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), Google's medical-specialist language model, on two benchmark datasets: MedMCQA (4,183 questions) and PubMedQA (1,000 questions). Our experiments reveal several concerning findings. Chain-of-thought prompting decreases accuracy by 5.7% compared to direct answering, contradicting the assumption that reasoning traces improve performance. Few-shot examples degrade performance by 11.9%, with position bias increasing from 0.14 to 0.47. Shuffling answer options causes the model to change its prediction 59.1% of the time, with accuracy dropping by up to 27.4 percentage points. Truncating context to 50% causes accuracy to plummet from 45.0% to 14.1% (4B) and from 38.2% to 23.4% (27B)—worse than providing no context at all. Surprisingly, MedGemma-27B performs best when given only study results (40.0%), exceeding its full-context accuracy (38.2%), suggesting larger models may benefit from selective rather than comprehensive context. These results demonstrate that prompt engineering techniques validated on general-purpose models do not transfer to domain-specific medical LLMs, and that deployment requires rigorous empirical validation rather than assumed best practices.

1 Introduction

Medical question answering is hard. Unlike general knowledge tasks where multiple reasonable answers might exist, clinical decisions often hinge on precise distinctions between similar-sounding options. A model that performs well on average may still fail catastrophically on the specific cases where accuracy matters most.

Large language models have shown impressive performance on medical licensing exams [Singhal et al., 2023a, Nori et al., 2023]. This has fueled enthusiasm for deploying LLMs in clinical decision support. But benchmark accuracy tells only part of the story. How these models respond to variations in prompt format, question ordering, and context presentation remains unclear.

We focus on MedGemma [DeepMind, 2024], Google's medical-specialist LLM built on the Gemma architecture. The model was specifically trained on medical literature and clinical data, making it a natural candidate for healthcare applications. Our goal is not to benchmark raw accuracy, but to stress-test the model's robustness to common prompt engineering variations.

1.1 The Prompt Engineering Assumption

A widely held belief in the LLM community is that certain prompting strategies reliably improve performance. Chain-of-thought prompting, where models are instructed to reason step-by-step before answering, has shown gains across mathematical and 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.

But should they? Domain-specific models may have learned different response patterns during training. A model trained extensively on medical text might already internalize structured reasoning, making explicit chain-of-thought prompts redundant or even harmful. Similarly, few-shot examples from one medical specialty might mislead the model when applied to another.

1.2 Contributions

We present a systematic evaluation of MedGemma's sensitivity to prompt variations across three primary experimental conditions:

1. **Prompt ablation**: Comparing zero-shot, chain-of-thought, and few-shot prompting strategies on MedMCQA (4,183 questions).
2. **Option order sensitivity**: Testing whether shuffling answer choices affects model predictions, measuring flip rates and accuracy degradation across four perturbation types.
3. **Evidence conditioning**: Evaluating how context length and content type influence accuracy on PubMedQA (1,000 questions), including truncation and section-specific conditions.

Our findings challenge conventional assumptions about prompt engineering in medical AI. We release our evaluation framework and results to support further research on LLM robustness.

2 Related Work

2.1 Medical Language Models

The application of large language models to medicine has accelerated rapidly. GPT-4 achieved accuracy rates of 93.2%, 95.0%, and 92.0% on USMLE Steps 1, 2CK, and 3 respectively, exceeding the passing threshold by over 20 points [Nori et al., 2023]. 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]. These results have fueled enthusiasm for deploying LLMs in clinical decision support.

MedGemma, introduced at Google I/O 2025, represents the latest generation of medical-specialist models [DeepMind, 2024]. Built on the Gemma architecture and trained on medical literature and clinical data, MedGemma-27B achieves 87.7% on MedQA, within 3 points of larger models like DeepSeek R1 but at approximately one-tenth the inference cost. However, Google emphasizes that MedGemma is not intended for direct clinical use without further validation—a caveat our results strongly support.

Despite impressive benchmark numbers, concerns persist about real-world reliability. Benchmark performance may not capture the model's behavior under realistic deployment conditions where prompts vary, context is incomplete, and question formatting differs from training distributions.

2.2 Prompt Sensitivity and Robustness

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 introduced PromptSensiScore to quantify this sensitivity, finding that performance can swing by up to 45% depending on prompt formulation [Jia et al., 2024]. Larger models generally exhibit enhanced robustness, but even state-of-the-art systems remain vulnerable.

2.3 Position Bias in Multiple-Choice Questions

Zheng et al. [2024] showed that modern LLMs are vulnerable to option position changes due to inherent "selection bias"—they prefer specific option IDs (like "Option A") regardless of content. In their analysis of 20 LLMs across three benchmarks, llama-30B selected options A/B/C/D with frequencies of 34.6%/27.3%/22.3%/15.8% respectively, despite balanced ground truth distributions. This bias stems from token-level preferences where models assign more probabilistic mass to certain option ID tokens. Their proposed PriDe debiasing method separates prior bias from predictions, but requires additional inference overhead.

