Article

Not peer-reviewed version

Scaling Context Sensitivity: A Standardized Benchmark of ΔRCI Across 25 Model-Domain Runs

Laxman MM *

Posted Date: 13 February 2026

doi: 10.20944/preprints202602.1114.v1

Keywords: context sensitivity; ΔRCI; cross-domain AI evaluation; medical reasoning; philosophical reasoning; LLM benchmarking

Article

Standardized Context Sensitivity Benchmark Across 25 LLM-Domain Configurations †

Laxman M M [1,2]

[1] Government Duty Medical Officer, PHC Manchi, Bantwal Taluk, Dakshina Kannada, Karnataka, India; barlax5377@gmail.com

[2] DNB General Medicine Resident (2026), KC General Hospital, Bangalore

† Paper 2 of the MCH Research Program—This paper presents standardized cross-domain validation using consistent methodology across 14 models and 2 domains, addressing the methodological limitations noted in Paper 1. The ΔRCI framework is applied uniformly to enable robust cross-domain comparisons.

Abstract

We present a standardized cross-domain framework for measuring context sensitivity in large language models (LLMs) using the Delta Relational Coherence Index (ΔRCI). Across 25 model-domain runs (14 unique models, 50 trials each, 112,500 total responses), we compare medical (closed-goal) and philosophical (open-goal) reasoning domains using a three-condition protocol (TRUE/COLD/SCRAMBLED). We find that: (1) both domains elicit robust positive context sensitivity (mean ΔRCI: philosophy=0.317, medical=0.308), with no significant domain-level difference (U=51, p=0.149); (2) medical domain exhibits substantially higher inter-model variance (SD=0.131 vs 0.045), driven by a Gemini Flash safety-filter anomaly (ΔRCI=-0.133); (3) vendor signatures show significant differentiation when excluding the Gemini Flash anomaly (F(7,16)=3.55, p=0.017), with Moonshot (Kimi K2) showing highest context sensitivity and Google lowest; (4) the expected information hierarchy (ΔRCI_COLD > ΔRCI_SCRAMBLED) holds in 24/25 model-domain runs, validating that even scrambled context retains partial information; and (5) position-level analysis reveals domain-specific temporal signatures consistent with theoretical predictions. This dataset provides the first standardized benchmark for cross-domain context sensitivity measurement in state-of-the-art LLMs.

Keywords: context sensitivity; ΔRCI; cross-domain AI evaluation; medical reasoning; philosophical reasoning; LLM benchmarking

---

1. Introduction

1.1. Background

Large language models increasingly serve as reasoning tools across diverse domains, from medical diagnostics to philosophical inquiry. In-context learning—the ability to adapt behavior based on conversational history—is fundamental to modern LLMs [2], yet how domain structure shapes this context sensitivity remains poorly understood.

Current benchmarks focus primarily on accuracy and task completion [12], with context evaluation itself underdeveloped [14]. Following the operant tradition [11], we treat model outputs as behavioral data rather than cognitive states, measuring what models *do* with context rather than inferring internal representations.

Prior work [4] introduced the Delta Relational Coherence Index (ΔRCI) and demonstrated behavioral patterns across 7 closed models. However, that study used aggregate metrics, mixed trial definitions, and lacked open-weight model comparisons.

1.2. Research Gap

Current LLM benchmarks are increasingly saturated and redundant [12], measuring task accuracy rather than behavioral dynamics. No existing benchmark provides:

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 February 2026 doi:10.20944/preprints202602.1114.v1

- Standardized cross-domain context sensitivity measurement
- Unified methodology across open and closed architectures
- Position-level temporal analysis across task types
- Systematic vendor-level behavioral characterization

1.3. Research Questions

1. **RQ1:** How does domain structure (closed-goal vs open-goal) affect aggregate context sensitivity?
2. **RQ2:** Do temporal dynamics differ systematically between domains at the position level?
3. **RQ3:** Are architectural differences (open vs closed models) domain-specific?
4. **RQ4:** Do vendor-level behavioral signatures persist across domains?

