Preprints.org

Article

Not peer-reviewed version

The Structure and Trajectory of Context Sensitivity in Large Language Models: Content-Order Decomposition and Variance Dissociation

Laxman M. M. *

Posted Date: 16 March 2026

doi: 10.20944/preprints202603.1116.v1

Keywords: context sensitivity; large language models; ΔRCI, model coherence hypothesis; contentorder decomposition; variance analysis, RAG systems, clinical AI safety

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 March 2026 doi:10.20944/preprints202603.1116.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

The Structure and Trajectory of Context Sensitivity in Large Language Models: Content-Order Decomposition and Variance Dissociation

Laxman M M

[1] Government Duty Medical Officer, PHC Manchi; barlax5377@gmail.com
[2] Government Duty Medical Officer, PHC Manchi
[3] DNB General Medicine Resident (2026), KC General Hospital, Bangalore

Abstract

Context sensitivity in large language models (LLMs) is typically treated as a single dimension — models either "use context" or they do not. We challenge this view by decomposing context sensitivity into two measurable, independent dimensions: structure and trajectory. Using the Relational Coherence Index (RCI) framework and TRUE/SCRAMBLED/COLD experimental conditions, we analysed the conservation-validated model subset from Paper 6 ($N = 8$ Medical, $N = 6$ Philosophy) using TRUE/SCRAMBLED/COLD experimental conditions. Content-Order decomposition demonstrates that medical reasoning is 61% content-driven while philosophical reasoning is 59% order-driven (Mann-Whitney U = 45, $p = 0.0047$, Cohen's $d = 1.59$). Both content and order increase ΔRCI but have *opposite* effects on variance: content amplifies variance 4–6× while order suppresses it to ∼30% of baseline. This variance machinery is domain-invariant ($p = 0.463$; $p = 0.867$), with domain specificity residing entirely in ΔRCI. Exploration Arc analysis reveals complete domain separation: philosophy expands conceptual diversity over conversation (mean Arc = 15.2) while medical responses remain stable (mean Arc = 1.7), with zero overlap between domains. A pilot analysis of Context Utilization Depth (CUD) is reported in Supplementary Material S1; three of four tested models show full recency dominance (CUD = 1), with Llama 4 Maverick as the sole exception (82%→98% K-curve). These findings provide a structural account of how LLMs process conversation, with direct implications for RAG design, prompt engineering, and clinical AI safety.

Keywords: context sensitivity; large language models; ΔRCI, model coherence hypothesis; content-order decomposition; variance analysis, RAG systems, clinical AI safety

---

Paper 7 of the MCH Research Program — This paper introduces a structural decomposition of context sensitivity into two independent dimensions: structure (Content-Order decomposition) and trajectory (Exploration Arc). Using TRUE/SCRAMBLED/COLD data from 14 LLMs across Medical and Philosophy domains, we demonstrate that the variance machinery governing context stability is domain-invariant, while sensitivity (ΔRCI) is domain-specific. Domain specificity resides entirely in ΔRCI, not in variance components. A pilot analysis of Context Utilization Depth (CUD) is reported in Supplementary Material S1.

1. Introduction

1.1. The Problem: Context as a Monolith

When researchers and practitioners ask whether a large language model "uses context", they implicitly treat context sensitivity as a single property. A model either leverages prior conversation or

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 March 2026 doi:10.20944/preprints202603.1116.v1

it does not. This binary framing has shaped evaluation benchmarks, RAG architectures, and clinical AI safety assessments alike.

Yet the empirical reality is more complex. A model may use the *content* of previous exchanges — the specific facts, diagnoses, or arguments established — while being insensitive to the *order* in which those exchanges occurred. Conversely, a model may encode temporal structure (recency, narrative arc) without retaining factual detail. These are not the same capability, and conflating them obscures both strengths and failure modes.

The present paper introduces a two-dimensional decomposition of context sensitivity:

1. **Structure (Content-Order):** Within context utilisation, how much is driven by informational content versus sequential order? And how do these components affect response stability?
2. **Trajectory (Exploration Arc):** Does response diversity expand, contract, or remain stable as conversation progresses?

A pilot analysis of a third dimension — Depth (CUD: how far back does the model actually look?) — is reported in Supplementary Material S1. Full CUD analysis across all 14 models is reserved for future work.

1.2. The MCH Research Program

This paper is the seventh in the Model Coherence Hypothesis (MCH) Research Program, a systematic empirical investigation of how LLMs process conversational context. Papers 1–2 [4,5] introduced the ΔRCI metric and validated universal context sensitivity across 25 model-domain configurations. Paper 3 [6] demonstrated domain-specific temporal dynamics, including P30 task enablement. Paper 4 [7] established the entanglement framework connecting ΔRCI to response variance ($r = 0.76$, $p = 2.37 \times 10^{-68}$). Paper 5 [8] identified stochastic incompleteness as a clinical AI safety risk. Paper 6 [9] formalised the conservation constraint $\Delta\text{RCI} \times \text{Var\_Ratio} \approx K(\text{domain})$, re-embedded all responses using a standardised response–response alignment method, and validated K across Medical ($N = 8$) and Philosophy ($N = 6$) models; it is currently in progress, extending this validation to three additional domains (Legal, Technical, and Applied Ethics) across 20+ models.

The three conditions used throughout the program are:

- **TRUE:** Full coherent 29-message history, followed by P30 query
- **COLD:** No history; single-turn P30 query only
- **SCRAMBLED:** Same 29 messages as TRUE, but randomised in order, followed by P30 query

The key metric $\Delta\text{RCI} = \overline{\text{RCI}_{\text{TRUE}}} - \overline{\text{RCI}_{\text{COLD}}}$ measures the net benefit of context over baseline, where RCI is mean pairwise cosine similarity of response embeddings (all-MiniLM-L6-v2, 384D) [13].

