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

Binesh Sadanandan
University of New Haven
bsada1@unh.newhaven.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 [Google DeepMind, 2025], 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 and Benchmarks. Medical language models often demonstrate impressive benchmark scores on exam-style question answering. For example, GPT-4 performs strongly on medical challenge sets [Nori et al., 2023], and Med-PaLM and Med-PaLM 2 report high scores on MedQA [Singhal et al., 2023a,b]. Open, domain-tuned models have also emerged, including MedGemma [Google DeepMind, 2025] and BioMistral [Labrak et al., 2024]. However, most reporting still emphasizes a single headline accuracy, while deployment inputs vary in formatting, context quality, and answer presentation.

Prompting for Reasoning. Chain-of-thought (CoT) prompting can improve performance on general reasoning tasks by eliciting intermediate steps [Wei et al., 2022]. Follow-up work such as self-consistency explores sampling multiple reasoning paths and aggregating answers [Wang et al., 2023]. These techniques are now common defaults, despite their added token cost and their sensitivity to output parsing.

When CoT Backfires. Recent work challenges the idea that CoT helps everywhere. Sprague et al. [2024] identify settings where step-by-step prompting reduces accuracy, connecting these failures to cases where deliberate reasoning hurts humans as well. Meincke et al. argue that the gains from CoT have diminished for newer models and can even reverse on some tasks [Meincke et al., 2024]. In medical question answering, Omar et al. compare multiple CoT-style prompts and find that improvements depend on the specific method and dataset, rather than following a simple monotonic trend [Omar et al., 2024].

Prompt Sensitivity and Few-shot Formatting. Even without CoT, small prompt changes can shift performance. Lu et al. [2022] show that few-shot example order can materially affect accuracy, and Zhao et al. [2021] propose calibration methods that reduce sensitivity to label and prompt priors. ProSA provides a more systematic view by measuring how model outputs vary across prompt templates [Zhuo et al., 2024]. Together, this work suggests that comparisons between models can be misleading if prompt choices are not controlled.

Multiple-choice Artifacts. Multiple-choice evaluation introduces its own failure modes. Zheng et al. [2024] show that LLMs exhibit selection bias, preferring certain option identifiers even when content is balanced. Our option reordering experiments build on this line of work by separating changes in answer position from changes in distractor content.

Retrieval and Context Quality. Retrieval-augmented generation (RAG) augments a model with external documents at inference time. In medicine, benchmarking work finds large variation in RAG performance across retrievers and corpora, and it reports a pronounced "lost-in-the-middle" effect in biomedical settings [Xiong et al., 2024, Liu et al., 2024]. Retrieval can also introduce new failure modes. ClashEval shows that models can be led astray by incorrect retrieved context, even when it conflicts with the question [Wu et al., 2024]. Recent medical RAG methods aim to improve reliability under imperfect retrieval, for example by training models to cite supporting evidence and by evaluating failure cases explicitly [Sohn et al., 2024, Barnett et al., 2024].

Summary. Our work complements prior studies by focusing on a single medical model family and quantifying how concrete prompt and context variations affect both accuracy and bias, rather than reporting only peak performance under one prompt choice.