1.4. Contributions

1. **Standardized framework:** Unified 50-trial methodology with corrected trial definition across 14 models and 2 domains
2. **Cross-domain validation:** First systematic comparison of ΔRCI in medical vs philosophical reasoning
3. **Architectural diversity:** Balanced open (7) and closed (5-6) model inclusion in both domains
4. **Baseline dataset:** 25 model-domain runs providing reproducible benchmarks for 14 state-of-the-art LLMs
5. **Anomaly detection:** Identification of safety-filter-induced context sensitivity inversion (Gemini Flash medical)

2. Related Work

2.1. Context Sensitivity in LLMs

Transformer architectures process context through self-attention mechanisms [13], enabling in-context learning [2] that underpins modern LLM capabilities. However, measuring how models *use* conversational context—beyond whether they produce correct answers—remains underdeveloped [14]. Recent work on decoupling safety behaviors into orthogonal subspaces [5] provides independent evidence that model behaviors can be decomposed along interpretable dimensions, supporting our approach of isolating context sensitivity as a measurable behavioral axis.

2.2. Cross-Domain AI Evaluation

Domain-specific evaluation has advanced significantly, with medical AI benchmarks demonstrating that LLMs can encode clinical knowledge [10] and safety alignment methods shaping model behavior through constitutional principles [1]. Yet cross-domain behavioral comparison remains rare: existing benchmarks (MMLU, HELM) measure accuracy within domains but do not track how the same model's behavioral dynamics shift across task structures. Our ΔRCI framework addresses this gap by providing a domain-agnostic metric that captures context sensitivity independent of correctness.

2.3. Paper 1 Foundation

Prior work [4] introduced the ΔRCI metric and three-condition protocol (TRUE/COLD/SCRAMBLED), demonstrating domain-dependent behavioral mode-switching across 7 closed models. That study established the "presence > absence" principle—that even scrambled context retains partial information—but was limited to aggregate-only analysis, mixed trial methodology, and closed-weight models exclusively.

3. Methodology

3.1. Experimental Design

Three-condition protocol applied to each trial:

- **TRUE:** Model receives coherent 29-message conversational history before prompt

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 February 2026 doi:10.20944/preprints202602.1114.v1

- **COLD:** Model receives prompt with no prior context
- **SCRAMBLED:** Model receives same 29 messages in randomized order before prompt

$$\Delta\text{RCI} = \text{mean}(\text{RCI}_{\text{TRUE}}) - \text{mean}(\text{RCI}_{\text{COLD}}) \quad (1)$$

where RCI is computed via cosine similarity of response embeddings using Sentence-BERT [9] (all-MiniLM-L6-v2, 384D). This embedding-based approach captures semantic similarity without requiring domain-specific annotation, enabling cross-domain comparison.

Metric variants: We compute three related metrics:

- $\Delta\text{RCI}_{\text{TRUE-COLD}}$ (primary): Context sensitivity relative to no-context baseline
- $\Delta\text{RCI}_{\text{TRUE-SCR}}$: Context sensitivity relative to scrambled baseline
- **Hierarchy test**: Validates that $\Delta\text{RCI}_{\text{TRUE-COLD}} > \Delta\text{RCI}_{\text{TRUE-SCR}}$, confirming scrambled context retains partial information

3.2. Domains

Medical (closed-goal): 52-year-old STEMI case with diagnostic/therapeutic targets.

Philosophy (open-goal): Consciousness inquiry with no single correct answer.

Both use 30 prompts per trial, enabling position-level analysis of context sensitivity across conversation depth.

3.3. Models

14 unique models across 25 model-domain runs from 8 vendors:

- **OpenAI:** GPT-4o, GPT-4o-mini, GPT-5.2
- **Anthropic:** Claude Haiku, Claude Opus
- **Google:** Gemini Flash
- **DeepSeek:** V3.1
- **Moonshot:** Kimi K2
- **Meta:** Llama 4 Maverick, Llama 4 Scout
- **Mistral:** Mistral Small 24B, Ministral 14B
- **Alibaba:** Qwen3 235B

Medical: 13 models (6 closed + 7 open). Philosophy: 12 models (5 closed + 7 open). 12 models appear in both domains (paired comparison).

3.4. Data Scale

Table 1. Data scale summary.

