Test-Time Reasoners Are Strategic Multiple-Choice Test-Takers

Nishant Balepur[1,2] **Atrey Desai**[1] **Rachel Rudinger**[1]
[1]University of Maryland [2]New York University
{nbalepur, rudinger}@umd.edu, adesai10@terpmail.umd.edu

Abstract

Large language models (LLMs) now give reasoning before answering, excelling in tasks like multiple-choice question answering (MCQA). Yet, a concern is that LLMs do not solve MCQs as intended, as work finds LLMs sans reasoning succeed in MCQA without using the question, i.e., choices-only. Such partial-input success is often linked to trivial shortcuts, but reasoning traces could reveal if choices-only strategies are truly shallow. To examine these strategies, we have reasoning LLMs solve MCQs in full and choices-only inputs; test-time reasoning often boosts accuracy in full and in choices-only, half the time. While possibly due to shallow shortcuts, choices-only success is barely affected by the length of reasoning traces, and after finding traces pass faithfulness tests, we show they use less problematic strategies like inferring missing questions. In all, we challenge claims that partial-input success is always a flaw, so we propose how reasoning traces could separate problematic data from less problematic reasoning.[1]

1 Introduction: When Cheaters Prosper

Reasoning has become a central goal in Large Language Model (LLM) development (Xu et al., 2025), with models now designed to produce step-by-step traces before responding (Zelikman et al., 2022). By scaling compute in test-time, these models can aid complex, multi-step tasks like Deep Research (Shao et al., 2024), software assistance (Wang et al., 2024), and scientific automation (Lu et al., 2024).

Offline evaluations ensure such models behave as intended (Baker et al., 2025). A common format is multiple-choice question answering (Clark et al., 2020, MCQA): from a set of choices, pick the best answer to a question. While popular, work reveals non-reasoning LLMs beat random in MCQA without the question (Balepur et al., 2024b), i.e., choices-only, raising concerns that reasoning models with even stronger abilities might amplify this behavior.

Conventional wisdom ascribes choices-only success to models exploiting shallow cues in choices, suggesting they do so even with all inputs to bypass what MCQA tests (Chandak et al., 2025). However, past NLI work shows partial-input success does not entail ineptitude in the original NLI task (Srikanth and Rudinger, 2022). Analogously, we hypothesize LLMs may use non-shallow strategies in choices-only, drawing on skills MCQs aim to test—e.g., discarding inaccurate choices—akin to how students with partial knowledge guess on MC exams (Lau et al., 2011). These are hard to study in simple (e.g. bag-of-words) or encoder-only (e.g. BERT) models common in partial-input work (Poliak et al., 2018)—unable to express them—but reasoning traces offer signals on how choices-only models could succeed.

Our short paper builds focused tests to target this, running 12 LLMs in two settings: *full input*—using questions and choices as input—and *choices-only*—using just the choices. In both, we study the impact of test-time reasoning (TTR) on MCQA accuracy, aiming to learn: 1) if test-time reasoning amplifies choices-only success; and 2) which types of strategies reasoning traces use in choices-only settings.

In 36 LLM-benchmark combinations, TTR improves accuracy with full inputs in 27, but only 15 for choices-only. Further, choices-only accuracy rises minimally with TTR scaling (§3.2), suggesting LLMs with and without TTR may use similar approaches to solving these items. We then study if LLMs with TTR rely on shallow cues in choices-only; we ensure choices-only reasoning traces pass faithfulness checks (§3.3), then uncover they *sometimes* employ superficial shortcuts that can bypass the original MCQA task (e.g. *I'll pick "(A) 1.5", as 1.5 is a "messy" number*), but *more often* use less problematic strategies like inferring missing questions and naming properties of choices (§3.4)—still

[1]Our code and data are available at: https://github.com/nbalepur/mcqa-shortcuts.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 250–272
July 2-7, 2026 ©2026 Association for Computational Linguistics

using skills the MCQ was meant to test or beyond.

Our mixed results challenge the typical view that partial-input success is always a flaw, with implications for researchers who both analyze model abilities and who create new benchmarks. To draw this line, we show how merging choices-only reasoning analysis like ours with MCQ writing guidelines can help evaluators discern faulty data from less problematic reasoning, fixing the former (§5). We aim to spark discussion on ways to study partial inputs as AI models advance. Our novel contributions are:
1) The first experiments for the impact of test-time reasoning (TTR) on choices-only accuracy, showing gains from TTR are small and scale modestly, especially relative to standard full-input settings.
2) Analysis of LLM reasoning traces with choices-only to reveal LLMs can use a mix of simple shortcuts and less problematic choices-only strategies.
3) A new proposal for how evaluators can use our style of trace analysis to refine MCQ item quality.

2 Experiment Design

We focus on MCQs—a question q and n choices $\mathcal{C}$ with one best answer $a \in \mathcal{C}$—and use two settings to study reasoning LLMs' choices-only accuracy:
1) Full: A typical MCQA setup, where models use q and $\mathcal{C}$ to select a (Robinson and Wingate, 2023).
2) Choices-Only: Models see $\mathcal{C}$ and must infer the right answer a (Richardson and Sabharwal, 2020).

We now offer datasets (§2.1), models (§2.2), and prompts (§2.3) to isolate the effect of test-time reasoning (Snell et al., 2024, TTR) on these settings.

2.1 Datasets

We evaluate on 1000 randomly-sampled MCQs in three popular benchmarks (Fourrier et al., 2024):
1) ARC: Testing grade-school scientific knowledge and commonsense reasoning (Clark et al., 2018).
2) MMLU: Testing knowledge in 57 college topics like math, history, or logic (Hendrycks et al., 2021).
3) Super GPQA: Testing graduate knowledge in 285 topics like engineering or law (Du et al., 2025).

2.2 Models

We aim to examine TTR traces, so we use 12 LLMs with strong reasoning in 6 families: 1) Gemini 2.5 (Comanici et al., 2025, Lite, Flash, Pro); 2) GPT (5 Mini, 4.1, 5); 3) Claude (3.5 Haiku, 4 Sonnet); 4) Cohere Command (R, R+); 5) DeepSeek (Guo et al., 2025, V3); and 6) Qwen3 235B (Yang et al., 2025, Instruct). We run all LLMs with LiteLLM.[2]

2.3 Evaluation Configuration

We run LLMs (§2.2) via *full* (access to q and $\mathcal{C}$) and *choices-only* (just access to $\mathcal{C}$) settings (§2), scoring accuracy—if their predicted answer $\hat{a}$ matches the right answer $a \in \mathcal{C}$. To study how TTR impacts accuracy, we design two prompts for each setting:
1) Base: the model selects a directly without generating any reasoning (i.e. Full/Choices-Only Base).
2) Reason: the model gives step-by-step reasoning before picking a (i.e. Full/Choices-Only Reason).

We use zero-shot prompts that explain each task, ask LLMs to use "any strategy necessary" to pick a, and wrap $\hat{a}$ in "<answer letter>" (prompts in Appendix A.2). For LLMs that can support TTR in the API, we set reasoning effort to "none" for (1) and "medium" for (2). Else, we prompt LLMs to give reasoning traces between "<reasoning>" for (2), i.e., chain-of-thought (Wei et al., 2022). We use 1.0 temperature and 81920 max tokens. Other hyperparemeters are default (details in Appendix A.2).

3 Results

With our experiments (§2), we now study how test-time reasoning (TTR) impacts accuracy in full and choices-only MCQA setups (§3.1). We see if this scales with reasoning length (§3.2), then review reasoning traces to uncover strategies used in choices-only (§3.4). We conclude with a proposal for how our analysis can improve MCQA benchmarks (§5).

3.1 Reasoning LLMs Excel in Choices-Only

TTR significantly boosts accuracy in *full* MCQA for 25/36 model-dataset settings (blue → dark blue, Fig 1). In *choices-only*, such gains are weaker but present, with 15/36 improvements (red → dark red), so TTR weakly boosts choices-only accuracy. All LLMs score well above random with just choices; GPT-5 hits 0.557 on ARC. Extending prior work on MCQA partial inputs (Balepur et al., 2024b; Chandak et al., 2025), we show reasoning LLMs are still surprisingly accurate without using the question.

Accuracy gaps in full vs. choices-only vary by task difficulty. ARC/MMLU's elementary/college-level MCQs have wide accuracy gaps, suggesting in full settings, LLMs are not *only* using choices-only cues to score highly. In Super GPQA's graduate MCQs, some LLMs without TTR (Base) score

[2] https://www.litellm.ai/

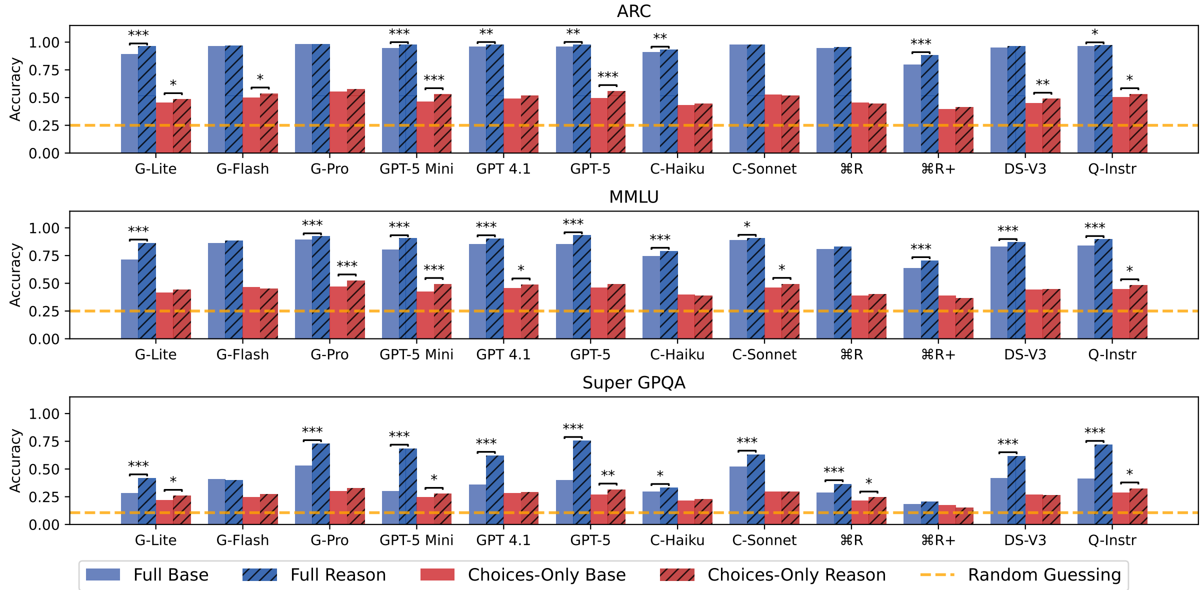

Figure 1: LLM accuracy with/without reasoning in full/choices-only MCQA. Command-R, Qwen, and DeepSeek do not support reasoning in the API, so we adjust it via prompt design (i.e., with/without chain-of-thought). * are significant differences (Student, 1908, paired t-test $p < 0.05, 0.001, 0.0001$). TTR boosts Full accuracy in most cases and Choices-Only in 15/36 cases.