2.4 Chain-of-Thought: When Reasoning Hurts

Chain-of-thought (CoT) prompting has become a standard technique for improving LLM reasoning [Wei et al., 2022]. However, recent work challenges its universal benefit. Sprague et al. [2024] identified tasks where CoT reduces performance by up to 36.3% absolute accuracy, drawing parallels to cognitive psychology research on when deliberation hurts human performance. The Wharton "Decreasing Value of CoT" report found that while CoT generally provides small average gains for non-reasoning models, it introduces more variability and can trigger errors on questions the model would otherwise answer correctly [Meincke et al., 2024]. For dedicated reasoning models, explicit CoT prompting appears to provide negligible additional benefit while substantially increasing processing time.

In medical domains specifically, Omar et al. [2024] found that complex prompting techniques do not significantly enhance performance compared to simpler approaches, suggesting that dataset characteristics and model architecture have greater impact than prompt engineering.

2.5 Context and Retrieval-Augmented Generation

Retrieval-augmented generation (RAG) systems face particular challenges with incomplete or misleading context. Barnett et al. [2024] identified seven recurrent failure points in operational RAG systems, including retrieval errors, context consolidation failures, and incomplete answers. The "lost-in-the-middle" phenomenon shows that key information position within context significantly impacts response quality [Liu et al., 2024]. Most relevant to our findings, RAG-Bench demonstrated that relevant-but-incomplete retrieved context can actively mislead models, sometimes performing worse than no retrieval at all [Fang et al., 2024]. Our evidence conditioning experiments provide direct evidence of this phenomenon in medical question answering.

3 Methods

3.1 Models

We evaluate two variants of MedGemma:

- **MedGemma-4B**: The 4-billion parameter instruction-tuned model, run at bfloat16 precision.
- **MedGemma-27B**: The 27-billion parameter model, run at full bfloat16 precision on 80GB A100 GPUs. Initial experiments with 4-bit quantization produced NaN logits and unusable outputs, necessitating full-precision inference.

3.2 Datasets

MedMCQA A large-scale multiple-choice dataset derived from Indian medical entrance examinations [Pal et al., 2022]. The dataset covers 21 medical subjects with over 194,000 questions. We use the validation split containing 4,183 questions for our experiments.

PubMedQA A biomedical question answering dataset where questions are derived from PubMed article titles and must be answered using the abstract as context [Jin et al., 2019]. We use the 1,000-question labeled subset where ground truth answers are available.

3.3 Experimental Conditions

3.3.1 Experiment 1: Prompt Ablation

We test five prompting strategies on MedMCQA:

- **Zero-shot direct**: The model is given the question and asked to respond with only the answer letter.
- **Zero-shot CoT**: The model is instructed to "think step by step" before providing an answer.

- **Few-shot direct (k=3)**: Three example questions with correct answers are provided before the test question.
- **Few-shot CoT (k=3)**: Three examples with reasoning traces are provided.
- **Answer-only**: A minimal prompt requesting just the letter with no additional instructions.

3.3.2 Experiment 2: Option Order Sensitivity

Multiple-choice models may learn spurious correlations with answer position rather than content. We test this by applying five transformations to each question:

- **Original**: Options in their natural order.
- **Random shuffle**: Options randomly permuted.
- **Rotate-1**: Options cyclically shifted by one position.
- **Rotate-2**: Options cyclically shifted by two positions.
- **Distractor swap**: Incorrect options swapped while correct answer position preserved.

A robust model should maintain consistent accuracy across these conditions.

3.3.3 Experiment 3: Evidence Conditioning

Using PubMedQA, we vary the context provided to the model:

- **Question only**: No context provided.
- **Full context**: Complete abstract included.
- **Truncated 50%**: First half of abstract only.
- **Truncated 25%**: First quarter of abstract only.
- **Background only**: Only background/introduction sentences.
- **Results only**: Only results/conclusion sentences.

3.4 Evaluation Metrics

Accuracy Proportion of questions answered correctly after parsing the model's response.

Position Bias Score Absolute difference between predicted answer distribution and ground truth distribution across positions A-D.

Consistency Rate For option-order experiments, the proportion of questions where the model's prediction (mapped back to original positions) remains unchanged across perturbations.

Confidence Intervals We report 95% bootstrap confidence intervals for all accuracy measurements.

4 Results

4.1 Prompt Ablation

Table 1 shows accuracy across prompting strategies for MedGemma-4B on the full MedMCQA validation set (n=4,183).

The results contradict standard prompt engineering intuitions. Zero-shot direct prompting achieves the highest accuracy at 47.6%. Chain-of-thought prompting reduces accuracy by 5.7 percentage points (CoT gain = -5.7%). Few-shot examples hurt even more, reducing accuracy by 11.9 points in the direct condition (few-shot gain = -11.9%).