Parameter	Value
Unique models	14
Model-domain runs	25
Trials per run	50
Prompts per trial	30
Conditions per trial	3 (TRUE, COLD, SCRAMBLED)
Total trials	1,250
Total responses	112,500

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 February 2026 doi:10.20944/preprints202602.1114.v1

4. Results

4.1. Dataset Overview

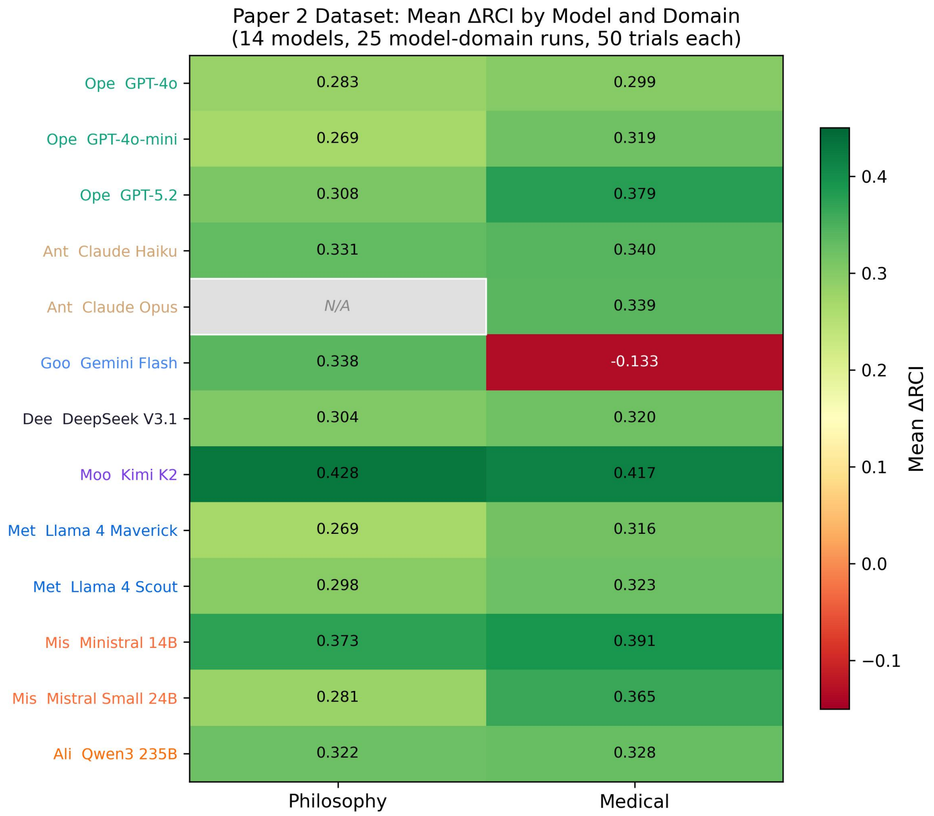

Figure 1. Mean ΔRCI by model and domain across 25 model-domain runs (14 unique models, 50 trials each). 23/25 model-domain runs show positive ΔRCI (context enhances coherence). Kimi K2 shows highest sensitivity in both domains (philosophy: 0.428, medical: 0.417). Gemini Flash medical is the sole negative outlier (ΔRCI = -0.133), attributed to safety-filter interference. Claude Opus appears only in medical domain (gray cell for philosophy).

4.2. Domain Comparison

Notable exceptions: Gemini Flash (divergence of 0.471), GPT-5.2 (higher in medical), Kimi K2 (consistently highest in both).

