When Chain-of-Thought Backfires: Evaluating Prompt Sensitivity in Medical Language Models

Author Name
Institution
email@institution.edu

Abstract

Large language models are increasingly deployed in medical settings, yet their sensitivity to prompt formatting remains poorly characterized. We evaluate MedGemma (4B and 27B variants) on MedMCQA (4,183 questions) and PubMedQA (1,000 questions). Our experiments reveal concerning findings: chain-of-thought prompting *decreases* accuracy by 5.7% compared to direct answering; few-shot examples degrade performance by 11.9% while increasing position bias from 0.14 to 0.47; shuffling answer options causes the model to change predictions 59.1% of the time with accuracy dropping up to 27.4 percentage points; and truncating context to 50% causes accuracy to plummet below the no-context baseline. These results demonstrate that prompt engineering techniques validated on general-purpose models do not transfer to domain-specific medical LLMs.

1 Introduction

Large language models have achieved impressive performance on medical licensing exams, with GPT-4 exceeding passing thresholds by over 20 points [Nori et al., 2023] and Med-PaLM 2 reaching 86.5% on MedQA [Singhal et al., 2023b]. These results have fueled enthusiasm for deploying LLMs in clinical decision support. However, benchmark accuracy tells only part of the story—how models respond to variations in prompt format, question ordering, and context presentation remains poorly understood, despite being critical for real-world deployment where inputs are rarely formatted identically to benchmark conditions.

We focus on MedGemma [DeepMind, 2024], Google's medical-specialist LLM built on the Gemma architecture. A widely held belief in the LLM community is that certain prompting strategies reliably improve performance. Chain-of-thought prompting, which instructs models to reason step-by-step before answering, has shown consistent gains on mathematical and logical reasoning tasks [Wei et al., 2022]. Few-shot learning, where examples are provided in-context, helps models understand desired output formats [Brown et al., 2020]. These techniques are often treated as "best practices" that should transfer across domains and model families.

But should they? Domain-specific models may have internalized different patterns during training. A model trained extensively on medical literature might already encode structured clinical reasoning, making explicit chain-of-thought prompts redundant or even counterproductive. Similarly, few-shot examples drawn from one medical specialty might prime the model with concepts that are irrelevant or misleading for questions in other specialties.

We present a systematic evaluation of MedGemma's sensitivity to prompt variations across three experimental conditions. First, we conduct a prompt ablation study comparing zero-shot, chain-of-thought, and few-shot strategies on 4,183 MedMCQA questions, measuring both accuracy and position bias. Second, we test option order sensitivity by shuffling answer choices and measuring how often the model changes its prediction—a direct test of whether responses reflect semantic understanding or superficial position cues. Third, we evaluate evidence conditioning on 1,000 PubMedQA questions, systematically varying context completeness to understand how partial information affects accuracy. Our findings challenge conventional assumptions about prompt engineering in medical AI and have important implications for clinical deployment.

2 Related Work

Medical Language Models. The application of large language models to medicine has accelerated rapidly. Med-PaLM was the first AI system to surpass the 60% passing mark on USMLE-style questions, and Med-PaLM 2 subsequently achieved 86.5% on MedQA [Singhal et al., 2023a,b]. MedGemma, introduced at Google I/O 2025, achieves 87.7% on MedQA [DeepMind, 2024]. However, Google emphasizes that MedGemma requires further validation before clinical use—a caveat our results strongly support.

Prompt Sensitivity. The fragility of LLM predictions to prompt variations is well-documented in general domains. Lu et al. [2022] demonstrated that few-shot example ordering significantly affects performance. The ProSA framework found performance can swing by up to 45% depending on prompt formulation [Jia et al., 2024]. These findings suggest that apparent model capabilities may be artifacts of prompt formatting rather than genuine understanding.

Position Bias. Zheng et al. [2024] showed that modern LLMs exhibit "selection bias"—they prefer specific option IDs regardless of content. In their analysis, llama-30B selected options A/B/C/D with frequencies of 34.6%/27.3%/22.3%/15.8% despite balanced ground truth. This bias stems from token-level preferences where models assign more probability mass to certain option tokens.