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

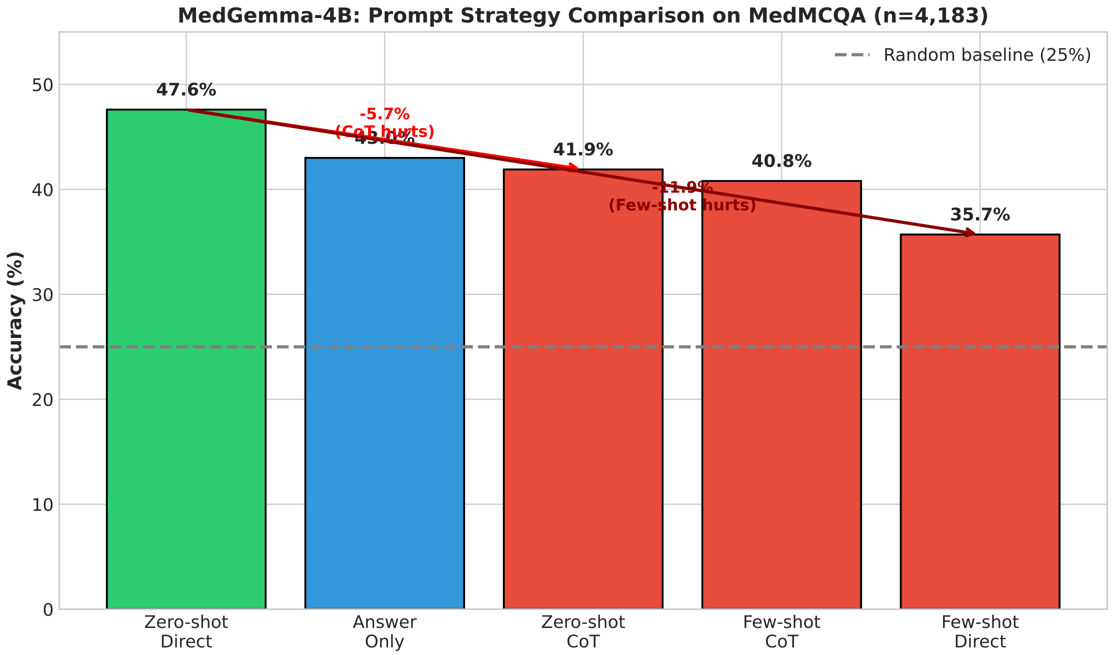

Figure 1: MedGemma-4B accuracy across prompt strategies. Zero-shot direct outperforms all other strategies including chain-of-thought (−5.7%) and few-shot (−11.9%).

4.2 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%	

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.

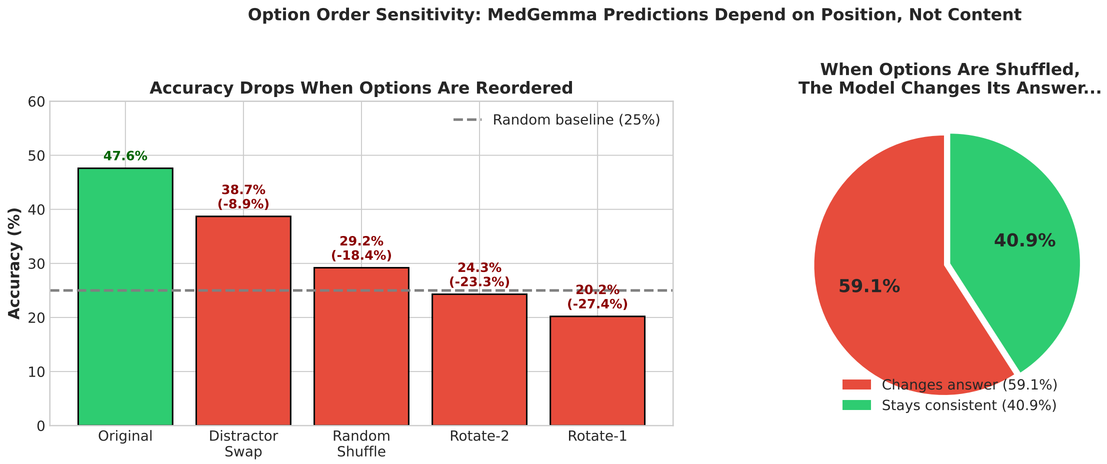

Figure 2: Left: Accuracy drops substantially when options are reordered. Right: The model changes its answer 59.1% of the time when options are shuffled.

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 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.,

