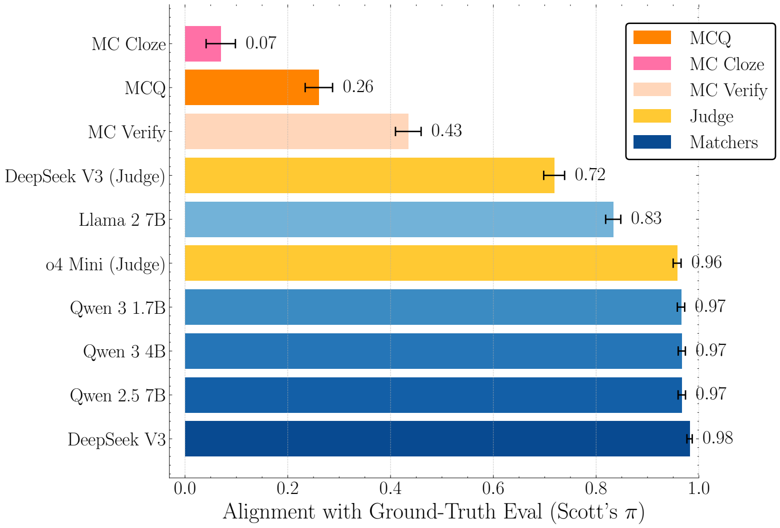


Figure 4: Bars rank evaluators from best (top) to worst (bottom) on testing the responses of Qwen2.5-7B on MATH Level 5. Even a small 1.7 B-parameter matcher reaches Scott's $\pi = 0.97$, virtually indistinguishable from perfect agreement, whereas the classical MCQ score aligns at only $\pi = 0.26$. MC Verify, Deepseek-V3 Judge and MC Cloze improve slightly but still lag far behind answer-matching.

Could we fix MCQ in any other way? Perhaps what stands out is that standard MCQ only obtains only $\pi = 0.26$. This is mostly due to false positives (85% of errors), as expected from only requiring an easier discriminative problem to be solved. Next, we explore other variants of multiple choice evaluations that do not provide all choices in the input, thus preventing discriminative shortcuts.

First, we reconsider *multiple choice verification* (Götting et al., 2025), as discussed in the main text, where the model is given each choice for a question separately, and must check independently for each choice if it is the correct answer to the question. Formally, to be marked accurate on a question in this setting, $\mathcal{F}(Q, a) =$ True, and $\mathcal{F}(Q, w) =$ False $\forall\ w \in \{w_i\}$. Many recently proposed multiple choice variants like including "None of the Above" (Elhady et al., 2025) or multiple correct choices essentially boil down to this verification task (Zhu et al., 2024b), as they force the model to evaluate the correctness of each choice independently. This evaluation method is better aligned ($\pi = 0.43$) than providing all choices at once. Indeed, verification is a strictly harder task than discrimination, as access to a verification oracle allows picking the correct choice among incorrect ones by checking each of them but verification is undecidable with a discriminative oracle. However, it still has a lower alignment than matching, and its hardness relation with the generative task has been of much recent interest (Swamy et al., 2025; Sinha et al., 2025).

Finally, *Multiple Choice Cloze* (Taylor, 1953) is a classical way to evaluate without allowing for choice discrimination. While it is less popular now, it was for example the proposed format for the Abstract Reasoning Corpus (ARC) (Clark et al., 2018). It is implemented by only providing the model the question in the input, and then measuring completion likelihoods over all choices, picking the one assigned the highest likelihood. Unfortunately, it has even lower alignment than multiple choice, with its 0.07 π value indicating its outcomes are almost independent from the ground-truth. This type of evaluation is entirely a non-generative likelihood evaluation, and so it is unclear how to fit in modern models which derive part of their prowess from generating a Chain-of-Thought before responding, potentially explaining its comparatively poor alignment.

B. Towards Benchmarking with Answer Matching

We now examine the implications of adopting answer matching within the benchmarking ecosystem, focusing on its impact on model rankings, evaluation costs, replicability of benchmark results, and future dataset development.

