Overwhelmed by Choice: Studying LLM Decision Making at Scale

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly deployed in applications that require decision-making over massive candidate spaces, including reranking, medical coding, and tool orchestration. However, most prior analyses of failures under large-scale inputs in LLMs primarily focus on retrieval-oriented failures, such as lost-in-the-middle behavior and context-length scaling effects, while decision-making settings are typically evaluated using relatively small candidate sets. Unlike retrieval-oriented settings that primarily require locating relevant information, large-option decision-making additionally requires repeated comparison and prioritization across many competing candidates. Our controlled analyses suggest that retrieval-oriented explanations alone cannot fully explain degradation under large candidate sets. Instead, we identify two systematic failure patterns in large-scale candidate comparison. First, we observe *gold-margin collapse*, where the score gap between the correct answer and the strongest distractor progressively shrinks as the candidate set scales, driven primarily by weakening confidence in the correct answer rather than increasingly competitive distractors. Second, predictions become increasingly dominated by earlier candidate evaluations under large option sets, consistent with controlled gold-insertion experiments showing that candidate influence progressively decreases with later positions. Motivated by these findings, we evaluate restructuring strategies based on hierarchical partitioning and permutation inference, improving robustness under large candidate settings by roughly 20 percentage points at $N = 160$ on both HotpotQA and MIMIC. Overall, our findings suggest that large-option degradation reflects fundamental weaknesses in large-scale candidate comparison beyond standard long-context retrieval failures.

1 Introduction

Large language models (LLMs) are increasingly deployed in real-world settings that require decision-making over massive candidate spaces. Examples include large-scale reranking in retrieval systems (Zhu et al., 2023), medical coding and diagnosis classification (Johnson et al., 2016; Singh et al., 2025), large taxonomy prediction, and agentic systems navigating extensive tool registries (Lei et al., 2025).

Despite the growing importance of large-scale candidate selection, most existing MCQ evaluations involve only a small number of candidate options, typically fewer than ten (Hendrycks et al., 2020; Chollet, 2019; Rein et al., 2023). While such settings have become standard for measuring reasoning ability, they may substantially underestimate the challenges of scalable decision-making (Zhang and Nguyen, 2025), where models must reliably distinguish among hundreds or even thousands of competing candidates.

Figure 1 provides an overview of the large-option degradation setting and the comparison-level failure patterns studied in this work. We systematically evaluate LLM behavior under large candidate spaces across diverse decision-making domains, including reranking, medical diagnosis, long-context classification, and information extraction (Thakur et al., 2021; Johnson et al., 2016; Cohan et al., 2018; Wang et al., 2020). Across datasets, prompting strategies, and model scales, we observe that decision quality deteriorates consistently as the number of candidate options increases.

A natural hypothesis is that this degradation primarily arises from standard long-context inference failures, including noise and interference effects in long-context reasoning (Hong et al., 2025; Shi et al., 2023) and positional biases in transformer attention (Hsieh et al., 2024; Liu et al., 2024; Wu et al., 2025). However, prior studies have largely focused

on retrieval-oriented settings, where the primary challenge is locating relevant information within long contexts. In contrast, large-option decision-making additionally requires reliable comparison, suppression, and prioritization across many competing candidates. Through controlled analysis, we find that existing long-context interference and transformer positional-bias explanations alone cannot fully account for the degradation observed under large candidate settings.

Instead, we identify two consistent comparison-level failure patterns. First, we observe *gold-margin collapse*, where the score gap between the correct answer and the strongest distractor progressively shrinks as candidate sets scale. Importantly, this collapse is driven primarily by weakening confidence in the correct answer rather than increasingly competitive distractors. Second, model predictions become increasingly dominated by earlier candidate evaluations under large option sets, consistent with controlled gold-insertion experiments showing that candidate influence progressively decreases with later positions. Together, these findings suggest that large-option degradation reflects not only long-context retrieval challenges, but also fundamental weaknesses in large-scale candidate comparison.

Motivated by these observations, we investigate two scalable inference-time interventions: hierarchical partitioning and permutation-based inference. Hierarchical partitioning decomposes large candidate spaces into smaller local comparison stages, reducing effective comparison complexity, while permutation-based inference mitigates early-option dominance by aggregating predictions across shuffled candidate orders. Despite their simplicity, both methods substantially improve robustness under large candidate settings. At $N = 160$, they improve accuracy by roughly 20 percentage points over standard single-pass inference on both HotpotQA and MIMIC.

Overall, our contributions are summarized as follows:

- We systematically characterize large-scale candidate comparison failures in LLM decision-making across diverse domains.
- We show that large-option degradation cannot be fully explained by standard long-context retrieval failures, highlighting an important distinction between retrieval-oriented reasoning and large-scale candidate comparison.
- We identify two consistent failure patterns in large-scale candidate comparison: *gold-margin collapse*, where confidence in the correct answer progressively weakens as candidate sets scale, and *early-option dominance*, where predictions become increasingly dominated by earlier candidate evaluations.
- We demonstrate that restructuring strategies, including hierarchical partitioning and permutation inference, substantially improve robustness under large candidate settings.

2 Large-Scale Candidate Comparison

2.1 Experimentation Details

Problem Description We study large-scale multiple-choice decision-making, where a model receives a task instruction I, a query q, and a candidate set

$$\mathcal{O} = \{(1, o_1), (2, o_2), \ldots, (N, o_N)\},$$

where exactly one indexed candidate corresponds to the correct answer. Each candidate is associated with a unique index used for answer generation. We focus exclusively on the single-answer setting and do not consider tasks with multiple correct options.

Task Description We evaluate models across diverse MCQ tasks spanning similarity reranking, scientific classification, medical diagnosis, and long-context reasoning. More information about dataset statistics is provided in Appendix Table 2.

We use publicly available datasets and pretrained models under their original research licenses and usage terms.

Output Generation We evaluate open-source models from the Qwen3, Gemma, Llama, and Phi families under varying candidate set sizes

$$N \in \{5, 10, 20, 40, 80, 160\}.$$

Unless otherwise specified, models are evaluated under standard zero-shot prompting. We additionally explore several inference-time reasoning settings to examine whether large-option degradation persists under stronger reasoning configurations.

Evaluation We evaluate decision accuracy under varying candidate set sizes. To investigate the mechanisms underlying large-option degradation,

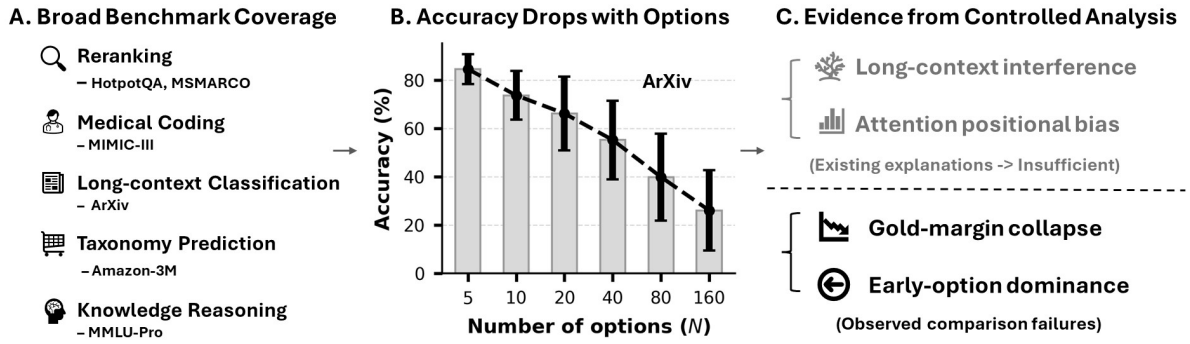


Figure 1: **Overview of large-option degradation under scalable candidate comparison.** Controlled analyses suggest that standard long-context explanations alone cannot fully explain the observed degradation, revealing additional comparison-level failure patterns under large candidate sets.

we further analyze model confidence dynamics, candidate interaction behavior, and the potential roles of context length and attention-based positional effects under large candidate spaces.

2.2 Accuracy Degradation at Large N

Main Results Figure 2 shows that accuracy consistently decreases as the number of options increases across all tasks and model families. This trend generalizes across additional datasets (Appendix 9), suggesting that large-option degradation reflects a general limitation in scalable candidate comparison rather than a dataset-specific phenomenon.

System-2 Reasoning We additionally evaluate several inference-time reasoning strategies, including chain-of-thought, self-critique, few-shot prompting, and pairwise comparison. As shown in Figure 3, although some methods improve absolute accuracy, all approaches continue to degrade substantially as the number of candidates increases. These results suggest that stronger reasoning alone does not fully eliminate large-option degradation. Additional details are provided in Appendix A.3.

3 Analysis

To understand why performance degrades under large candidate sets, we analyze model confidence dynamics, candidate interaction behavior, and the potential effects of context length and attention-based positional biases during candidate comparison.

3.1 Gold-Margin Collapse Under Large Candidate Sets

We first analyze the autoregressive sequence scores assigned to candidate options as the number of candidates increases. Specifically, we define the score of an option as the summed token-level log-probability of generating its corresponding answer index in the format `<Answer> idx </Answer>`.

We define the *gold margin* (GM) as the score difference between the correct option and the highest-scoring distractor, and the *distractor mass* as the log-sum-exp aggregated score across all distractor options.

Figure 4 (a,b) shows that the gold margin decreases consistently as the number of options increases, eventually becoming negative in some large-option regimes. Figure 4(c) further shows that while distractor scores remain relatively stable, the gold score decreases sharply. This suggests that degradation is driven primarily by collapsing confidence in the correct answer rather than increasingly strong distractors.

3.2 Context Length Alone Does Not Fully Explain Large-Option Degradation

To isolate the effect of context length, we increase prompt length while holding candidate comparison structure fixed.

Table 1 shows that randomly increasing irrelevant context length within the prompt does not produce a consistent degradation trend across tasks or models, unlike the degradation patterns observed under candidate-set scaling (Figure 2).

Similar trends are observed under distractor-verbosity controls (Appendix A.5), where substantially longer options also fail to consistently reproduce the degradation induced by candidate-set expansion.