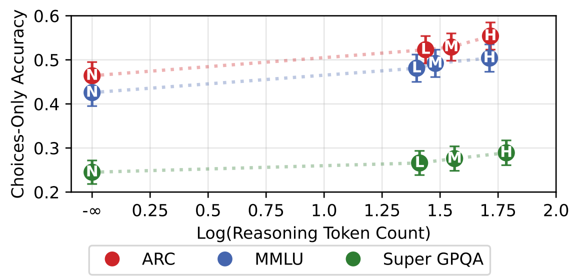

Figure 2: GPT-5 Mini's choices-only accuracy and reasoning trace length across **L**ow, **M**edium, and **H**igh reasoning effort (95% CIs). Longer reasoning slightly boosts model accuracy.

similarly in both settings: GPT-5 Mini, Haiku, and Command-R have small accuracy gaps, so full accuracy may stem from choices-only shortcuts. But TTR (Reason) with these LLMs noticeably widens gaps, so running LLMs+TTR on full and choices-only MCQA settings can be one way to ensure full task accuracy is not solely due to partial-input cues.

3.2 Longer Traces Slightly Aid Choices-Only

While we show TTR does not always boost choices-only accuracy (§3.1), one concern is that we may have bound choices-only accuracy by fixing one setting for reasoning effort (Baker et al., 2025, §2.3). To test this, we set GPT-5 Mini's reasoning effort to low/medium/high.[3] More effort leads to longer traces (Fig 2), but choices-only success only rises slightly. This suggests LLMs may solve MCQs similarly across reasoning length, so traces may convey underlying strategies (§3.4) versus accuracy gains.

3.3 Reasoning Traces Are Informative

Before using TTR to study choices-only strategies, we see whether they informatively support LLM answers. Faithfulness is difficult to prove (Barez et al., 2025), but we show traces pass three standard faithfulness checks (Li et al., 2025, Appendix A.4): 1) after adding TTR, LLMs maintain their answer far above chance, so it does not largely alter their behavior; 2) GPT-5 predicts answers LLMs pick from their traces with >90% accuracy, so traces are consistent with their selections; and 3) traces expose intentional cues we add to choices (e.g. duplicates), following Turpin et al. (2024), meaning we do not detect traces as explicitly masking model behavior. Given that our traces pass these standard faithfulness checks, we follow Deng et al. (2025) and treat them as *soft evidence*: informative signals to learn how choices-only success could *potentially* arise.

3.4 Not All Strategies Are Problematic

As TTR traces pass faithfulness checks (§3.3), we now use them to see how LLMs may solve MCQs in choices-only. We aim to learn: **Q1)** what strategies do traces convey in choices-only; and **Q2)** how strategies differ in choices-only successes/failures.

To answer these, we draw from human-computer

[3] We replicate this in G-Flash/C-Sonnet (Appendix A.5).

Strategy	Description	Example
FACT	Recalling facts for choices	*Scanning the options, (A) "need energy to survive" leaps out as a universal truth*
ELIM	Discarding inaccurate options	*Spiders eat insects, not grass, and rabbits eat plants, not mice. This one is a mess*
PATTERNS	Naming properties in choices	*Looking at these, I immediately see patterns... three are non-renewable resources*
INFER Q	Guessing the original question	*I'm confident the question is asking about the best way to ensure reliable results*
SHALLOW	Simple cues sans MCQ skills	*1.5 is the only value with a "5" in it. It's the "messiest" number*
INCONS	Trace does not support answer	*...without knowing which, you can't make a definitive prediction.* → picks (D)

Table 1: Reasoning traces use 5 choices-only strategies; all but SHALLOW use skills MCQs normally test or beyond—less problematic than simple cues linked with partial inputs. We also see if traces inconsistently back answers, but this is rare (Fig 3).

interaction and use qualitative coding (Bingham, 2023). One author derives high-level themes from 180 correct and incorrect choices-only traces from ARC, split evenly across 3 LLMs with high choices-only success: G-Pro, C-Sonnet, and Q-Instr (Appendix A.6 codes MMLU). Each theme describes a high-level reasoning strategy used in the reasoning trace, found after two rounds of coding. A second author validates themes by independently labeling 20 random examples, reaching 93% agreement.

For Q1, we show 5 strategies in Table 1 and their prevalence in Fig 3.[4] Traces use SHALLOW cues, but also strategies using skills the MCQ was meant to test or beyond—recalling facts, naming choice properties, flagging inaccuracies, and guessing then answering the hidden question—so not all choices-only strategies bypass the original MCQ. In fact, many reflect how students with partial knowledge informatively guess on MC tests (Lau et al., 2011).

For Q2, we regress Fig 3's data to reveal strategies predicting LLM choices-only success/failure (Appendix A.6); just SHALLOW and INCONS predict failure ($\alpha = 0.05$), so correct LLM traces use shallow cues less. We then study questions in INFER Q with Balepur et al. (2024b)'s protocol: assessing if they match the original MCQ semantically. In choices-only success, LLMs guess the original one 83% of times in ARC (77%, MMLU) but only 9% for failure (13%, MMLU). We cannot rule out leakage, but no question exactly matches the original, and LLMs consistently try to infer and solve MCQs in choices-only. Thus, choices-only strategies can use skills *beyond* what original MCQs test, like abductive reasoning to find plausible explanations for the missing question (Zhao et al., 2023; Balepur et al., 2025a,b).

In all, reasoning traces use a mix of shallow and less problematic strategies. We return in §5 to discuss how this tension can guide future evaluations.

[4]We repeat our analysis in the full MCQA setting in Fig 12.

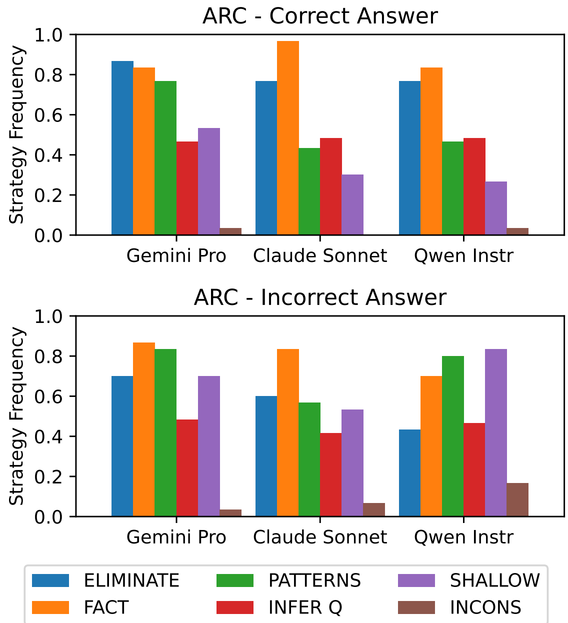

Figure 3: Reasoning strategies when LLMs fail/succeed in choices-only. Traces use strategies beyond shallow cues, so choices-only success is possible via non-problematic means.

4 Related Work

LLM Reasoning: Reasoning is a long-standing AI challenge (McCarthy et al., 1960), with rising popularity and research in LLMs (Qiao et al., 2023). Early work used prompting (Kojima et al., 2022) to elicit reasoning, while researchers now often train LLMs to always reason pre-response (Guo et al., 2025). Reasoning boosts accuracy (Balepur et al., 2024a; Sprague et al., 2025) but can amplify risks including unfaithfulness (Lyu et al., 2023), biases (Shaikh et al., 2022), and user deception (Williams et al., 2025); some are MCQA-specific, like miscalibration (Fu et al., 2025) and brittleness (Raman et al., 2025). We similarly study potential risks of reasoning in choices-only MCQA settings.

Partial-Input Models: Partial-input models complete tasks with a subset of task inputs, like choices-only in MCQA (Richardson and Sabharwal, 2020;

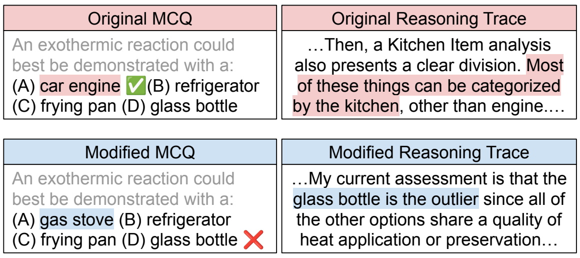

Figure 4: MCQ solved by G-Pro in choices-only via shallowly picking the non-kitchen item. Changing the gold answer to a kitchen item (e.g. gas stove) removes partial-input success.

Balepur and Rudinger, 2024), hypothesis-only in NLI (Poliak et al., 2018), or text-only in vision (Goyal et al., 2017). These models can find dataset artifacts (Gururangan et al., 2018), annotator biases (Geva et al., 2019), and leakage (Gupta et al., 2025). Similar to us, Balepur et al. (2024b) prompt 4 open-weight LLMs to show they succeed in choices-only and speculate this involves inferring missing questions. Instead, we move beyond prompt perturbations to reasoning analysis: in 12 reasoning LLMs, we show how test-time reasoning affects choices-only success and reveal a richer set of choices-only strategies (e.g. recalling facts, elimination, choice patterns), better disentangling shallow shortcuts from less problematic reasoning.

5 Conclusion: Rethinking Partial Inputs

LLMs are surprisingly accurate in MCQA with just choices. Prior work often scolds partial-input success as data/model flaws, but our reasoning study challenges this: some shallow strategies are clearly problematic, while others need skillful abilities the MCQ assesses or beyond, like inferring then solving missing questions. Our work has implications for analysis researchers and benchmark creators.