1.3. Motivation from Prior Work

Papers 1–5 established: (i) the information hierarchy TRUE > SCRAMBLED > COLD holds universally across 25 configurations [5]; (ii) the conservation constraint K is domain-specific and stable within domains (CV < 0.20) [7,8]; (iii) the P30 spike — anomalous elevation of context utilisation at the final position — is present across virtually all models in both conditions [6]; (iv) Llama models exhibit extreme divergent entanglement (Var_Ratio up to 7.46) at the P30 summarization position [7,8].

What Papers 1–5 did not address: *why* does K differ across domains? And what structural properties of context processing drive that difference? The present paper answers these questions by decomposing ΔRCI into its content and order components, measuring how each component affects both sensitivity and stability, and characterising the temporal trajectory of contextual exploration.

1.4. Related Work

Transformer self-attention mechanisms [14] enable in-context learning that underlies modern LLM capabilities [2]. Position effects in context utilization have been documented by Liu et al. [11], who demonstrated a "lost in the middle" phenomenon. The role of ordering in model responses has been studied by Pezeshkpour & Hruschka [12] in multiple-choice settings. Retrieval-Augmented

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 March 2026 doi:10.20944/preprints202603.1116.v1

Generation (RAG) systems [10] depend critically on assumptions about which aspects of retrieved context models actually utilise. Clinical AI safety [1,15] requires precise understanding of how medical context is processed in high-stakes settings. Concurrent with the present work, Goldfeder et al. [3] have proposed Superhuman Adaptable Intelligence (SAI) as a framework for domain-specific AI performance; the ΔRCI framework directly addresses their call for empirical measurement of rapid domain adaptation.

2. Methods

2.1. Data Sources and Models

All data were collected under the MCH Research Program using the TRUE/SCRAMBLED/COLD protocol. Table 1 summarises datasets used in this paper.

Source	Models	Domain	Trials	Conditions
Paper 2 (alignment scores)	14 LLMs, 25 runs	Medical & Philosophy	50	TRUE, SCRAMBLED, COLD (P30 spike confirmation in 4 additional closed-API models, Section 3.3)
Paper 6 (re-embedded, Medical)	8 LLMs	Medical (P30)	50	TRUE, SCRAMBLED, COLD
Paper 6 (re-embedded, Philosophy)	6 LLMs	Philosophy (P30)	50	TRUE, SCRAMBLED, COLD
CUD pilot (Supp. S1)	4 LLMs	Med & Phil	50/K-level	K ∈ {1,5,10,15,20,29}

Table 1. Data sources for Paper 7 analyses. CUD pilot data are reported in Supplementary Material S1.

Both content-order decomposition (Figure 1) and variance decomposition (Figure 3) are reported for the **conservation-validated subset** from Paper 6: $N = 8$ Medical (Gemini Flash 2.0, DeepSeek V3.1, Kimi K2, Llama 4 Maverick, Llama 4 Scout, Ministral 14B, Mistral Small 24B, Qwen3 235B) and $N = 6$ Philosophy (Claude Haiku, Gemini Flash 2.0, GPT-4o, GPT-4o-mini, DeepSeek V3.1, Llama 4 Maverick). Using this subset ensures that content-order, variance, and K decomposition analyses are internally consistent across all figures. These models have SCRAMBLED response text stored (required for variance re-embedding via all-MiniLM-L6-v2) and verified conservation products ($\Delta RCI \times Var\_Ratio \approx K$) from Paper 6 [9]. Models outside this subset either lack stored SCRAMBLED response text (closed-API models collected under an earlier experiment script) or are absent from the conservation product validation.

2.2. Context Utilization Depth (CUD) — Pilot

CUD is defined as the minimum K (number of most recent context messages) required to achieve $\geq$90% of full-context ΔRCI (see Supplementary Material S1 for full definition and results). Four models were piloted: DeepSeek V3.1, Gemini Flash 2.0, Qwen3 235B, and Llama 4 Maverick. Full CUD analysis across all 14 models is deferred to future work; a summary of pilot results is noted in Section 3.1.

2.3. Content-Order Decomposition

The SCRAMBLED condition enables direct decomposition of ΔRCI into content and order components:

$$\Delta RCI_{content} = \overline{RCI_{SCRAMBLED}} - \overline{RCI_{COLD}} \quad (1)$$

$$\Delta RCI_{order} = \overline{RCI_{TRUE}} - \overline{RCI_{SCRAMBLED}} \quad (2)$$

$$\text{Content Fraction} = \frac{\Delta RCI_{content}}{\Delta RCI_{TRUE}} \quad (3)$$

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 March 2026 doi:10.20944/preprints202603.1116.v1

Content Fraction is computed at the final conversation position (P30 for Medical, P15 for Philosophy) using pre-stored alignment scores across 50 trials.
Variance decomposition uses three ratios, each computed across 50 independent trials:

$$\mathrm{VR}_{\mathrm{Content}} = \frac{\mathrm{Var}(\text{SCRAMBLED embeddings})}{\mathrm{Var}(\text{COLD embeddings})} \tag{4}$$

$$\mathrm{VR}_{\mathrm{Order}} = \frac{\mathrm{Var}(\text{TRUE embeddings})}{\mathrm{Var}(\text{SCRAMBLED embeddings})} \tag{5}$$

$$\mathrm{VR}_{\mathrm{Total}} = \frac{\mathrm{Var}(\text{TRUE embeddings})}{\mathrm{Var}(\text{COLD embeddings})} \tag{6}$$

All variance computations used all-MiniLM-L6-v2 embeddings (384D) [13]. Content Fraction (Figure 1) and variance decomposition (Figures 3, 4) are both computed for the conservation-validated subset: $N = 8$ Medical and $N = 6$ Philosophy models with stored SCRAMBLED response text, ensuring internal consistency across Figures 1–4. Domain comparisons used Mann-Whitney U; all between-group tests are two-tailed.