Overall, these results suggest that context length alone cannot fully explain the stable degradation

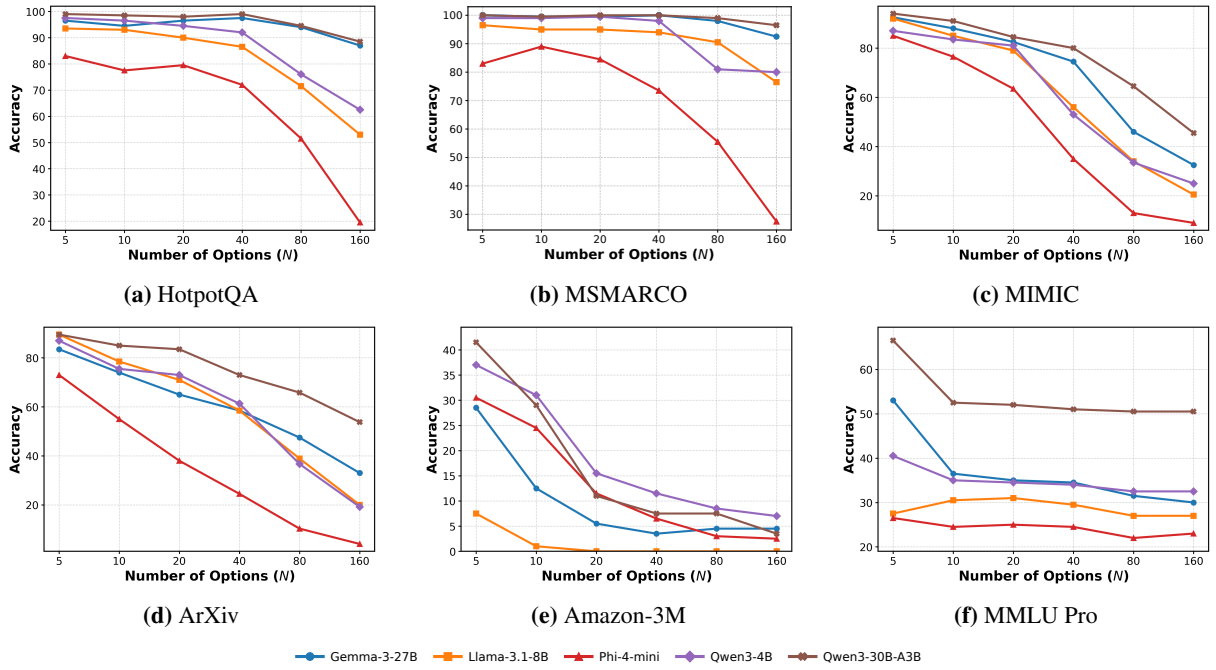


Figure 2: **Accuracy degradation visualization in the large N regime.** As the number of options increases, accuracy consistently declines across all datasets.

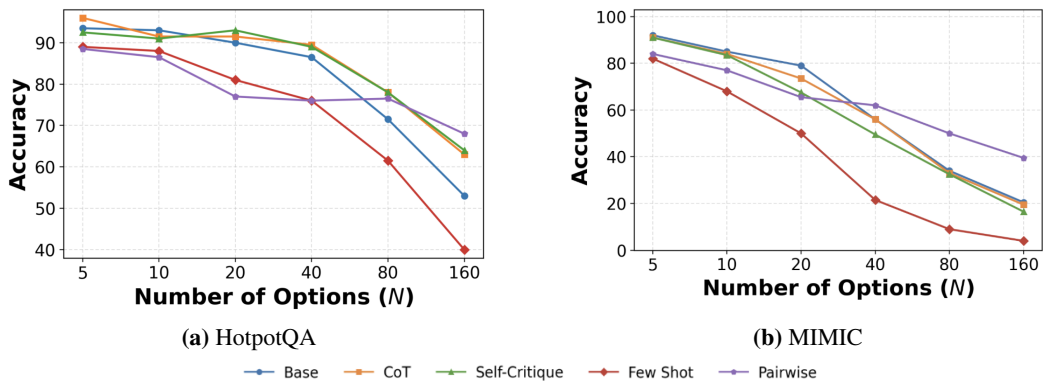


Figure 3: **System-2 reasoning under large candidate sets (Llama-3.1-8B).** All reasoning strategies degrade substantially as the number of options increases.

observed under candidate-set scaling.

3.3 Attention Distribution Does Not Directly Reflect Prediction Decisions

Prior work on *lost-in-the-middle* and subsequent studies on long-context reasoning suggest that positional retrieval behavior may be related to attention allocation patterns in transformer models. Motivated by these findings, we investigate whether prediction behavior in large-option decision making similarly correlates with attention allocation across candidate positions.

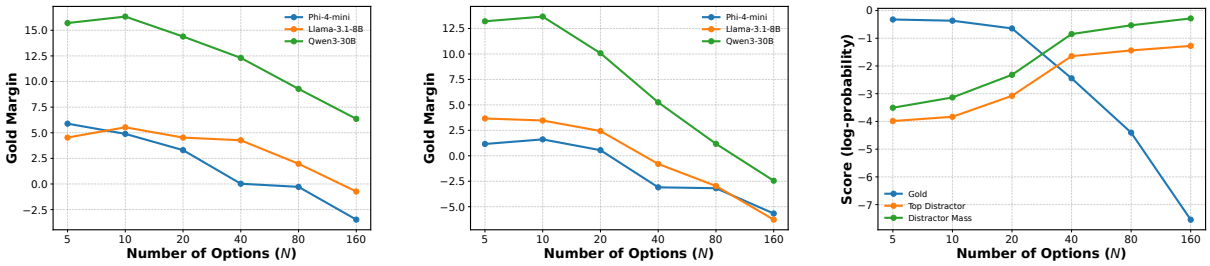
Figure 5 compares attention distributions and empirical prediction-position distributions on HotpotQA with $N = 80$.

While attention allocation patterns vary substantially across models, prediction distributions remain consistently front-skewed. Similar inconsistencies are also observed across different candidate-set sizes and datasets (Appendix A.6).

These results suggest that attention allocation patterns in transformer models do not directly translate into model prediction behavior under large candidate sets.

3.4 Early-Option Dominance in Large-Option Comparison

Previous analyses already showed that model predictions become front-skewed under large candidate settings (Figure 5(a,c)). Additional analyses in Appendix A.9 further show that this trend becomes progressively stronger as the number of candidate



(a) HotpotQA **(b)** MIMIC **(c)** Score dynamics (MIMIC)

Figure 4: **Gold margin across models and score dynamics for Llama-3.1-8B as the number of options increases.** As the number of options grows, the gold margin decreases while the gold score collapses substantially faster than the distractor score.

Task	Model	Added Tokens (Δ Accuracy)							Avg. Δ
		Base	+235	+470	+940	+1880	+3760	+7520	
HotPotQA	Phi-4-mini	72.00	+7.00	+7.00	+5.50	+7.00	+6.00	+2.50	+6.43
	Llama-3.1-8B	86.50	+0.50	+2.00	+0.50	+1.50	+0.50	+1.50	+1.07
	Qwen3-30B	99.00	-0.50	-0.50	-0.50	-0.50	-0.50	+0.00	-0.43
MIMIC	Phi-4-mini	35.00	-5.50	-3.50	-3.50	-3.00	-1.50	-5.00	-3.57
	Llama-3.1-8B	56.00	-3.50	-6.00	-5.50	-8.50	-7.50	-8.50	-6.50
	Qwen3-30B	80.00	-2.50	-1.00	-1.00	-4.50	-3.00	-3.00	-2.50

Table 1: **Accuracy change under increasing irrelevant prefix tokens.** The first column shows base accuracy (no added tokens), and subsequent columns report the change relative to the base. Added tokens are constructed by repeating a fixed block of task-irrelevant text.

options increases and persists under reversed-index controls.

To further examine how early-option dominance affects final prediction accuracy, we conduct controlled gold-position insertion experiments while keeping the distractor set fixed. Figure 6 shows that model behavior strongly depends on the insertion region of the gold answer. Early-positioned gold answers are recovered substantially more reliably, while later-positioned gold answers become progressively harder to recover as the number of candidates grows. This effect is also reflected in the probability-distribution view (Appendix A.8), where prediction mass becomes progressively less responsive to later gold insertions as the option set increases.

To further characterize the comparison dynamics underlying early-option dominance, we conduct a fine-grained gold-insertion experiment. For each example, we first identify the strongest distractor before gold insertion, which we refer to as the *pre-anchor*, and then insert the gold answer at different positions. We report two persistence-based measures: *anchor persistence*, which measures whether the final prediction remains the pre-anchor after gold insertion, and *retained anchor rate*, which measures whether the pre-anchor remains the highest-scoring distractor even when the final prediction changes.

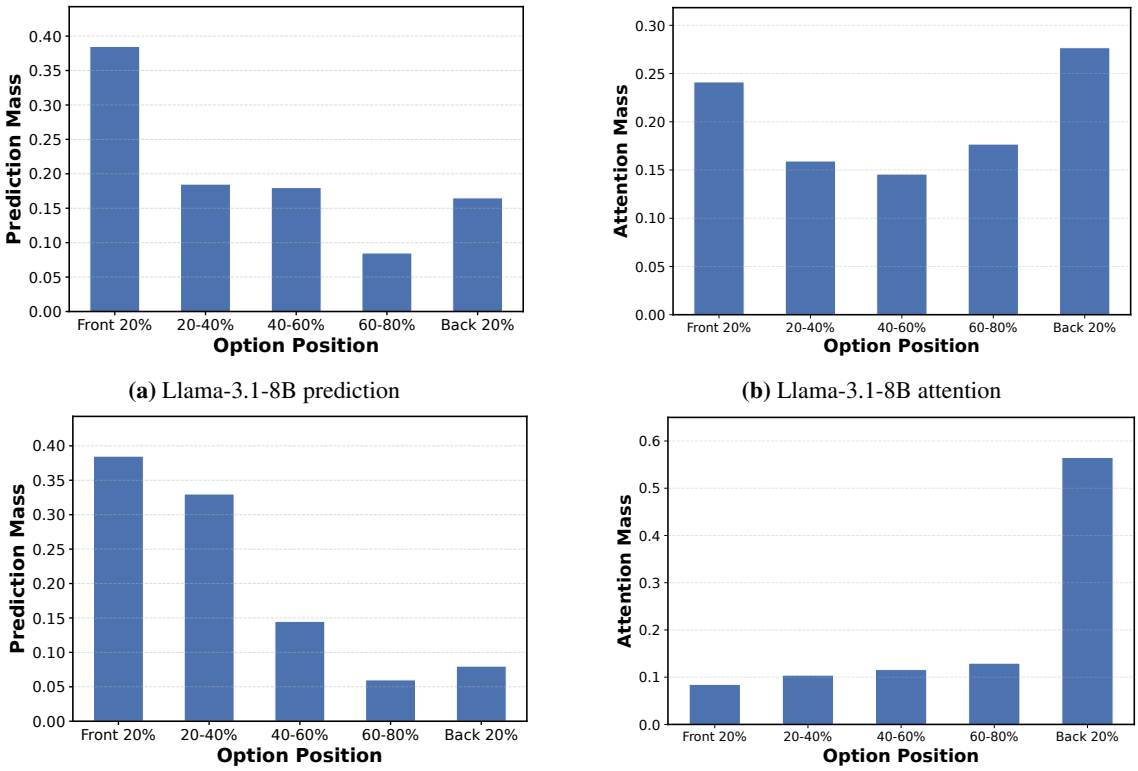
Figure 7 shows that later-positioned gold answers progressively lose their ability to overturn earlier distractor preferences. As the gold answer is inserted later in the sequence, gold probability steadily decreases, weakening its suppressive effect on the pre-anchor distractor. As a result, anchor persistence increases and earlier distractor preferences become more likely to survive as the final prediction. Despite minor edge-position effects, the overall trend remains consistent across datasets and models (Appendix A.10).