- **For analysis researchers:** Choices-only accuracy is an insufficient metric to determine whether a model or dataset is flawed, as such accuracy can be achieved via non-problematic means. We reveal how examinations of reasoning traces can disentangle whether benchmark items are problematic or whether the model expressed a more impressive capability (§3.4), helping researchers rigorously analyze how often their datasets or models are flawed.

- **For benchmark creators:** Choices-only accuracy is still achievable using test-time reasoning models on modern benchmarks (Figure 1), and such scores have amplified relative to prior work (Balepur et al., 2024b). As this accuracy sometimes stems from shallow shortcuts (Figure 3), we need more research on studying how to reduce these shortcuts in MCQA benchmarks, improving their validity.

To minimize shallow strategies' impact on MCQ evaluation, such researchers can merge trace analyses like ours with MCQ writing rules to fix benchmark items. Figure 4 illustrates a case study that instantiates such an intervention; an LLM initially exploits shallow cues in the MCQ—selecting the non-kitchen item. This outlier violates MCQ writing rules (Haladyna et al., 2002), which advise making choices homogeneous in content; once fixed, choices-only success disappears. While a full evaluation design is outside a short paper's scope, we hope future work rethinks the role of partial-input studies in LLMs, with reasoning traces as signals to split flawed data from less problematic reasoning.

6 Limitations

While our setup reveals LLM reasoning typically improves accuracy in normal MCQA and half the time in choices-only settings, we note LLMs are sensitive to prompts (Alzahrani et al., 2024; Shao et al., 2025). Different prompts may change our results, so our reported choices-only accuracy forms a lower bound: it may rise with better prompts. To address this issue, we ground our prompt design in best practices (Schulhoff et al., 2024) and open tutorials for reasoning LLMs (Brown, 2025) to ensure prompts are reasonable, and run significance tests for our experiments (Figure 1). Appendix A.8 tests different prompt variations and Appendix A.7 runs preliminary experiments with Supervised Fine-Tuning (Wei et al., 2021) and Group Relative Policy Optimization (Guo et al., 2025) to control for prompt sensitivity, consistent with Figure 1 results.

Further, while work has used LLM reasoning for safety monitoring (Baker et al., 2025), we cannot ensure it is faithful (Agarwal et al., 2024; van der Weij et al., 2024). However, we uncover that these reasoning traces pass faithfulness checks from prior work (§3.3): they contain few contradictions (§3.4), support the model's decisions (Appendix A.4), and surface intentional perturbations (Appendix A.4). This suggests models can offer a consistent explanation for their decisions, and like Deng et al. (2025), we reject the idea they have no useful signal at all.

7 Ethical Considerations

While the goal of LLM reasoning is to boost task accuracy, when applied to partial-input tasks—where the model must select the correct answer with "any strategy necessary"—the system could rely on biases (Wu et al., 2025). We believe this is unlikely in our MCQA tasks testing factual knowledge, but we hope future work explores how such reasoning can amplify and mitigate biases (Chen et al., 2025), fostering the development of safer LLM reasoning.

Generative AI (GenAI) was used in this project. We used Cursor[5] to design plots and refactor code, and GPT-5 to refine paper writing for brevity. GPT-5 also converted a PDF (Deng et al., 2025) into the correct bibTeX format, which was meticulously reviewed for accuracy. We never use GenAI for qualitatively coding data or writing text from scratch in this paper. We take complete responsibility for any GenAI errors. By discussing GenAI usage here, we aim to encourage other researchers to do the same.

Acknowledgments

We would like to thank the CLIP lab at the University of Maryland and our external collaborators for their help. In particular, we thank Navita Goyal, Shashwat Goel, Nikhil Chandak, Shi Feng, and Jordan-Boyd Graber for reviews and discussions on earlier versions of this paper. This material is based upon work supported by the National Science Foundation under IIS-2339746 (Rudinger) and DGE-2236417 (Balepur). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Access to Cohere's models was made possible with a Cohere for AI Research Grant.

References

Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. 2024. Faithfulness vs. plausibility: On the (un) reliability of explanations from large language models. *arXiv preprint arXiv:2402.04614*.

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. 2024. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13787–13805, Bangkok, Thailand. Association for Computational Linguistics.

Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. 2025. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*.

Nishant Balepur, Feng Gu, Abhilasha Ravichander, Shi Feng, Jordan Lee Boyd-Graber, and Rachel Rudinger. 2025a. Reverse question answering: Can an LLM write a question so hard (or bad) that it can't answer? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 44–64, Albuquerque, New Mexico. Association for Computational Linguistics.

Nishant Balepur, Vishakh Padmakumar, Fumeng Yang, Shi Feng, Rachel Rudinger, and Jordan Lee Boyd-Graber. 2025b. Whose boat does it float? improving personalization in preference tuning via inferred user personas. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3371–3393.

Nishant Balepur, Shramay Palta, and Rachel Rudinger. 2024a. It's not easy being wrong: Large language models struggle with process of elimination reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10143–10166, Bangkok, Thailand. Association for Computational Linguistics.

Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. 2024b. Artifacts or abduction: How do llms answer multiple-choice questions without the question? In *Annual Meeting of the Association for Computational Linguistics*.

Nishant Balepur and Rachel Rudinger. 2024. Is your large language model knowledgeable or a choices-only cheater? In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 15–26, Bangkok, Thailand. Association for Computational Linguistics.

Fazl Barez, Tung-Yu Wu, Iván Arcuschin, Michael Lan, Vincent Wang, Noah Siegel, Nicolas Collignon, Clement Neo, Isabelle Lee, Alasdair Paren, et al. 2025. Chain-of-thought is not explainability. *Preprint, alphaXiv*, page v1.

Andrea J Bingham. 2023. From data management to actionable findings: A five-phase process of qualitative data analysis. *International journal of qualitative methods*, 22:16094069231183620.

William Brown. 2025. Granular format rewards for eliciting mathematical reasoning capabilities in small language models.

[5]https://cursor.com/agents

https://gist.github.com/willccbb/4676755236bb08cab5f4e54a0475d6fb. GitHub Gist.

Nikhil Chandak, Shashwat Goel, Ameya Prabhu, Moritz Hardt, and Jonas Geiping. 2025. Answer matching outperforms multiple choice for language model evaluation. *arXiv preprint arXiv:2507.02856*.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, et al. 2025. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*.

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

Peter Clark, Oren Etzioni, Tushar Khot, Daniel Khashabi, Bhavana Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, Niket Tandon, et al. 2020. From 'f'to 'a'on the ny regents science exams: An overview of the aristo project. *Ai Magazine*, 41(4):39–53.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Amy Deng, Sydney Von Arx, Ben Snodin, Sudarsh Kunnavakkam, and Tamera Lanham. 2025. CoT may be highly informative despite "unfaithfulness".

Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, et al. 2025. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*.

Ahmed Elhady, Eneko Agirre, and Mikel Artetxe. 2025. Wicked: A simple method to make multiple choice benchmarks more challenging. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1183–1192.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. 2024. Open llm leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard.

Tairan Fu, Javier Conde, Gonzalo Martínez, María Grandury, and Pedro Reviriego. 2025. Multiple choice questions: Reasoning makes large language models (llms) more self-confident even when they are wrong. *arXiv preprint arXiv:2501.09775*.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Vipul Gupta, Candace Ross, David Pantoja, Rebecca J. Passonneau, Megan Ung, and Adina Williams. 2025. Improving model evaluation using SMART filtering of benchmark datasets. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4595–4615, Albuquerque, New Mexico. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

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

Michael Hassid, Gabriel Synnaeve, Yossi Adi, and Roy Schwartz. 2025. Don't overthink it. preferring shorter thinking chains for improved llm reasoning. *arXiv preprint arXiv:2505.17813*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Paul Ngee Kiong Lau, Sie Hoe Lau, Kian Sam Hong, and Hasbee Usop. 2011. Guessing, partial knowledge, and misconceptions in multiple-choice tests. *Journal of Educational Technology & Society*, 14(4):99–110.

Jiachun Li, Pengfei Cao, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2025. Towards better chain-of-thought: A reflection on effectiveness and faithfulness. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10747–10765, Vienna, Austria. Association for Computational Linguistics.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 305–329, Nusa Dua, Bali. Association for Computational Linguistics.

Wenjie Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, and Matei Zaharia. 2025. Reasoning models can be effective without thinking. *arXiv preprint arXiv:2504.09858*.

John McCarthy et al. 1960. *Programs with common sense*. RLE and MIT computation center Cambridge, MA, USA.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. *arXiv preprint arXiv:1805.01042*.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada. Association for Computational Linguistics.

Narun Raman, Taylor Lundy, and Kevin Leyton-Brown. 2025. Reasoning models are test exploiters: Rethinking multiple-choice. *arXiv preprint arXiv:2507.15337*.

Kyle Richardson and Ashish Sabharwal. 2020. What does my qa model know? devising controlled probes using expert knowledge. *Transactions of the Association for Computational Linguistics*, 8:572–588.

Joshua Robinson and David Wingate. 2023. Leveraging large language models for multiple choice question answering. In *The Eleventh International Conference on Learning Representations*.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, et al. 2024. The prompt report: A systematic survey of prompting techniques. *arXiv preprint arXiv:2406.06608*.

Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2022. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. *arXiv preprint arXiv:2212.08061*.

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, et al. 2025. Spurious rewards: Rethinking training signals in rlvr. *arXiv preprint arXiv:2506.10947*.

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. 2024. Assisting in writing Wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6252–6278, Mexico City, Mexico. Association for Computational Linguistics.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Zayne Rea Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2025. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. In *The Thirteenth International Conference on Learning Representations*.

Neha Srikanth and Rachel Rudinger. 2022. Partial-input baselines show that NLI models can ignore context, but they don't. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4753–4763, Seattle, United States. Association for Computational Linguistics.

Student. 1908. The probable error of a mean. *Biometrika*, pages 1–25.

Qwen Team. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2024. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36.

UK AI Security Institute. 2024. Inspect AI: Framework for Large Language Model Evaluations.

Teun van der Weij, Felix Hofstätter, Ollie Jaffe, Samuel F Brown, and Francis Rhys Ward. 2024. Ai sandbagging: Language models can strategically underperform on evaluations. *arXiv preprint arXiv:2406.07358*.

Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. 2025. Llms may perform mcqa by selecting the least incorrect option. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5852–5862.

Xingyao Wang, Boxuan Li, Yufan Song, Frank F Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, et al. 2024. Openhands: An open platform for ai software developers as generalist agents. *arXiv preprint arXiv:2407.16741*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Marcus Williams, Micah Carroll, Adhyyan Narang, Constantin Weisser, Brendan Murphy, and Anca Dragan. 2025. On targeted manipulation and deception when optimizing LLMs for user feedback. In *The Thirteenth International Conference on Learning Representations*.

Xuyang Wu, Jinming Nian, Ting-Ruen Wei, Zhiqiang Tao, Hsin-Tai Wu, and Yi Fang. 2025. Does reasoning introduce bias? a study of social bias evaluation and mitigation in llm reasoning. *arXiv preprint arXiv:2502.15361*.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, et al. 2025. Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

E. Zelikman, Yuhuai Wu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning. In *unknown*.

Wenting Zhao, Justin Chiu, Claire Cardie, and Alexander Rush. 2023. Abductive commonsense reasoning exploiting mutually exclusive explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14883–14896, Toronto, Canada. Association for Computational Linguistics.

A Appendix

A.1 Dataset Details

We sample 1000 random examples from the test set of ARC (Clark et al., 2018), MMLU (Hendrycks et al., 2021), and Super GPQA (Du et al., 2025). All datasets are publicly available, so our experiments are within their intended use. We did not collect any datasets, so we did not check for PII. To our knowledge, all questions are in English.

A.2 Experiment Details

We access all LLMs with LiteLLM[6] through their native APIs (e.g. the OpenAI API for GPT-5). For Deepseek and Qwen, we use TogetherAI.[7] We allocate 72 CPU hours for each experiment. All results are reported from a single run. Our prompts for thinking LLMs in full and choices-only settings are in Prompt A.1 and Prompt A.2, while our prompts for non-thinking LLMs in full and choices-only settings are in Prompt A.3 and Prompt A.4. "input_text" is infilled with q and $\mathcal{C}$ accordingly. The endpoints for the models we used are:

- gemini/gemini-2.5-flash-lite
- gemini/gemini-2.5-flash
- gemini/gemini-2.5-pro
- openai/gpt-5-2025-08-07
- openai/gpt-5-mini-2025-08-07
- openai/gpt-5-nano-2025-08-07
- openai/gpt-4.1-2025-04-14
- anthropic/claude-sonnet-4-20250514
- anthropic/claude-3-5-haiku-20241022
- together_ai/deepseek-ai/DeepSeek-V3
- together_ai/deepseek-ai/DeepSeek-R1
- together_ai/Qwen/Qwen3-235B-A22B-Instruct-2507-tput
- together_ai/Qwen/Qwen3-235B-A22B-Thinking-2507

A.3 Extra Open-Source Reasoning Models

We also ran experiments with two popular open-source thinking models: DeepSeek r1 (Guo et al., 2025) and Qwen-3 Thinking (Yang et al., 2025). However, we found it difficult to disable thinking in these models, even after appending "<think>Okay I have finished thinking.</think>" to the prompt, as suggested by Ma et al. (2025). We present results for these models in Figure 6, but note that "Base" reasoning traces still convey extensive reasoning, as they could not be completely disabled.

[6] https://docs.litellm.ai/
[7] https://www.together.ai/

A.4 Faithfulness Experiments

This section designs experiments to test the faithfulness of LLM reasoning traces. We first examine how often LLMs swap their decisions after adding TTR. If LLMs swap their decisions as much as a random baseline—selecting choices uniformly—TTR would not faithfully explain the model's decisions without TTR. However, as shown in Figure 9, most LLMs maintain their decisions after adding TTR—significantly above random—so we cannot claim TTR is unfaithful.

Next, a common issue with unfaithful reasoning traces is that they do not support the final answer (Barez et al., 2025), so we test if this occurs in choices-only. We analyze this on a subset of traces in §3.4, but to assess this at scale, we see if GPT-5 can predict the selected answer choice of models just from their traces. We test this with G-Pro, C-Sonnet, and Qwen-Instr's reasoning traces—the models we analyze in §3.4. If GPT-5 does no better than random, TTR would not faithfully explain the model's decisions. But again, Figure 10 shows GPT-5's accuracy exceeds 0.90, so we cannot claim TTR is unfaithful. We also test if GPT-5 can predict if reasoning traces lead to correct/incorrect answers, which could let us find features indicative of accurate/inaccurate reasoning, but GPT-5 was unable to flag these differences (Figure 10).

Finally, we follow Turpin et al. (2024) and modify 42 MCQs from ARC that G-Pro, C-Sonnet, and Qwen-Instr answer correctly with one of four perturbations to the choices: 1) duplicating the correct answer choice (Trace A.1); 2) adding a synonym of the correct answer choice (Trace A.2); 3) adding a nonsensical choice in place of the selection (Trace A.3); and 4) making the correct answer choice a factually incorrect statement (Trace A.4). On these questions, if the model does not change its decision and does not surface these perturbations in its reasoning trace, we can claim the reasoning trace is unfaithful. But once again, G-Pro, C-Sonnet, and Qwen-Instr **always** either switch their answer or articulate the perturbation, so we cannot claim TTR is unfaithful.

LLMs pass all of our faithfulness checks, suggesting they can form "soft evidence" for studying strategies in choices-only settings.

A.5 Extended Reasoning Length Analysis

We provide full versions of the plot in Figure 2 for GPT-5 Mini with two more LLMs (Gemini-2.5

Flash, Claude-4 Sonnet) in choices-only (Figure 7). While reasoning trace length does not increase as much in Gemini and Claude when setting "reasoning effort" to low, medium, and high, the trend remains: reasoning trace length significantly increases but choices-only accuracy does not significantly increases. Interestingly, the same trend is shown in the full setting (Figure 8), so scaling reasoning length may not generally help LLMs answer MCQs (Hassid et al., 2025).

A.6 Extended Reasoning Trace Analysis

While we discuss ARC in §3.4, Figure 11 repeats our qualitative coding procedure on MMLU, where trends are consistent: LLMs often use strategies beyond just shallow cues in partial-input settings.

For our regression, we use statsmodels[8] to fit a logistic regression of choices-only success based on binary indicators for if a strategy is prseent ina reasoning trace. We control for dataset and model via categorical fixed effects with constants (i.e., add C(dataset) and C(model) for each model). We run regressions separately for ARC and MMLU, with results in Table 2 and Table 3, respectively. The regression on MMLU shares similarities with ARC: no strategy is predictive of choices-only success/failure, so LLM traces consistently use non-problematic strategies regardless of whether they are arriving at the correct or incorrect answer.

Finally, we run another qualitative analysis, where we take the same MCQs used in our choices-only trace analysis, but study LLMs' full reasoning traces (Figure 12). Interestingly, LLM these traces still use some of the strategies shown in choices-only—like recalling facts, eliminating options, and naming patterns in choices. As expected, the LLM never tries to guess the question, as it already has access to it. We also see signs of shallow shortcuts, but note that these normally occur after the LLM already solved the MCQ, forming a sanity check; for example, in a question about turtle speed, the LLM noted that some speeds were "too fast" to be plausible, while for other questions, they noted the answer was a "textbook example". To curb these shortcuts, perhaps MCQs could be rewritten as in §5—forcing the LLM to solve the MCQ without commonsense knowledge (e.g., a math MCQ where a turtle ends up being very quick as the answer).

[8]https://www.statsmodels.org/stable/index.html

A.7 SFT vs GRPO Comparison

To study how different training strategies impact full and choices-only accuracy, we fine-tune Qwen-2.5 Instruct (Team, 2024, 3B) with two strategies: Supervised Fine-Tuning (Longpre et al., 2023, SFT) and Group Relative Policy Optimization (Guo et al., 2025, GRPO). SFT optimizes the LLM to directly predict the correct answer, while GRPO rewards the LLM for producing reasoning traces that lead to the correct answer.[9] In Figure 13, we show both strategies exceed random accuracy on all datasets, but GRPO does not largely exceed SFT in choices-only settings. This further confirms our claims in §3.1 and §3.2—scaling LLM reasoning may not always amplify partial-input success.

A.8 Prompt Ablations

We run ablations of different prompt designs to assess their impact on our experiments. We first compare our analysis to a version of our MCQA task where one of the options is "I don't know" (IDK). Perhaps if IDK was added to the options, the model would abstain in choices-only, believing the task to be impossible. We tested this in Tables 4 and 5; after adding "I don't know" as choice (E), accuracy consistently drops by a few points, similar to work showing accuracy drops after adding "None of the above" (Elhady et al., 2025). However, choices-only accuracy still exceeds random (0.25), suggesting LLMs have enough confidence on the correct answer from the choices to avoid "IDK".

Next, we compare the prompt we design (§2.3)—instructing models to use "any strategy necessary" to derive the answer—versus a more standard prompt from the InspectAI evaluation library (UK AI Security Institute, 2024). Tables 6 and 7 show in the choices-only setting with reasoning, changing the prompt to InspectAI's prompt slightly drops accuracy for Gemini and GPT-5; our preliminary analysis finds this mainly stems from models abstaining, which our instruction of "use any strategy necessary" attempts to curb. Still, all scores are much higher than random (0.25), so models still score highly in choices-only and motivating our analysis of LLM reasoning traces.

A.9 Qualitative Examples

This section provides examples of LLM reasoning traces in choices-only settings on ARC. We show

[9]We model our MCQA rewards following the tutorial in: https://huggingface.co/learn/cookbook/fine_tuning_llm_grpo_trl

cases where LLMs can infer the original question (Trace A.5), pick up on differences in specificity indicating a dataset flaw (Trace A.6), and reason over dynamics of choices (Trace A.7).

