No Question, No Passage, No Problem: Investigating Artifact Exploitation and Reasoning in Multiple-Choice Reading Comprehension

Anonymous Author(s)

Abstract

Large Language Models (LLMs) can achieve above majority baseline performance on natural language processing (NLP) tasks even when deprived of parts of the input, raising concerns that benchmarks reward artifacts rather than reasoning. Prior work has demonstrated this phenomenon in multiple-choice QA and natural language inference, but not in multiple-choice reading comprehension (MCRC), where both a passage and question are integral to the task. We study MCRC under a stricter ablation, removing both passage and question to leave only the answer options. Despite this severe ablation, models consistently exceed majority baselines across five benchmarks. To probe how such accuracy arises, we introduce two reasoning-based strategies: Process-of-Elimination, which iteratively discards distractors, and Abductive Passage Inference, which infers a context to justify an option. Both strategies closely track choices-only accuracy, suggesting that strong performance reflects genuine reasoning procedures rather than dataset artifacts alone.

1 Introduction

Reading comprehension (RC) has long served as a core test of language understanding for humans and machines [3, 42, 44]. For humans, reading enables knowledge acquisition and reasoning [5, 25]; in NLP, RC has become a natural proxy for evaluating model competencies [3, 46]. Open-ended RC, however, is costly to grade and often subjective, complicating large-scale, reliable evaluation [17, 19].

Multiple-choice formats mitigate these issues by fixing a candidate set and enabling efficient, objective scoring [6, 22]. As a result, multiple-choice reading comprehension (MCRC) plays a central role in LLM evaluation, pairing RC's cognitive depth with practical scoring [40].

Yet improved scores may reflect dataset artifacts: superficial cues that allow success without genuine comprehension [12]. Partial-input studies show above-chance performance when critical components are withheld (e.g., hypothesis-only in NLI; passage- or question-only in RC/VQA) [18, 24, 36, 41, 47]. However, training dedicated partial-input models is impractical, motivating inference-time probes. Balepur et al. propose partial-input prompting, showing that LLMs can exceed majority baselines with choices only, and advance Abductive Question Inference as an alternative explanation for such gains [2].

We extend partial-input prompting to MCRC under a stricter ablation: we remove the question and passage and the model receives only the answer options, removing two thirds of the intended input. All evaluations are zero-shot and use closed-source LLMs common in practice. To probe how accuracy arises, we test two reasoning strategies: Process-of-Elimination (PoE), which iteratively discards distractors, and Abductive Passage Inference (API), which synthesizes a plausible passage and then answers against it. Both closely track choices-only performance, indicating that elevated partial-input accuracy need not stem solely from brittle artifacts; models appear able to organize

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

option-set signals into usable structure. This motivates broader study of reasoning strategies under ablation to separate shallow shortcuts from genuine inference in modern LLMs.

2 Related Works

2.1 Dataset artifacts

Benchmarks can contain shortcuts that let models succeed without true comprehension, inflating headline scores [8, 47]. Such artifacts arise from annotation habits and templates [13, 16, 33] and, increasingly, from quirks in synthetic data [49]. Evidence spans many tasks in which systems perform well with only a subset of the input, such as hypothesis-only in NLI and passage-/question-only in RC and VQA, indicating that option sets or prompts can leak label information [15, 18, 24, 36, 41, 43].

2.2 Probing and detecting artifacts

Partial-input testing removes components (e.g., passage or question) to measure residual signal [24, 36]. Contrast sets and controlled perturbations provide complementary stress tests by minimally editing inputs or labels; robust models should flip predictions when semantics flip, yet often do not [10, 11, 14, 21, 30, 32, 45]. Mitigation attempts, which are adversarial or debiasing objectives, revised collection protocols, and context alterations, show mixed effectiveness and dataset sensitivity [4, 9, 27, 37, 43]. For modern LLMs, partial-input prompting turns artifact diagnosis into an inference-time probe and already yields above-majority choices-only accuracy in MCQA [2]; our work transfers this probe to stricter MCRC ablations.