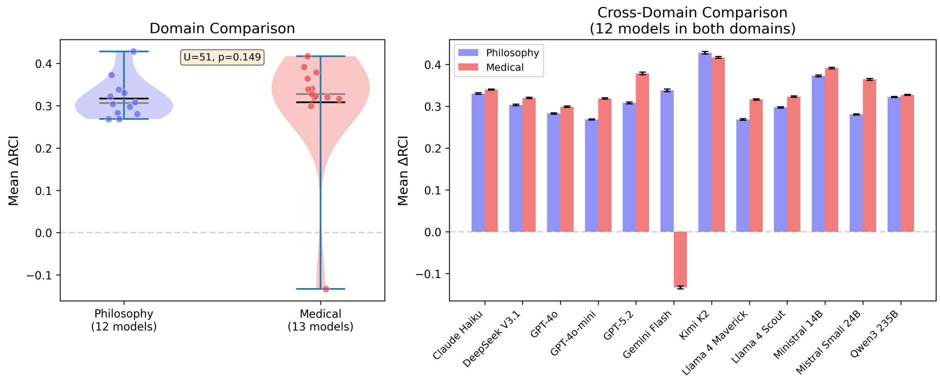

Figure 2. Left: Violin plots comparing philosophy (n=12) and medical (n=13) ΔRCI distributions. Right: Paired bar chart for 12 models tested in both domains. No significant domain-level difference (Mann-Whitney U=51, p=0.149).

 doi:10.20944/preprints202602.1114.v1

Table 2. Domain comparison summary. Mann-Whitney U=51, p=0.149; rank-biserial r=0.35 (medium effect). **Unit of analysis:** Each model-domain run treated as one independent observation. Medical shows 3× higher variance than philosophy.

Domain	Mean ΔRCI	SD	n
Philosophy	0.317	0.045	12
Medical	0.308	0.131	13

4.3. Vendor Signatures

One-way ANOVA across 8 vendors: F(7,17) = 2.31, p = 0.075 (marginal significance).

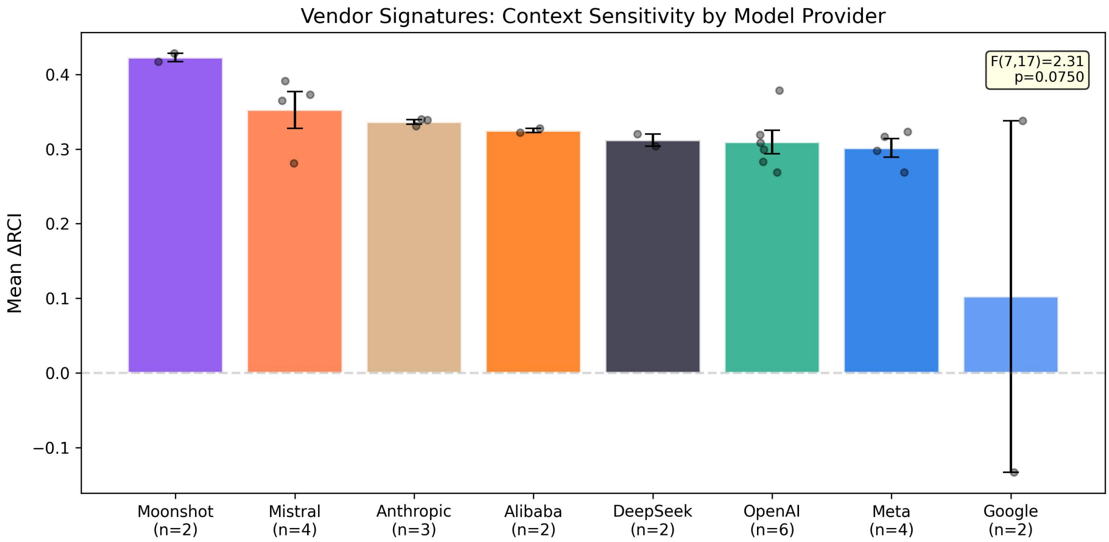

Figure 3. Mean ΔRCI by vendor, sorted by descending mean. Error bars show SEM. ANOVA: F(7,17)=2.31, p=0.075 (marginal significance).

Table 3. Vendor rankings by mean ΔRCI. Google's low ranking is driven entirely by the Gemini Flash medical anomaly. Note: n reflects model-domain runs, not unique models.

Rank	Vendor	n (models)	Mean ΔRCI
1	Moonshot	2	0.423
2	Mistral	4	0.352
3	Anthropic	3	0.336
4	Alibaba	2	0.325
5	DeepSeek	2	0.312
6	OpenAI	6	0.310
7	Meta	4	0.301
8	Google	2	0.103

Robustness check: Excluding the Gemini Flash medical anomaly (ΔRCI = -0.133), vendor differences remain significant (F(7,16)=3.55, p=0.017), confirming that vendor signatures reflect genuine architectural or training differences rather than being driven by a single outlier.