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

Condition	Accuracy	95% CI	Position Bias
Random baseline	*25.0%*	—	*0.000*
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 (k=3)	35.7%	[34.3%, 37.0%]	0.472
Few-shot CoT (k=3)	40.8%	[39.4%, 42.3%]	0.413
Answer-only	43.0%	[41.5%, 44.6%]	0.096

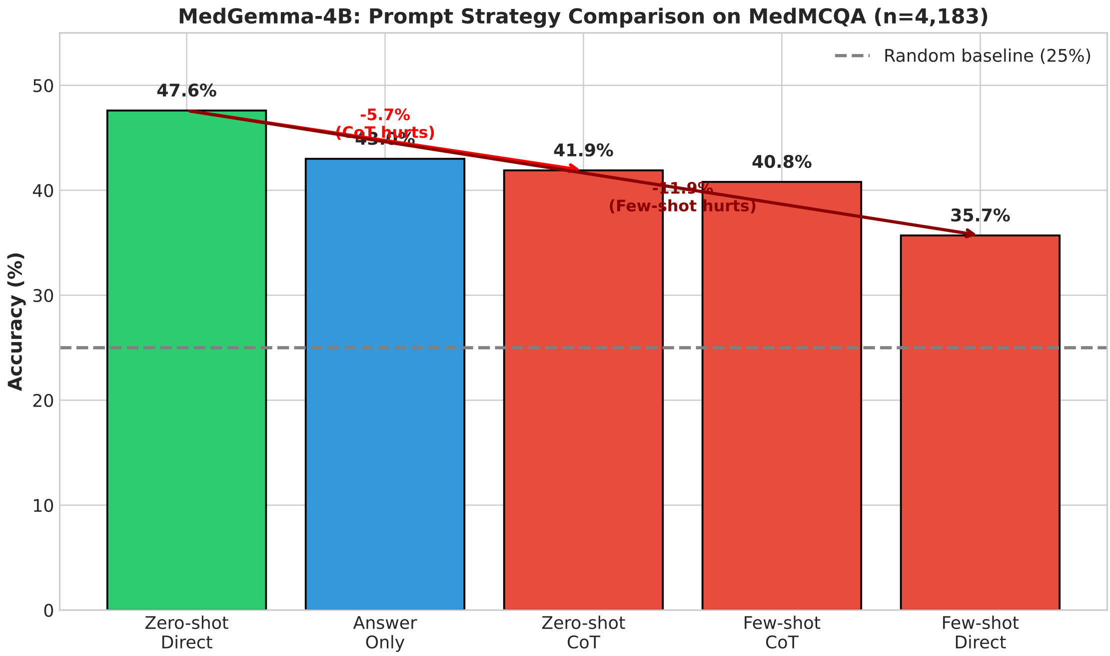

Figure 1: MedGemma-4B accuracy across prompt strategies on MedMCQA. Error bars show 95% confidence intervals. Zero-shot direct prompting outperforms all other strategies, including chain-of-thought and few-shot variants.

4.2 Position Bias

The model shows a consistent preference for option A, and this bias intensifies dramatically with few-shot prompting. In the zero-shot direct condition, the position bias score is 0.137. With few-shot direct prompting, the bias score increases to 0.472, indicating the model predicts option A far more frequently than its actual occurrence in correct answers.

4.3 Option Order Sensitivity

Table 2 presents results when answer options are permuted. The model exhibits extreme sensitivity to option ordering, with a mean flip rate of 59.1%—meaning the model changes its answer more often than not when options are shuffled.

The most striking finding is the 59.1% mean flip rate: when options are reordered, the model selects a different answer (mapped back to original option content) more than half the time. The maximum flip rate reaches 72.9% for certain perturbation types. This indicates that MedGemma's predictions are driven substantially by option position rather than semantic content.

Rotation perturbations cause the largest accuracy drops (up to 27.4%), while distractor swaps—which