2.3 Reasoning in MCRC

Multiple-choice reasoning involves integrating evidence while suppressing distractors; even humans benefit from elimination strategies [38, 39]. In LLMs, Process-of-Elimination (PoE) prompting changes decision dynamics and can improve full-input accuracy [1, 29], while sensitivity to option ordering suggests that choice-set structure itself shapes predictions [35]. Abductive Passage Inferenceapproaches ask models to hypothesize latent explanations and then answer conditioned on them, improving reliability on reasoning tasks [23]. Closest to our setup, Balepur et al. show that Abductive Question Inference can match choices-only performance in MCQA, implying that high partial-input scores need not stem solely from brittle artifacts [2].

3 Choices-Only Evaluation

3.1 Task and Input

We formulate our target problem as a zero-shot multiple-choice reading comprehension task. Each instance consists of a passage P, a question Q about the passage, and a fixed set of four candidate answers $C = \{A, B, C, D\}$. The model must select exactly one option from this set.

In our partial input ablations, we evaluate on full-input and choices-only prompts. The full-input condition, in which the model receives $P + Q + C$, serves as the reference point. In the choices-only condition, we omit both the passage and the question, providing only C, eliminating about two-thirds of the original signal as opposed to prior work that ablates either P or Q alone (removing roughly half the input).

Full prompt design and structure can be found in Appendix A.1.

3.2 Datasets

We evaluate our ablation prompts across four established passage-based MCRC benchmarks, chosen for their varied domain focus, difficulty, and community usage.

QuALITY (Easy / Hard). QuALITY is a long-form reading comprehension dataset featuring passages with average token lengths of roughly 5,000. We report results separately on the Easy and Hard splits: the Easy subset contains questions answerable with minimal inference, while the Hard split tests deeper reasoning across the entirety of lengthy passages [7].

RACE High. RACE consists of English reading comprehension exams used in Chinese middle and high schools [26]. We focus exclusively on the High School subset in our evaluations.

ReClor. ReClor is a reading comprehension and logical reasoning benchmark extracted from the GMAT and LSAT exams [48]. We report results on the development set, as the gold labels are not public for the test set.

LogiQA 2.0. LogiQA 2.0 is another reading comprehension and logical reasoning benchmark constructed from Chinese Civil Service Examination questions [28]. We report results on the development set, as the gold labels for the test set are not public.

3.3 Models

We evaluate three state-of-the-art closed-source LLMs spanning varying architectures and scales: gpt-3.5-turbo-0125 and gpt-4o-2024-08-06, which represent popular, general models, as well as o3-2025-04-16 as a high-end reasoning model. Each model is accessed through the OpenAI API with a temperature of 0.0 and max_tokens set to None, and all other parameters left at their default values. We run three independent replicates per model and ablation, reporting the average accuracy across runs.

4 Hypothesis Evaluation

This section introduces two hypothesized mechanisms that may allow large language models to exceed majority-class performance even when both the passage P and question Q are withheld in multiple-choice reading comprehension (MCRC) benchmarks.

4.1 Process-of-Elimination Reasoning

Rationale. Even when deprived of semantic context, an LLM might still exploit world knowledge, stylistic cues, or answer-set regularities to discard unlikely distractors in a stepwise fashion. If the model can reliably isolate and remove implausible options, it could converge on the correct answer without ever reconstruing the missing passage or question. Our approach is inspired by prior work on Process-of-Elimination (PoE) prompting [1, 29], but importantly, these studies did not examine elimination under partial-input conditions, where artifact exploitation can be revealed most directly. Full prompt design and structure can be found in Appendix A.2

4.1.1 Abductive Passage Inference

Rationale. A more ambitious mechanism posits that a language model can generate a short passage whose content privileges one of the candidate answers, and then resolve the multiple-choice item against this synthetic context. This aligns with the classical notion of abduction as inference to the best explanation [34], but here instantiated in generative form: the model attempts to hypothesize a missing passage that would render one option most plausible. Success in this setting would suggest a capacity for substantive generative reasoning beyond elimination or question reconstruction.