Chain-of-Thought Limitations. Recent work challenges CoT's universal benefit. Sprague et al. [2024] identified tasks where CoT reduces accuracy by up to 36.3%, paralleling cognitive psychology research on when deliberation hurts human performance. The Wharton report found that CoT introduces more variability and can trigger errors on questions models would otherwise answer correctly [Meincke et al., 2024]. In medical domains specifically, Omar et al. [2024] found complex prompting does not significantly outperform simpler approaches.

RAG Challenges. Retrieval-augmented generation systems face challenges with incomplete context. RAG-Bench demonstrated that relevant-but-incomplete retrieved context can actively mislead models, sometimes performing worse than no retrieval [Fang et al., 2024]. The "lost-in-the-middle" phenomenon shows that information position within context significantly impacts response quality [Liu et al., 2024].

3 Methods

3.1 Models and Datasets

We evaluate MedGemma-4B, the 4-billion parameter instruction-tuned variant at bfloat16 precision, and MedGemma-27B, the 27-billion parameter model requiring full bfloat16 precision on 80GB A100 GPUs. Initial experiments with 4-bit quantization on the 27B model produced NaN logits—a notable finding suggesting that quantization techniques validated on general models may not transfer to medical-specialist architectures.

We use two standard medical QA benchmarks. MedMCQA [Pal et al., 2022] contains questions from Indian medical entrance examinations across 21 subjects; we use the 4,183-question validation split. PubMedQA [Jin et al., 2019] contains research questions derived from PubMed titles that must be answered using abstracts; we use the 1,000-question labeled subset.

3.2 Experimental Conditions

Experiment 1: Prompt Ablation. We test five prompting strategies on MedMCQA: (1) zero-shot direct, presenting the question and requesting only the answer letter; (2) zero-shot CoT, adding "think step by step"; (3) few-shot direct, providing three example Q&A pairs; (4) few-shot CoT, providing three examples with reasoning traces; and (5) answer-only, a minimal prompt with no instructions.

Experiment 2: Option Order Sensitivity. We apply five transformations to each question: original order, random shuffle, rotate-1 (cyclic shift by one), rotate-2 (shift by two), and distractor swap (exchange incorrect options while preserving correct answer position). We measure the flip rate—how often the model changes its answer when options are reordered.

Experiment 3: Evidence Conditioning. On PubMedQA, we vary context: question-only (no context), full abstract, truncated 50%, truncated 25%, background-only (introduction sentences), and results-only (conclusion sentences). This tests how context completeness and type affect accuracy.

3.3 Metrics

We report accuracy with 95% bootstrap confidence intervals (1,000 iterations). Position bias is computed as the absolute difference between predicted and ground truth answer distributions across positions A-D. For option-order experiments, we compute the flip rate: the proportion of questions where the model's prediction changes when options are reordered.

4 Results

4.1 Prompt Ablation

Table 1 shows accuracy across prompting strategies. Zero-shot direct achieves the highest accuracy at 47.6%, while chain-of-thought *reduces* accuracy by 5.7 percentage points. Few-shot examples cause an even larger degradation of 11.9%, while simultaneously increasing position bias from 0.137 to 0.472—indicating the model learns spurious patterns from examples rather than useful formats.

Table 1: Prompt ablation results on MedMCQA (n=4,183). Random baseline is 25%.

Condition	**Accuracy**	**95% CI**	**Pos. Bias**
Zero-shot direct	47.6%	[46.1%, 49.1%]	0.137
Zero-shot CoT	41.9%	[40.4%, 43.3%]	0.275
Few-shot direct	35.7%	[34.3%, 37.0%]	0.472
Few-shot CoT	40.8%	[39.4%, 42.3%]	0.413
Answer-only	43.0%	[41.5%, 44.6%]	0.096

4.2 Option Order Sensitivity

Table 2 reveals extreme sensitivity to option ordering. The mean flip rate is 59.1%—the model changes its answer more often than not when options are shuffled. Rotation perturbations cause the largest accuracy drops (up to 27.4%), while distractor swaps show smaller impact (−8.9%), confirming that position rather than distractor content drives fragility.

Table 2: Option order sensitivity on MedMCQA (n=4,183). Random baseline is 25%.