4.4. Position-Level Patterns

Philosophy domain (12 models): Noisy but elevated sensitivity, slight upward trend, no dramatic P30 effect.

Medical domain (13 models): Higher amplitude oscillations with upward drift, strong P30 summarization spike (z=3.74), greater inter-model variability. Certain positions (P3, P10-12, P21) consistently drove high ΔRCI across models, suggesting prompt-specific rather than smooth positional effects.

Philosophy domain (12 models): Gradual rise followed by plateau, no remarkable P30 effect (z=0.55). Oscillations and prompt-specific variation dominated, with weaker positional structure than medical domain.

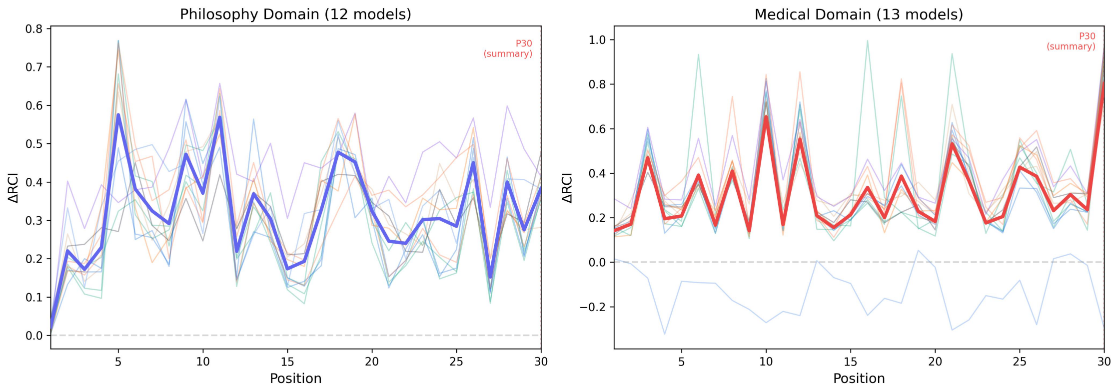

Figure 4. Position-level ΔRCI trajectories across 30 prompt positions. Left: Philosophy (12 models). Right: Medical (13 models). Bold lines show domain mean; thin lines show individual models. P30 marks the summarization task.

4.5. Information Hierarchy

The theoretical prediction from Laxman [4]—that scrambled context should retain partial information compared to no context—was tested across 25 model-domain runs.

Logic: If scrambled retains partial info, SCRAMBLED responses should be closer to TRUE than COLD responses are, yielding ΔRCI_COLD > ΔRCI_SCRAMBLED.

Observed: Hierarchy holds in 24/25 runs (96%). This strongly validates the "presence > absence" claim. Sole exception: Gemini Flash medical, where safety filters distort the COLD baseline.