Full prompt design and structure can be found in Appendix A.3

5 Results

In Figure 1, we observe a consistent pattern in the zero shot setting across all MCRC benchmarks. Full-input remains strongest for every model and dataset. For choices-only, this is a stronger ablation than prior MCQA studies that remove only one component, so one would expect a sharper drop. Yet, despite ablating both the passage and the question, the choices only prompt still attains accuracy that is clearly above the majority baseline in the vast majority of model and dataset combinations. The ordering by model capacity that appears under full input also appears under ablation, with o3 above gpt-4o above gpt-3.5-turbo in nearly every panel, which indicates that the competencies that drive full input gains also transfer to settings where only the options are available.

Our two hypothesis prompts closely track the choices only condition. In Figures 2 and 3, Abductive Passage Inference and Process of Elimination all lie in a narrow band around choices only on every dataset. In several panels the abductive variants slightly exceed choices only, while on others they are essentially indistinguishable. This parity is important. If choices only success were driven primarily by brittle lexical artifacts, one would expect large divergences when the prompt requires additional reasoning structure. Instead, the abductive procedures preserve the same level, which suggests that the information that supports decisions from options alone is robust to how it is organized at inference time.

6 Analysis

The primary empirical surprise is the strength of choices only in zero shot, even though two thirds of the input is removed. A natural first interpretation is artifact use. In curated multiple choice items, options are not arbitrary strings. They encode topical constraints, entailment relations, role structure, and stylistic signatures that distinguish correct answers from foils. If a model exploits these surface regularities, it can outperform a majority baseline without reading the passage or the question.

Our subsequent probes refine this explanation. Abductive Passage Inference reaches similar accuracy to the choices-only prompt while explicitly requiring the model to construct and then use a short hypothesized context. This shows that a large share of the choices only accuracy can be reproduced by a reasoning procedure that is coherent with the task definition, namely, infer a small set of latent contexts that make the options jointly coherent, then select the option that best fits those contexts. In this view, choices only is not only a test of artifact sensitivity, it is also a test of how well a model can aggregate constraints that are implicit in the option set into a decision. The persistence of capacity ordering under ablation supports this interpretation, stronger models carry richer priors and better calibrated judgments about option plausibility, and these capabilities help both when a passage is present and when it is absent.

Process of Elimination can sit near choices only under partial input because it leverages the same plausibility signals while reducing noise in the comparison. Many option sets contain one answer that is a weak outlier under broad background knowledge, and removing that competitor increases the effective separation among the remaining options and moves the ranking toward the ordering that a choices only scorer would already favor. Relative comparisons among options also expose small inconsistencies that are not tied to any specific passage but still track general world and linguistic knowledge. An option that is slightly less compatible with most plausible readings will be screened out, which leaves a smaller set that is easier to judge with the same cues used by choices only. The shift from four competing options to a tighter set further reduces variance in the final choice. Small spurious features have less chance to dominate once an obviously weak option is gone, so the decision aligns with the stable part of the model's prior over plausible answers. These effects do not require strong assumptions about memorization. They rely on signals present in the options themselves and on calibrated priors about what answers tend to look like, which is why we believe Process of Elimination tracks choices only in the severe ablation setting.

7 Conclusion

Despite removing two thirds of the input in a zero-shot regime, large language models achieved strong choices-only accuracy across benchmarks. By testing Process-of-Elimination and Abductive Passage Inference hypothesis, we showed that reasoning-based strategies can reproduce choices-only performance, suggesting that models can actively reason with limited input information rather than relying solely on superficial cues.

8 Limitations and Future Work

Our experiments are limited to proprietary GPT-series models, which prevents deeper white-box analyses of token-level logits or attention patterns. Future work can replicate our ablations on open-weight and Mixture-of-Experts models.

In addition, LLM performance is sensitive to prompt design and hyperparameters [31]. We used simple zero-shot prompts with default settings, leaving open the possibility that tuning could further raise choices-only accuracy or alter the relative strengths of reasoning strategies.