Perturbation	**Accuracy**	**Drop**
Original	47.6%	—
Random shuffle	29.2%	−18.4%
Rotate-1	20.2%	−27.4%
Rotate-2	24.3%	−23.3%
Distractor swap	38.7%	−8.9%
Mean flip rate	59.1%	

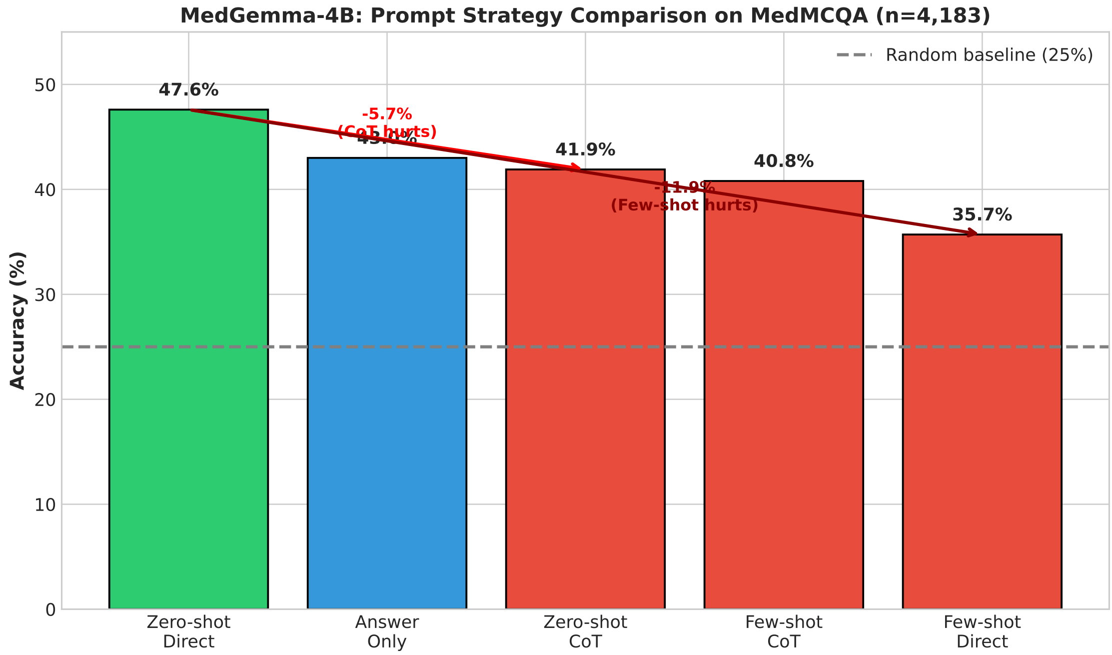

Figure 1: MedGemma-4B accuracy across prompt strategies. Zero-shot direct outperforms all other strategies including chain-of-thought (−5.7%) and few-shot (−11.9%).

4.3 Evidence Conditioning

Table 3 shows context substantially affects PubMedQA performance. Most critically, truncated context performs *worse* than no context: 50% truncation yields 14.1% (4B) and 23.4% (27B), far below question-only baselines of 36.7% and 31.0%. This indicates partial context actively misleads rather than simply providing less information.

Surprisingly, MedGemma-27B achieves its best performance with results-only context (40.0%), which *exceeds* its full-context accuracy (38.2%). The 27B model also underperforms 4B on most conditions, suggesting scale does not guarantee robustness.

Table 3: Evidence conditioning on PubMedQA (n=1,000). Random baseline is 33.3%.