These results suggest that progressive early-option dominance reduces effective candidate exploration under large candidate sets, contributing to the observed accuracy degradation under large-option scaling.

4 Analysis-Driven Intervention

4.1 Overview

Motivated by the observed gold-margin collapse and early-option dominance, we evaluate two inference-time interventions that restructure the comparison process: (1) Hierarchical Partitioning, which performs staged candidate selection over small groups, and (2) Permutation-based Inference, which aggregates predictions across multiple candidate orderings. We additionally include Chain-of-



(a) Llama-3.1-8B prediction

(b) Llama-3.1-8B attention

(c) Phi-4-mini prediction

(d) Phi-4-mini attention

Figure 5: **Prediction-position and attention distributions on HotpotQA with $N = 80$.** Both models exhibit strongly front-skewed prediction distributions. However, their attention allocation patterns differ substantially: Phi-4-mini allocates most attention mass to later option positions, while Llama-3.1-8B exhibits a comparatively more balanced attention distribution. These discrepancies suggest that attention allocation does not directly correspond to prediction behavior during answer selection.

Thought (CoT) as a strong reasoning-based baseline. Detailed implementation settings are provided in Appendix A.11.

4.2 Intervention Results

As shown in Figure 8(a)(b), both Hierarchy and Permutation consistently outperform Single-Pass inference, with gains increasing as the number of options grows. At $N = 160$, both methods improve accuracy by roughly 20 percentage points on both HotpotQA and MIMIC. In contrast, although CoT improves absolute accuracy, it still degrades substantially at larger option sizes.

4.3 Mechanism Analysis

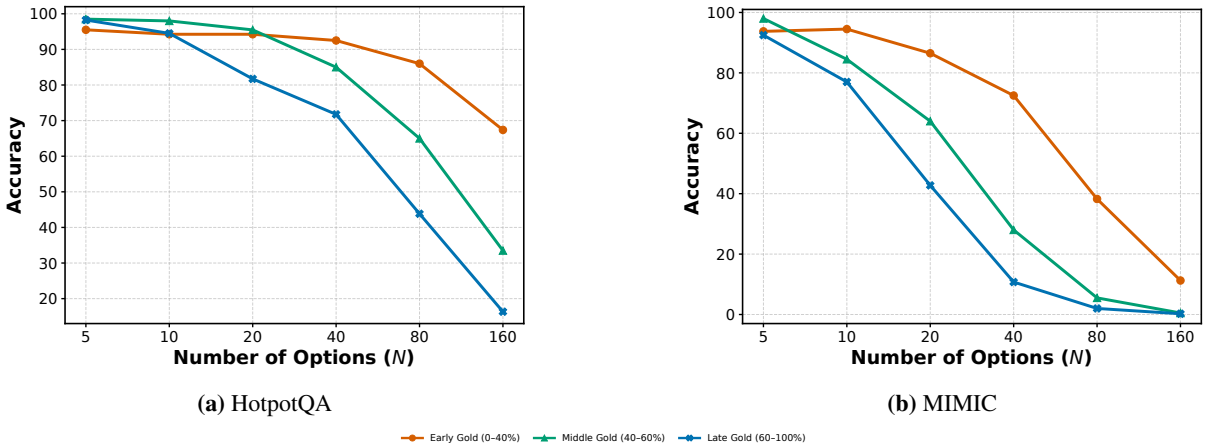
To analyze the source of improvement, we evaluate model behavior under the final decision stage of each intervention.

Figure 8(c)(d) shows that Hierarchy and Permutation maintain substantially larger gold margins and weaker early-position bias under large candidate settings compared with Single-Pass inference. In contrast, although CoT improves overall reasoning quality, it does not substantially mitigate gold-margin collapse or early-option dominance. Interestingly, we find that Permutation exhibits relatively lower final-round gold margins at small N. This occurs because many easy examples are already resolved through early agreement across permutations and therefore do not enter the final comparison stage, leaving a more difficult remaining comparison pool (Figure 22(b)).

These results suggest that restructuring the comparison process directly alleviates the key failure modes identified in our analysis. Additional analyses on intervention-specific behaviors and trade-offs are provided in Appendix A.12.

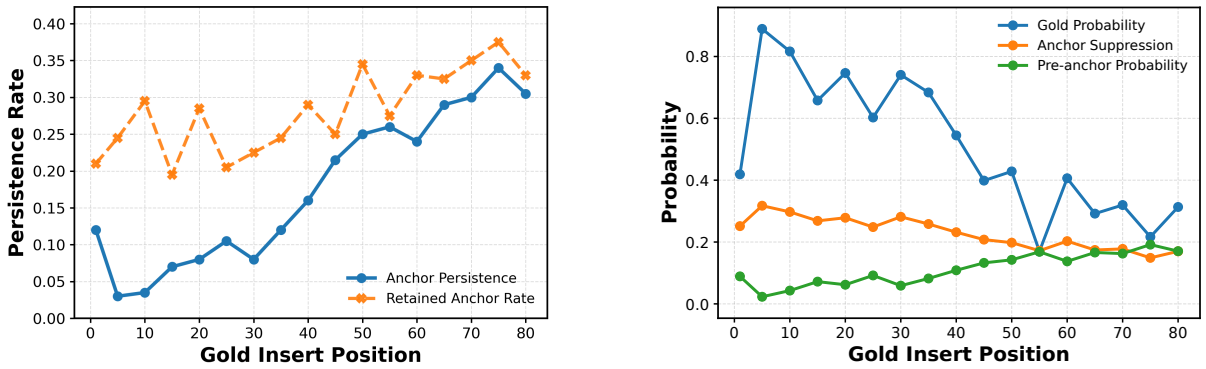
5 Related Work

MCQ-based Benchmarking Traditional MCQ benchmarks (Hendrycks et al., 2020; Rein et al., 2023; Zhang et al., 2024) typically evaluate LLMs using only a small number of answer choices. Recent studies show that increasing candidate difficulty through larger option sets, distractor augmentation, or stronger reasoning requirements substantially reduces model performance (Wang et al., 2024; Zhang and Nguyen, 2025; Elhady et al.,



(a) HotpotQA **(b)** MIMIC

Figure 6: **Controlled gold-position intervention on Llama-3.1-8B-Instruct.** Accuracy under different gold insertion regions on HotpotQA and MIMIC. As the number of options increases, later-positioned gold answers become progressively harder to recover.



(a) Persistence measures ($N = 80$) **(b)** Probability dynamics ($N = 80$)

Figure 7: **Revision dynamics under large candidate sets on HotpotQA (Llama-3.1-8B).** (a) Anchor persistence measures whether the final prediction remains the pre-anchor after gold insertion, while retained anchor rate measures whether the pre-anchor remains the highest-scoring distractor. (b) Probability dynamics after gold insertion, including gold probability, pre-anchor probability, and anchor suppression.

2025). Our work instead focuses on scalable candidate comparison under large-option decision-making settings.

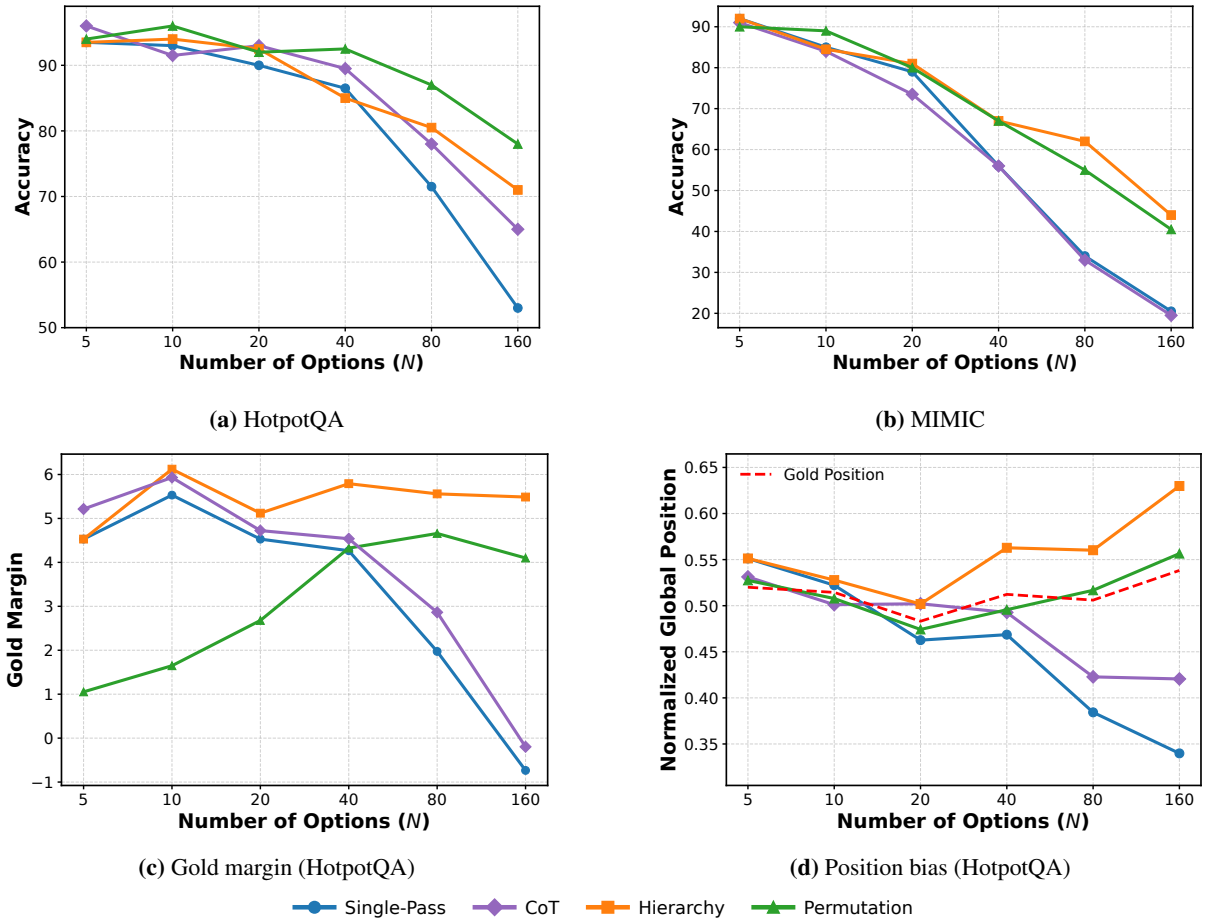
Long-Context Reasoning and Positional Bias Prior work has shown that LLMs struggle under long-context settings due to context interference(Hong et al., 2025; Shi et al., 2023; Du et al., 2025; Lee et al., 2025), retrieval degradation (Liu et al., 2024), and positional biases (Hsieh et al., 2024; Wu et al., 2025). These studies primarily focus on retrieval-oriented long-context behavior, whereas our work studies scalable candidate comparison under large-option decision-making settings.