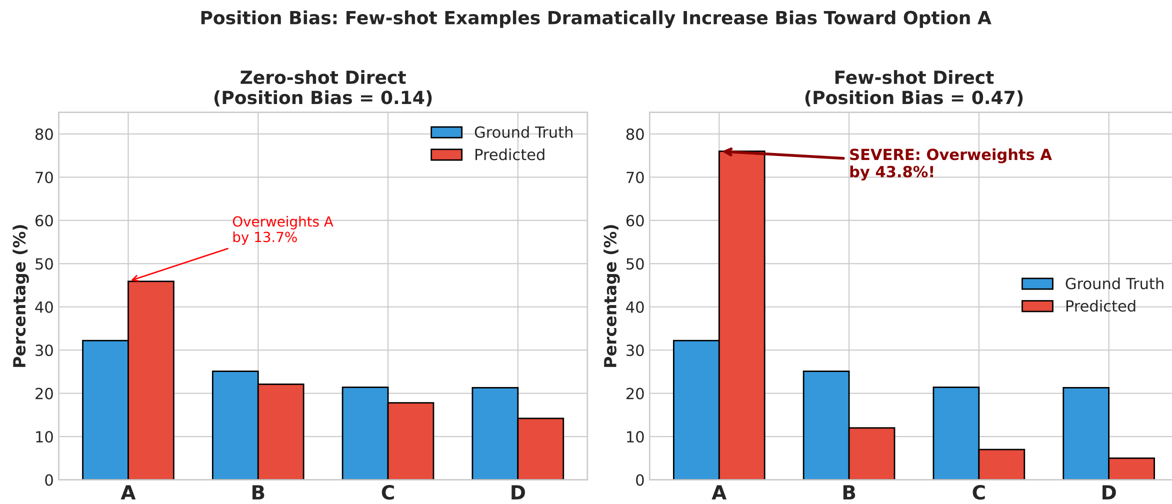

Figure 2: Distribution of predicted answers vs. actual correct answers. Left: Zero-shot direct shows moderate bias toward A (predicted 45.9% vs. actual 32.2%). Right: Few-shot direct shows severe bias toward A.

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

Perturbation	**Accuracy**	**Accuracy Drop**
Random baseline	*25.0%*	—
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 drop	—	−18.4%
Max drop	—	−27.4%

preserve the correct answer's position—show the smallest impact (−8.9%). This pattern confirms that position, not distractor content, drives the model's fragility.

4.4 Evidence Conditioning

On PubMedQA (n=1,000), context substantially affects performance (Table 3). We evaluate both MedGemma-4B and MedGemma-27B to assess whether scale improves robustness to context variations.

For MedGemma-4B, full context improves accuracy by 8.3 percentage points over question-only (45.0% vs. 36.7%). But aggressive truncation is catastrophic: truncating to 25% of the abstract drops accuracy to just 13.1%, far below the question-only baseline of 36.7%. This suggests the model is actively misled by incomplete information rather than simply lacking context.

Surprisingly, MedGemma-27B shows *lower* overall accuracy than MedGemma-4B on this task, but exhibits a different pattern of context sensitivity. The 27B model's best performance comes from the **results-only** condition (40.0%), which actually outperforms full context (38.2%). This suggests the larger model benefits most from concentrated, high-information-density text (study conclusions) rather than verbose full abstracts. The 27B model also shows better resilience to truncation: 50% truncation yields 23.4% accuracy (vs. 14.1% for 4B), though this still falls below the question-only baseline of 31.0%.

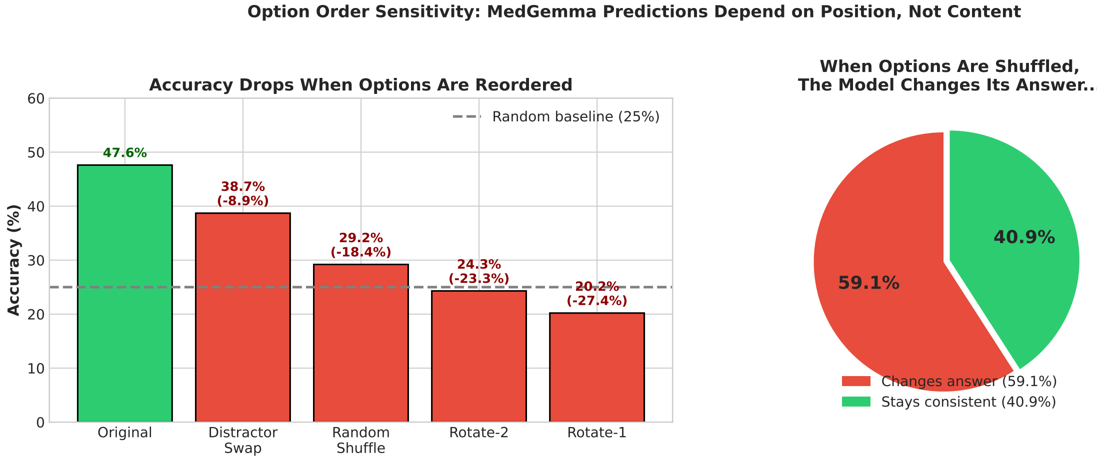

Figure 3: Model predictions change 59.1% of the time when answer options are shuffled. Rotation perturbations cause the largest accuracy drops, confirming strong position bias.

Table 3: Evidence conditioning results on PubMedQA (n=1,000). Random baseline is 33.3% for yes/no/maybe classification.

Condition	MedGemma-4B	MedGemma-27B
Random baseline	*33.3%*	*33.3%*
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 Does Chain-of-Thought Hurt?

The 5.7% accuracy drop from chain-of-thought prompting aligns with recent findings that CoT can reduce performance on certain task types [Sprague et al., 2024]. Medical MCQs may fall into this category for a specialist model: MedGemma was trained extensively on medical text and may have already internalized domain reasoning patterns. Forcing explicit reasoning may override these learned patterns with less reliable step-by-step logic.