Condition	**MedGemma-4B**	**MedGemma-27B**
Question only	36.7%	31.0%
Full context	45.0%	38.2%
Truncated 50%	14.1%	23.4%
Truncated 25%	13.1%	18.6%
Background only	26.5%	19.8%
Results only	41.7%	**40.0%**

5 Discussion

5.1 Why Chain-of-Thought Hurts

The 5.7% accuracy drop from CoT prompting aligns with recent findings that deliberation can reduce performance on certain tasks [Sprague et al., 2024]. MedGemma was trained extensively on medical text

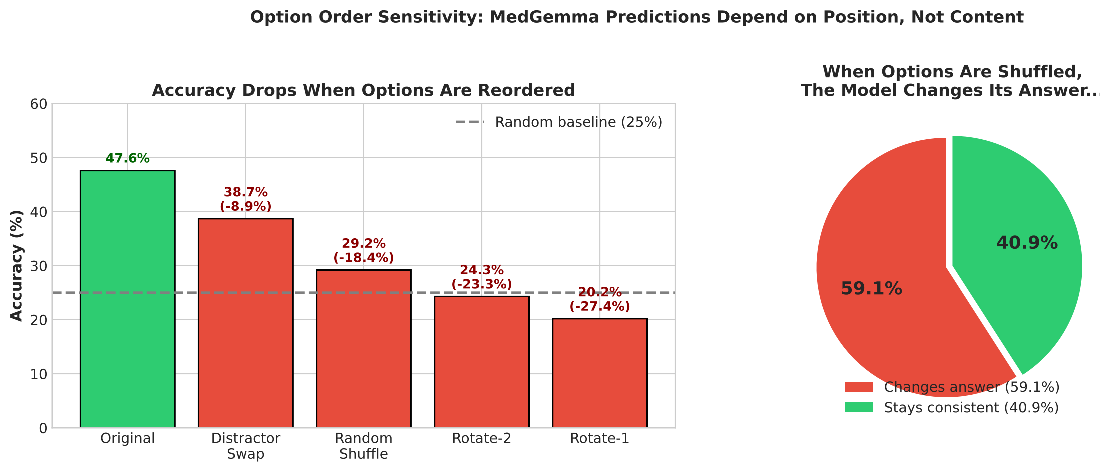

Figure 2: Left: Accuracy drops substantially when options are reordered. Right: The model changes its answer 59.1% of the time when options are shuffled.

and may have internalized domain reasoning patterns; forcing explicit step-by-step logic may override these learned patterns with less reliable deliberation.

Case-level analysis reveals the mechanism: CoT prompting changed answers on 1,262 of 4,183 questions, hurting 750 (direct correct, CoT wrong) while helping only 512—a net loss of 238 questions. The predominant failure pattern involves verbose reasoning (90.7% exceeded 500 characters) where longer chains create opportunities for errors to compound. We also observed self-contradiction (25.6% contained hedge words introducing conflicting logic) and confident wrong conclusions (11.1% stated "therefore" before incorrect answers).

The characteristic failure mode: the model correctly identifies relevant medical concepts early in reasoning, considers alternatives, then talks itself into the wrong answer. In one case involving organophosphate poisoning, CoT correctly identified the condition and atropine's role as antidote, then continued deliberating and selected neostigmine—which would worsen the condition.

5.2 The 59% Flip Rate Problem

MedGemma changes its answer 59.1% of the time when options are shuffled, far exceeding random noise. The maximum flip rate of 72.9% for certain perturbations means that for nearly three-quarters of questions, answers depend more on option position than content. This magnitude exceeds typical findings [Zheng et al., 2024], suggesting medical-specialist training may not mitigate—and could exacerbate—position bias.

For clinical applications, this fragility is unacceptable. A diagnostic support system that changes recommendations based on option ordering provides no reliable signal to clinicians and undermines the premise of using AI to assist medical decision-making.

5.3 Partial Context Actively Misleads

Truncating context to 50% yields 14.1% accuracy while no context achieves 36.7%—a 22.6 point gap indicating partial context actively misleads the model. This has direct implications for RAG systems in medical applications: incomplete retrieval may be worse than no retrieval, confirming findings from RAG-Bench [Fang et al., 2024].

Interestingly, results-only context (41.7% for 4B, 40.0% for 27B) nearly matches or exceeds full context, while background-only achieves just 26.5%/19.8%. Both models benefit from conclusions rather than methodological background, suggesting RAG systems should prioritize high-information-density content.