Decision-Making Degradation and Mitigation Prior work has studied factors affecting LLM decision making, including distractor competition (Elhady et al., 2025; Amanlou et al., 2026), token-selection bias (Zheng et al., 2023), and inference-time mitigation strategies such as multi-agent reasoning, retrieval augmentation, and debiasing (Parekh et al., 2025; Gan and Sun, 2025; Zheng et al., 2026). Our work differs in focusing specifically on scalable candidate comparison under large-option decision-making settings.

6 Conclusion

We study large-option decision making in LLMs and show that performance degrades consistently as the number of candidate options increases across diverse domains and tasks.

Through controlled analyses, we find that this degradation cannot be fully explained by standard retrieval-oriented long-context failures alone. Instead, our results suggest that large-option settings



(a) HotpotQA **(b)** MIMIC

(c) Gold margin (HotpotQA) **(d)** Position bias (HotpotQA)

Figure 8: **Effect of restructuring the comparison process under large candidate sets (Llama-3.1-8B-Instruct).** Hierarchy and Permutation consistently outperform Single-Pass, with larger gains as the number of options increases. Restructuring also increases the gold margin and reduces early-position bias.

introduce an additional challenge in scalable candidate comparison, where models progressively struggle to maintain confidence in the correct answer and become increasingly dominated by earlier candidate evaluations.

Motivated by these findings, we evaluate restructuring strategies that improve robustness under large candidate settings by mitigating comparison collapse and candidate-order sensitivity.

Overall, our findings highlight scalable candidate comparison as an important and underexplored limitation of current LLM systems. As language models are increasingly deployed in reranking, large-scale classification, and agentic decision-making pipelines, improving robustness under massive candidate spaces may become increasingly critical for reliable real-world deployment.

7 Limitations

Our study focuses primarily on inference-time behavior under large candidate spaces and does not provide a formal theoretical characterization of the observed comparison dynamics. While our controlled analyses suggest systematic comparison-level failure patterns, the exact internal mechanisms underlying these behaviors remain an open question.

In addition, our experiments mainly evaluate English-language benchmarks and text-only multiple-choice settings. Whether similar degradation patterns arise in multilingual, multimodal, or interactive decision-making environments remains for future work.

Finally, although our restructuring strategies substantially improve large-option performance, they introduce additional inference cost due to repeated or hierarchical comparisons. More efficient scalable comparison methods remain an important direction for future research.

Our analysis also includes decision-making settings such as medical diagnosis and reranking systems. Although our work primarily studies model

behavior rather than deployment, failures under large candidate spaces may introduce risks if such systems are applied without appropriate human oversight.

References

Mohammad Amanlou, Erfan Shafiee Moghaddam, Yasaman Amou Jafari, Mahdi Noori, Farhan Farsi, and Behnam Bahrak. 2026. Knight: Knowledge graph-driven multiple-choice question generation with adaptive hardness calibration. *arXiv preprint arXiv:2602.20135*.

François Chollet. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. 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 615–621.

Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Babu Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz, Eliu A Huerta, and Hao Peng. 2025. Context length alone hurts LLM performance despite perfect retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23281–23298, Suzhou, China. Association for Computational Linguistics.

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, Vienna, Austria. Association for Computational Linguistics.

Tiantian Gan and Qiyao Sun. 2025. Rag-mcp: Mitigating prompt bloat in llm tool selection via retrieval-augmented generation. *arXiv preprint arXiv:2505.03275*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *CoRR*, abs/2009.03300.

Kelly Hong, Anton Troynikov, and Jeff Huber. 2025. Context rot: How increasing input tokens impacts llm performance. Chroma Research Report.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long T. Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. 2024. Found in the middle: Calibrating positional attention bias improves long context utilization. In *Findings of the Association for Computational Linguistics: ACL*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Celine Lee, Alexander M Rush, and Keyon Vafa. 2025. Critical thinking: Which kinds of complexity govern optimal reasoning length? In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1038–1060, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Fei Lei, Yibo Yang, Wenxiu Sun, and Dahua Lin. 2025. Mcpverse: An expansive, real-world benchmark for agentic tool use. *CoRR*, abs/2508.16260.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Tanmay Parekh, Kartik Mehta, Ninareh Mehrabi, Kai-Wei Chang, and Nanyun Peng. 2025. DiCoRe: Enhancing zero-shot event detection via divergent-convergent LLM reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20560–20582, Suzhou, China. Association for Computational Linguistics.

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

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Ujjwal Singh, Sarthak Deshwal, Nitish Dube, and Arjun Sharma. 2025. Max-eval-11: A comprehensive benchmark for evaluating large language models on full-spectrum icd-11 medical coding. *medRxiv*, pages 2025–10.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the 35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A Massive General Domain Event Detection Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in*

Natural Language Processing (EMNLP), pages 1652–1671, Online. Association for Computational Linguistics.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Xinyi Wu, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. 2025. On the emergence of position bias in transformers. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Yue Zhang and Nhan Nguyen. 2025. Rethinking mcq benchmarks: Mandatory reasoning evaluation reveals significant performance drops in large language models. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*.

Ziyin Zhang, Lizhen Xu, Zhaokun Jiang, Hongkun Hao, and Rui Wang. 2024. Multiple-choice questions are efficient and robust LLM evaluators. *CoRR*, abs/2405.11966.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors. *CoRR*, abs/2309.03882.

Jinquan Zheng, Jia Yuan, Jiacheng Yao, Chenyang Gu, Pujun Zheng, and Guoxiu He. 2026. Mitigating selection bias in large language models via permutation-aware grpo. *arXiv preprint arXiv:2603.21016*.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *CoRR*, abs/2308.07107.

A Appendix

A.1 Dataset Statistics

For each dataset, we randomly sample 200 evaluation instances to construct the large-option benchmark setting.

Table 2 summarizes the datasets used in our experiments, including task types, average option lengths, and total prompt lengths at N=160.

A.2 Additional Datasets for Accuracy Degradation

To further validate the generality of the observed degradation, we evaluate a broader set of datasets beyond the main text. As shown in Figure 9, accuracy consistently declines as the number of options increases across diverse tasks. This trend holds across different domains, indicating that the degradation is not dataset-specific.

Dataset	Task	Avg. Option Tokens	Prompt Tokens at N=160
MSMARCO	Reranking	79.8	12,763.4
HotpotQA	Multi-hop Reranking	90.2	14,434.8
MIMIC-3	Medical Diagnosis	116.4	18,641.5
arXiv	Long-context Classification	8.8	1,413.9
Amazon-3M	Taxonomy Prediction	13.9	2,229.9
MMLU-Pro	Knowledge Reasoning	11.2	2,202.6

Table 2: **Dataset statistics.** We report task type, average option length, and total prompt length at N=160.

A.3 Additional Implementation Details and Results for System-2 Reasoning

Additional Implementation Details. We utilized the following common System-2 inference-time strategies to test our hypothesis on accuracy degradation at scale: **(1) Chain-of-thought (CoT) Reasoning**, where the model develops a reasoning chain before selecting a response, **(2) Self-Critique**, where the model makes a choice and critiques it before selecting a response, **(3) Few-shot Prompting**, where the model is provided sample questions and answers in its context, and **(4) Pairwise Comparison**, where the model is repeatedly shown two options at a time and performs an aggregation to select a final answer.

Additional Results. We provide additional System-2 reasoning results on Qwen3-4B in Figure 10. Similar to the main-text observations on Llama-8B, all reasoning strategies continue to exhibit substantial degradation as the number of options increases, despite improving absolute accuracy in smaller candidate settings.

These results further support our conclusion that increasing reasoning depth alone does not fundamentally resolve failures in scalable candidate comparison.

A.4 Irrelevant Contents

For the context length experiments, we use a fixed template of neutral text (47 tokens) that is semantically unrelated to the task. This text contains no information relevant to the question or answer choices, and is repeated to control the total input length. The template is shown below.

This is additional background text used

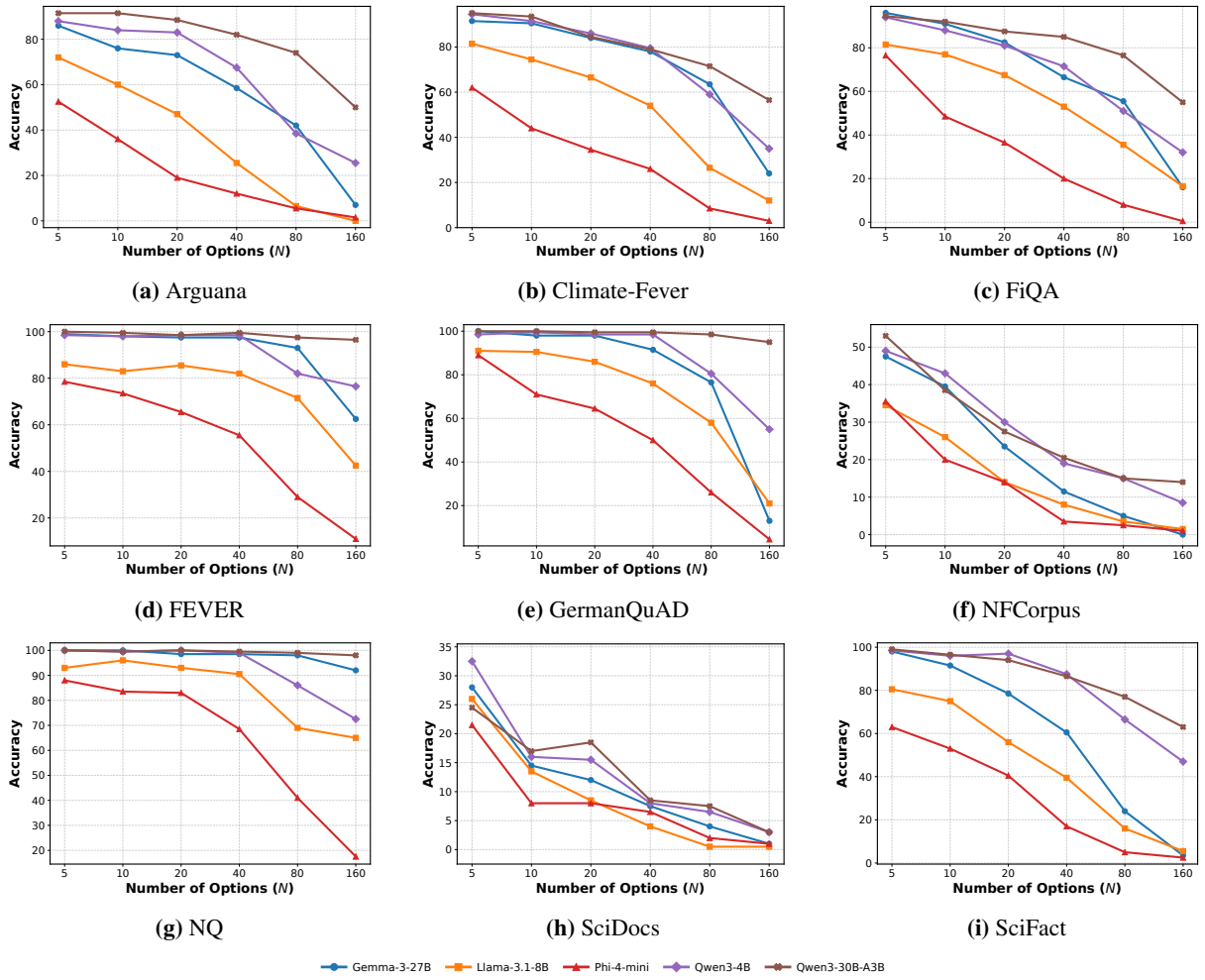


Figure 9: **Accuracy degradation across additional datasets.** The degradation trend remains consistent across a wide range of tasks, supporting the generality of our findings beyond the main benchmarks.

to increase the length of the input. It does not contain any information relevant to the question or answer choices. The purpose of this text is solely to vary the context length without affecting the decision process.