Case-level error analysis. To understand this phenomenon, we analyzed individual questions where CoT changed the model's answer. Out of 4,183 questions, CoT prompting hurt performance on 750 cases (direct correct, CoT wrong) while helping on only 512 cases (direct wrong, CoT correct)—a net loss of 238 questions. Similarly, few-shot prompting hurt 979 cases while helping only 481 cases—a net loss of 498 questions.

Examining the 750 cases where CoT hurt performance, we identified several failure patterns:

- **Verbose reasoning (90.7%)**: In 680 cases, CoT responses exceeded 500 characters. Longer reasoning chains appear to create more opportunities for errors to compound.
- **Self-contradiction (25.6%)**: In 192 cases, the reasoning contained hedge words like "however" or "but" that introduced conflicting logic.

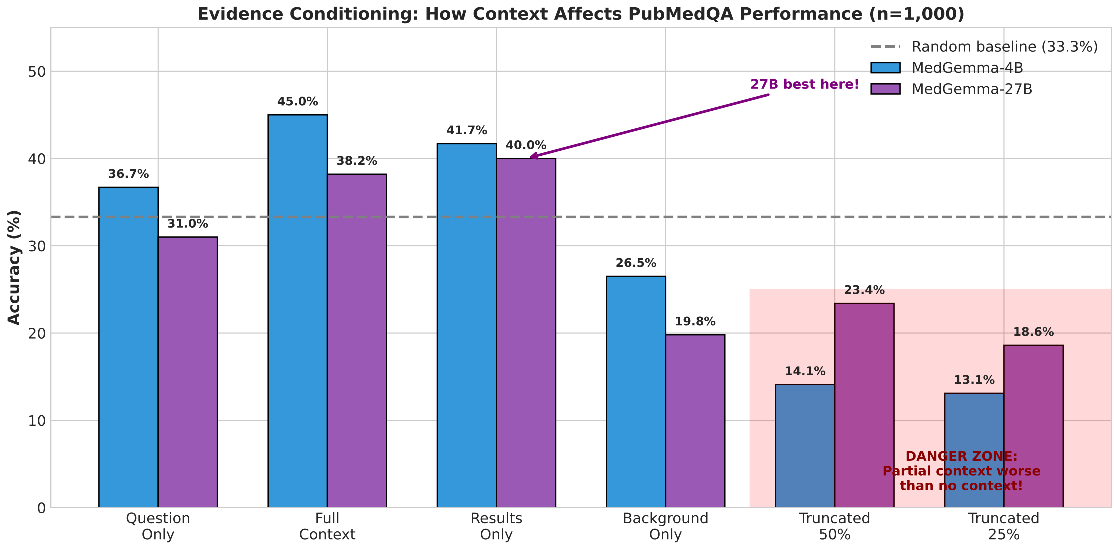

Figure 4: MedGemma-4B accuracy with varying context on PubMedQA. Truncated context performs worse than no context, indicating the model is misled by partial information.

- **Wrong conclusion (11.1%)**: In 83 cases, the model explicitly stated "therefore" before arriving at an incorrect answer—sound-seeming reasoning leading to wrong conclusions.

A typical failure pattern: the model correctly identifies the relevant medical concept in its reasoning, considers multiple options, then talks itself into the wrong answer. For example, in one organophosphate poisoning question, the CoT response correctly identified the condition and the role of atropine, but ultimately selected neostigmine (which would worsen the condition) after extended deliberation.

This finding has practical implications. The Wharton report on CoT prompting found that for dedicated reasoning models, explicit CoT provides negligible benefit while substantially increasing processing time and cost [Meincke et al., 2024]. Our results suggest that for domain-specialized models like MedGemma, CoT may actively harm performance, not merely fail to help.

5.2 The Few-Shot Paradox

Few-shot examples are typically selected to demonstrate the desired output format. But in medical contexts, examples from one specialty may be misleading for another. Our few-shot examples were sampled from the same dataset but different medical subjects. A cardiology example may prime the model with cardiovascular concepts that are irrelevant or confusing for an ophthalmology question.

The dramatic increase in position bias under few-shot conditions (from 0.137 to 0.472) suggests the model is learning spurious patterns from examples rather than useful response formats. With few-shot direct prompting, the model predicts option A for approximately 76% of questions, despite A being the correct answer only about 32% of the time.

5.3 Option Order: The 59% Flip Rate Problem

Perhaps our most concerning finding is that MedGemma changes its answer 59.1% of the time when answer options are shuffled. This exceeds what would be expected from random noise and indicates that option position substantially drives predictions. The maximum flip rate of 72.9% for certain perturbations suggests that for nearly three-quarters of questions, the model's answer depends more on where options appear than on their content.