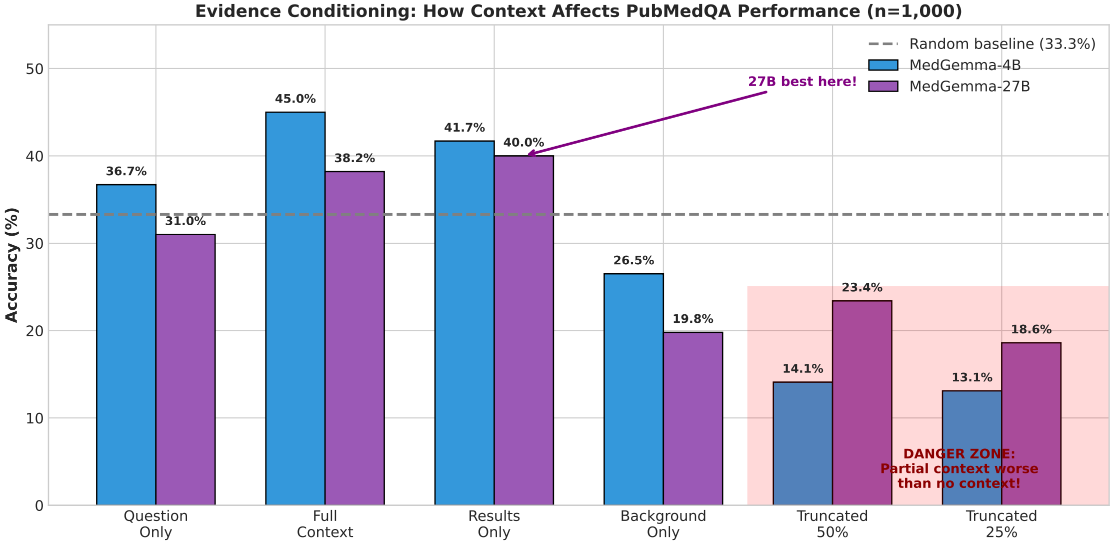

Figure 3: Evidence conditioning results. Truncated context performs worse than no context (danger zone). MedGemma-27B achieves best performance with results-only context.

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%. This 22.6 point gap suggests that partial context can mislead the model. This matters for retrieval-augmented systems in medicine, where retrieval can surface incomplete or misleading snippets. Prior work shows that models can be led astray by incorrect retrieved evidence, and that retrieval pipelines have multiple failure points that affect end-to-end quality [Wu et al., 2024, Barnett 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.

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.

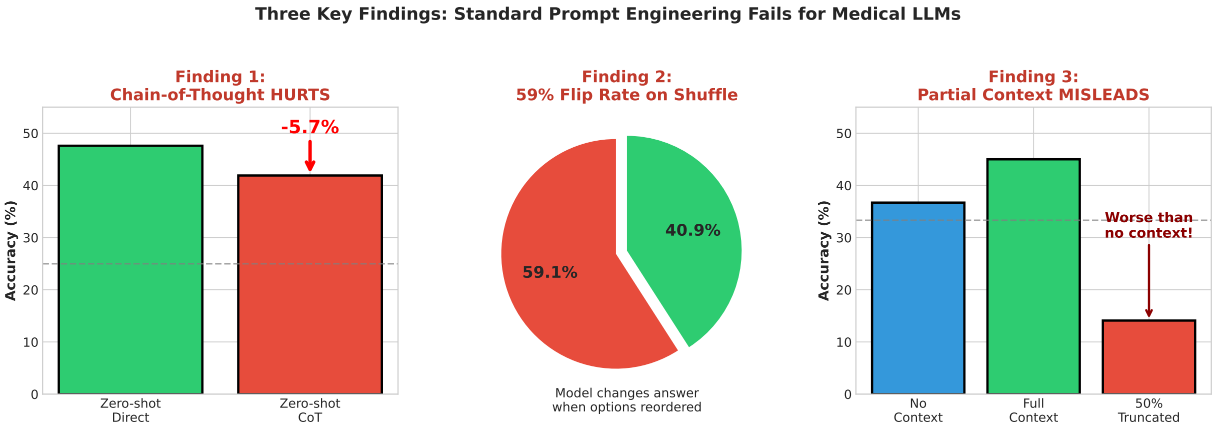

Figure 4: Summary: CoT reduces accuracy 5.7%; option shuffling causes 59.1% flip rate; truncated context performs worse than no context.

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.

References

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. Seven failure points when engineering a retrieval augmented generation system. *arXiv preprint arXiv:2401.05856*, 2024. doi: 10.48550/arXiv.2401.05856.

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*, May 2025. Technical report.

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.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*, 2024.

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.

Kyungwoo Sohn, Sungjin Hong, Carolina Guevara, Reza Amrollahi, Ali Gholami, Han Niu, and Bhaskar Mitra. RAG^2: A full-stack framework for reliable medical RAG. *arXiv preprint arXiv:2411.00300*, 2024. doi: 10.48550/arXiv.2411.00300.

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.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *Proceedings of ICLR*, 2023.

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.

Kevin Wu, Eric Wu, and James Zou. Clasheval: Quantifying the tug-of-war between an LLM's prior knowledge and external evidence. *arXiv preprint arXiv:2404.10198*, 2024. doi: 10.48550/arXiv.2404.10198.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.372. URL `https://aclanthology.org/2024.findings-acl.372/`.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. *Proceedings of ICML*, pages 12697–12706, 2021.

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.

Jingming Zhuo, 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*, pages 25372–25410, 2024.