A.5 Distractor Verbosity Control

To further examine whether prompt length alone explains large-option degradation, we compare candidate-set expansion using distractor options with systematically different verbosity levels.

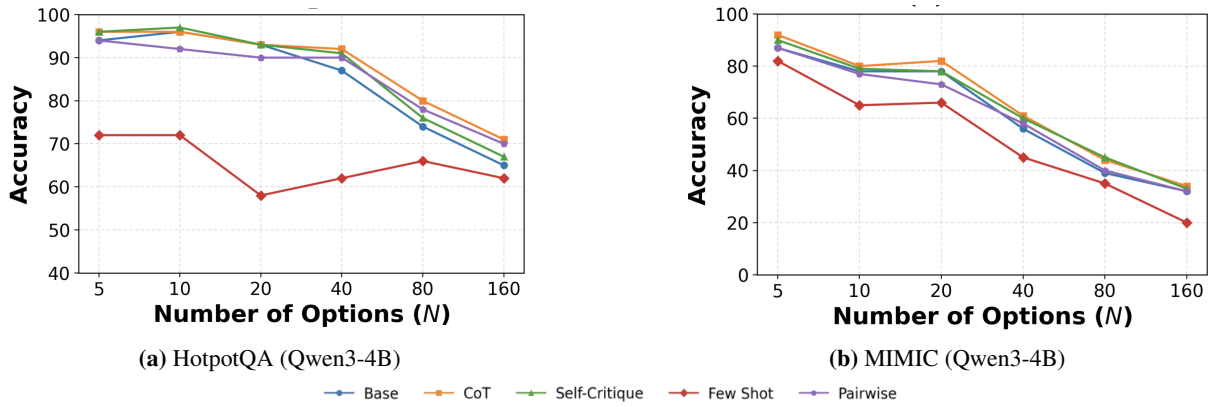
Figure 11 shows that although long-option distractors substantially increase average prompt length, the resulting degradation trends remain substantially less consistent than those induced by increasing the number of candidate options. These results further support that context length alone cannot fully explain large-option degradation.

A.6 Attention-Prediction Discrepancies Across Candidate Sizes and Datasets

We further examine attention allocation and prediction-position distributions across different candidate-set sizes and datasets.

Figures 12 and 13 show that attention distributions vary substantially across models, datasets, and candidate-set sizes. In several settings, models allocate increasing attention mass toward later option positions as the number of candidates increases. However, prediction distributions remain consistently front-skewed even under these changing attention patterns.

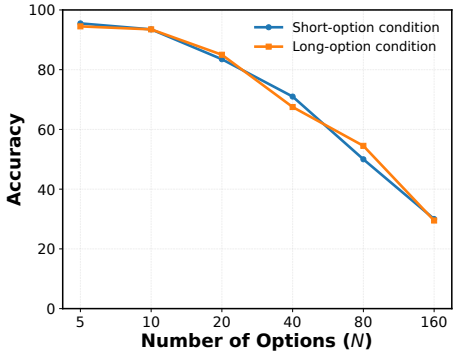
These observations further support that prediction behavior under large-option decision making cannot be directly explained by attention allocation patterns alone.



(a) HotpotQA (Qwen3-4B)

(b) MIMIC (Qwen3-4B)

Figure 10: **Additional System-2 reasoning results on Qwen3-4B.** All reasoning strategies continue to degrade substantially as the number of options increases.



(a) Verbosity control

N	Short-option	Long-option
5	3688	3866
10	3990	4382
20	4597	5426
40	5810	7518
80	8240	11693
160	13120	20075

(b) Average prompt token length

Figure 11: **Distractor verbosity control under candidate-set expansion (MIMIC, Llama-3.1-8B).** Although longer distractor options substantially increase total context length, they do not consistently reproduce the degradation induced by increasing the number of candidate options.

A.7 Prediction Positional Bias Analysis

We further provide additional analyses on prediction-position bias under large candidate sets, including average prediction-position trends, fine-grained position distributions, and reversed-index controls.

Figure 14 shows that model predictions progressively shift toward earlier candidate positions as the number of candidate options increases, despite uniformly distributed gold-answer positions. This trend remains consistent across datasets and model families, indicating increasingly strong early-position preference under large-option settings.

To further quantify this effect, Table 3 reports five-bin prediction distributions at $N = 160$. Predictions remain heavily concentrated toward earlier candidate regions for most models, with smaller models exhibiting particularly strong front-position concentration.

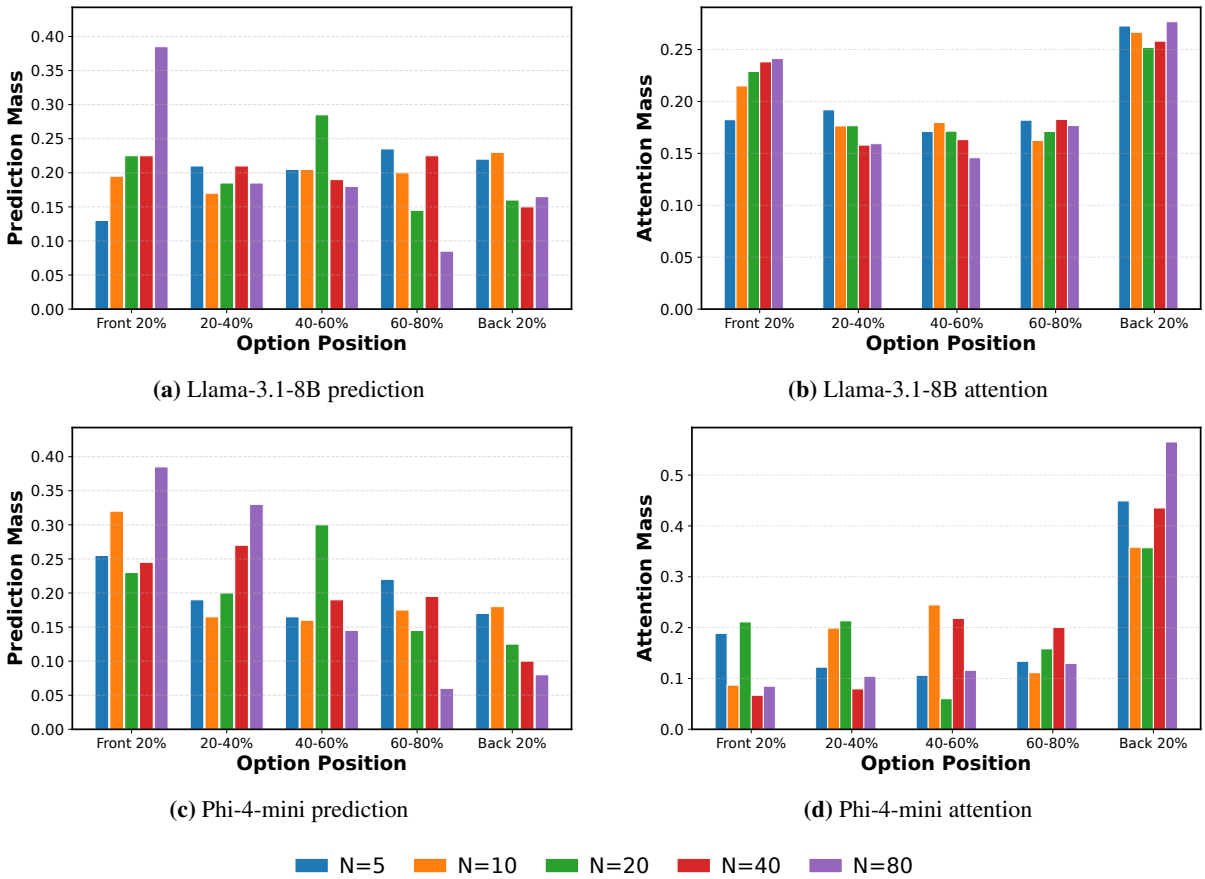
We further perform a reversed-index control experiment to distinguish positional ordering effects from potential token-level index priors.

Under this setting, earlier candidates are assigned larger index tokens while later candidates receive smaller indices. Figure 15 shows that similar early-position trends persist even under reversed indexing across both HotpotQA and MIMIC.

These findings suggest that the observed positional bias is driven primarily by candidate ordering rather than superficial preferences for specific answer indices.

A.8 Additional Probability Dynamics under Controlled Gold Insertion

Figure 16 further visualizes the corresponding probability distributions under controlled gold-position intervention. As the number of options increases, prediction mass becomes progressively less responsive to the gold insertion region and increasingly concentrates toward earlier candidates.



(a) Llama-3.1-8B prediction

(b) Llama-3.1-8B attention

(c) Phi-4-mini prediction

(d) Phi-4-mini attention

Figure 12: **Prediction-position and attention distributions across candidate-set sizes on HotpotQA.** Prediction distributions become increasingly concentrated toward earlier candidate positions as the number of options increases. However, attention allocation patterns vary substantially across models and candidate-set sizes, with several settings exhibiting increased attention concentration toward later option positions.

A.9 Early-Option Dominance Generalization

We further evaluate whether the early-option dominance generalizes across tasks and models using the arXiv dataset.

Figure 17 shows that predictions progressively shift toward earlier candidate positions as the number of options increases across models. We additionally evaluate a reversed-index control on arXiv, where answer indices are assigned in reverse order while preserving candidate positions. The same early-position preference remains under reversed indexing, further confirming that the observed behavior is driven by candidate-order effects rather than superficial preferences for smaller answer indices.

Figure 17 shows that predictions progressively shift toward earlier candidate positions as the number of options increases across models.

Figure 18 shows controlled gold-insertion results across models. For each model, we keep the distractor set fixed and vary only the insertion position of the gold answer. Across models, predictions become progressively less sensitive to the gold insertion position as the number of options increases.

Figure 19 shows anchor persistence across models under large candidate sets. Across all models, later-positioned gold answers become progressively less effective at overturning earlier candidate preferences.

Figure 20 shows probability revision dynamics across models. The probability assigned to the inserted gold generally decreases as the insertion position moves later, while the original anchor retains larger probability mass. Stronger models may exhibit a mild rebound at the final insertion positions, but the overall trend still indicates decreasing revision ability for later-arriving evidence.