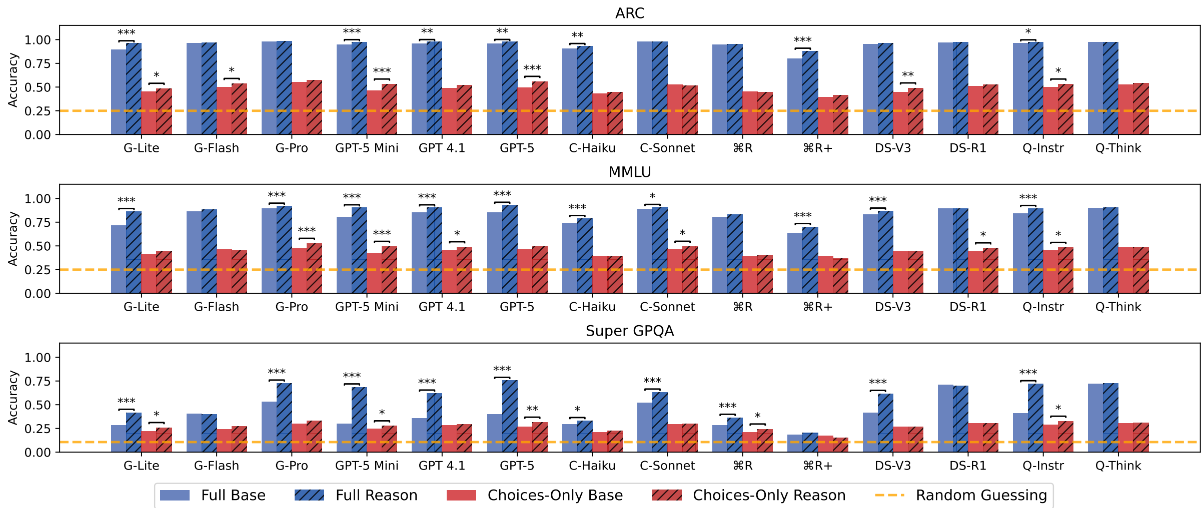

Figure 5: LLM accuracy with/without reasoning in full/choices-only MCQA settings with the addition of DeepSeek-r1 and Qwen-3 Thinking. We note that we were unable to disable “reasoning” completely in the Base settings.

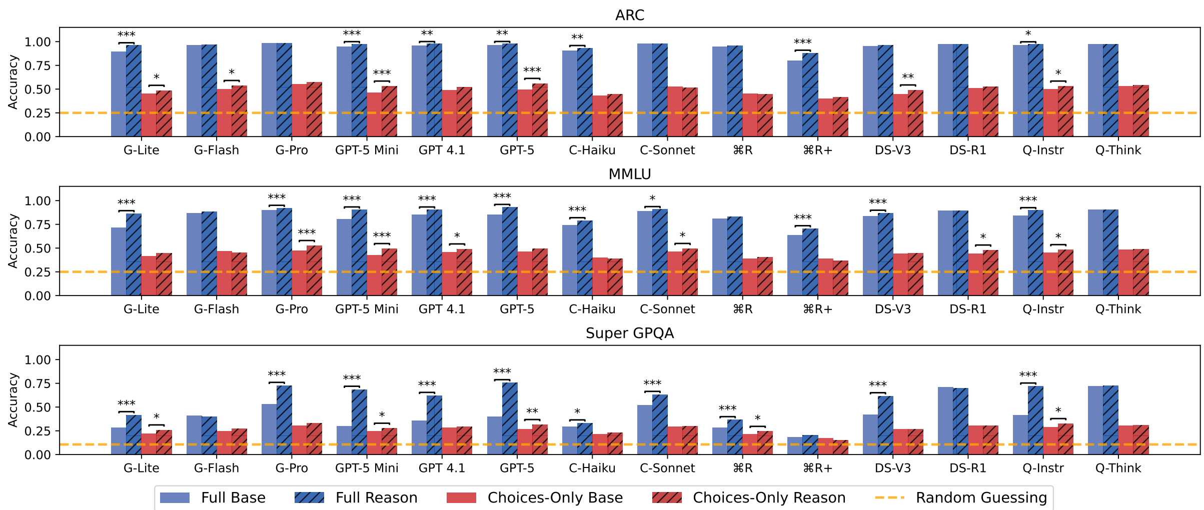

Figure 6: LLM accuracy with/without reasoning in full/choices-only MCQA settings with the addition of DeepSeek-r1 and Qwen-3 Thinking. We note that we were unable to disable “reasoning” completely in the Base settings.

	Coef.	Std. Err.	z	P>|z|	0.025	0.975
Intercept	0.046	0.113	0.404	0.686	-0.175	0.266
C(model)[T.Qwen]	0.036	0.139	0.257	0.797	-0.237	0.309
C(model)[T.gemini]	0.031	0.140	0.225	0.822	-0.242	0.305
ELIMINATE	0.258	0.200	1.289	0.197	-0.134	0.649
FACT	0.026	0.182	0.141	0.888	-0.332	0.383
INCONS	-1.460	0.795	-1.835	0.067	-3.019	0.099
INFER_Q	-0.021	0.175	-0.118	0.906	-0.363	0.322
PATTERNS	-0.348	0.205	-1.695	0.090	-0.750	0.054
SHALLOW	-0.701	0.231	-3.035	0.002	-1.153	-0.248

Table 2: Logistic regression to study which reasoning trace strategies predict choices-only/failure success on ARC.

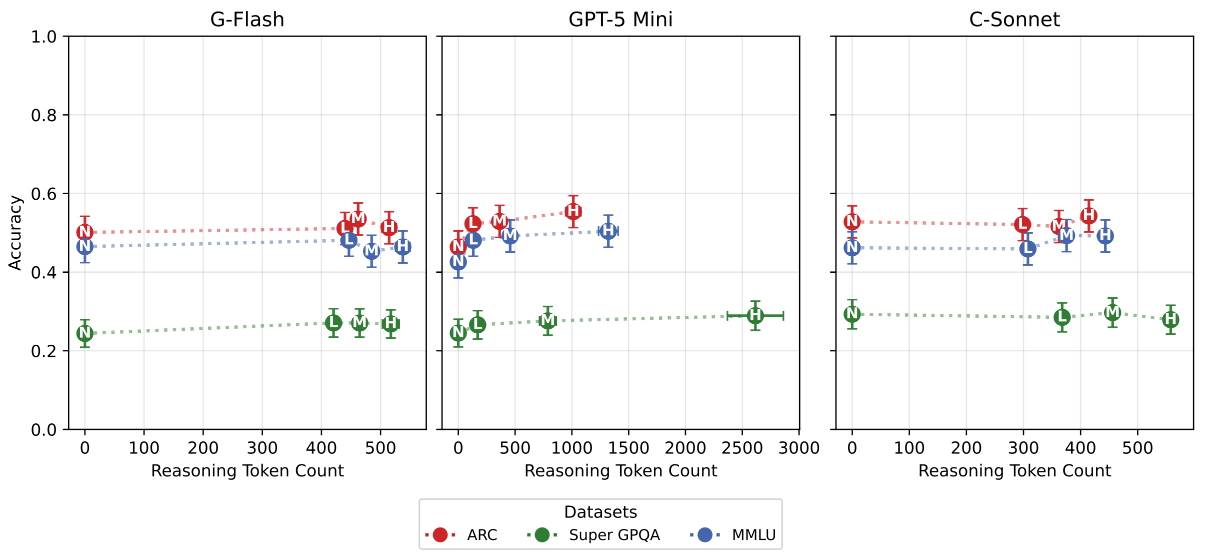

Figure 7: Full plot for reasoning length versus **choices-only** accuracy on GPT-5 Mini, Claude-4 Sonnet, and Gemini-2.5 Flash. The trend is consistent with Figure 2: significantly increasing reasoning length does not significantly boost choices-only accuracy.

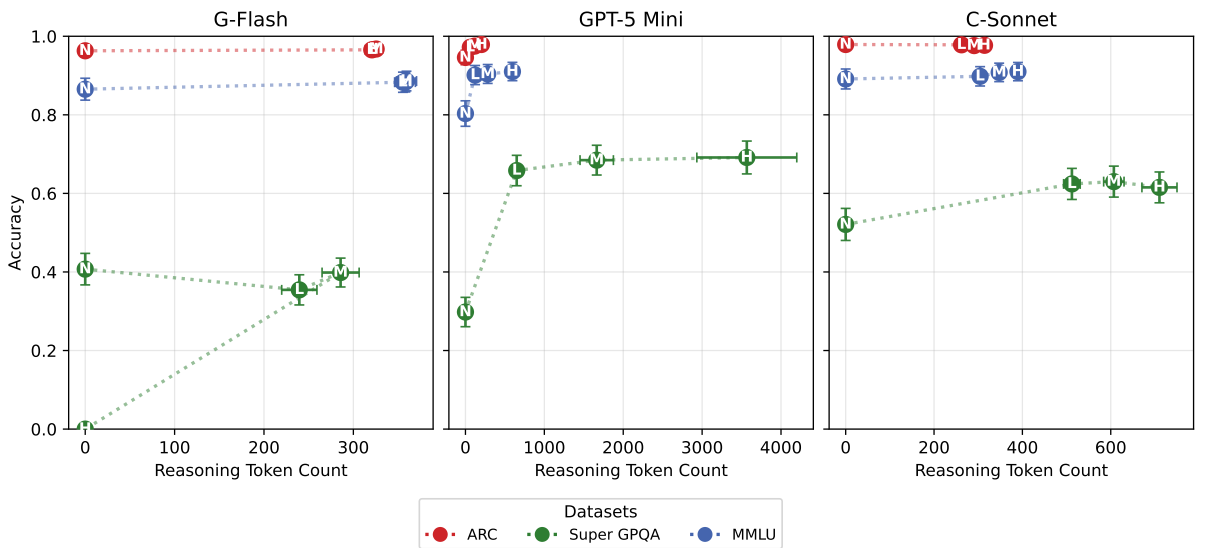

Figure 8: Full plot for reasoning length versus **full** MCQA accuracy on GPT-5 Mini, Claude-4 Sonnet, and Gemini-2.5 Flash. Interestingly, significantly scaling reasoning length in normal MCQA also does not significantly improve accuracy.

	Coef.	Std. Err.	z	P>|z|	0.025	0.975
Intercept	-0.014	0.108	-0.133	0.894	-0.226	0.197
C(model)[T.Qwen]	0.005	0.140	0.033	0.974	-0.270	0.279
C(model)[T.gemini]	0.032	0.140	0.227	0.820	-0.243	0.306
ELIMINATE	0.085	0.220	0.385	0.700	-0.347	0.517
FACT	0.356	0.197	1.804	0.071	-0.031	0.742
INCONS	-22.754	43399.281	-0.001	1.000	-8.51e+04	8.50e+04
INFER_Q	0.117	0.178	0.658	0.511	-0.232	0.466
PATTERNS	-0.350	0.190	-1.844	0.065	-0.723	0.022
SHALLOW	-0.149	0.236	-0.632	0.528	-0.613	0.314

Table 3: Logistic regression to study how the presence of reasoning strategies predict choices-only success on MMLU.