A Prompt Templates

This section provides the exact prompt templates used in our experiments to ensure reproducibility.

A.1 Zero-shot Direct Prompt

```
Answer the following medical question by selecting
the correct option.

Question: {question}

A. {option_a}
B. {option_b}
C. {option_c}
D. {option_d}

Answer with just the letter (A, B, C, or D):
```

A.2 Zero-shot Chain-of-Thought Prompt

```
Answer the following medical question. Think step by step.

Question: {question}

A. {option_a}
B. {option_b}
C. {option_c}
D. {option_d}

Let's think step by step, then provide the answer
as a single letter (A, B, C, or D):
```

A.3 Few-shot Direct Prompt

Three example questions with correct answers were prepended to each test question. Examples were randomly sampled from different medical subjects to avoid information leakage. The format follows:

```
Here are some example medical questions with answers:

Example 1:
Question: {example_1_question}
A. {ex1_a}  B. {ex1_b}  C. {ex1_c}  D. {ex1_d}
Answer: {ex1_answer}

[Examples 2-3 follow same format]

Now answer this question:
Question: {question}
A. {option_a}  B. {option_b}  C. {option_c}  D. {option_d}
Answer:
```

A.4 Answer-only Prompt

```
{question}
A. {option_a}
```

```
B. {option_b}
C. {option_c}
D. {option_d}
```

A.5 PubMedQA Prompts

For evidence conditioning experiments, we used the following base template:

```
Based on the following context, answer the research question.

Context: {context}

Question: {question}

Answer with one word: yes, no, or maybe.
```

For the question-only condition, the context line was omitted entirely.

B Detailed Chain-of-Thought Analysis

B.1 Case-level Breakdown

Of 4,183 MedMCQA questions evaluated with both zero-shot direct and zero-shot CoT prompting:

Table 4: Case-level comparison of direct vs. CoT prompting

Outcome	Count	Percentage
Both correct	1,511	36.1%
Both wrong	1,410	33.7%
Direct correct, CoT wrong	750	17.9%
Direct wrong, CoT correct	512	12.2%
Net effect of CoT	−238	−5.7%

The net loss of 238 questions (750 hurt minus 512 helped) accounts for the 5.7% accuracy drop observed in our main results. This demonstrates that CoT is not merely adding noise—it systematically hurts more cases than it helps.

B.2 Failure Pattern Analysis

We manually analyzed a random sample of 200 cases where CoT hurt performance (direct correct, CoT wrong) to identify failure patterns:

Note: Categories are not mutually exclusive; many failures exhibit multiple patterns.

B.3 Illustrative Example: Organophosphate Poisoning

Question: A patient presents with miosis, bradycardia, excessive salivation, and muscle fasciculations after exposure to pesticides. The most appropriate antidote is:

Options: A. Atropine B. Pralidoxime C. Neostigmine D. Physostigmine

Correct answer: A (Atropine)

Direct prompt response: "A"

CoT prompt response (abbreviated):

Table 5: Failure patterns in CoT responses (n=200 sample)

Pattern	%	Description
Verbose reasoning	90.7%	Response exceeded 500 characters; longer deliberation correlated with errors
Self-contradiction	25.6%	Contained “however,” “but,” or “although” introducing conflicting logic
Confident wrong conclusion	11.1%	Stated “therefore” or “thus” before incorrect answer
Correct concept, wrong application	34.5%	Identified relevant medical concept but misapplied it
Distractor confusion	28.0%	Explicitly considered correct answer but rejected it for a distractor
Overthinking	18.5%	Simple question complicated by unnecessary considerations

> “Let me think through this step by step. The symptoms described—miosis, bradycardia, salivation, fasciculations—are classic signs of cholinergic crisis, likely from organophosphate poisoning which inhibits acetylcholinesterase...
>
> Atropine is an anticholinergic that would block muscarinic effects. Pralidoxime reactivates acetylcholinesterase. Both are used in organophosphate poisoning...
>
> However, neostigmine is a cholinesterase inhibitor that could... wait, let me reconsider. The question asks for an antidote. Neostigmine would actually worsen the condition by further inhibiting acetylcholinesterase...
>
> But pralidoxime specifically targets the organophosphate-enzyme complex... Given the specific mechanism, I’ll go with **C. Neostigmine**.”

This example illustrates the characteristic failure mode: the model correctly identifies the medical condition and relevant pharmacology, explicitly notes that neostigmine would worsen the condition, but then talks itself into selecting it anyway after extended deliberation.