A.10 Additional Revision Dynamics Examples

We observe similar revision dynamics on ArXiv dataset across additional model families and scales. Figure 21 provides additional examples using Llama-3.2-3B and Qwen3-4B. Despite differences

Model	$N = 160$	
	HotpotQA	MIMIC
Phi-4-mini	0.81 / 0.10 / 0.02 / 0.01 / 0.01	0.69 / 0.14 / 0.06 / 0.06 / 0.03
Llama-3.1-8B	0.48 / 0.14 / 0.10 / 0.13 / 0.13	0.81 / 0.13 / 0.04 / 0.01 / 0.01
Qwen3-30B-A3B	0.26 / 0.15 / 0.21 / 0.20 / 0.20	0.46 / 0.19 / 0.14 / 0.14 / 0.09

Table 3: **Five-bin prediction-position distributions at $N = 160$.** Values indicate the proportion of predictions falling into each position region (front to back).

in model scale and architecture, both models exhibit the same overall trend: later-position gold answers progressively lose the ability to overturn earlier candidate preferences. Interestingly, stronger models occasionally exhibit mild probability recovery near the final insertion positions, suggesting limited late-stage corrective ability despite strong early-anchor dominance.

A.11 Additional Implementation Details

Prompting Strategies. We evaluate three inference strategies that differ in how the model is asked to evaluate the candidate set. These strategies allow us to (1) benchmark the LLM's ability in the MCQ reasoning tasks and (2) attempt to mitigate the accuracy decline at large N.

Single-Pass Choice Inference. Our first strategy is a direct inference baseline in which the model receives the full task context together with all N candidate options in a single prompt, and is asked to select exactly one answer. This setting evaluates the model's ability to discriminate the correct option from all distractors under a flat large-option comparison structure, and serves as the standard reference baseline throughout our experiments.

Hierarchical Candidate Partitioning. Our second strategy decomposes the large candidate set into a sequence of smaller group-based comparisons. At each round, we partition the current candidate set into groups of size K, ask the model to select one candidate from each group, and carry the selected candidates into the next round. This process is repeated until only one candidate remains. Formally, at round t, we partition the active candidate set $\mathcal{O}^{(t)}$ into groups $\{\mathcal{O}_1^{(t)}, \ldots, \mathcal{O}_{G_t}^{(t)}\}$, select

$$\hat{o}_g^{(t)} = \arg\max_{o \in \mathcal{O}_g^{(t)}} p(o),$$

and construct the next candidate set as

$$\mathcal{O}^{(t+1)} = \{\hat{o}_1^{(t)}, \hat{o}_2^{(t)}, \ldots, \hat{o}_{G_t}^{(t)}\}.$$

The final prediction is the remaining candidate after repeated filtering. This hierarchical construction reduces the effective comparison set at each decision stage and tests whether large-option decision reliability improves when flat large-scale comparison is decomposed into smaller local comparisons.

Permutation-Based Candidate Aggregation. Our third strategy reduces sensitivity to candidate ordering by evaluating multiple shuffled views of the same candidate set. For each shuffled view v, we randomly permute the candidate order and ask the model to select one candidate:

$$\hat{o}_v = \arg\max_{o \in \pi_v(\mathcal{O})} p(o),$$

where $\pi_v(\mathcal{O})$ denotes the v-th shuffled candidate ordering. We then collect the selected candidates from all shuffled views into a reduced candidate set

$$\mathcal{C} = \{\hat{o}_1, \hat{o}_2, \ldots, \hat{o}_V\},$$

and perform one final inference step over $\mathcal{C}$:

$$\hat{o} = \arg\max_{o \in \mathcal{C}} p(o).$$

This strategy preserves candidates selected under different candidate orderings and reduces the influence of any single ordering on the final decision.

A.12 Mechanism Analysis of Intervention

We further analyze how hierarchical partitioning and permutation-based inference improve decision quality under large candidate sets.

While hierarchical partitioning addresses the limitations in scalable candidate comparison by reducing the effective decision complexity, permutation-based inference mitigates positional bias by reducing sensitivity to candidate ordering.

Both methods aim to alleviate the comparison bottleneck identified in our analysis, but operate through distinct mechanisms.

Hierarchical Partitioning. Hierarchical partitioning decomposes a large candidate set into smaller subsets and performs local selection within each group before a final comparison stage. This reduces comparison complexity at each step, enabling more reliable local decisions under large candidate spaces.

As shown in Figure 22(a), this design introduces a key trade-off: the probability that the gold candidate survives to the final round decreases as the number of options increases. This suggests that hierarchical partitioning shifts the dominant failure mode from **unreliable large-scale comparison** to **irreversible early elimination**, where mistakes made during local selection cannot be corrected in later stages.

Permutation-Based Inference. While hierarchical partitioning reduces comparison complexity, permutation-based inference instead focuses on mitigating positional bias by reducing sensitivity to candidate ordering.

Permutation-based inference aggregates predictions across multiple reordered views of the same candidate set. To better understand its behavior, we decompose the final accuracy as:

$$\mathrm{Acc} = P_{\text{only}} + P_{\text{multi}} \cdot \mathrm{Acc}_{\text{final}},$$

where P_{only} denotes cases where the gold candidate survives aggregation and is the **only remaining candidate**, while P_{multi} corresponds to cases where the gold candidate survives but multiple candidates remain, requiring a non-trivial final comparison.

As shown in Figure 22(b), the gap between gold survival and gold-only selection widens as the number of options increases. Although the gold candidate is increasingly preserved after aggregation, it more frequently coexists with competing candidates, shifting the task from **unambiguous selection** to **comparison-driven decision making**. In this regime, final-stage comparison quality becomes the dominant factor determining overall accuracy. Permutation-based inference improves robustness in this setting by reducing sensitivity to sequential candidate ordering.

Unified Interpretation. Taken together, these results reveal that the two methods address distinct but complementary challenges.

Hierarchical partitioning improves decision reliability by reducing comparison complexity, while permutation-based inference mitigates positional bias by aggregating across different orderings.

Together, they improve decision quality by addressing both comparison difficulty and ordering sensitivity.

A.13 Additional Experimental Results

Empirical Validation Across Models. To further validate the robustness of hierarchical partitioning and permutation-based inference, we evaluate both methods across different model scales.

As shown in Figure 23, both hierarchical grouping and permutation-based inference consistently mitigate accuracy degradation as the number of options increases. This trend holds across models of different capacities, including Qwen3-30B and Phi-4-mini, indicating that the effectiveness of these strategies is not tied to model scale.

While single-pass inference suffers from severe performance drops at large N, both methods improve robustness by either reducing comparison complexity (hierarchical grouping) or alleviating ordering sensitivity (permutation-based inference).

Effect of Group Size and Number of Permutations. We further analyze how hyperparameters influence the effectiveness of both methods, focusing on group size for hierarchical partitioning and the number of shuffled views for permutation-based inference.

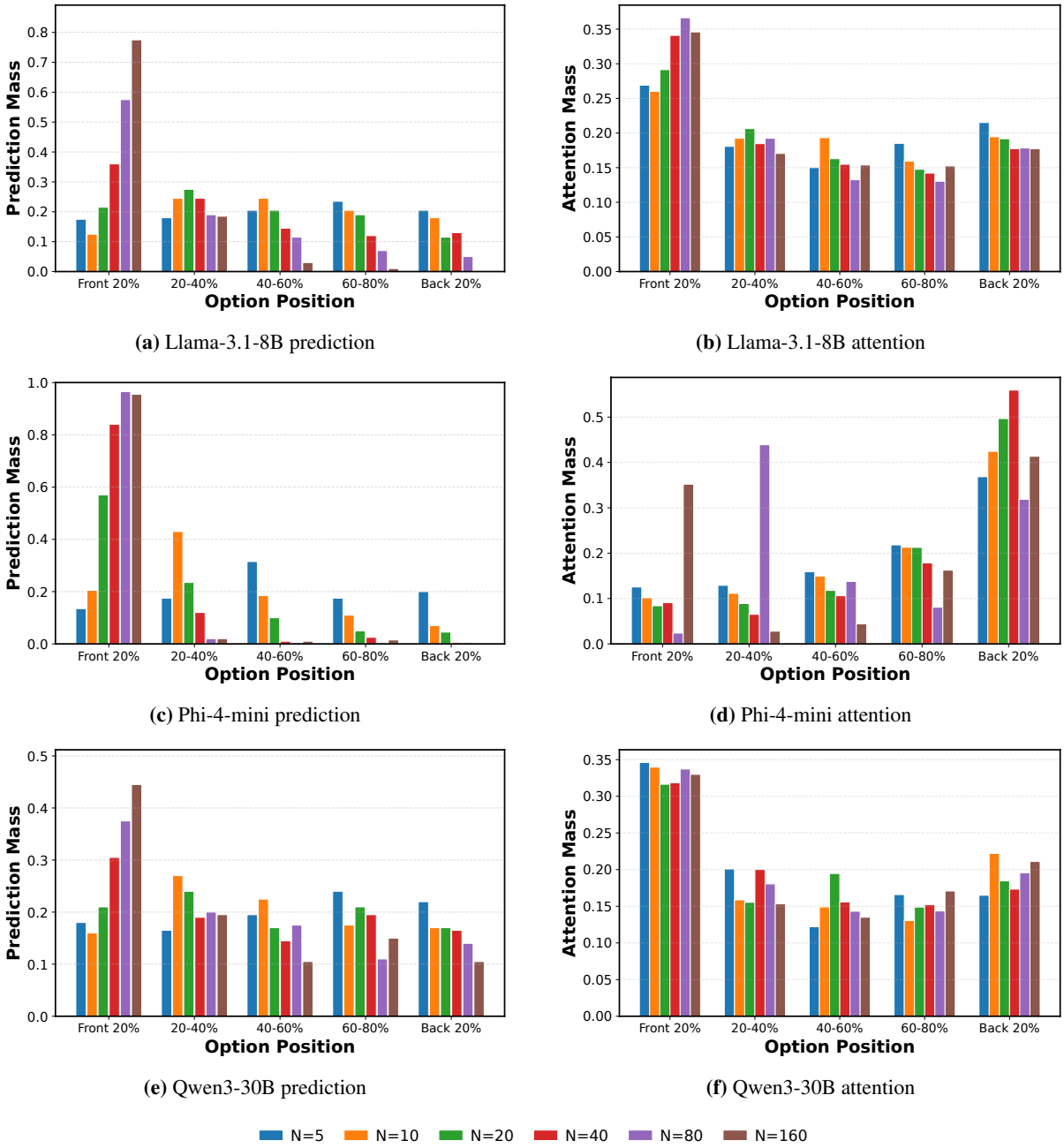
As shown in Figure 24(a), larger group sizes generally lead to better performance for hierarchical partitioning across all values of N. In particular, moderate-to-large group sizes (e.g., Group-10 and Group-20) consistently outperform smaller groups, suggesting that excessively small groups may suffer from insufficient candidate diversity in early stages, leading to suboptimal local selections.

However, all configurations still exhibit performance degradation as the number of options increases, indicating that hierarchical partitioning cannot fully eliminate the comparison bottleneck, but only mitigates it.