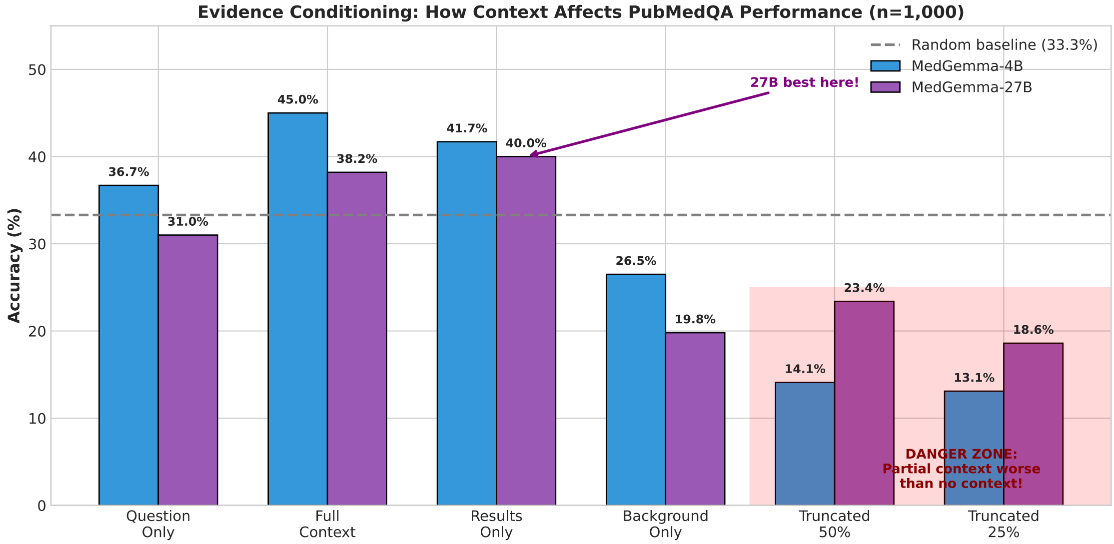

Figure 3: Evidence conditioning results. Truncated context performs worse than no context (danger zone). MedGemma-27B achieves best performance with results-only context.

5.4 Scale and Robustness

MedGemma-27B underperforms 4B on evidence conditioning (38.2% vs 45.0% with full context), demonstrating that medical benchmark performance does not scale uniformly with model size. However, 27B shows a different pattern: its best performance comes from results-only (40.0%), exceeding full-context accuracy. This "less is more" finding suggests larger models may be more susceptible to distraction from verbose context but respond well to concentrated information. For deployment, this implies larger models may require selective rather than comprehensive retrieval strategies.

6 Conclusion

Our evaluation reveals that standard prompt engineering techniques do not reliably improve—and may actively harm—medical question answering performance. Chain-of-thought decreases accuracy by 5.7% while increasing position bias; few-shot examples decrease accuracy by 11.9% while tripling position bias; shuffling options causes 59.1% flip rates with accuracy drops up to 27.4 percentage points; and truncated context performs worse than no context.

For practitioners deploying medical LLMs, we recommend: (1) default to zero-shot direct prompting until evidence justifies added complexity; (2) test option order sensitivity before deployment and consider averaging across orderings or using debiasing techniques [Zheng et al., 2024]; (3) validate retrieval completeness for RAG systems, as incomplete context can be worse than none; and (4) for larger models, prefer selective retrieval of high-density information over comprehensive retrieval.

The extreme sensitivity to prompt variations raises fundamental questions about what benchmark accuracy measures. Before deploying medical LLMs, rigorous empirical validation on specific use cases is essential—assumed best practices from general-purpose models do not transfer.

Acknowledgments

We thank the MedGemma team at Google for releasing open model weights. Experiments were conducted on NVIDIA A100 GPUs.