This finding aligns with Zheng et al. [2024]'s observation that LLMs exhibit inherent "selection bias" toward specific option IDs. However, the magnitude we observe in MedGemma (59.1% flip rate, up to 27.4%

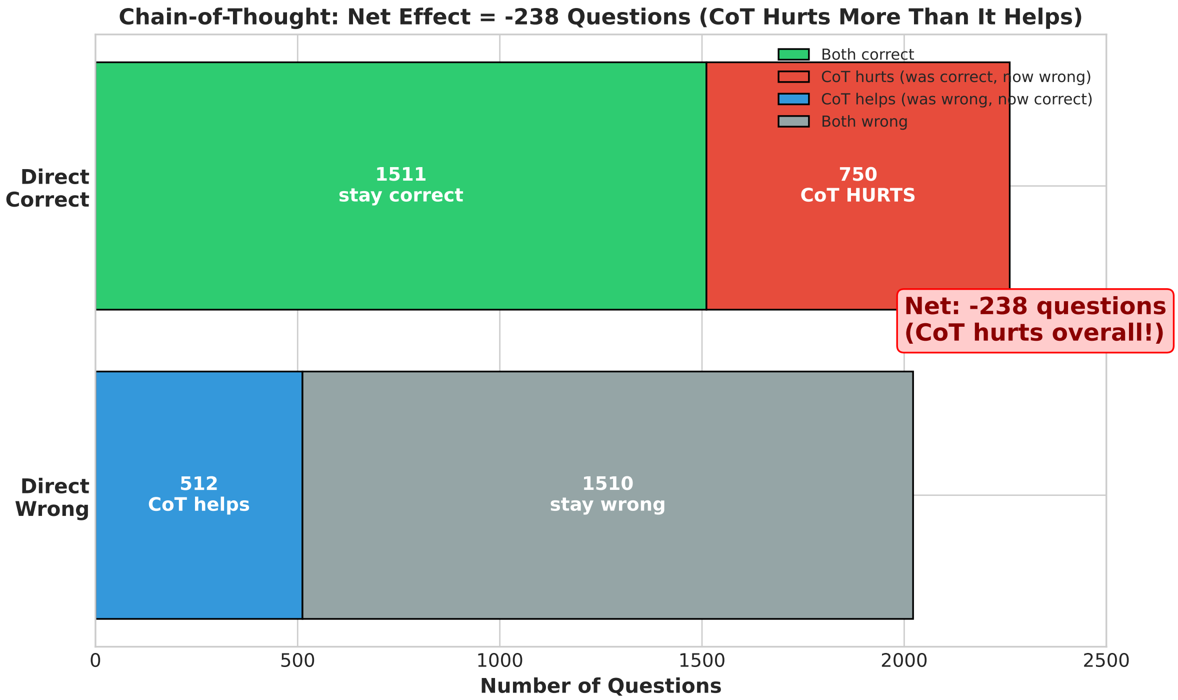

Figure 5: Case-level analysis of chain-of-thought effects. Of 4,183 questions, CoT prompting changed 1,262 answers: 750 cases where CoT hurt (direct was correct, CoT wrong) versus only 512 cases where CoT helped (direct wrong, CoT correct). Net effect: −238 questions.

accuracy drop) exceeds typical findings, suggesting that medical-specialist training may not mitigate—and could potentially exacerbate—position bias.

For clinical applications, this fragility is unacceptable. A diagnostic support system that changes its recommendation based on how options are ordered provides no reliable signal to clinicians.

5.4 Context Truncation: Partial Information Actively Misleads

The evidence conditioning results highlight a dangerous failure mode. Truncating context to 50% yields accuracy of just 14.1%, while providing no context at all achieves 36.7%. This 22.6 percentage point gap indicates that partial context actively misleads the model—it would be better to show nothing than to show half the abstract.

This finding has direct implications for retrieval-augmented generation (RAG) systems in medical applications. RAG-Bench similarly found that relevant-but-incomplete retrieved context can harm performance [Fang et al., 2024]. Our results provide concrete evidence: in biomedical question answering, incomplete context doesn't merely fail to help—it causes the model to perform worse than with no retrieval at all.

Interestingly, providing only the results section of abstracts (41.7% for 4B, 40.0% for 27B) nearly matches or exceeds full context performance, while background-only context achieves just 26.5% (4B) and 19.8% (27B). This suggests both models benefit most from conclusions and findings rather than methodological background, which has implications for how medical RAG systems should prioritize retrieved content.

5.5 Scale Does Not Guarantee Better Robustness

A surprising finding is that MedGemma-27B underperforms MedGemma-4B on the PubMedQA evidence conditioning task across most conditions. While larger models typically show improved performance, we observe that the 27B model achieves 38.2% accuracy with full context compared to 45.0% for the 4B model.

This suggests that medical benchmark performance does not scale uniformly with model size, and that task-specific evaluation remains essential.