2.4. Exploration Arc

Exploration Arc is defined as the ratio of response diversity at P30 to P1:

$$\mathrm{Arc} = \frac{\mathrm{Var}(\text{TRUE embeddings at P30})}{\mathrm{Var}(\text{TRUE embeddings at P1})} \tag{7}$$

Arc > 5.0 defines an *Exploration Zone* (diversity expands over conversation); Arc < 3.0 defines a *Stable/Convergent Zone*.

3. Results

3.1. Context Utilization Depth: Pilot Note

CUD pilot data (Supplementary Figure S1) show that three of four tested models are fully recency-dominant (CUD = 1), with $\Delta\mathrm{RCI}_{\mathrm{truncated}}(K = 1) \geq 100\%$ of $\Delta\mathrm{RCI}_{\mathrm{TRUE}}$ in both domains. Llama 4 Maverick is the sole exception: K-curves rise monotonically from 82% at $K = 1$ to 98% at $K = 29$ in the Medical domain, indicating genuine integration of distal context (CUD > 1). Philosophy CUD was undefined for all four models, suggesting that philosophical reasoning context provides insufficient coherence signal for depth measurement. Full CUD analysis across all 14 models is reserved for future work.

3.2. Content-Order Decomposition of ΔRCI

Figure 1 shows the Content Fraction for all models across both domains. Medical domain models cluster at 61.3% $\pm$ 6.7% (mean $\pm$ SD), indicating that informational content drives the majority of context benefit in clinical reasoning. Philosophy domain models cluster at 41.2% $\pm$ 16.7%, indicating that sequential order drives 59% of context benefit in open-goal reasoning.

The domain difference is significant (Mann-Whitney U = 45, $p = 0.0047$, Cohen's $d = 1.59$), with a large effect size. This dissociation is consistent across model families and vendors within each domain.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 March 2026 doi:10.20944/preprints202603.1116.v1

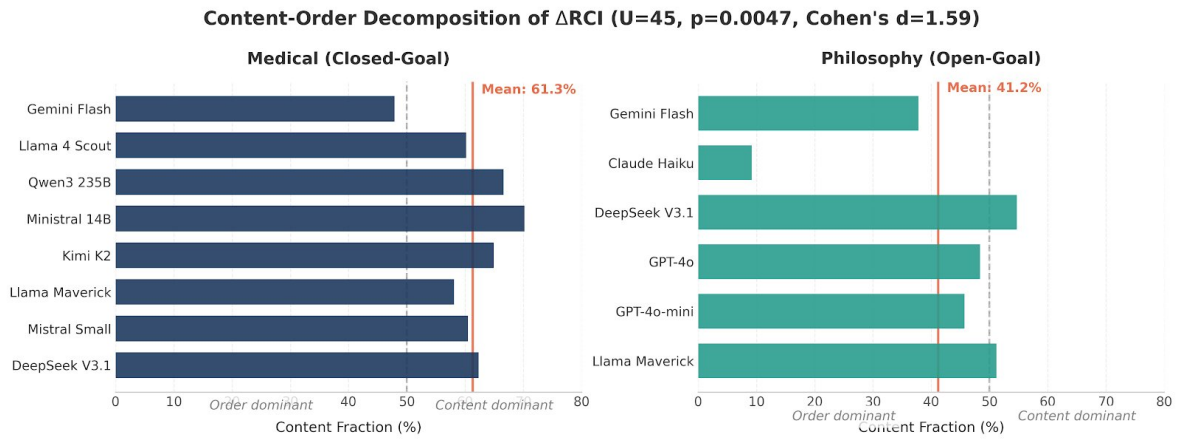

Figure 1. Content-Order Decomposition of ΔRCI. Medical domain (left, $N = 8$) and Philosophy domain (right, $N = 6$) shown by model. Content Fraction = proportion of ΔRCI attributable to informational content (SCRAMBLED – COLD). Mean values: Medical 61.3%, Philosophy 41.2%. Mann-Whitney U = 45, $p = 0.0047$, Cohen's $d = 1.59$.

3.3. The P30 Spike: Content and Order Both Contribute

The P30 spike (first reported in Paper 3 [6]) is decomposed in Figure 2. All 8/8 conservation-validated models show significant SCRAMBLED–COLD elevation ($z > 1.96$), confirming that the P30 spike is driven by informational content — the specific clinical facts, management decisions, and learning points accumulated across the conversation — and not merely by the temporal position of the P30 query. The spike was also confirmed in four additional closed-API models outside the conservation subset (Claude Haiku, GPT-4o, GPT-4o-mini); GPT-5.2 showed a weaker but present effect ($z = +1.88$).