References

[1] Nishant Balepur, Shramay Palta, and Rachel Rudinger. It's not easy being wrong: Large language models struggle with process of elimination reasoning, 2024. URL https://arxiv.org/abs/2311.07532.

[2] Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. Artifacts or abduction: How do llms answer multiple-choice questions without the question?, 2024. URL https://arxiv.org/abs/2402.12483.

[3] Razieh Baradaran, Razieh Ghiasi, and Hossein Amirkhani. A survey on machine reading comprehension systems, 2020. URL https://arxiv.org/abs/2001.01582.

[4] Yonatan Belinkov, Adam Poliak, Stuart M. Shieber, Benjamin Van Durme, and Alexander M. Rush. On adversarial removal of hypothesis-only bias in natural language inference, 2019. URL https://arxiv.org/abs/1907.04389.

[5] Reese Butterfuss, Jasmine Kim, and Panayiota Kendeou. Reading comprehension, 01 2020. URL https://oxfordre.com/education/view/10.1093/acrefore/9780190264093.001.0001/acrefore-9780190264093-e-865.

[6] Nikhil Chandak, Shashwat Goel, Ameya Prabhu, Moritz Hardt, and Jonas Geiping. Answer matching outperforms multiple choice for language model evaluation, 2025. URL https://arxiv.org/abs/2507.02856.

[7] Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers, 2021. URL https://arxiv.org/abs/2105.03011.

[8] Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. Shortcut learning of large language models in natural language understanding, 2023. URL https://arxiv.org/abs/2208.11857.

[9] Yanai Elazar, Bhargavi Paranjape, Hao Peng, Sarah Wiegreffe, Khyathi Raghavi, Vivek Srikumar, Sameer Singh, and Noah A. Smith. Measuring and improving attentiveness to partial inputs with counterfactuals, 2024. URL https://arxiv.org/abs/2311.09605.

[10] Shi Feng, Eric Wallace, and Jordan Boyd-Graber. Misleading failures of partial-input baselines, 2019. URL https://arxiv.org/abs/1905.05778.

[11] Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets, 2020. URL https://arxiv.org/abs/2004.02709.

[12] Matt Gardner, William Merrill, Jesse Dodge, Matthew E. Peters, Alexis Ross, Sameer Singh, and Noah A. Smith. Competency problems: On finding and removing artifacts in language data, 2021. URL https://arxiv.org/abs/2104.08646.

[13] Mor Geva, Yoav Goldberg, and Jonathan Berant. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets, 2019. URL https://arxiv.org/abs/1908.07898.

[14] Max Glockner, Vered Shwartz, and Yoav Goldberg. Breaking nli systems with sentences that require simple lexical inferences, 2018. URL https://arxiv.org/abs/1805.02266.

[15] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, 2017. URL https://arxiv.org/abs/1612.00837.

[16] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data, 2018. URL https://arxiv.org/abs/1803.02324.

[17] Owen Henkel, Adam Boxer, Libby Hills, and Bill Roberts. Can large language models make the grade? an empirical study evaluating llms ability to mark short answer questions in k-12 education, 2024. URL https://arxiv.org/abs/2405.02985.

[18] Christine Herlihy and Rachel Rudinger. Mednli is not immune: Natural language inference artifacts in the clinical domain, 2021. URL https://arxiv.org/abs/2106.01491.

[19] Andrea Horbach, Itziar Aldabe, Marie Bexte, Oier Lopez de Lacalle, and Montse Maritxalar. Linguistic appropriateness and pedagogic usefulness of reading comprehension questions. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1753–1762, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.217/.

[20] Alexander Hoyle, Rupak Sarkar, Pranav Goel, and Philip Resnik. Natural language decompositions of implicit content enable better text representations, 2025. URL https://arxiv.org/abs/2305.14583.

[21] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems, 2017. URL https://arxiv.org/abs/1707.07328.