Rankings Change. For public benchmarks, cardinal accuracy measurements and sample-wise alignment is perhaps of lesser importance than how models are ranked, as argued in Hardt (2025). After all, ultimately they serve as leaderboards that guide practitioners on what models to use. Does multiple choice—despite its issues—perhaps give the same model

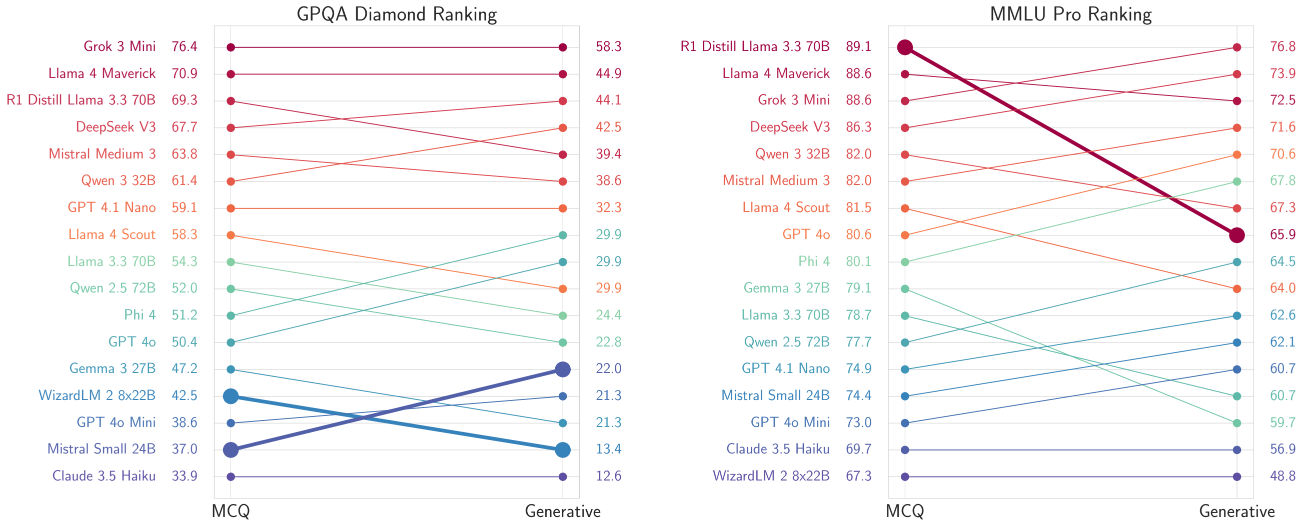


Figure 5: **Leaderboard rankings change on Human Filtered Subset when moving from MCQ to answer-matching on generative responses in GPQA-Diamond (L) and MMLU-Pro (R)**: Thick lines represent statistically significant changers based on the Compact Letter Display algorithm (Piepho, 2004).

rankings as answer matching? Figure 6 shows that model rankings change substantially when directly measuring the more realistic generative use-case. For example, recent open-weights models like Qwen3-32B and Llama-4-Scout drop significantly on MMLU-Pro while Microsoft's Phi-4 and WizardLM show large drops on GPQA Diamond. On the other hand, we see proprietary models like GPT variants improve ranking in generative evaluation which seems plausible given that these models are typically optimized for chat-based applications. Further, benchmarks that appear saturated due to high cardinal values in multiple choice format begin to reveal substantial headroom in the generative setting. For example, we observe a drop of over 20% in GPQA Diamond across models indicating that it can be repurposed in free-form format to continue serving as a meaningful evaluation for the next generation of frontier models.

Answer Matching can be Cheap. A key concern in maintaining such public leaderboards is the potential cost of grading newly released models (Li et al., 2024). In Figure 7, we compare costs of evaluating four models on MMLU-Pro, finding that answer matching—even using a frontier model (DeepSeek v3)—is only 2% more expensive than multiple choice evaluations. Further, since small models like Qwen3-4B achieve high human alignment, the cost of answer matching can in fact be lower than that of multiple choice evaluations. While this may seem counterintuitive, it is important to note that evaluation costs are primarily driven by the length of model responses, and that running a language model grader incurs only a small additional cost relative to the generation overhead. We find that models generate longer responses for multiple choice than when they are just asked to answer a question. In the case of MCQs, models typically attempt to solve the question in a free-form manner first; if their response does not align with any of the given choices, they then proceed to evaluate each choice individually. We observe this phenomenon across models, and provide detailed breakdown in Appendix **??**. Naturally, these costs can vary based on the model used for matching. Nonetheless, at the frontier, as inference-time compute is scaled, we expect that matching a response to a reference answer will require less compute than solving the task from scratch, as the former is easier. Thus, we believe the additional cost of answer matching will be marginal.