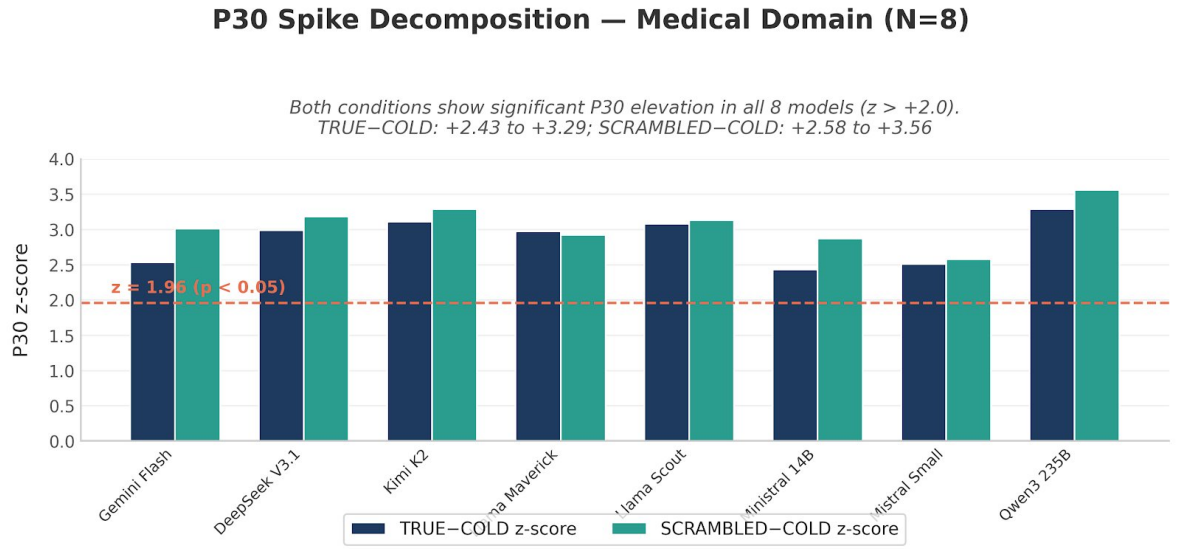

Figure 2. P30 Spike Decomposition — Medical Domain. TRUE–COLD z-scores (dark blue) and SCRAMBLED–COLD z-scores (teal) for 8 models (conservation-validated subset). Dashed orange line marks $z = 1.96$ (significance threshold). All 8/8 models exceed the significance threshold in both conditions, confirming that the P30 spike is not purely an order effect — informational content (present in SCRAMBLED) is sufficient to drive it. TRUE–COLD z-scores: +2.43 to +3.29; SCRAMBLED–COLD z-scores: +2.58 to +3.56. The spike was also confirmed in four additional closed-API models outside the conservation subset (Claude Haiku, GPT-4o, GPT-4o-mini; GPT-5.2 showed a weaker effect at $z = +1.88$).

3.4. Variance Decomposition: Content Destabilizes, Order Stabilizes

Figure 3 presents the central finding of this paper. The variance decomposition reveals opposing mechanisms:

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 March 2026 doi:10.20944/preprints202603.1116.v1

Content effect (Panel A): VR_Content = Var(SCRAMBLED) / Var(COLD). Medical: 4.22 ± 1.20; Philosophy: 5.69 ± 2.73. Content *amplifies* variance 4–6×. Domain difference: $p = 0.463$ (NS).

Order effect (Panel B): VR_Order = Var(TRUE) / Var(SCRAMBLED). Medical: 0.32 ± 0.07; Philosophy: 0.32 ± 0.08. Order *suppresses* variance to ∼30% of the content-inflated baseline. Domain difference: $p = 0.867$ (NS).

Net effect (Panel C): The large amplification from content and the strong suppression from order approximately cancel, yielding near-neutral total variance — but through opposing mechanisms, not through absence of effect.

Critically, both VR_Content and VR_Order are statistically indistinguishable between Medical and Philosophy domains. Domain specificity in context sensitivity resides entirely in ΔRCI, not in the variance machinery.

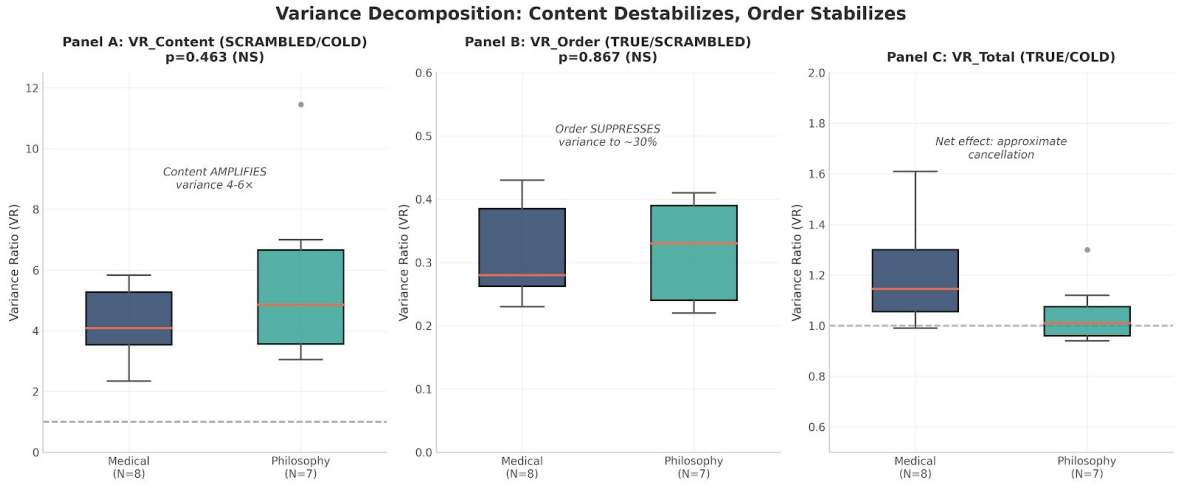

Figure 3. Variance Decomposition: Content Destabilizes, Order Stabilizes. Panel A: VR_Content (SCRAMBLED/COLD) — content amplifies variance 4–6×, domain difference $p = 0.463$ (NS). Panel B: VR_Order (TRUE/SCRAMBLED) — order suppresses variance to ∼30% of SCRAMBLED baseline, domain difference $p = 0.867$ (NS). Panel C: VR_Total (TRUE/COLD) — net approximate cancellation. $N = 8$ Medical, $N = 6$ Philosophy.