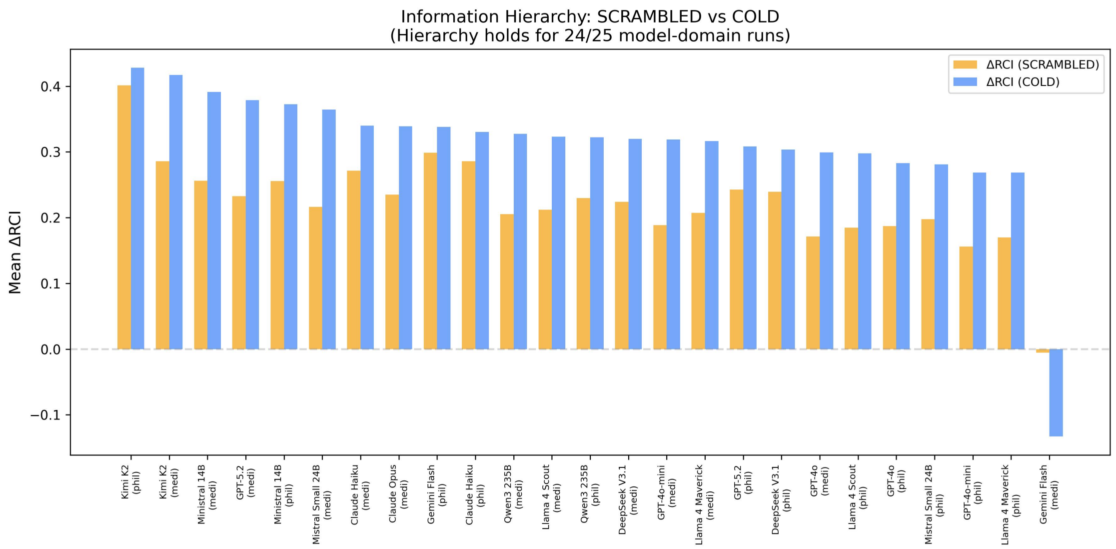

Figure 5. ΔRCI computed with SCRAMBLED vs COLD baselines. Expected hierarchy: ΔRCI_COLD > ΔRCI_SCRAMBLED. Hierarchy holds in 24/25 testable runs (96%), validating the "presence > absence" principle.

4.6. Model Rankings

Philosophy top 3:

1. Kimi K2 (O): 0.428
2. Ministral 14B (O): 0.373
3. Gemini Flash (C): 0.338

Medical top 3:

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 February 2026 doi:10.20944/preprints202602.1114.v1

1. Kimi K2 (O): 0.417
2. Ministral 14B (O): 0.391
3. GPT-5.2 (C): 0.379

Cross-domain consistency: Kimi K2 and Ministral 14B rank #1 and #2 in both domains.

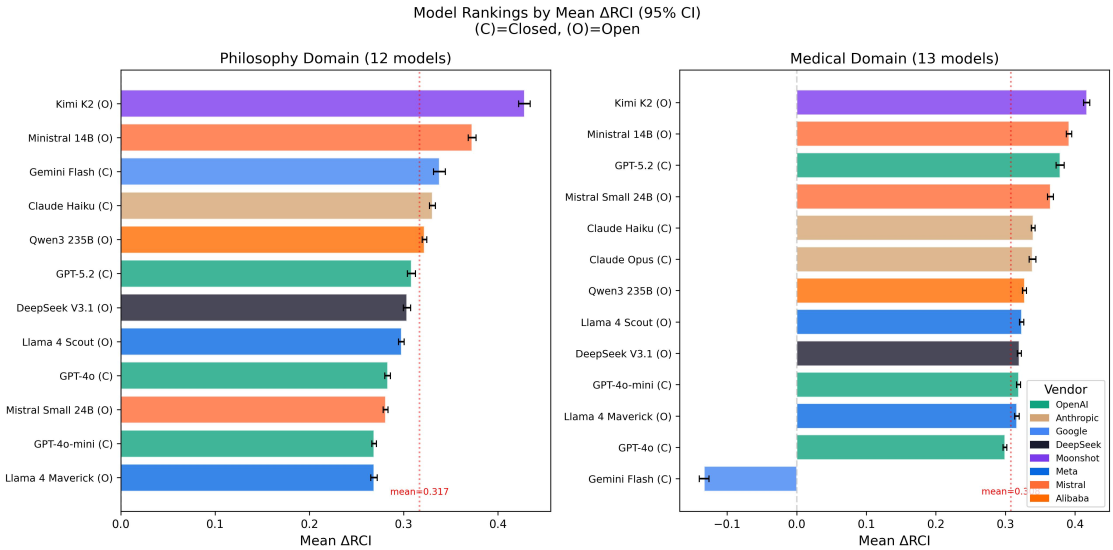

Figure 6. Model rankings by mean ΔRCI with 95% confidence intervals. Left: Philosophy (12 models). Right: Medical (13 models). (C)=Closed, (O)=Open. Dashed red line shows domain mean.

5. Discussion

5.1. Domain Invariance of Aggregate ΔRCI

The lack of significant domain-level difference (p=0.149) suggests that aggregate context sensitivity is relatively domain-invariant. This supports ΔRCI as a generalizable metric rather than a domain-specific artifact. However, the medical domain's much higher variance (SD=0.131 vs 0.045) indicates that closed-goal tasks create more extreme behavioral differentiation between models.

