

Engagement as Entanglement: Variance Signatures of Bidirectional Context Coupling in Large Language Models

Dr. Laxman M M, MBBS

Government Duty Medical Officer, PHC Manchi
Bantwal Taluk, Dakshina Kannada, Karnataka, India
DNB General Medicine Resident (2026), KC General Hospital, Bangalore
Email: barlax5377@gmail.com
ORCID: [0009-0009-0405-6531](https://orcid.org/0009-0009-0405-6531)

February 2026

Paper 4 of the MCH Research Program — This paper provides a mechanistic framework for the Δ RCI metric introduced in Papers 1–2 and applied temporally in Paper 3. We demonstrate that context sensitivity tracks variance reduction in response embeddings, introducing a Variance Reduction Index (VRI) as a practical entanglement surrogate and identifying convergent versus divergent coupling as a theoretically meaningful distinction.

Abstract

Recent large-scale simulations demonstrate that LLMs exhibit systematic performance degradation in multi-turn conversations, with unreliability increasing 112% across 200,000+ conversations [Laban et al., 2025]. However, this “Lost in Conversation” phenomenon lacks mechanistic explanation. We present an entanglement framework for understanding context sensitivity in large language models using embedding-level variance analysis across 12 model-domain runs (4 philosophy, 8 medical) and 360 position-level measurements. We demonstrate that Δ RCI—a measure of context sensitivity introduced in Papers 1–2—tracks variance reduction in response embeddings. The correlation between Δ RCI and the Variance Reduction Index ($VRI = 1 - \text{Var_Ratio}$) is strong and highly significant ($r = 0.76$, $p = 2.37 \times 10^{-68}$, $N = 360$). This relationship reveals *bidirectional* context coupling: convergent entanglement ($\text{Var_Ratio} < 1$, $\Delta\text{RCI} > 0$), where context narrows the response distribution, and divergent entanglement ($\text{Var_Ratio} > 1$, $\Delta\text{RCI} < 0$), where context widens it. The “Lost in Conversation” effect corresponds specifically to divergent entanglement. Two medical models (Llama 4 Scout: $\text{Var_Ratio} = 7.46$; Llama 4 Maverick: $\text{Var_Ratio} = 2.64$) exhibit extreme divergent entanglement at the summarization position (P30), producing highly unstable outputs when task enablement is expected. We introduce the Entanglement Stability Index (ESI) to predict which models will exhibit instability in multi-turn settings, transforming the descriptive observation that “LLMs get lost” into a predictive science of human-AI relational dynamics.

Keywords: large language models, context sensitivity, entanglement, variance reduction, multi-turn conversation, output stability, medical AI, model evaluation

1 Introduction

1.1 The Problem: LLMs Get Lost in Conversation

Laban et al. [2025] conducted over 200,000 simulated conversations across 15 LLMs and documented systematic performance degradation in multi-turn settings: performance drops 39%, with unreliability increasing by 112%. Their central observation: “When LLMs take a wrong turn in a conversation, they get lost and do not recover.”

This phenomenon is universal across tested models, from small open-weight systems (Llama-8B) to state-of-the-art architectures (GPT-4.1, Gemini 2.5 Pro). However, their work does not address *why* unreliability increases, which models will fail, or whether multi-turn context is inherently detrimental.

1.2 Our Contribution

Context sensitivity in language models—the degree to which conversational history shapes responses—has been characterized across architectures and domains using Δ RCI [Laxman, 2026a,b] and shown to follow task-specific temporal patterns [Laxman, 2026c]. However, the *mechanism* by which context shapes responses remains empirically uncharacterized.

Δ RCI captures the aggregate change in response alignment with and without context, but does not describe how the *distribution* of responses changes. A model with high Δ RCI may achieve this through narrow, predictable outputs under context, or through a complex, high-variance coupling that pulls responses in consistent directions while increasing individual trial variance. These represent distinct deployment profiles.