3.5. Conservation Product K Decomposition

Figure 4 decomposes the conservation product $K = \Delta\text{RCI} \times \text{Var\_Ratio}$ into its content and order constituents. K_total differs significantly between domains (Medical: 0.429, CV = 0.17; Philosophy: 0.301, CV = 0.17; Mann-Whitney U = 46, $p = 0.003$, Cohen's $d = 2.06$). However, K_content and K_order do not differ between domains ($p = 0.852$ and $p = 0.282$, both NS).

This dissociation identifies the mechanistic source of domain-level K differences: ΔRCI is higher in medical than philosophical reasoning (reflecting closed-goal convergence), while the variance machinery operates identically across domains. K is latent in the domain prior; context activates it but does not create it.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 March 2026 doi:10.20944/preprints202603.1116.v1

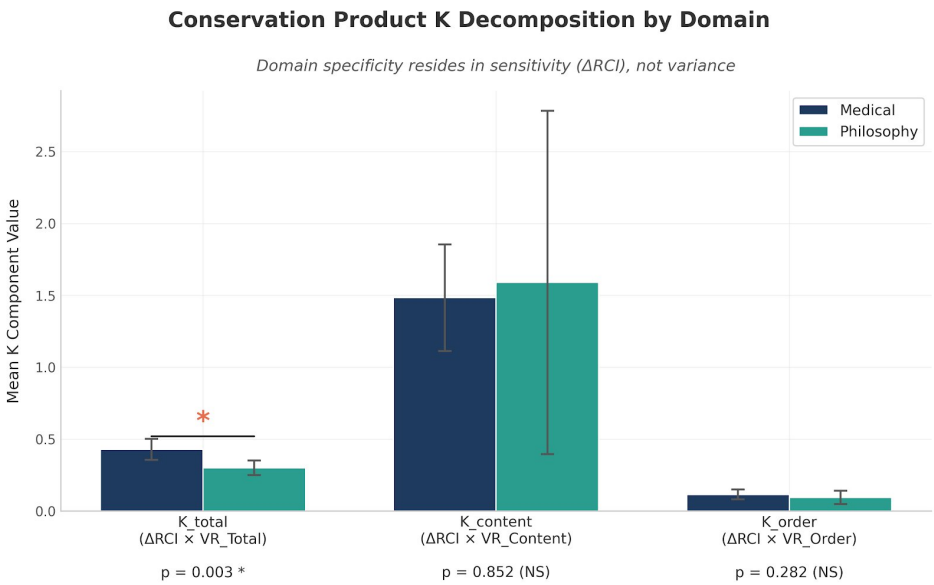

Figure 4. Conservation Product K Decomposition by Domain. K_total (ΔRCI × VR_Total): Medical 0.429, Philosophy 0.301, $p = 0.003$ (*). K_content (ΔRCI × VR_Content): $p = 0.852$ (NS). K_order (ΔRCI × VR_Order): $p = 0.282$ (NS). Domain specificity in K resides in ΔRCI, not in variance components.

3.6. The Sensitivity-Stability Dissociation

Figure 5 summarises the fundamental asymmetry: content increases ΔRCI *and* amplifies variance (destabilising); order increases ΔRCI further *and* suppresses variance (stabilising). The net effect of a coherent conversation (TRUE condition) is therefore: maximal sensitivity combined with moderate stability. Sensitivity is domain-dependent; the variance machinery is domain-invariant.

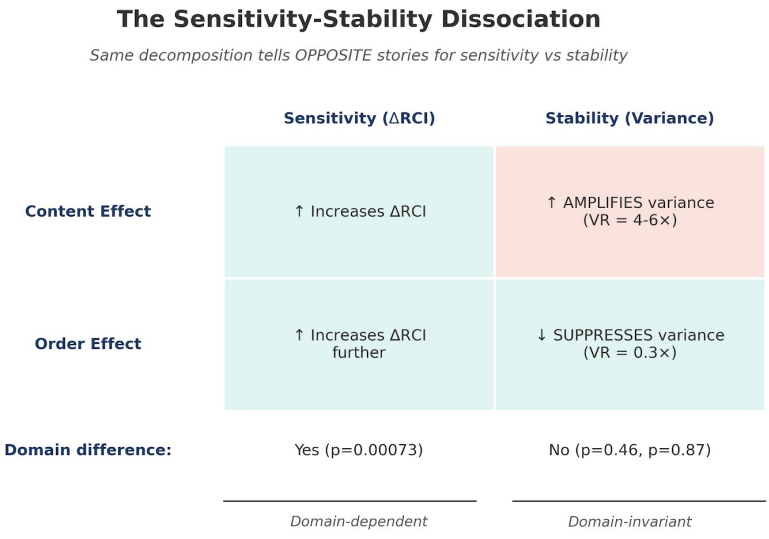

Figure 5. The Sensitivity-Stability Dissociation. Content and order effects on sensitivity (ΔRCI, left column) versus stability (Variance, right column). Both content and order increase ΔRCI, but have opposite effects on variance. The variance machinery is domain-invariant; the sensitivity effect is domain-dependent.