Figure 24(b) shows the effect of the number of permutations. Increasing the number of shuffled views improves robustness at moderate values, but does not lead to consistent gains at larger N. In fact, excessive permutations can introduce noise in aggregation, leading to slight performance drops in some cases.

These results highlight that both methods involve trade-offs: hierarchical partitioning benefits from sufficiently large group sizes to preserve strong candidates, while permutation-based infer-

ence achieves the best performance with a moderate number of diversified views.



(a) Llama-3.1-8B prediction

(b) Llama-3.1-8B attention

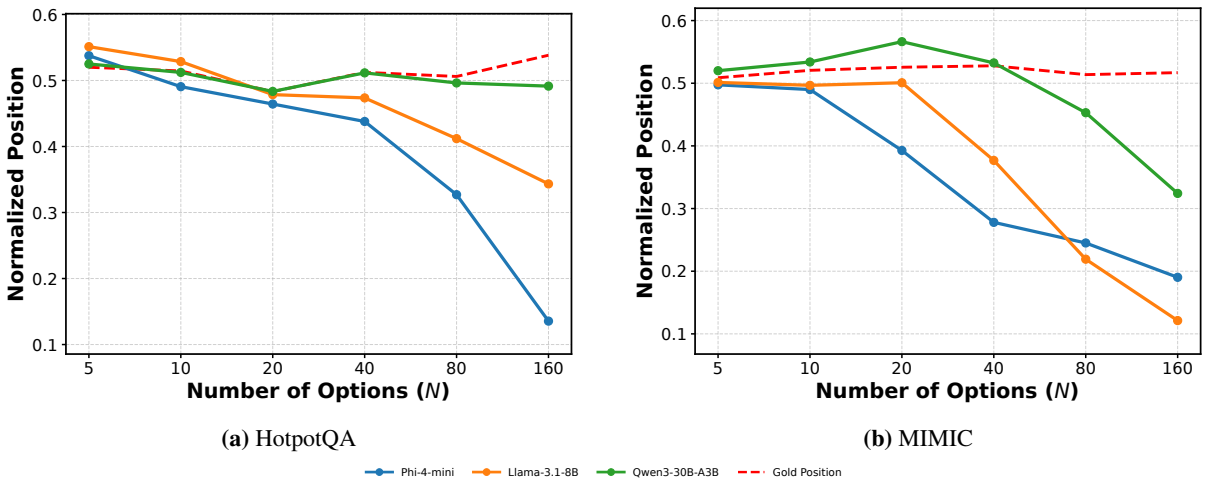
(c) Phi-4-mini prediction

(d) Phi-4-mini attention

(e) Qwen3-30B prediction

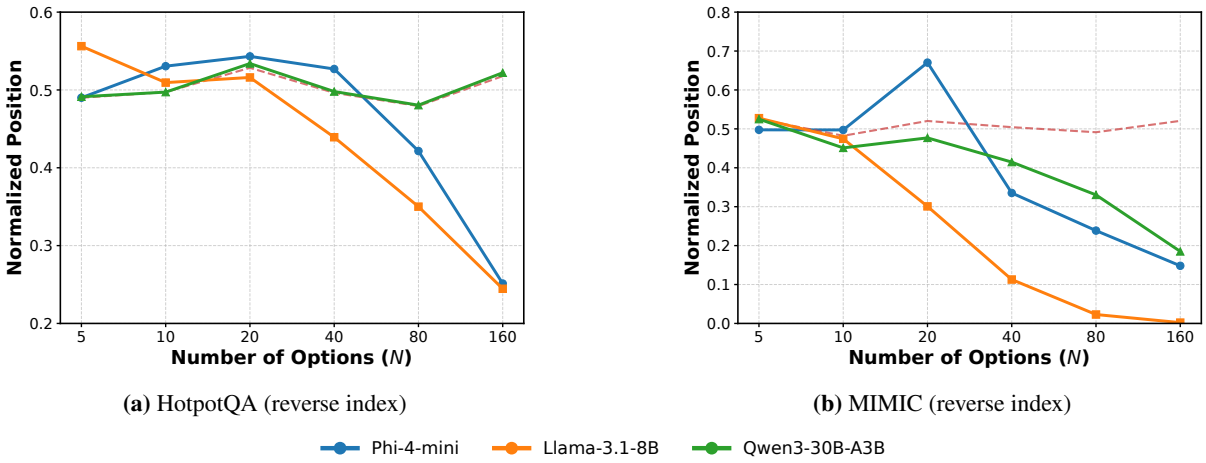
(f) Qwen3-30B attention

Figure 13: **Prediction-position and attention distributions across candidate-set sizes on ArXiv.** Prediction distributions remain consistently concentrated toward earlier candidate positions across models and candidate-set sizes, despite substantial variation in attention allocation patterns.



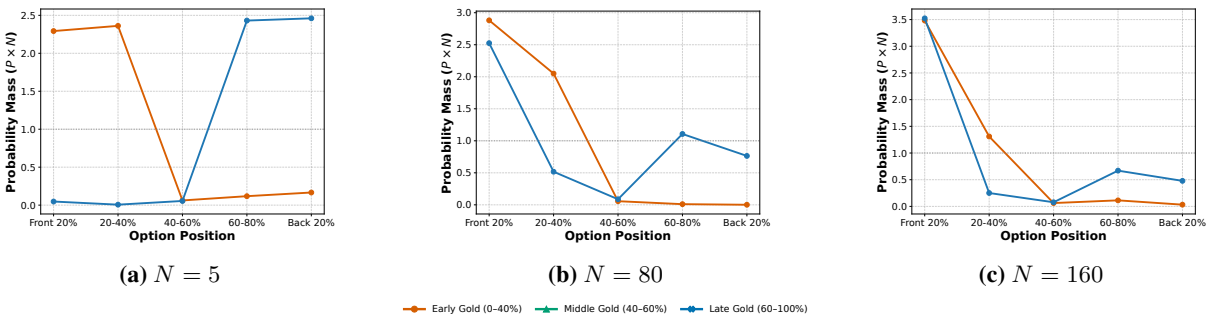
(a) HotpotQA **(b)** MIMIC

Figure 14: **Average prediction position under candidate-set scaling.** As the number of candidate options increases, model predictions progressively shift toward earlier candidate positions across datasets and model families, while the gold-answer distribution remains stable.



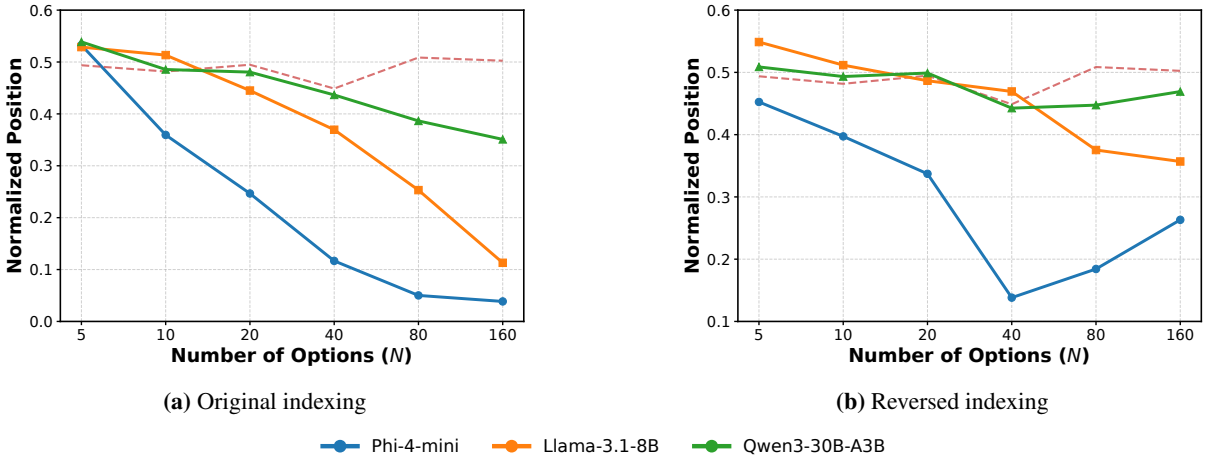
(a) HotpotQA (reverse index) **(b)** MIMIC (reverse index)

Figure 15: **Early-position bias persists under reversed index assignment.** Even when smaller answer indices are assigned to later candidates, model predictions continue to increasingly concentrate toward earlier candidate positions as the number of options grows.



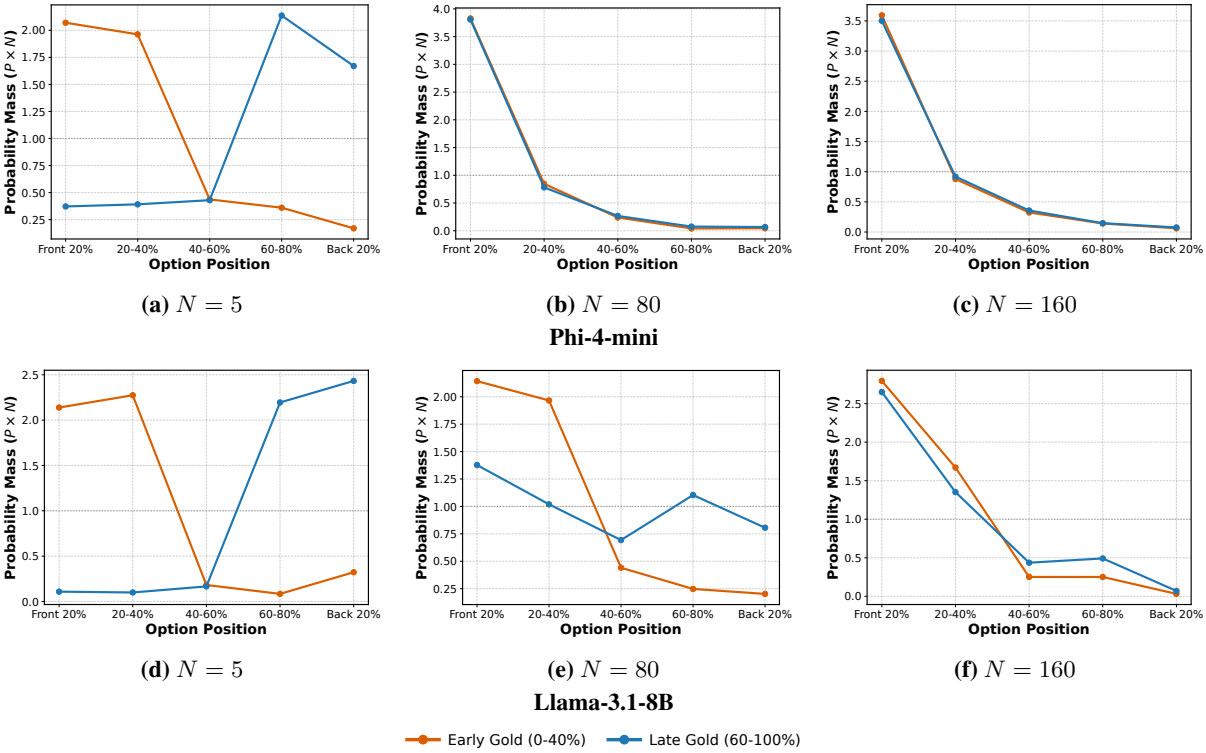
(a) $N = 5$ **(b)** $N = 80$ **(c)** $N = 160$

Figure 16: **Probability distributions under controlled gold-position intervention.** Probability mass progressively concentrates toward earlier candidates as the option set grows.