[22] Yiming Ju, Yuanzhe Zhang, Zhixing Tian, Kang Liu, Xiaohuan Cao, Wenting Zhao, Jinlong Li, and Jun Zhao. Enhancing multiple-choice machine reading comprehension by punishing illogical interpretations. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3641–3652, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.295. URL https://aclanthology.org/2021.emnlp-main.295/.

[23] Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. Maieutic prompting: Logically consistent reasoning with recursive explanations, 2022. URL https://arxiv.org/abs/2205.11822.

[24] Divyansh Kaushik and Zachary C. Lipton. How much reading does reading comprehension require? a critical investigation of popular benchmarks, 2018. URL https://arxiv.org/abs/1808.04926.

[25] Panayiota Kendeou, Paul van den Broek, Annemarie Helder, and Josefin Karlsson. A cognitive view of reading comprehension: Implications for reading difficulties. *Learning Disabilities Research & Practice*, 29(1):10–16, 2014. doi: 10.1111/ldrp.12025.

[26] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations, 2017. URL https://arxiv.org/abs/1704.04683.

[27] Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. Wanli: Worker and ai collaboration for natural language inference dataset creation, 2022. URL https://arxiv.org/abs/2201.05955.

[28] Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang. Logiqa 2.0—an improved dataset for logical reasoning in natural language understanding. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 31:2947–2962, July 2023. ISSN 2329-9290. doi: 10.1109/TASLP.2023.3293046. URL https://doi.org/10.1109/TASLP.2023.3293046.

[29] Chenkai Ma and Xinya Du. Poe: Process of elimination for multiple choice reasoning, 2023. URL https://arxiv.org/abs/2310.15575.

[30] R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference, 2019. URL https://arxiv.org/abs/1902.01007.

[31] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.759. URL https://aclanthology.org/2022.emnlp-main.759/.

[32] John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp, 2020. URL https://arxiv.org/abs/2005.05909.

[33] Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. Don't blame the annotator: Bias already starts in the annotation instructions, 2024. URL https://arxiv.org/abs/2205.00415.

[34] Charles Sanders Peirce. *Collected Papers of Charles Sanders Peirce, Volume 1*. Harvard University Press, 1974.

[35] Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions, 2023. URL https://arxiv.org/abs/2308.11483.

[36] Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference, 2018. URL https://arxiv.org/abs/1805.01042.

[37] Abhilasha Ravichander, Joe Stacey, and Marek Rei. When and why does bias mitigation work? In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9233–9247, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.619. URL https://aclanthology.org/2023.findings-emnlp.619/.

[38] Kenneth Royal and Mari-Wells Hedgpeth. A novel method for evaluating examination item quality. *International Journal of Psychological Studies*, 7:17–17, 02 2015. doi: 10.5539/ijps.v7n1p17.

[39] Kenneth D. Royal and Myrah R. Stockdale. The impact of 3-option responses to multiple-choice questions on guessing strategies and cut score determinations. *Journal of Advances in Medical Education & Professionalism*, 5(2):84–89, 2017.

[40] Andreas Säuberli and Simon Clematide. Automatic generation and evaluation of reading comprehension test items with large language models, 2024. URL https://arxiv.org/abs/2404.07720.

[41] Krunal Shah, Nitish Gupta, and Dan Roth. What do we expect from multiple-choice qa systems?, 2020. URL https://arxiv.org/abs/2011.10647.

[42] Chenglei Si, Ziqing Yang, Yiming Cui, Wentao Ma, Ting Liu, and Shijin Wang. Benchmarking robustness of machine reading comprehension models, 2021. URL https://arxiv.org/abs/2004.14004.

[43] Neha Srikanth and Rachel Rudinger. Partial-input baselines show that nli models can ignore context, but they don't, 2022. URL https://arxiv.org/abs/2205.12181.

[44] Saku Sugawara, Pontus Stenetorp, and Akiko Aizawa. Benchmarking machine reading comprehension: A psychological perspective, 2021. URL https://arxiv.org/abs/2004.01912.

[45] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing nlp, 2021. URL https://arxiv.org/abs/1908.07125.