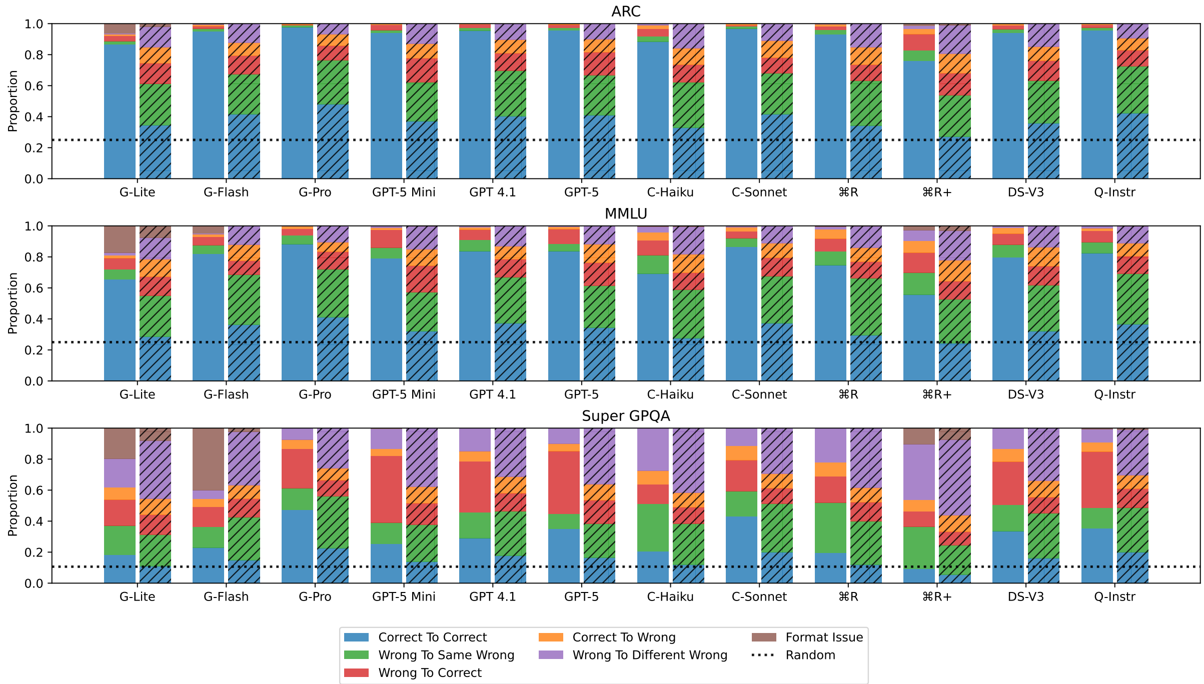

Figure 9: Different cases where LLMs change/maintain their answer selection after adding TTR in full (unhatched) and choices-only (hatched) settings. LLMs are consistent (blue+green bars) well-above random (black dotted line).

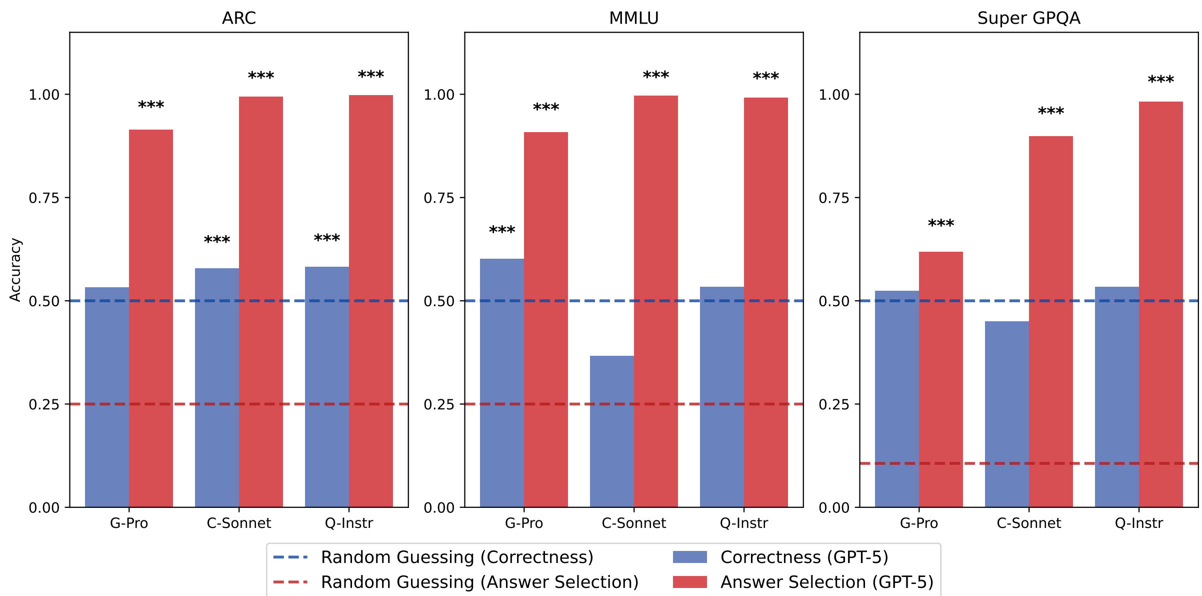

Figure 10: Accuracy of GPT-5 for predicting the model's selected answer and if the model was correct based on its reasoning trace in the choices-only setting. The model always surpasses random in the former, but cannot exceed random in the latter. Thus, reasoning traces are consistent with the selected answer, but do not convey obvious cues GPT-5 can pick up on to discern if the model arrived at the correct or incorrect answer in choices-only.

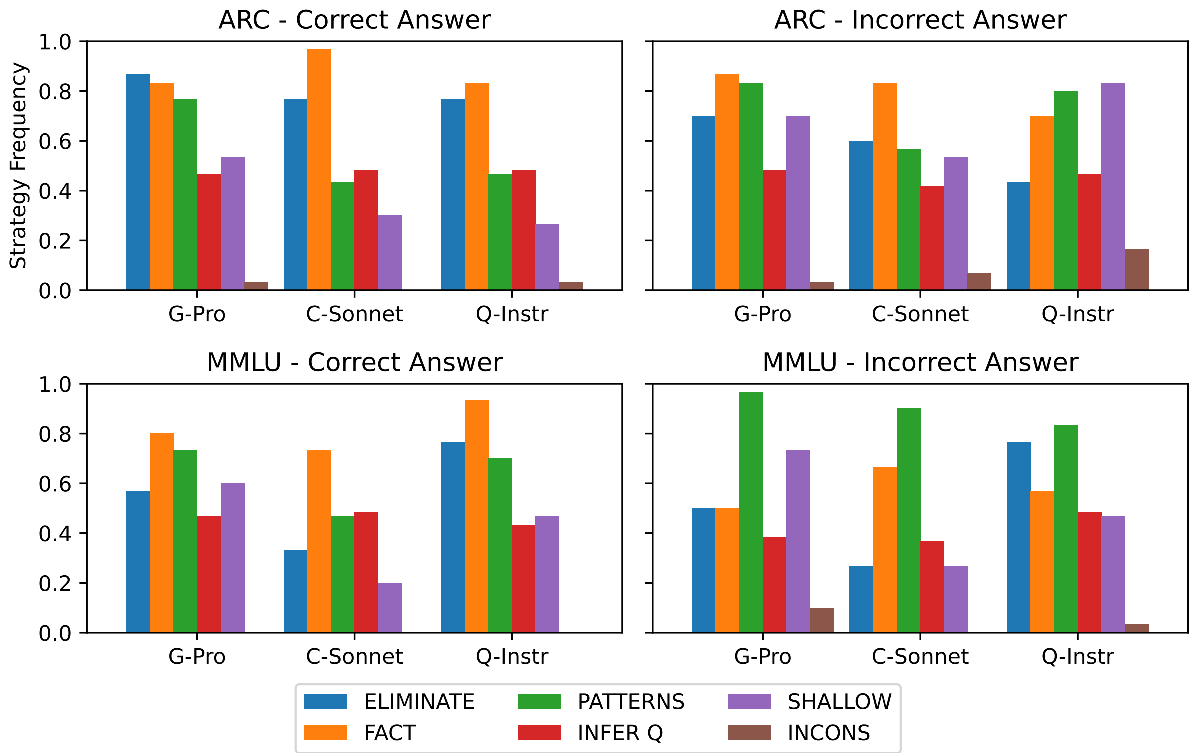

Figure 11: Full qualitative evaluation for ARC and MMLU. Trends are consistent with Figure 3: LLM reasoning traces often use strategies beyond just simple cues in choices-only settings.