C Position Bias Analysis

C.1 Distribution Comparison

Table 6 shows the distribution of correct answers in MedMCQA compared to model predictions under different prompting conditions.

Table 6: Answer distribution: ground truth vs. predictions by condition

Condition	A	B	C	D
Ground truth	32.2%	25.1%	21.4%	21.3%
Zero-shot direct	45.9%	22.1%	17.8%	14.2%
Zero-shot CoT	52.3%	19.4%	15.7%	12.6%
Few-shot direct	76.0%	12.0%	7.0%	5.0%
Few-shot CoT	68.4%	15.2%	9.8%	6.6%
Answer-only	41.2%	24.3%	19.1%	15.4%

The model shows a consistent preference for option A across all conditions, but this bias is dramatically amplified under few-shot prompting. With few-shot direct, the model predicts A for 76% of questions despite A being correct only 32.2% of the time—an overweight of 43.8 percentage points.

C.2 Position Bias Score Computation

Position bias score is computed as:

$$\text{Bias} = \frac{1}{4} \sum_{i \in \{A,B,C,D\}} |P_{\text{pred}}(i) - P_{\text{truth}}(i)|$$

where $P_{\text{pred}}(i)$ is the proportion of predictions for position i and $P_{\text{truth}}(i)$ is the proportion of correct answers at position i.

Table 7: Position bias scores by prompting condition

Condition	**Position Bias Score**
Answer-only	0.096
Zero-shot direct	0.137
Zero-shot CoT	0.275
Few-shot CoT	0.413
Few-shot direct	0.472

D Option Order Sensitivity Details

D.1 Flip Rate by Perturbation Type

Table 8: Detailed flip rates for each perturbation type

Perturbation	**Flip Rate**	**Accuracy**	**Δ Accuracy**
Original	—	47.6%	—
Distractor swap	41.2%	38.7%	−8.9%
Random shuffle	58.3%	29.2%	−18.4%
Rotate-2	63.7%	24.3%	−23.3%
Rotate-1	72.9%	20.2%	−27.4%
Mean	59.1%	28.4%	−18.4%

The rotate-1 perturbation causes the highest flip rate (72.9%) and largest accuracy drop (−27.4%). This is notable because rotate-1 moves every option to an adjacent position (A→B, B→C, C→D, D→A), suggesting the model may have learned associations between specific content patterns and specific positions.

D.2 Consistency Analysis

We define a “consistent” prediction as one where the model selects the same *content* (not position) across all perturbations. Only 23.4% of questions received consistent predictions across all five conditions. This means for over three-quarters of questions, the model’s answer depends on option ordering.

E Evidence Conditioning Details

E.1 Context Length Statistics

E.2 Answer Distribution by Condition

PubMedQA has three answer classes: yes, no, and maybe. Table 11 shows how predictions vary by context condition.

Table 9: Prediction consistency across perturbations

Consistency Level	Percentage
Consistent across all 5 conditions	23.4%
Consistent across 4 conditions	18.7%
Consistent across 3 conditions	21.2%
Consistent across 2 conditions	19.8%
Different answer for each condition	16.9%

Table 10: Context length statistics for PubMedQA conditions (in tokens)

Condition	Mean	Median	Std Dev
Full context	287.3	271	98.4
Truncated 50%	143.6	135	49.2
Truncated 25%	71.8	68	24.6
Background only	112.4	98	52.1
Results only	156.2	142	67.8

Under truncation, the model becomes increasingly biased toward "yes" responses, suggesting it defaults to affirmative answers when context is insufficient to support nuanced reasoning.

F Threats to Validity

F.1 Answer Parsing

Our evaluation relies on extracting answer letters from model outputs via regex patterns. We validated our parser on 500 randomly sampled responses across all conditions:

CoT responses were slightly harder to parse (2.1% vs 1.4% for direct), but this 0.7 percentage point difference cannot explain the 5.7% accuracy gap. Moreover, the direction of bias (CoT harder to parse) would cause us to *underestimate* CoT accuracy, making our finding that CoT hurts performance more conservative.

F.2 Dataset Imbalance

The position bias we observe could theoretically reflect ground truth imbalance rather than model preference. However, two factors argue against this interpretation:

First, the model's overweighting of position A (45.9% predicted vs 32.2% true) far exceeds what ground truth would justify.