[46] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks, 2015. URL https://arxiv.org/abs/1502.05698.

[47] Sarah Wiegreffe and Ana Marasović. Teach me to explain: A review of datasets for explainable natural language processing, 2021. URL https://arxiv.org/abs/2102.12060.

[48] Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. Reclor: A reading comprehension dataset requiring logical reasoning, 2020. URL https://arxiv.org/abs/2002.04326.

[49] Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. Large language model as attributed training data generator: A tale of diversity and bias, 2023. URL https://arxiv.org/abs/2306.15895.

A Prompt Templates

Below we reproduce all zero-shot prompt templates used in our experiments.

A.1 Full Input and Choices Only

Our full-input prompt is structured as follows:

Full Input Prompt

Passage: $\mathcal{P}$
Question: $\mathcal{Q}$
Choices:\n(A) c_a \n(B) c_b \n(C) c_c \n(D) c_d
Answer:

And our choices-only prompt is structured as follows:

Choices-Only Prompt

Choices:\n(A) c_a \n(B) c_b \n(C) c_c \n(D) c_d
Answer:

A.2 Process-of-Elimination

Prompt design. We implement PoE as a two-round dialogue. In the first round the model receives the four options and is instructed to name the single least plausible choice. The remaining three options are then re-presented, and the model is asked to identify the most plausible answer among them.

Our evaluation of Process-of-Elimination relies on a two-step prompt, detailed below:

PoE – Step 1: Eliminate One Choice

Choices:\n(A) c_a \n(B) c_b \n(C) c_c \n(D) c_d
Incorrect answer:

PoE – Step 2: Answer Among Remaining

Choices:\n(A) c_a \n(B) c_b \n(C) c_c
Answer:

A.3 Abductive Passage Inference

Prompt design. Our formulation of Abductive Passage Inference (API) builds on recent abductive prompting strategies in multiple-choice reasoning [2], as well as broader evidence that LLMs benefit from explicitly verbalizing latent content [20]. API proceeds in two stages. In the first stage, the model is given the four answer options and asked to compose a passage that could plausibly accompany those options in a standard MCRC item. In the second stage, this generated passage is embedded into a new prompt alongside the same four options, and the model is asked to select an answer letter.

API – Step 1: Infer Passage

Choices:\n(A) c_a \n(B) c_b \n(C) c_c \n(D) c_d
Infer a passage:

API – Step 2: Answer w/ Inferred Passage

Passage: $\widehat{\mathcal{P}}$
Choices:\n(A) c_a \n(B) c_b \n(C) c_c \n(D) c_d
Answer:

Scoring. In our evaluation, instances where the model fails to provide a valid answer, such as deferring, refusing, or producing outputs that cannot be mapped to a choice, are assigned a default score of 0.25, reflecting the expected accuracy of a uniform random guess among four answer options.

B Additional Figures

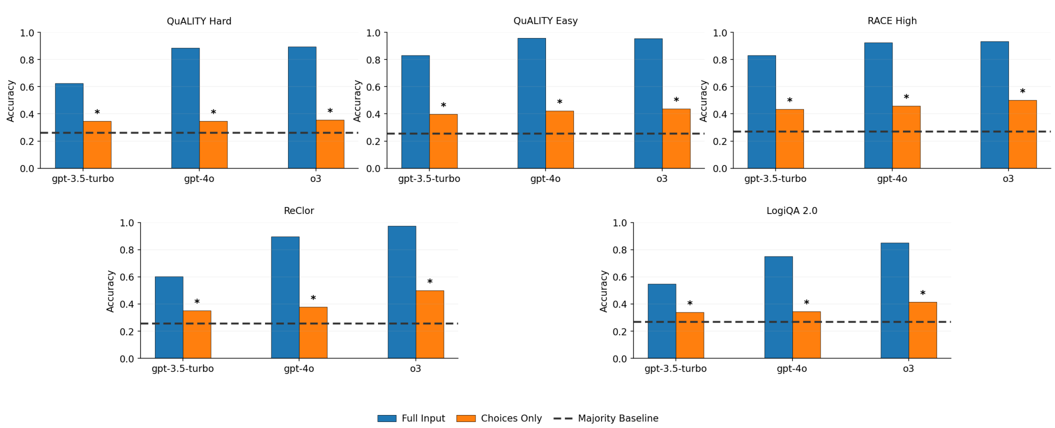

Figure 1: Full-input versus choices-only accuracy. An asterisk above a bar indicates accuracy significantly above the dataset's majority class baseline at $p < 0.05$ using a two sample t test across runs.