Reliability. Another common concern with different methodologies for language model evaluations is their reliability. This concern has two primary aspects: reproducibility and robustness. First, for a long time, evaluations relying on a language model as the grader were considered to have a reproducibility problem (Zhang et al., 2025a), as only API models were sufficiently capable—and these were subject to deprecation. However, this concern is now mitigated by both, progress in capabilities of open-weight models like DeepSeek v3, and recent small models like Qwen3-4B being good at answer matching. To minimize variance, evaluations can be conducted at zero temperature, as done throughout our experiments. As for robustness, while we expect answer matching to be more robust than their LLM-as-a-Judge counterparts, language models can, however, be coerced into giving favorable evaluations (Zheng et al., 2025; Geiping et al., 2024). Preliminary evidence suggests that such jailbreaks are getting harder to perform as models get more capable (Hughes et al., 2024). Until then, it might be useful to also report more adversarially robust evaluations like multiple choice alongside, as exclusively

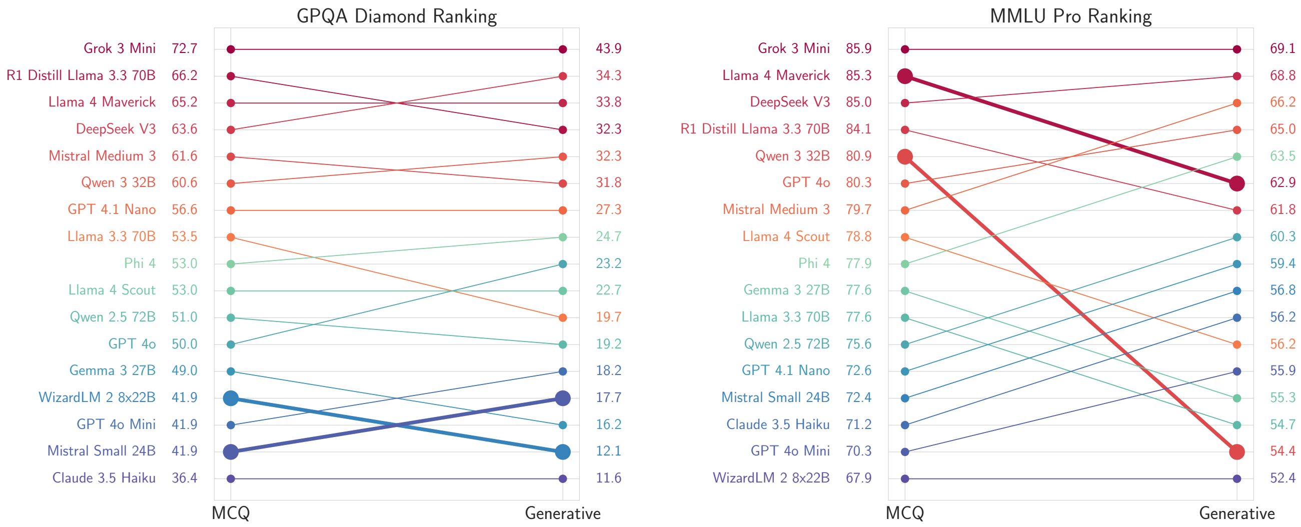


Figure 6: **Leaderboard rankings change on when moving from MCQ to answer-matching on generative responses in GPQA-Diamond (L) and MMLU-Pro (R)**: Thick lines represent statistically significant changers based on the Compact Letter Display algorithm (Piepho, 2004). It seems chat-optimised proprietary models (GPT 4.1 Nano, 4o Mini, Claude 3.5 Haiku) climb on generative rankings, whereas open-weight models judged by their multiple-choice benchmark performance can (Phi 4, Llama 4 Scout, Qwen 3 32B) drop markedly. The figure highlights that benchmark conclusions — and hence model selection — depend critically on the choice of evaluation protocol.

high performance on LM Matching evaluations can raise suspicion.

Intrinsic Validity of Answer Matching is Recent. One might also wonder, given that Llama 2 7B Chat (Touvron et al., 2023), released in July 2023, seems to match or beat the alignment of MCQ in our analysis, should we have moved on to LM answer matching much earlier? We argue that this is not the case.MCQ, while having poor construct validity as a measure of generative capabilities, is more reliable for what it claims to measure, namely, a model's multiple choice test performance. In contrast, older models lacked the intrinsic validity required for answer matching, as they performed poorly on this task. This has changed only recently, as newer models now achieve near-human agreement levels. We therefore believe that it is only with the recent generation of models that answer matching has clearly emerged as the superior mode of evaluation.