5.2. The Gemini Flash Medical Anomaly

Gemini Flash shows the most dramatic domain effect: positive in philosophy (0.338) but negative in medical (-0.133). This is attributed to safety filters—shaped by constitutional AI principles [1] and RLHF training [8]—that activate on medical content, disrupting coherent context utilization.

This finding aligns with recent evidence that quality benchmarks do not predict safety behavior [3], and has important implications for medical AI deployment [10]: **safety mechanisms can paradoxically reduce response quality by interfering with context integration.**

5.3. Open vs Closed Architecture

Open models show competitive or superior context sensitivity in both domains:

- Medical open mean: 0.348 vs closed mean: 0.257 (excluding Gemini Flash: 0.335)
- Philosophy open mean: 0.325 vs closed mean: 0.306

This suggests that open-weight models, despite generally smaller parameter counts, can achieve comparable context sensitivity.

5.4. Vendor Clustering

The vendor effect becomes significant (p=0.017) when excluding the Gemini Flash anomaly, suggesting that organizational-level design decisions—training data, RLHF procedures [8], safety tuning [1]—create genuine behavioral signatures rather than noise. Moonshot's consistent dominance and Google's safety-filter-driven anomaly represent the extremes.

 doi:10.20944/preprints202602.1114.v1

5.5. Information Hierarchy Validation

The near-universal confirmation of the expected hierarchy (ΔRCI_COLD > ΔRCI_SCRAMBLED in 24/25 runs) is a significant methodological validation. It confirms that scrambled context retains partial information—even disrupted conversational structure provides extractable signal. This validates the three-condition protocol as a well-ordered measurement framework and confirms the "presence > absence" principle [4] at scale.

5.6. Limitations

1. **Single scenario per domain:** One medical case (STEMI) and one philosophical topic (consciousness)
2. **Embedding model ceiling:** all-MiniLM-L6-v2 [9] may not capture all semantic distinctions
3. **Temperature fixed at 0.7:** Other settings may yield different patterns
4. **Claude Opus:** Medical only (absent from philosophy); recovered data lacks response text
5. **Position-level noise:** 50 trials provide limited statistical power for 30-position analysis

5.7. Future Directions

Several extensions would strengthen cross-domain comparisons:

- **Token limit variation:** Testing whether max_tokens (currently 1024) affects context sensitivity differently across domains
- **Multilingual prompts:** Extending ΔRCI measurement to non-English conversations to assess language-specific effects
- **Quantifying openness:** Since philosophy is open-goal, measuring response entropy could provide a continuous "openness" metric for domain characterization, enabling more precise comparisons than the binary closed/open classification
- **Temperature sweeps:** Systematic variation of temperature (0.0–1.0) to map the context sensitivity–randomness tradeoff

6. Conclusions

This study establishes a standardized cross-domain framework for measuring context sensitivity in LLMs. Across 14 models and 112,500 responses, we find that:

1. **Context sensitivity is robust and positive** for nearly all models in both domains (23/25 runs)
2. **Domain structure shapes variance, not mean:** Medical and philosophical domains yield similar average ΔRCI but dramatically different inter-model spread
3. **Safety mechanisms can invert context sensitivity:** Gemini Flash medical anomaly demonstrates deployment-critical risk
4. **Open models compete with closed:** No systematic architectural disadvantage for open-weight models
5. **Vendor signatures are detectable:** Organizational design choices create significant and consistent behavioral patterns ($p=0.017$ excluding anomalies)

This dataset and methodology—building on the ΔRCI framework [4] and addressing gaps in current LLM evaluation [12,14]—provide the foundation for deeper analyses of temporal dynamics (Paper 3) and information-theoretic mechanisms (Paper 4).

Data Availability Statement: All experimental data and analysis code are available at: https://github.com/LaxmanNandi/MCH-Research.

Acknowledgments: This research builds on human-AI collaborative methodology established in Paper 1 [4]. AI systems (Claude, ChatGPT, DeepSeek) assisted with data analysis, visualization, and manuscript preparation. The framework, findings, and interpretations remain the author's sole responsibility.

 doi:10.20944/preprints202602.1114.v1

Appendix A. Complete Per-Model Statistics

Table A1. Complete per-model statistics for all 25 model-domain runs (50 trials each).