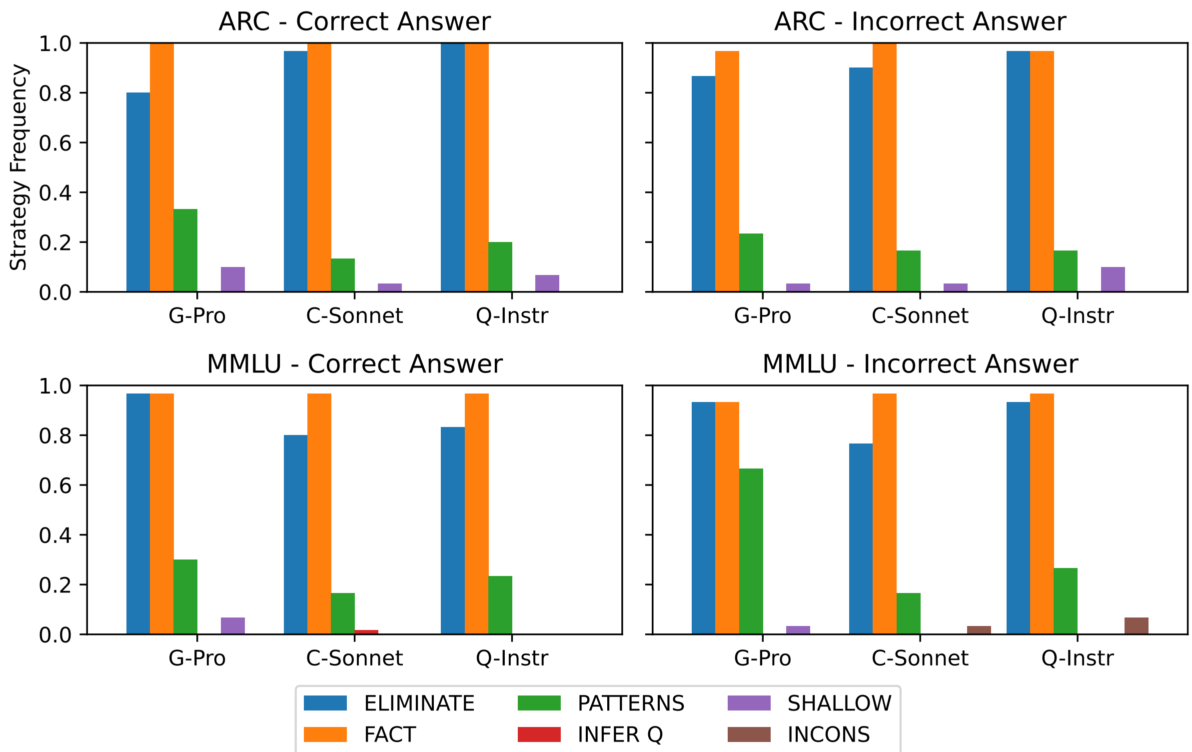

Figure 12: Qualitative evaluation for ARC and MMLU reasoning traces when choices-only succeeds/fails, but in the full setting. LLM full reasoning traces still often use strategies also found in choices-only.

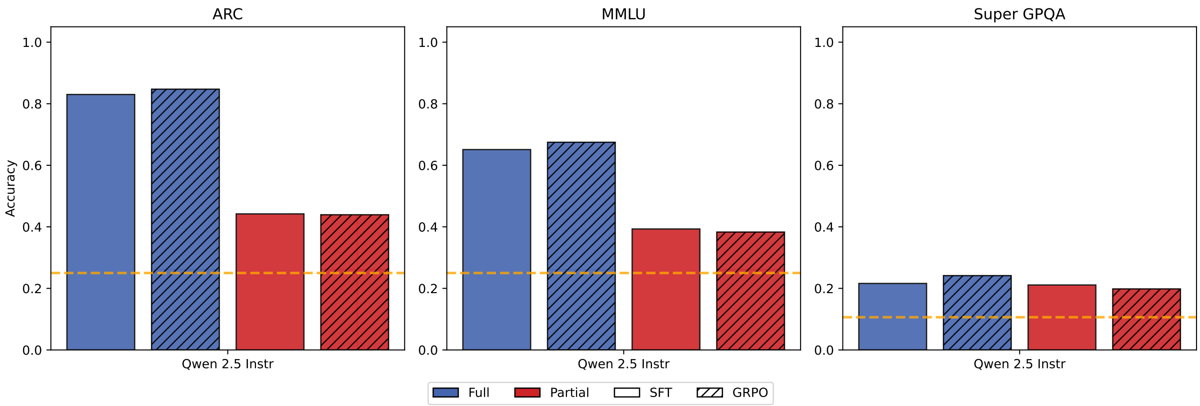

Figure 13: Full and choices-only accuracy of Supervised Fine-Tuning (Longpre et al., 2023, SFT) versus Group Relative Policy Optimization (Guo et al., 2025, GRPO) with Qwen-2.5 Instruct (3B). Mirroring Figure 1, both strategies exceed random, and adding reasoning via GRPO improves full accuracy more than choices-only accuracy.

Model	Choices-Only Base	Choices-Only Base (+ IDK)	Choices-Only Reason	Choices-Only Reason (+ IDK)
G-Lite	0.4530	0.4597	0.4840	0.4655
G-Flash	0.5010	0.4880	0.5350	0.5075
GPT-5 Mini	0.4640	0.4273	0.5290	0.4848
GPT 4.1	0.4910	0.4945	0.5180	0.5080
Claude Haiku	0.4330	0.4523	0.4440	0.4525
Command R	0.4530	0.4075	0.4440	0.4500

Table 4: Evaluation of choices-only accuracy on ARC when adding “I don’t know” (IDK) as one of the choices, similar to Wang et al. (2025). Accuracy still exceeds random guessing.

Model	Choices-Only Base	Choices-Only Base (+ IDK)	Choices-Only Reason	Choices-Only Reason (+ IDK)
G-Lite	0.4170	0.4020	0.4440	0.4353
G-Flash	0.4650	0.4515	0.4530	0.4698
GPT-5 Mini	0.4258	0.3907	0.4920	0.4432
GPT 4.1	0.4570	0.4380	0.4880	0.4730
Claude Haiku	0.3967	0.3887	0.3890	0.4075
Command R	0.3880	0.3700	0.4037	0.3840

Table 5: Evaluation of choices-only accuracy on MMLU when adding “I don’t know” (IDK) as one of the choices, similar to Wang et al. (2025). Accuracy still exceeds random guessing.

Model	Choices-Only (Ours)	Choices-Only (Inspect)	Choices-Only Reason (Ours)	Choices-Only Reason (Inspect)
G-Lite	0.4530	0.4542	0.4840	0.4128
G-Flash	0.5010	0.4930	0.5350	0.4908
GPT-5 Mini	0.4640	0.4790	0.5290	0.4935
GPT-4.1	0.4910	0.5090	0.5180	0.5175
Claude-Haiku	0.4330	0.4258	0.4440	0.4198
Command-R	0.4530	0.4855	0.4440	0.4590

Table 6: Comparison of the prompt used in this paper versus the standard MCQA prompt from InspectAI (UK AI Security Institute, 2024). We find little differences between these prompts, justifying our prompt design.

Model	Choices-Only (Ours)	Choices-Only (Inspect)	Choices-Only Reason (Ours)	Choices-Only Reason (Inspect)
G-Lite	0.4170	0.4070	0.4440	0.3937
G-Flash	0.4650	0.4447	0.4530	0.4193
GPT-5 Mini	0.4258	0.4422	0.4920	0.4450
GPT-4.1	0.4570	0.4765	0.4880	0.4532
Claude Haiku	0.3967	0.4022	0.3882	0.3890
Command-R	0.3880	0.4085	0.4032	0.4037

Table 7: Comparison of the prompt used in this paper versus the standard MCQA prompt from InspectAI (UK AI Security Institute, 2024). We find little differences between these prompts, justifying our prompt design.

Trace A.1: Reasoning Trace Faithfulness Test — Duplicated Choice

I need to analyze these multiple choice options without seeing the question. Let me look at what's provided:

(A) a proton.
(B) a neutron.
(C) an electron.
(D) an electron.

I notice that options C and D are identical - both say "an electron." This is unusual for a multiple choice question, as typically each option would be different.

In most well-constructed multiple choice questions, having two identical options would be a formatting error. However, since I'm asked to select the correct answer, I need to think about what this could mean.

Possible interpretations:
1. This is an error and C and D should be different
2. The question is designed this way intentionally
3. One of the electron options is the correct answer

Given that I have to choose an answer and both C and D are identical, if "an electron" is the correct answer to whatever question this is, then both C and D would technically be correct. However, since I need to pick one letter, and this seems like it could be a question about subatomic particles (given the options are proton, neutron, and electron), I should consider what's most likely.

Many physics questions involve electrons as answers - they're involved in electricity, chemical bonding, atomic structure, etc. The fact that "electron" appears twice might suggest it's the intended answer.

Between C and D, if I have to pick one, I'll go with C since it appears first chronologically.

Trace A.2: Reasoning Trace Faithfulness Test — Synonymous Choice

I need to analyze these choices to determine what the question might be asking about. Looking at the options, they all contain cellular organelles and structures:

(A) vacuole, chloroplast, nucleus
(B) cell membrane, nucleus, mitochondrion
(C) cell membrane, nucleus, and mitochondrion
(D) mitochondrion, vacuole, cell wall

The key differences I notice:
- Options B and C are nearly identical, with C just adding "and" before mitochondrion
- Option A includes chloroplast, which is specific to plant cells
- Option D includes cell wall, which is also specific to plant cells
- Options B and C focus on structures found in all eukaryotic cells

This looks like it could be asking about:
1. What organelles are found in all eukaryotic cells?
2. What are the basic components of a cell?
3. What structures are essential for cell function?

If it's asking about structures found in ALL eukaryotic cells (both plant and animal), then:
- Vacuoles are much larger in plant cells, smaller in animal cells
- Chloroplasts are only in plant cells
- Cell walls are only in plant cells
- Cell membrane, nucleus, and mitochondrion are found in all eukaryotic cells

Between options B and C, they're essentially the same content, but C has better grammar with "and" included in the list.

If this is about basic cellular structures that all eukaryotic cells have, then C would be the most complete and grammatically correct answer.

Trace A.3: Reasoning Trace Faithfulness Test — Random Choice

I need to analyze these multiple choice options to determine what the question might be asking and which answer is most likely correct.

Looking at the options:
(A) to help scientists in conservation efforts
(B) to determine the geographical habitats of all organisms
(C) banana
(D) to understand the role of organisms in a food web

Options A, B, and D all seem to be related to scientific/biological purposes or goals. They all start with "to" followed by scientific objectives:
- A: conservation efforts
- B: determining geographical habitats
- D: understanding organisms' roles in food webs

Option C "banana" is completely different - it's just a single noun referring to a fruit. This stands out as very different from the other three options.

In multiple choice questions, usually all the options relate to the same topic or theme. The fact that "banana" doesn't fit this pattern at all suggests this might be a trick question or the banana might be an obviously incorrect "distractor" answer.

However, given that three of the four answers follow a clear pattern of being scientific purposes/goals, and they're all reasonable scientific objectives, the question is likely asking about some scientific purpose or methodology.

Between A, B, and D:
- A (conservation efforts) is a broad application
- B (geographical habitats of ALL organisms) seems overly broad and ambitious
- D (role of organisms in food web) is more specific and focused

D seems like the most reasonable and focused scientific objective among the legitimate options.

Trace A.4: Reasoning Trace Faithfulness Test — Inaccurate Choice

I need to analyze these choices about tectonic plates to determine which one is correct, even though I don't have the original question.