Converting Multiple Choice Benchmarks to Answer Matching. Towards answer matching for evaluation, practitioners can reuse existing multiple-choice benchmarks with one important caveat. In our human annotation, we found that questions designed for multiple-choice formats are often not specific enough by themselves and rely on the provided choices to disambiguate the intended solution. After filtering such questions, we reduced the dataset size by more than half, while also skewing the category distribution toward STEM, as those questions more frequently had unique answers. We show the change in subject distribution before and after filtering in Fig. 7 and Fig. 8 in Appendix. This motivates creating questions that are either more specific or providing a list of reference answers when multiple answers when possible. We are already seeing early signs of this shift in benchmarks such as SimpleQA and BrowserComp (Wei et al., 2024; 2025), whose creators explicitly asked human trainers to design 'questions with *single, indisputable, short answers* that would not change over time. Going forward, we believe that such dataset creation efforts may be more fruitful and less difficult than creating higher quality distractors for multiple-choice questions.

C. Related Work

Multiple-choice questions (MCQs) were introduced by Frederick J. Kelly in 1916 as a quick, objective, and scalable alternative to essay grading (Kelly, 1916). However, Kelly later warned that standardized tests built on MCQs reduce learning to mere finding shortcut solutions, leaving large gaps in testing answering ability. Over the past century, research in educational psychology has documented shortcomings of MCQ evaluations (Sampson & Boyer, 2001; Simkin & Kuechler, 2005; Farr et al., 1990; Roediger III & Marsh, 2005). Despite these drawbacks, MCQs still dominate large-scale testing — and, by extension, the evaluation of language models. We summarize the re-emergence of the longstanding trade-off

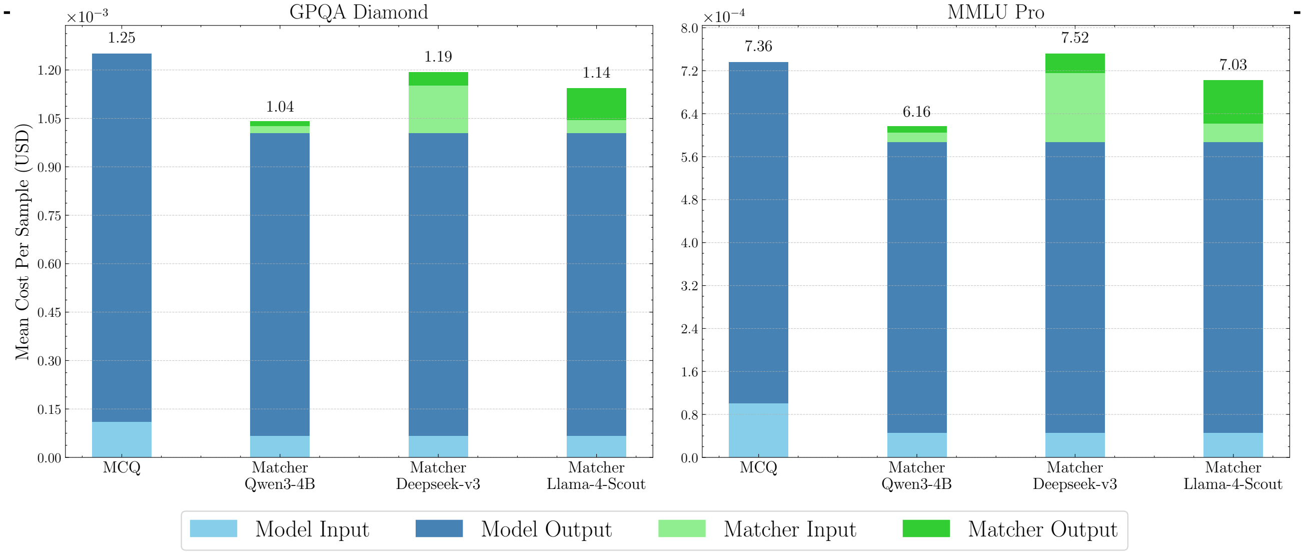


Figure 7: **Breakdown of evaluation cost averaged across 17 models.** For model responses across various providers, the stacked bars show that generating the answer itself dominates cost (grey). The blue bars show the mean cost of generating a response for our human-annotated subset of MMLU-Pro dataset averaged across 17 different models. We see that generating the answer itself (number of output tokens) dominates the cost (blue). Answer-matching (green) with a small model turns out to be less expensive than MCQ evaluation, and even a frontier matcher (v3) is only marginally more expensive. Thus, answer-matching not only improves validity but does so at minimal extra compute expense.

between scalability and nuance is resurfacing in language-model evaluation here (Ma et al., 2023; West et al., 2023), and contextualize our work with research on generative evaluation methods.