(a) Original indexing **(b)** Reversed indexing

Figure 17: **Early-position bias on arXiv under original and reversed indexing.** Predictions progressively concentrate toward earlier candidate positions as the number of options increases across models. The same trend persists under reversed index assignment, indicating that the observed bias is driven primarily by the option position rather than token-level index priors.



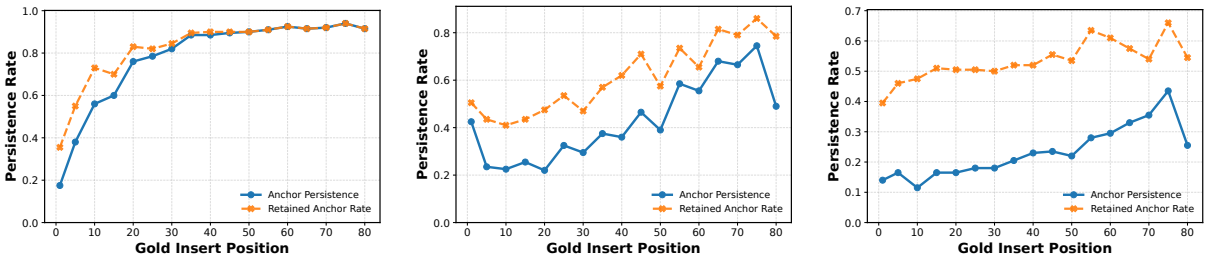
(a) $N = 5$ **(b)** $N = 80$ **(c)** $N = 160$

Phi-4-mini

(d) $N = 5$ **(e)** $N = 80$ **(f)** $N = 160$

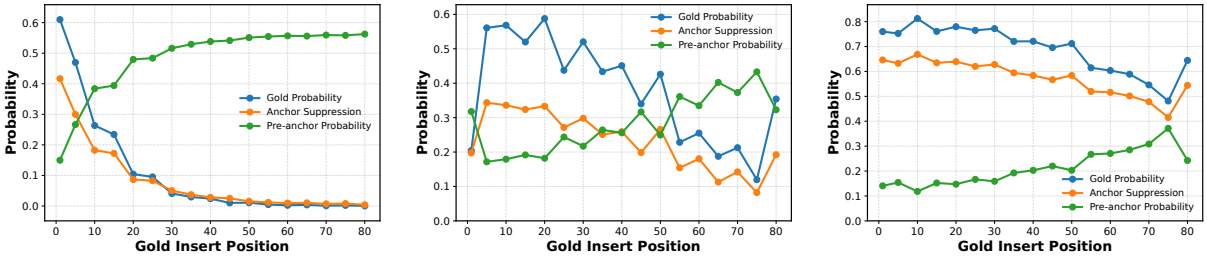
Llama-3.1-8B

Figure 18: **Controlled gold-position intervention on arXiv across models.** Each row corresponds to a model, while columns vary the candidate set size. As the number of options increases, predictions become progressively concentrated toward earlier candidate positions regardless of the true gold location.



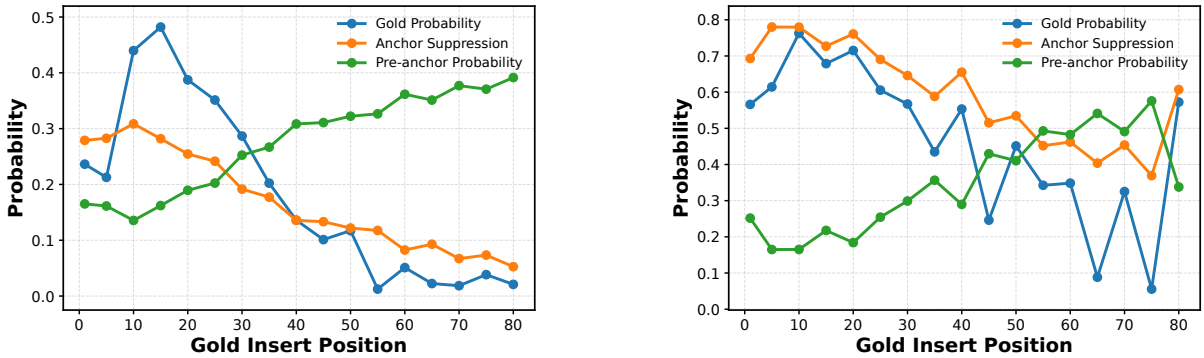
(a) Phi-4-mini **(b)** Llama-3.1-8B **(c)** Qwen3-30B-A3B

Figure 19: **Anchor persistence on arXiv across models.** Later-positioned gold answers become progressively less effective at overturning earlier candidate preferences.



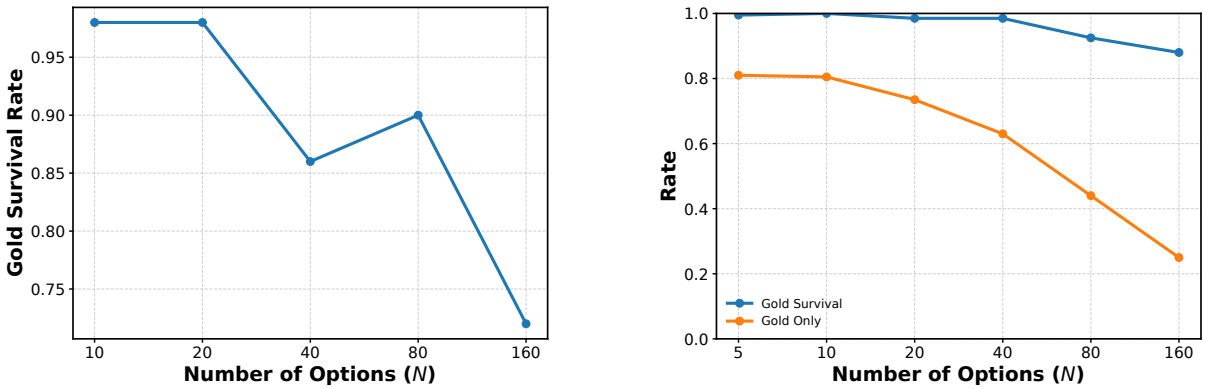
(a) Phi-4-mini **(b)** Llama-3.1-8B **(c)** Qwen3-30B-A3B

Figure 20: **Probability revision dynamics on arXiv across models.** Later-arriving gold answers progressively lose the ability to overturn earlier candidate preferences.



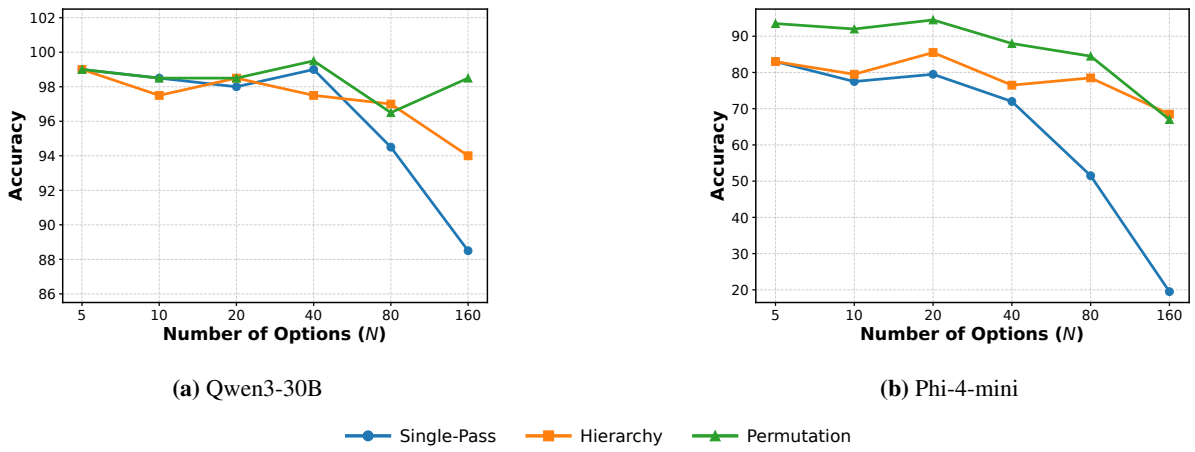
(a) Llama-3.2-3B **(b)** Qwen3-4B

Figure 21: **Additional sequential revision dynamics across models.** Across additional model families, later-arriving gold answers progressively lose the ability to revise earlier candidates under large-option comparison.



(a) Hierarchical partitioning **(b)** Permutation-based inference

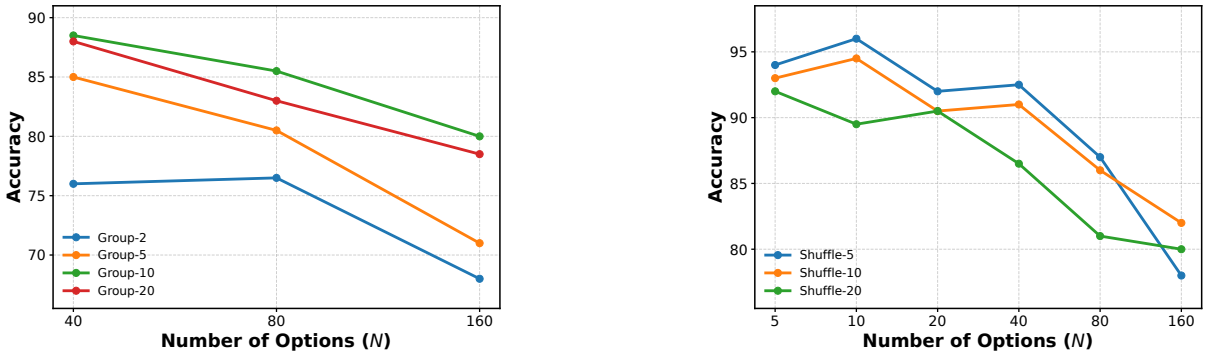
Figure 22: **Mechanisms of inference-time restructuring.** (a) Hierarchical partitioning reduces comparison difficulty through local candidate selection, but introduces irreversible early elimination errors as the number of options increases. (b) Permutation-based inference mitigates positional bias by aggregating predictions across reordered candidate views, improving robustness during final-stage comparison among surviving candidates.



(a) Qwen3-30B

(b) Phi-4-mini

Figure 23: **Generalization across model scales.** Hierarchical grouping and permutation-based inference consistently mitigate large-option degradation across both small and large models.



(a) Effect of group size

(b) Effect of number of permutations

Figure 24: **Effect of hyperparameters.** Larger group sizes improve hierarchical partitioning by preserving stronger candidates during local selection, while permutation-based inference benefits from a moderate number of shuffled views but exhibits diminishing returns at larger scales.