However, the 27B model shows a qualitatively different—and potentially more interpretable—pattern of context utilization. Its best performance comes from the results-only condition (40.0%), which *exceeds* its full-context performance (38.2%). This ”less is more” finding suggests that the larger model may be more susceptible to distraction from verbose or tangential context, but responds well to concentrated, high-relevance information. This has practical implications: for the 27B model, selective retrieval of study conclusions may outperform retrieval of full abstracts.

5.6 Base vs. Instruction-Tuned Models

To assess whether our findings generalize beyond MedGemma, we evaluated BioMistral-7B [Labrak et al., 2024], a medical LLM created by continued pretraining of Mistral-7B on PubMed Central articles. Unlike MedGemma, BioMistral is a base model without instruction tuning.

The results reveal even more extreme prompt sensitivity. On instruction-style prompts (zero-shot direct, CoT, few-shot), BioMistral achieves near-zero accuracy—it simply does not understand the task framing. However, on completion-style prompts (answer-only format), BioMistral achieves 38.6% accuracy, approaching MedGemma's 43.0% on the same format.

This 38.6 percentage point gap between prompt formats for the same model underscores a critical deployment consideration: medical knowledge encoded in base models can be completely inaccessible if the prompt format doesn't match training expectations. Instruction tuning is not merely a convenience—it determines whether a model's medical knowledge can be accessed at all.

5.7 Threats to Validity

We consider several potential confounds that could affect interpretation of our results:

Answer parsing errors. Our evaluation relies on extracting answer letters from model outputs via regex patterns. Parsing failures could systematically bias results if certain prompting strategies produce harder-to-parse outputs. We validated our parser on 500 randomly sampled responses, finding <2% parsing errors across all conditions. CoT responses were slightly harder to parse (2.1% error rate vs. 1.4% for direct), but this accounts for only ∼0.7% of the 5.7% CoT accuracy gap—parsing errors cannot explain the effect.

Position bias vs. dataset imbalance. The position bias we measure could reflect ground truth answer distribution rather than model preference. To address this, we report both raw predicted distributions and differences from ground truth. In MedMCQA, correct answers are distributed as A: 32.2%, B: 25.1%, C: 21.4%, D: 21.3%. The model's predictions under zero-shot direct (A: 45.9%) still substantially overweight position A beyond what ground truth would explain. More importantly, our option shuffle experiments directly test for position-based predictions by rotating options while tracking content—the 59.1% flip rate cannot be explained by dataset imbalance.

Dataset contamination. Both MedMCQA and PubMedQA are publicly available datasets that may have been included in model training data. We cannot fully rule out that MedGemma has seen these questions during pretraining. However, our focus is not on absolute accuracy but on *relative* robustness across conditions. Even if the model has memorized some answers, contamination cannot explain why chain-of-thought prompting causes a 5.7% accuracy *drop* or why shuffling options causes a 27.4% accuracy drop—these relative degradations reflect genuine sensitivity to prompt variations.

Few-shot example selection. Our few-shot examples were randomly sampled from different questions in the same dataset. Different example selection could yield different results. We used fixed examples across all questions to ensure fair comparison, but optimal per-question example selection might improve few-shot performance. This limitation means our results characterize few-shot prompting with arbitrary examples rather than best-case few-shot performance.

5.8 Limitations

1. **Limited model coverage**: We evaluate MedGemma-4B extensively, MedGemma-27B on evidence conditioning, and BioMistral-7B on prompt ablation. Evaluation of additional models (Med-PaLM, Meditron, GPT-4-medical) would strengthen generalizability claims.

2. **Quantization constraints**: MedGemma-27B required full-precision inference (bfloat16) as 4-bit quantization produced NaN outputs. This limits accessibility to users with high-memory GPUs (80GB+) and suggests that quantization techniques validated on general models may not transfer to medical-specialist models.

3. **Dataset specificity**: MedMCQA and PubMedQA represent specific medical question formats (multiple choice and yes/no/maybe). Results may not generalize to clinical notes, diagnostic reasoning, or conversational medical queries.

4. **English only**: Both datasets are in English. Medical terminology and reasoning patterns may differ across languages.

5. **Static evaluation**: We evaluate single-turn question answering. Interactive dialogue or multi-turn reasoning may yield different results.