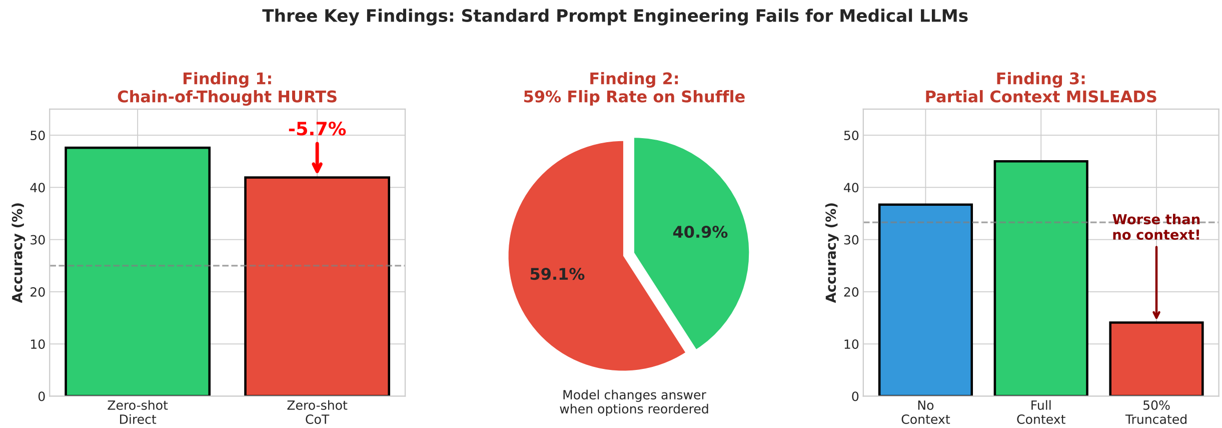

Figure 4: Summary: CoT reduces accuracy 5.7%; option shuffling causes 59.1% flip rate; truncated context performs worse than no context.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Google DeepMind. Medgemma: Medical language models. *Google AI Blog*, 2024. Technical report.

Yuxuan Fang et al. Rag-bench: A benchmark for retrieval-augmented generation. In *Proceedings of NeurIPS Datasets and Benchmarks Track*, 2024.

Jingming Jia, Shuang Xing, Zixuan Hu, Zhonghai Wang, Guangtao Zhai, and Xiao-Ping Zhang. Prosa: Assessing and understanding the prompt sensitivity of llms. In *Findings of the Association for Computational Linguistics: EMNLP*, 2024.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *Proceedings of EMNLP-IJCNLP*, pages 2567–2577, 2019.

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

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *Proceedings of ACL*, pages 8086–8098, 2022.

Lennart Meincke, Ethan R Mollick, Lilach Mollick, and Dan Shapiro. The decreasing value of chain of thought in prompting. Technical report, Wharton Generative AI Labs, 2024. SSRN 5285532.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

Rana Omar et al. A comparative evaluation of chain-of-thought-based prompt engineering techniques for medical question answering. *Computers in Biology and Medicine*, 2024.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. *Proceedings of the Conference on Health, Inference, and Learning*, pages 248–260, 2022.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023a.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023b.

Zayne Sprague, Jiasheng Pei, Akari Chaturvedi, Zeyao Lee, Nan Gao, Yige Chen, and Rui Zhang. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. *arXiv preprint arXiv:2410.21333*, 2024.

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

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. Spotlight paper.

A Detailed Analysis

A.1 Chain-of-Thought Failure Patterns

Of 4,183 questions, CoT changed 1,262 answers: 750 hurt vs. 512 helped. Failure patterns:

Verbose reasoning (90.7%): 680 cases exceeded 500 characters. Longer chains create compounding error opportunities.

Self-contradiction (25.6%): 192 cases contained “however” or “but” introducing conflicting mid-reasoning logic.

Confident wrong conclusions (11.1%): 83 cases stated “therefore” before incorrect answers.

A.2 Position Bias Details

MedMCQA ground truth: A: 32.2%, B: 25.1%, C: 21.4%, D: 21.3%. Zero-shot direct predicts A 45.9% (overweight 13.7%). Few-shot direct predicts A 76% despite 32% correct rate.

A.3 Threats to Validity

Parsing errors: <2% across conditions. CoT slightly harder (2.1% vs 1.4%), but cannot explain 5.7% gap.

Dataset imbalance: We report differences from ground truth. The 59.1% flip rate cannot be explained by imbalance.

Contamination: We focus on relative robustness. Memorized answers should be prompt-robust; large degradations indicate genuine sensitivity.

A.4 Limitations

Limited model coverage (MedGemma, BioMistral-7B); 27B requires 80GB GPUs; MCQ and yes/no/maybe formats only; English only; single-turn evaluation.