3.7. Llama P30 Anomaly Decomposed

The Llama P30 anomaly (first reported in Paper 4 as extreme divergent entanglement: Scout Var_Ratio = 7.46, Maverick = 2.64 [7]) is fully explained by the content-order decomposition (Figure 6). The anomaly is not a failure of order processing, but an extreme form of content sensitivity. The Llama architecture is maximally responsive to informational content accumulated across the conversation, particularly at synthesis positions.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 March 2026 doi:10.20944/preprints202603.1116.v1

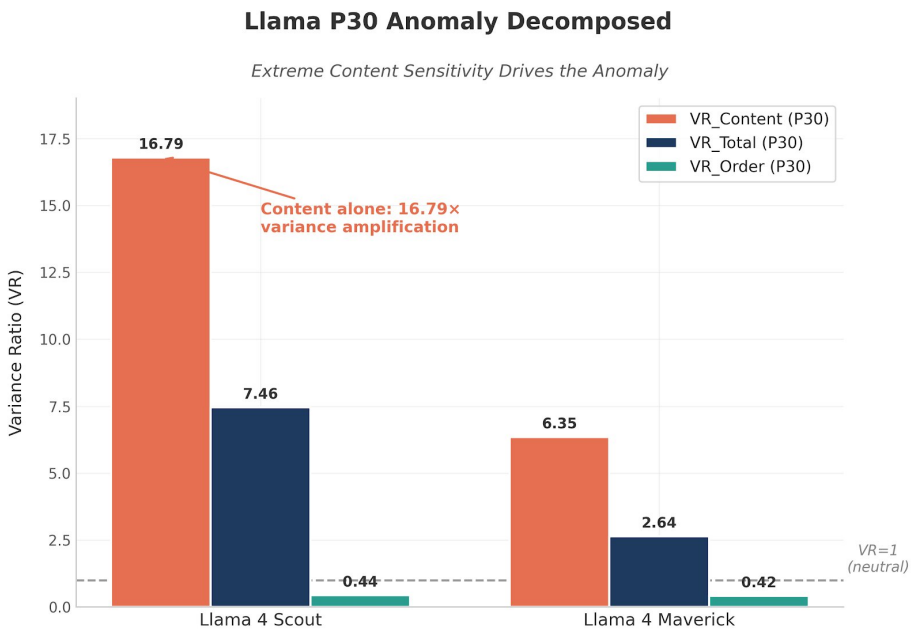

Figure 6. Llama P30 Anomaly Decomposed. Llama 4 Scout (left) and Llama 4 Maverick (right) show extreme VR_Content at P30 (Scout: 16.79×; Maverick: 6.35×), far exceeding the domain mean of 4–6×. VR_Order remains normal (Scout: 0.44; Maverick: 0.42). The anomaly is driven entirely by extreme content sensitivity at the summarization position.

3.8. Exploration Arc

Figure 7 reveals complete domain separation in Exploration Arc. All eight Medical models fall below Arc = 3.0 (mean 1.72 ± 0.68, range 0.75–2.95), consistent with a closed-goal domain where responses progressively integrate toward a clinical synthesis. All six Philosophy models fall above Arc = 5.0 (mean 15.23 ± 16.64, range 5.05–48.52), consistent with an open-goal domain where each exchange opens new directions of inquiry. The zero overlap between domains is confirmed: max Medical (2.95) $<$ min Philosophy (5.05). Domain membership fully predicts Arc category (Mann-Whitney U = 0.0, $p = 0.0007$, Cohen's $d = 1.15$).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 March 2026 doi:10.20944/preprints202603.1116.v1

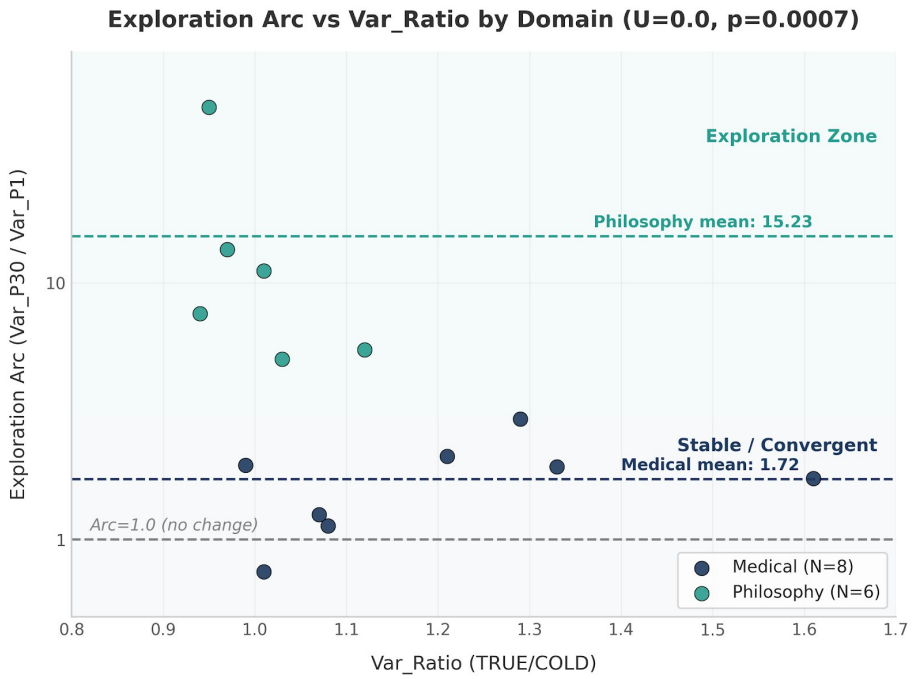

Figure 7. Exploration Arc vs VR_Total by Domain (log scale). Medical models ($N = 8$, dark blue) cluster in the Stable/Convergent Zone (Arc < 3.0, mean 1.72 ± 0.68, range 0.75–2.95). Philosophy models ($N = 6$, teal) cluster in the Exploration Zone (Arc > 5.0, mean 15.23 ± 16.64, range 5.05–48.52). Zero overlap between domains: max Medical (2.95) $<$ min Philosophy (5.05). Mann-Whitney U = 0.0, $p = 0.0007$, Cohen's $d = 1.15$.