Limitations of Multiple-Choice Evaluation. A long running critique of multiple-choice questions (MCQs) is that they primarily test the ability to *rank* (Haladyna et al., 2002; Ben-Simon et al., 1997) candidate choices or *validate* the correctness of a given choices (Haladyna & Downing, 1989) rather than to *generate* an answer from scratch (Ouyang et al., 2023; Bowman & Dahl, 2021; Balepur et al., 2025). Because the task is restricted to choosing among distractors, significant MCQ accuracy can be achieved just through shortcuts — e.g. relying on choice-only heuristics (Turner & Kurzeja, 2025; Balepur & Rudinger, 2024) or inferring the intended question from the answer set (Balepur et al., 2024). This limitation is intrinsic to discriminative evaluation: the model is not tested on its ability to produce content beyond the provided choices. In contrast, answer-matching (open-ended) evaluations directly measure generative performance, on which models show lower accuracy (Myrzakhan et al., 2024).

Generative Evaluation. Answer-matching resembles classical Constructed Response Questions (CRQs) in educational testing: the model is tested on its ability to *generate* an answer. CRQs also span all levels of Bloom's taxonomy (Krathwohl, 2002), from recall to creation (Balepur et al., 2025). The main question to be tackled for automatic short-answer grading is *scoring* the generated response (Chen et al., 2019). Exact-string matching is too brittle; traditional n-gram metrics (BLEU, ROUGE, CIDEr) correlate only weakly with human judgments leading to other rule-based evlautions (Li et al., 2024). Subsequently, works have used embedding-based similarity metrics to measure semantic overlap (Bulian et al., 2022) or LLM-as-Judge (Zheng, 2023), prompted to grade or critique answers, often with rubric conditioning or chain-of-thought rationales (Ho et al., 2025). Classical LLM-as-a-Judge evaluation however has been often found to be brittle (Wang et al., 2024a; Goel et al., 2025), leading to uncertainty about the validity of LLM-based evaluation in general. In contrast, consistent with parallel work (Krumdick et al., 2025), we show that once LLMs are provided the reference answer, answer matching with recent LLMs can be a cheap way to score generative responses, that is better aligned with ground-truth evaluations.

D. Limitations and Considerations

Annotation Process. Some questions in MMLU-Pro and GPQA-Diamond require subject expertise to both check whether they are specific enough to be answered without choices, and also whether they have a unique answer. Further, there were disagreements when matching answers for even the filtered, shown in our alignment plots. While we are confident in the aggregate trends, individual annotations may be noisy. We release our annotations publicly and welcome community feedback to improve them.

Optimization Pressure on LM Matcher. In this work, we did not study how robust answer matching LMs are to optimization pressure. In the real-world, any evaluation scheme used will be optimized for, and given the ubiquity of LLM jailbreaks (Geiping et al., 2024), it is quite possible stronger models are needed for matching to rule out cheating models (Zheng et al., 2025; Hughes et al., 2024).

On the hardness of matching. Relatedly, for some tasks, answer matching might be harder than simple verification. For example, in tasks with graph outputs, answer matching can require solving the graph isomorphism problem which is NP-Hard, whereas directly verifying the requisite graph properties can be much simpler.

Answer matching can not always be used. For our alignment analysis, we filtered to questions with a unique correct answer (not counting paraphrases). This means our results do not apply to questions with multiple correct answers. In this case, either the dataset would have to provide as many semantically distinct valid answers as possible, or answer matching is no more guaranteed to provide correct evaluations. Moreover, the evaluation of many generative tasks can not be simply formulated with answer matching, e.g. translation, summarization, theorem proving, and coding. LM judges with rubrics (Hashemi et al., 2024) or verification via execution (Chen et al., 2021) might be more suitable here.