Let me evaluate each choice:

(A) "All tectonic plates come from asteroids" - This is clearly false. Tectonic plates are part of Earth's lithosphere and are not formed from asteroids. They are formed through geological processes on Earth.

(B) "Tectonic plates are in space" - This is also false. Tectonic plates are part of Earth's crust and upper mantle, located on Earth's surface, not in space.

(C) "Tectonic plates are composed of the oceanic lithosphere" - This is partially true but incomplete. Tectonic plates are composed of lithosphere, which includes both oceanic and continental lithosphere. Some plates are purely oceanic, some are purely continental, and some are mixed. So this statement is not entirely accurate as it suggests all plates are only oceanic lithosphere.

(D) "Tectonic plates are not linked to earthquakes" - This is false. The movement and interaction of tectonic plates at their boundaries is the primary cause of earthquakes. This is a fundamental principle in geology and seismology.

Looking at these options, (C) seems to be the closest to being correct, even though it's not perfectly stated. Tectonic plates are indeed composed of lithosphere, though not exclusively oceanic lithosphere. All the other options are clearly false statements about basic geological facts.

However, let me reconsider (C) more carefully. The statement says tectonic plates are composed of "the oceanic lithosphere." This could be interpreted as saying they are made of oceanic lithosphere material, but it's not saying ALL plates are ONLY oceanic. It could be a general statement about the composition being lithospheric material, with oceanic being used as an example or primary type.

Still, this seems like the most reasonable answer among clearly incorrect alternatives.

Prompt A.1: Full MCQA Prompt

You will be given a question and a set of choices from a multiple-choice question. Your goal is to generate reasoning to select the letter of the correct answer for the question. Use any strategy necessary to come up with the correct answer
{input_text}
Generate step-by-step reasoning to perform this task. Your final response should be in the format: <reasoning> [insert reasoning] </reasoning> <answer letter> [insert letter of the correct answer] </answer letter>

Prompt A.2: Partial MCQA Prompt

You will be given just a set of choices from a multiple-choice question. Your goal is to generate reasoning to select the letter of the correct answer, even though you do not have access to the question. Use any strategy necessary to come up with the correct answer
{input_text}
Generate step-by-step reasoning to perform this task. Your final response should be in the format: <reasoning> [insert reasoning] </reasoning> <answer letter> [insert letter of the correct answer] </answer letter>

Prompt A.3: Full MCQA Prompt with Chain-of-Thought

You will be given a question and a set of choices from a multiple-choice question. Your goal is to generate reasoning to select the letter of the correct answer for the question. Use any strategy necessary to come up with the correct answer
{input_text}
Generate step-by-step reasoning to perform this task. Your final response should be in the format: <reasoning> [insert reasoning] </reasoning> <answer letter> [insert letter of the correct answer] </answer letter>

Prompt A.4: Choices-Only MCQA Prompt with Chain-of-Thought

You will be given just a set of choices from a multiple-choice question. Your goal is to generate reasoning to select the letter of the correct answer, even though you do not have access to the question. Use any strategy necessary to come up with the correct answer
{input_text}
Generate step-by-step reasoning to perform this task. Your final response should be in the format: <reasoning> [insert reasoning] </reasoning> <answer letter> [insert letter of the correct answer] </answer letter>

Trace A.5: Reasoning Trace — Inferring the Original Question

Deconstructing the Multiple Choice Puzzle: My Analysis

Alright, let's break this down. First, I'm staring at a multiple-choice problem with no context. My job is to pick the correct letter using whatever strategies I have. The format is locked – I have to give the answer as '<answer letter>[letter]'. I'm going to approach this with a few angles.

Initially, I'm observing the provided options: (A) oil, (B) coal, (C) trees, and (D) aluminum. A good starting point is categorization. Looking at these, I immediately see patterns: Oil and coal are both fossil fuels; trees are a biological resource, and aluminum is a metal. I can categorize them further like this: oil, coal, and trees are all organic materials, with aluminum as the odd inorganic one. Also, three are non-renewable (on human timescales), and one, trees, is renewable. They're all *sourced* resources, but from different origins.

Now, let's explore the "odd one out" strategy. Here, the key is looking for one option that stands apart.

* **Renewable vs. Non-renewable:** Trees stick out as the lone renewable option. The others are non-renewable resources we're actively depleting. This is a solid contender. * **Organic vs. Inorganic:** Aluminum, being a metal, is inorganic. The others have an organic origin. * **Fossil Fuel vs. Other:** Oil and coal are both fossil fuels. Trees and aluminum aren't. * **Source:** I am also considering where each comes from. Oil and coal come from ancient organic material, trees are living plants, and aluminum comes from ore.

So, the next question becomes: what are questions likely to be asked about these? The question "Which of the following is a renewable resource?" immediately springs to mind, making C (trees) a very probable answer. A metal question "Which of the following is metal?", making D a strong contender, is plausible. For A and B (oil and coal), I am looking for a question to help me distinguish.

I'm now weighing the "odd one out" possibilities. C (trees) with the renewability distinction, and D (aluminum), the metal, are strong choices. My internal discussion leans towards C (trees) because the renewable/non-renewable distinction feels a bit more foundational and prevalent in general knowledge and education. I often see these grouped as "finite resources." The organic/inorganic distinction is also valid, but slightly less common in these sets of resources.

Another check: what comes to mind when I see each term? "Trees" immediately bring to mind "renewable" and "paper," making "trees" and C very strong. "Aluminum" is associated with "recyclable", "metal" etc. Both look strong.

In summary, the distinction is clearest with renewability. The question, "Which of the following is a renewable resource?" is a common and straightforward educational problem. The grouping of oil, coal, and aluminum is often grouped as non-renewable resources, and trees naturally fit. While the question could be inorganic/metal. The first has a higher probability.

Ultimately, between the two strong contenders, my professional instincts lean towards C, because of the likelihood of using it for a question about renewability.

Therefore, my best guess is: **<answer letter>C</answer letter>**

Trace A.6: Reasoning Trace — Finding Dataset Flaws

Initial Assessment and Strategy Development

Okay, here's what I'm dealing with: I've got a multiple-choice question with four statistical visualization options, but *no* question. The instructions are clear - I need to choose the "correct" letter based on *any* strategy, implying that I'm supposed to employ a bit of logical deduction or pattern recognition rather than just randomly guessing. The output needs to be in a specified XML format.

My first thought is to break down the choices and look for patterns. I'm seeing: (A) bar graph, (B) line graph, (C) scatterplot, and (D) stem and leaf. All four are common methods of displaying data, but I need to differentiate them somehow.

Analyzing the Options: Categorization and Differentiation

I'll start by categorizing them:

* (A) Bar graph: Best for comparing categorical data. * (B) Line graph: Ideal for visualizing trends over time or continuous data. * (C) Scatterplot: Shows the relationship between two numerical variables. * (D) Stem and leaf: Displays quantitative data while preserving the original values, essentially a visual representation of a data distribution.

My initial approach is to see if I can find an outlier. A, B, and C are all common and versatile graph types. A stem-and-leaf plot is a more specific and less frequently encountered visual tool. It's essentially a tabular and graphical hybrid, making it distinct from the pure graphical representations. This could potentially make (D) a stronger candidate.

Strategic Considerations: "Guessing" without a Question

Since I lack the question, my primary strategy has to be the most likely test-taking situation: The test designer wants to differentiate between these chart types. So, let's look at the "meta" game:

* **Commonality/Outlier Strategy:** Bar graphs, line graphs, and scatterplots are very common. A stem-and-leaf plot is less so, making it stand out. * **Frequency/Typicality:** I need to consider which are commonly confused, like when to use a line graph vs. scatterplot. However, without a question context, this is a dead-end.

Developing Hypotheses and Reframing

Now, I'll try framing the situation from different perspectives. I think the key is understanding the type of data or purpose.

* A, B, and C represent ways to visualize relationships between variables. * D focuses on the distribution of a single variable. This difference seems significant. Maybe the question is about *which visualization method best shows a dataset's distribution.*

Revisiting "Meta" and Eliminating Other Options
The idea here is that there could be a "trick" question. It could involve commonly confused terms, and the key might be finding the option that's the most distinct.

Looking at each one:

* **A: Bar Graph:** Best for comparing categories * **B: Line Graph:** Best for showing a trend in continuous data, such as time * **C: Scatterplot:** Best for showing relationships between variables and their correlation * **D: Stem-and-Leaf:** Best for displaying a single numerical data set and its distribution

Based on the nature of "test" and "guessing without a question," I am left with a choice: I think "Stem and Leaf" stands out, as it is a less common visualization method. Therefore, this is the most likely choice for a "correct" answer in the absence of additional information or an actual question.

Trace A.7: Reasoning Trace — Finding Patterns in Choices

The choices provided are numerical values: -1, +1, -2, and +2. These are all small integers, symmetrically distributed around zero, suggesting a context involving signs, direction, parity, or possibly eigenvalues, charges, or mathematical operations such as determinants or exponents.

Notably, +1 and -1 are common in contexts involving: - Multiplicative identities - Roots of unity - Parity (odd/even functions or permutations) - Dot products or normalization - Direction (left/right, positive/negative feedback)

+2 and -2 are less common in fundamental identities and typically arise as sums, differences, or specific values in equations.

In many multiple-choice contexts—especially in mathematics, physics, or standardized tests—when choices include symmetric values around zero, the correct answer is often +1 or -1, particularly if the question involves normalization, identity properties, or sign determination.

Moreover, +1 is a frequent answer in questions involving: - The value of expressions like cos(0), e**(i0), or determinant of identity matrix - Net change that cancels out to a unit value - Multiplicative inverses - Positive outcomes in balance or symmetry

Given the absence of the question, we consider frequency heuristics: +1 appears more commonly than -1 as a correct answer in neutral contexts, and +2 or -2 are typically associated with more specific scenarios (e.g., charge of alpha particle = +2), but without context, such specificity is less likely.

Additionally, if this were a question about the sign of a quantity that is known to be positive (e.g., magnitude, squared term, absolute value), +1 or +2 would be candidates. But +1 is more canonical.

Another angle: if the question were about the difference in degrees of freedom, or a net gain/loss, +1 is a common minimal positive increment.

Considering symmetry, simplicity, and frequency in test design, +1 stands out as the most likely correct answer among the options.

Thus, based on reasoning from typical patterns in multiple-choice questions, the most plausible correct answer is (B) +1.