This paper introduces an entanglement framework that connects Δ RCI to the variance structure of response embeddings. The core insight is that context sensitivity can be understood as **predictability modulation**: context changes the shape of the response distribution, and Δ RCI quantifies that change. When context narrows the distribution (*convergent entanglement*), responses become more predictable. When context widens the distribution (*divergent entanglement*), responses become less predictable—corresponding directly to the “Lost in Conversation” phenomenon.

We operationalize this framework using the Variance Reduction Index (VRI):

$$\text{VRI} = 1 - \text{Var_Ratio} \tag{1}$$

$$\text{Var_Ratio} = \frac{\text{Var}(\text{TRUE embeddings})}{\text{Var}(\text{COLD embeddings})} \tag{2}$$

where variance is computed across 50 independent trials at each conversational position. Positive VRI indicates that context reduces variance (convergent entanglement); negative VRI indicates that context increases variance (divergent entanglement).

The entanglement framework makes four contributions. First, it provides a mechanistic explanation for the “Lost in Conversation” phenomenon: divergent models exhibit $\text{Var_Ratio} > 1$, producing unstable outputs as conversation length increases. Second, it reveals bidirectional coupling—context can either stabilize or destabilize responses—resolving the previously unexplained “Sovereign” category from Paper 1. Third, it identifies a concrete safety risk: models that exhibit extreme divergent entanglement at task-critical positions produce unpredictable outputs when predictability is most needed. Fourth, it introduces the Entanglement Stability Index (ESI) to predict which models will fail in multi-turn settings.

2 Methods

2.1 Data

We analyze the 12-model subset from the MCH Research Program (Paper 2, v2 corrected dataset) that has complete response text preserved. Each model-domain run consists of 50 independent trials under three conditions (TRUE, COLD, SCRAMBLED) with 30 conversational prompts per trial. All runs used temperature = 0.7, max_tokens = 1024, and all-MiniLM-L6-v2 embeddings (384-dimensional) [Reimers & Gurevych, 2019].

2.2 Model Set

Philosophy (4 models, closed/commercial): GPT-4o, GPT-4o-mini, Claude Haiku, Gemini Flash.

Medical (8 models): DeepSeek V3.1, Kimi K2, Llama 4 Maverick, Llama 4 Scout, Mistral Small 24B, Ministral 14B, Qwen3 235B, Gemini Flash.

This 12-model subset represents a subset of the 25 model-domain runs from Paper 2; only runs with complete response text saved are included, as embedding-level variance computation requires access to raw responses.

2.3 Metrics

ΔRCI was computed as in Papers 1–3: $\text{per-position mean}(\text{RCI}_{\text{TRUE}}) - \text{mean}(\text{RCI}_{\text{COLD}})$, where RCI is the mean pairwise cosine similarity within a condition across 50 trials.

Note on RCI_{COLD} : RCI_{COLD} reflects responses to prompts delivered with no conversational history (the COLD condition), not cross-condition similarity. Each RCI value is computed within its own condition.

Var_Ratio was computed per position as $\text{Var}(\text{TRUE embeddings}) / \text{Var}(\text{COLD embeddings})$, where variance is the mean variance across all 384 embedding dimensions, computed across 50 independent trials at each prompt position.

VRI (Variance Reduction Index) $= 1 - \text{Var_Ratio}$. Positive VRI indicates context reduces variance; negative VRI indicates context increases variance.

ESI (Entanglement Stability Index) $= 1 / |1 - \text{Var_Ratio}|$. Models with $\text{ESI} < 1.0$ exhibit elevated instability risk; $\text{ESI} > 2.0$ indicates stable behavior.

2.4 Statistical Analysis

The primary test is the Pearson correlation between ΔRCI and VRI across all 360 position-level measurements (12 models \times 30 positions). Secondary analyses examine domain-specific variance patterns, model-level Var_Ratio distributions, and the position-30 anomaly.