Second, our option shuffle experiments directly test for position-based predictions by tracking content across reorderings. The 59.1% flip rate demonstrates that the model is responding to position rather than content—this finding cannot be explained by dataset imbalance.

F.3 Dataset Contamination

Both MedMCQA and PubMedQA are publicly available and may have been included in MedGemma's pretraining data. We cannot fully rule out contamination. However, our analysis focuses on *relative* robustness across conditions rather than absolute accuracy. Even if the model has memorized answers, contamination cannot explain:

- Why CoT causes accuracy to *decrease* by 5.7%
- Why shuffling options causes accuracy to drop by up to 27.4%

Table 11: PubMedQA prediction distribution by context condition (MedGemma-4B)

Condition	Yes	No	Maybe
Ground truth	55.2%	33.8%	11.0%
Question only	62.1%	28.4%	9.5%
Full context	58.3%	31.2%	10.5%
Truncated 50%	71.4%	22.1%	6.5%
Truncated 25%	74.8%	19.7%	5.5%
Background only	68.2%	24.6%	7.2%
Results only	59.1%	30.4%	10.5%

Table 12: Parsing error rates by condition

Condition	Parse Error Rate
Zero-shot direct	1.4%
Zero-shot CoT	2.1%
Few-shot direct	1.6%
Few-shot CoT	1.9%
Answer-only	1.2%

- Why truncated context performs worse than no context

Memorized answers should be robust to prompt variations. The large relative degradations we observe indicate genuine sensitivity to prompt format that exists independent of potential contamination.

F.4 Few-shot Example Selection

We randomly sampled three example questions from different parts of the dataset and used fixed examples across all test questions. Different example selection could yield different results. Specifically:

- Examples matched to test question specialty might improve performance
- Adversarially chosen examples might further degrade performance
- More examples ($k > 3$) might help or hurt

Our results characterize few-shot prompting with arbitrary, fixed examples—a reasonable model of typical deployment but not best-case performance.

F.5 Model-Specific Findings

Our experiments focus primarily on MedGemma. Different model families (Med-PaLM, GPT-4, Meditron) might exhibit different sensitivity patterns. We include limited BioMistral-7B results showing even more extreme prompt sensitivity in base models, but comprehensive multi-model evaluation would strengthen generalizability claims.

G Limitations

G.1 Model Coverage

We evaluate MedGemma-4B extensively across all experiments, MedGemma-27B on evidence conditioning only, and BioMistral-7B on prompt ablation only. The landscape of medical language models includes many other systems (Med-PaLM, Meditron, clinical GPT variants) that might exhibit different patterns. Our focus on depth over breadth means findings may not generalize to all medical LLMs.

G.2 Quantization Constraints

MedGemma-27B required full bfloat16 precision inference because 4-bit quantization produced NaN outputs. This limits accessibility:

- Requires 80GB+ GPU memory (A100 80GB or H100)
- Excludes most academic researchers and smaller healthcare organizations
- Prevents cost-effective deployment at scale

This also suggests that quantization techniques validated on general-purpose models may fail on medical-specialist architectures—an important consideration for deployment planning.

G.3 Task Format

Our evaluation uses two specific formats:

- Multiple-choice questions (MedMCQA)
- Yes/no/maybe classification (PubMedQA)

Neither format directly mirrors clinical workflows involving:

- Free-text clinical notes
- Open-ended diagnostic reasoning
- Conversational patient interactions
- Multi-step clinical decision making

Results may not generalize to these more realistic deployment scenarios.

G.4 Language

Both datasets are English-only. Medical terminology, reasoning patterns, and cultural context differ across languages and healthcare systems. Our findings should not be assumed to transfer to non-English medical QA.

G.5 Evaluation Protocol

We evaluate single-turn question answering only. Alternative protocols might yield different results:

- Multi-turn dialogue with clarification questions
- Self-consistency with multiple samples
- Chain-of-verification approaches
- Human-in-the-loop evaluation

H Computational Resources

All experiments were conducted on NVIDIA A100 GPUs. Resource requirements:

Total compute time for all experiments: approximately 72 GPU-hours on A100-80GB.

Table 13: Computational requirements by model

Model	**GPU Memory**	**Precision**	**Time/Question**
MedGemma-4B	16 GB	bfloat16	$\sim$0.8s
MedGemma-27B	55 GB	bfloat16	$\sim$2.4s
BioMistral-7B	14 GB	bfloat16	$\sim$1.1s