3.9. Information Hierarchy Schematic

Figure 8 provides the conceptual framework underlying the three-condition design. A coherent conversation (TRUE) contains three additive information components: Baseline (domain prior, present in COLD), Content (semantic information, present in SCRAMBLED), and Order (temporal structure, recoverable only from TRUE). Scrambling preserves content while destroying order — enabling the surgical decomposition reported in this paper.

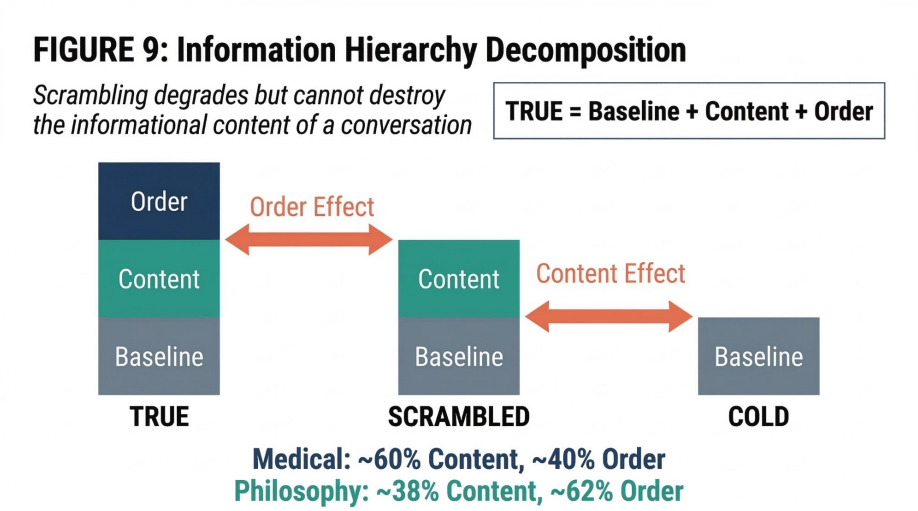

Figure 8. Information Hierarchy Decomposition. TRUE = Baseline + Content + Order. Scrambling degrades but cannot destroy the informational content of a conversation. Medical domain: ~60% Content, ~40% Order. Philosophy domain: ~38% Content, ~62% Order.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 March 2026 doi:10.20944/preprints202603.1116.v1

4. Discussion

4.1. Context Is Not a Monolith

The central finding of this paper is that context sensitivity cannot be meaningfully described by a single number. A model with high ΔRCI may be content-driven (Medical-type) or order-driven (Philosophy-type); these are structurally different forms of context utilisation with different implications for system design. The three-dimensional decomposition provides a complete fingerprint of how a model processes conversation.

4.2. The Variance Machinery Is Universal

Perhaps the most surprising finding is the domain invariance of variance effects. Content amplification (VR_Content $\sim$ 4–6×) and order suppression (VR_Order $\sim$ 0.30×) are statistically identical in Medical and Philosophy domains ($p = 0.463$ and $p = 0.867$). This suggests the mechanisms governing response stability are architectural rather than domain-specific. Domain specificity ($K_{\text{Medical}} > K_{\text{Philosophy}}$, $p = 0.003$) resides entirely in ΔRCI, which reflects domain-level response patterns embedded in model weights during training. The conservation product K is latent in the domain prior; context activates it but does not create it.

This finding is further supported by Paper 6's (in progress) baseline analysis: cross-vendor Medical cosine similarity > 0.97; cross-domain $\sim$0.18 [9]. The "Prior Voice" — domain-specific response patterns present before any context is provided — is now fully explained mechanistically. Domain determines both the baseline prior and the sensitivity to contextual perturbation.

4.3. Implications for RAG Design

Retrieval-Augmented Generation systems [10] typically optimise for content retrieval. The present findings suggest that order of retrieved content may be equally important in open-goal domains, but relatively less important in closed-goal domains. Specifically:

- For closed-goal tasks (clinical decision support, legal analysis, technical debugging): prioritise content completeness. Order effects contribute only $\sim$40% of context benefit; content arrangement matters less than factual coverage.
- For open-goal tasks (philosophical reasoning, creative generation, strategic planning): preserve narrative order. Order drives $\sim$62% of context benefit; shuffling retrieved chunks may destroy the primary mechanism of context utilisation.

The variance findings add a further dimension: content-heavy RAG designs will amplify response variability (VR_Content 4–6×), while order-preserving designs will suppress it (VR_Order $\sim$0.30×). Applications requiring stable, reproducible outputs (clinical AI) should favour order-preserving context assembly.

4.4. Implications for Clinical AI Safety

The P30 spike decomposition confirms that the task enablement effect [6] is driven by informational content. Wornow et al. [15] and Asgari et al. [1] identified reliability and consistency as central concerns for clinical AI deployment. The present findings provide a mechanistic account: content sensitivity (VR_Content 4–6×) means that different clinical histories will produce substantially different response distributions. Order sensitivity (VR_Order $\sim$0.30×) means that standardised history formats will suppress this variability. The Exploration Arc finding is also clinically relevant: medical models are uniformly convergent (Arc $\approx$ 1.72), meaning that longer clinical conversations produce more stable responses — a desirable property for clinical applications.