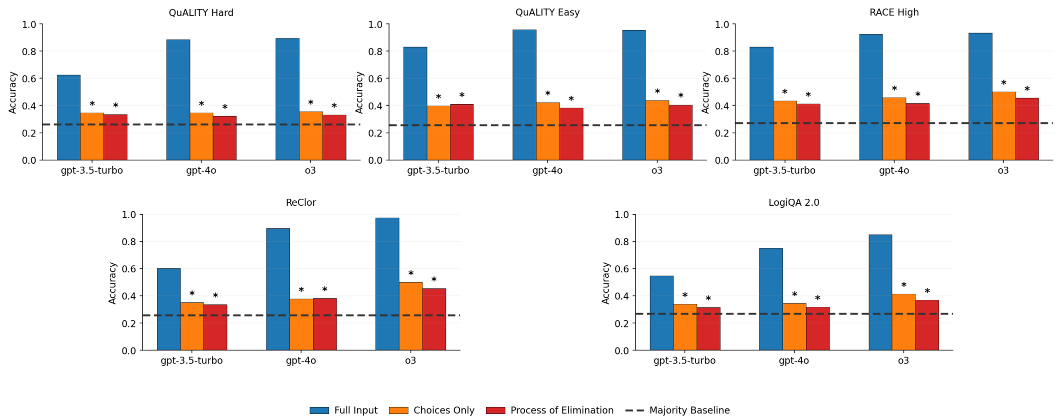

Figure 2: Full-input versus Process-of-Elimination accuracy. An asterisk above a bar indicates accuracy significantly above the dataset's majority class baseline at $p < 0.05$ using a two sample t test across runs.

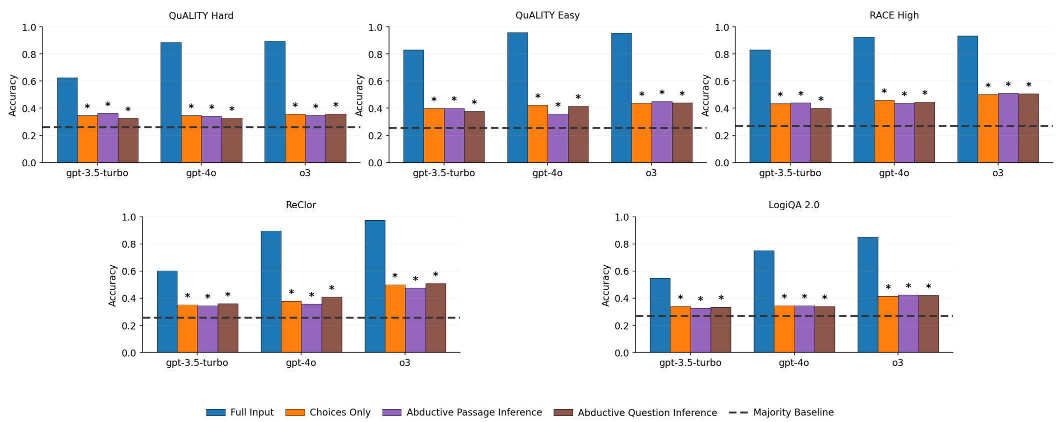

Figure 3: Full-input versus Abductive Passage Inference versus Abductive Question Inference accuracy. An asterisk above a bar indicates accuracy significantly above the dataset's majority class baseline at $p < 0.05$ using a two sample t test across runs.