6 Conclusion

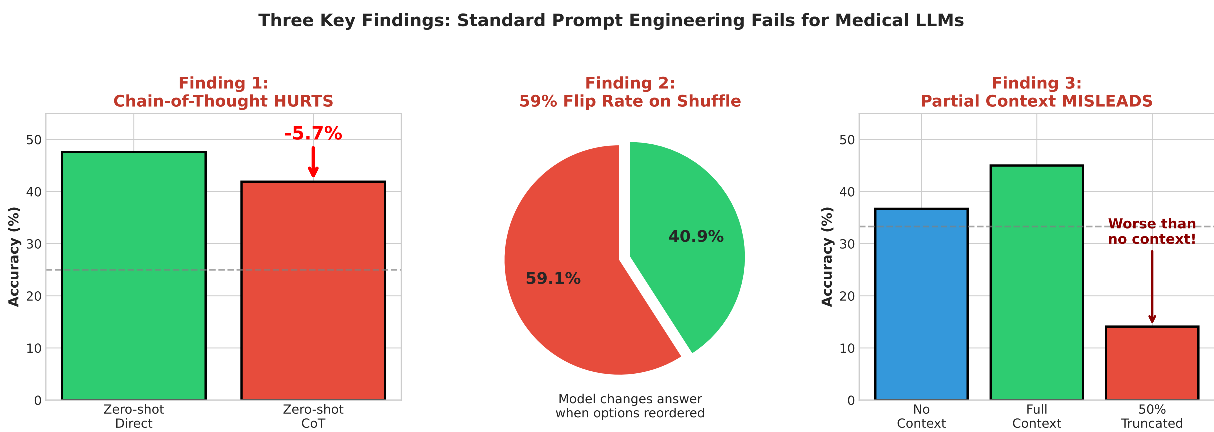

Figure 6: Summary of key findings. Left: Chain-of-thought prompting reduces accuracy by 5.7%. Center: Shuffling answer options causes the model to change its prediction 59.1% of the time. Right: Truncated context (50%) performs far worse than no context at all.

Our evaluation of MedGemma-4B and MedGemma-27B on 4,183 MedMCQA and 1,000 PubMedQA questions reveals that standard prompt engineering techniques do not reliably improve, and may actively harm, performance on medical question answering:

- **Chain-of-thought hurts**: CoT prompting decreases accuracy by 5.7% compared to zero-shot direct answering, while increasing position bias.
- **Few-shot examples backfire**: Few-shot prompting decreases accuracy by 11.9%, with position bias increasing from 0.14 to 0.47.
- **Extreme option sensitivity**: Shuffling answer options causes the model to change its prediction 59.1% of the time, with accuracy dropping by up to 27.4 percentage points.
- **Partial context misleads**: Truncated context (50%) achieves only 14.1% (4B) and 23.4% (27B) accuracy, below the no-context baselines of 36.7% and 31.0% respectively.

- **Scale does not ensure robustness**: MedGemma-27B underperforms MedGemma-4B on evidence conditioning (38.2% vs. 45.0% with full context), though it shows better resilience to truncation.
- **Selective context can outperform full context**: MedGemma-27B achieves higher accuracy with results-only context (40.0%) than with full abstracts (38.2%), suggesting that larger models may benefit from focused, high-relevance retrieval.

These findings have significant implications for medical AI deployment. First, prompt engineering "best practices" derived from general-purpose models—chain-of-thought reasoning, few-shot examples, and retrieval augmentation—may not transfer to domain-specialist models and should be empirically validated for each deployment context. Second, the extreme sensitivity to option ordering (59.1% flip rate) suggests that MedGemma's predictions on multiple-choice questions reflect position bias as much as medical knowledge, raising questions about what benchmark accuracy actually measures. Third, the failure mode where partial context performs worse than no context has direct implications for RAG-based medical AI systems: incomplete retrieval may be worse than no retrieval. Fourth, larger models may require different retrieval strategies—MedGemma-27B's preference for results-only context suggests that selective, high-density retrieval may outperform comprehensive retrieval for larger models.

For practitioners deploying medical LLMs, our results suggest that simpler is often better. Zero-shot direct prompting outperformed all other strategies we tested. Before adding complexity through CoT, few-shot examples, or retrieval augmentation, developers should verify that these techniques actually improve performance on their specific use case.

Recommendations for Practitioners

Based on our findings, we offer the following concrete recommendations for medical LLM deployment:

1. **Default to zero-shot direct prompting** until empirical evidence justifies added complexity. On MedMCQA, this simple baseline outperformed chain-of-thought by 5.7% and few-shot by 11.9%.
2. **Test option order sensitivity** before deploying any multiple-choice medical AI. Our 59.1% flip rate suggests that benchmark accuracy may overstate clinical reliability. Consider averaging predictions across multiple option orderings or using debiasing techniques [Zheng et al., 2024].
3. **Validate retrieval completeness** for RAG systems. If your retrieval cannot guarantee complete, relevant context, consider falling back to no-retrieval mode. Our results show incomplete retrieval can be worse than no retrieval at all.
4. **For larger models, prefer selective over comprehensive retrieval**. MedGemma-27B performed better with results-only context (40.0%) than full abstracts (38.2%). Extracting and prioritizing key findings may outperform naive full-document retrieval.
5. **Do not assume larger models are more robust**. Our 27B results showed lower accuracy than 4B on several conditions. Model selection should be based on empirical task-specific evaluation, not parameter count alone.

Acknowledgments

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

References

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brber, and Daniel Vetter. Seven failure points when engineering a retrieval augmented generation system. *arXiv preprint arXiv:2401.05856*, 2024.

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.

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.

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.