4.5. Connection to Concurrent Work

Goldfeder et al. [3] argue that AI systems must embrace domain specialisation and propose SAI as a framework for measuring rapid adaptation, raising the challenge of empirical measurement of domain-specific contextual performance. The ΔRCI framework directly addresses this: Content

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 March 2026 doi:10.20944/preprints202603.1116.v1

Fraction, Exploration Arc, and CUD together constitute a domain-sensitivity fingerprint characterising both the form and magnitude of contextual adaptation.

4.6. Limitations

1. **Two domains.** Content-Order decomposition was conducted on Medical and Philosophy domains only. Paper 6 [9] will test decomposition across three additional domains (Legal, Technical, Applied Ethics). Preliminary Legal data (2/5 models complete) confirms the information hierarchy TRUE > SCRAMBLED > COLD; full decomposition pending.
2. **CUD pilot scope (Supplementary S1).** CUD was measured in four models only, with Llama 4 Maverick as the sole model showing CUD > 1. Whether this reflects an architecture-specific property cannot be determined from four models; full CUD analysis across all 14 models is deferred to future work.
3. **Embedding model.** All variance computations used all-MiniLM-L6-v2. Robustness to alternative embeddings will be reported in Paper 6's robustness analysis.
4. **Philosophy Arc sample size.** Exploration Arc computed for $N = 6$ philosophy models. The zero-overlap finding is strong, but replication with larger samples is warranted.
5. **Single P30 synthesis prompt.** CUD and Arc are measured at P30. Whether these properties hold at intermediate positions and across different prompt types is an open question.
6. **Closed-source models.** GPT-4o, GPT-5.2, Claude Haiku, and Gemini Flash are closed-source; architectural interpretations remain descriptive.

5. Conclusion

Context sensitivity in LLMs is not a single property but a structured, measurable space with at least two independent dimensions. Structure (Content-Order decomposition) characterises what drives context benefit — content in closed-goal domains (Medical: 61%), order in open-goal domains (Philosophy: 59%), with the distinction significant ($p = 0.0047$, $d = 1.59$). Trajectory (Exploration Arc) characterises how response diversity evolves — convergent in medical, expansive in philosophical reasoning, with complete domain separation and zero overlap.

The most fundamental finding is the dissociation between sensitivity and stability: content and order both increase ΔRCI but have precisely opposite effects on variance. This variance machinery is domain-invariant — content amplifies 4–6× and order suppresses to ∼30% regardless of domain — while sensitivity is domain-specific. Domain specificity resides in ΔRCI, activated by context but latent in the domain prior.

These findings provide a structural account of how LLMs process conversation, with direct implications for RAG architecture, prompt engineering, and clinical AI safety. Context is not a monolith. Its anatomy matters.

Data Availability Statement: All analysis scripts, pre-registered hypotheses, and supplementary data are available at https://github.com/LaxmanNandi and https://osf.io/7954v/. OSF pre-registration: https://osf.io/dp8nj/ (March 6, 2026).

Conflicts of Interest: The author declares no competing interests.

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.

References

1. Asgari, E., et al. (2025). A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *npj Digital Medicine*, 8(1), 274. DOI: 10.1038/s41746-025-01776-y.
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.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 March 2026 doi:10.20944/preprints202603.1116.v1

3. Goldfeder, R., Wyder, T., LeCun, Y., & Shwartz-Ziv, R. (2026). AI must embrace specialization via Superhuman Adaptable Intelligence (SAI). *arXiv preprint arXiv:2602.23643*. Submitted February 27, 2026.
4. Laxman, M M. (2026a). Context curves behavior: Measuring AI relational dynamics with ΔRCI. *Preprints.org*. DOI: 10.20944/preprints202601.1881.v2.
5. Laxman, M M. (2026b). Scaling context sensitivity: Standardized benchmark across 25 LLM-domain configurations. *Preprints.org*. DOI: 10.20944/preprints202602.1114.v2.
6. Laxman, M M. (2026c). Domain-specific temporal dynamics of context sensitivity in large language models. *Preprints.org*. DOI: 10.20944/preprints202602.1674.v1.
7. Laxman, M M. (2026d). Engagement as entanglement: Variance signatures of bidirectional context coupling in large language models. *Preprints.org*. DOI: 10.20944/preprints202603.0055.v1.
8. Laxman, M M. (2026e). Stochastic incompleteness: A predictability taxonomy for clinical AI deployment. *Preprints.org*. DOI: 10.20944/preprints202602.2034.v1.
9. Laxman, M M. (2026f). Validating the conservation law across five domains: Legal, technical, and applied ethics. *Preprints.org*. In progress (data collection ongoing). OSF pre-registration: https://osf.io/dp8nj/.
10. Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.
11. Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12, 157–173. DOI: 10.1162/tacl_a_00638.
12. Pezeshkpour, P. & Hruschka, E. (2024). Large language models sensitivity to the order of options in multiple-choice questions. *Findings of ACL: NAACL 2024*, 2975–2984. DOI: 10.18653/v1/2024.findings-naacl.130.
13. Reimers, N. & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. *Proceedings of EMNLP 2019*. arXiv:1908.10084.
14. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30. arXiv:1706.03762.
15. Wornow, M., Xu, Y., Thapa, R., et al. (2023). The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1), 135. DOI: 10.1038/s41746-023-00879-8.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.