Model	Domain	Type	Mean ΔRCI	SD	95% CI
GPT-4o	Philosophy	Closed	0.283	0.011	±0.003
GPT-4o-mini	Philosophy	Closed	0.269	0.009	±0.002
GPT-5.2	Philosophy	Closed	0.308	0.015	±0.004
Claude Haiku	Philosophy	Closed	0.331	0.012	±0.003
Gemini Flash	Philosophy	Closed	0.338	0.022	±0.006
DeepSeek V3.1	Philosophy	Open	0.304	0.014	±0.004
Kimi K2	Philosophy	Open	0.428	0.022	±0.006
Llama 4 Maverick	Philosophy	Open	0.269	0.012	±0.003
Llama 4 Scout	Philosophy	Open	0.298	0.011	±0.003
Ministral 14B	Philosophy	Open	0.373	0.015	±0.004
Mistral Small 24B	Philosophy	Open	0.281	0.009	±0.003
Qwen3 235B	Philosophy	Open	0.322	0.009	±0.003
GPT-4o	Medical	Closed	0.299	0.010	±0.003
GPT-4o-mini	Medical	Closed	0.319	0.010	±0.003
GPT-5.2	Medical	Closed	0.379	0.021	±0.006
Claude Haiku	Medical	Closed	0.340	0.010	±0.003
Claude Opus	Medical	Closed	0.339	0.017	±0.005
Gemini Flash	Medical	Closed	-0.133	0.026	±0.007
DeepSeek V3.1	Medical	Open	0.320	0.010	±0.003
Kimi K2	Medical	Open	0.417	0.016	±0.004
Llama 4 Maverick	Medical	Open	0.316	0.012	±0.003
Llama 4 Scout	Medical	Open	0.323	0.011	±0.003
Ministral 14B	Medical	Open	0.391	0.014	±0.004
Mistral Small 24B	Medical	Open	0.365	0.015	±0.004
Qwen3 235B	Medical	Open	0.328	0.010	±0.003

References

1. Bai, Y., Jones, A., Ndousse, K., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. *arXiv:2212.08073*.
2. Brown, T., Mann, B., Ryder, N., et al. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33. arXiv:2005.14165.
3. Datasaur. (2025). LLM Scorecard 2025. https://datasaur.ai/blog-posts/llm-scorecard-22-8-2025.
4. Laxman, M M. (2026). Context Curves Behavior: Measuring AI Relational Dynamics with ΔRCI. *Preprints.org*. DOI: 10.20944/preprints202601.1881.v2.
5. Mou, X., et al. (2025). Decoupling Safety into Orthogonal Subspace. *arXiv:2510.09004*.
6. Nguyen, T., et al. (2025). A Framework for Neural Topic Modeling with Mutual Information. *Neurocomputing*. DOI: 10.1016/j.neucom.2025.130420.
7. NIH PMC. (2025). Empirically derived evaluation requirements for responsible deployments of AI in safety-critical settings. *npj Digital Medicine*. DOI: 10.1038/s41746-025-01784-y.
8. Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35. arXiv:2203.02155.
9. Reimers, N. & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of EMNLP 2019*. arXiv:1908.10084.
10. Singhal, K., Azizi, S., Tu, T., et al. (2023). Large Language Models Encode Clinical Knowledge. *Nature*, 620, 172-180.
11. Skinner, B. F. (1957). *Verbal Behavior*. Copley Publishing Group.
12. Subramani, N., Srinivasan, R., & Hovy, E. (2025). SimBA: Simplifying Benchmark Analysis. *Findings of EMNLP 2025*. DOI: 10.18653/v1/2025.findings-emnlp.711.
13. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30. arXiv:1706.03762.
14. Xu, Y., et al. (2025). Does Context Matter? ContextualJudgeBench for Evaluating LLM-based Judges. *Proceedings of ACL 2025*. DOI: 10.18653/v1/2025.acl-long.470.

 doi:10.20944/preprints202602.1114.v1