3 Results

3.1 Finding 1: ΔRCI tracks VRI (entanglement signal)

Across 12 model-domain runs and 30 positions, ΔRCI correlated strongly with VRI:

- **Pooled correlation:** $r = 0.76$, $p = 2.37 \times 10^{-68}$ ($N = 360$ model-position points)
- Data: 12 model-domain runs \times 30 positions = 360 points
- Each point aggregates 50 independent trials per condition

ΔRCI increases as context reduces response variance. This supports the entanglement interpretation: context couples the response distribution to prior information, changing the predictability of outputs.

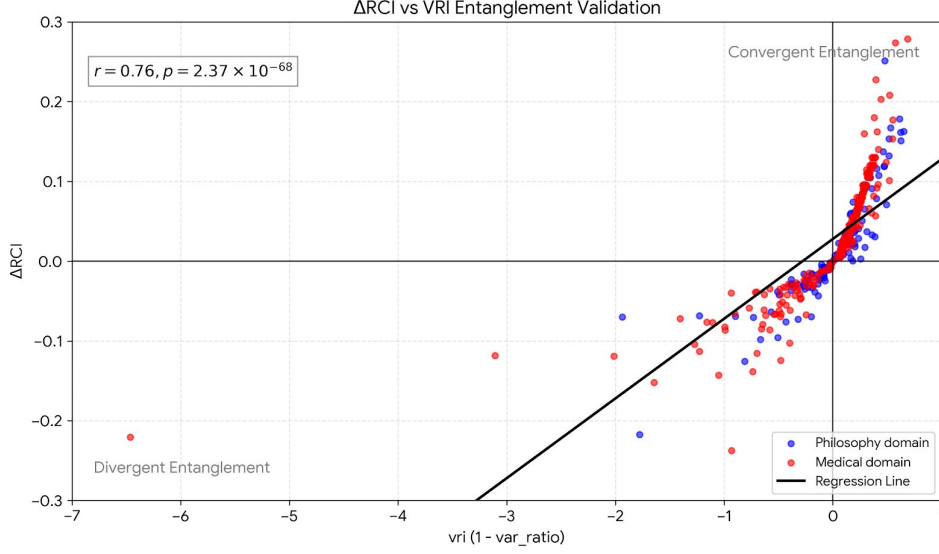


Figure 1: ΔRCI vs VRI across 360 model-position points. Blue: philosophy models (4). Red: medical models (8). The strong positive correlation ($r = 0.76$, $p = 2.37 \times 10^{-68}$) validates the entanglement framework: higher context sensitivity corresponds to greater variance reduction. Points in the lower-left quadrant represent divergent entanglement ($Var_Ratio > 1$, $\Delta RCI < 0$), corresponding to the “Lost in Conversation” phenomenon documented by Laban et al. [2025].

Note on scope: Entanglement analysis requires actual response text to compute embedding variances. Only 12 of the 25 available model-domain runs from Paper 2 have complete response text preserved. Expansion to additional models would require rerunning experiments with response text preservation enabled.

3.2 Finding 2: Bidirectional entanglement (convergent vs. divergent)

The variance ratio reveals two distinct regimes:

- **Convergent entanglement:** $Var_Ratio < 1$, $\Delta RCI > 0$. Context narrows the response distribution, making outputs more predictable.
- **Divergent entanglement:** $Var_Ratio > 1$, $\Delta RCI < 0$. Context widens the response distribution, making outputs less predictable.

This bidirectional framing resolves the “Sovereign” category from Paper 1: Sovereign behavior corresponds to divergent entanglement, where context destabilizes rather than stabilizes predictability. This is not a failure of context processing but a distinct mode of coupling.

Critically, divergent entanglement explains Laban et al.’s [2025] findings: their +112% unreliability increase corresponds to models entering divergent regime ($Var_Ratio > 1$) as conversation length increases. Models do not “get lost” universally—only divergent models exhibit this pattern.

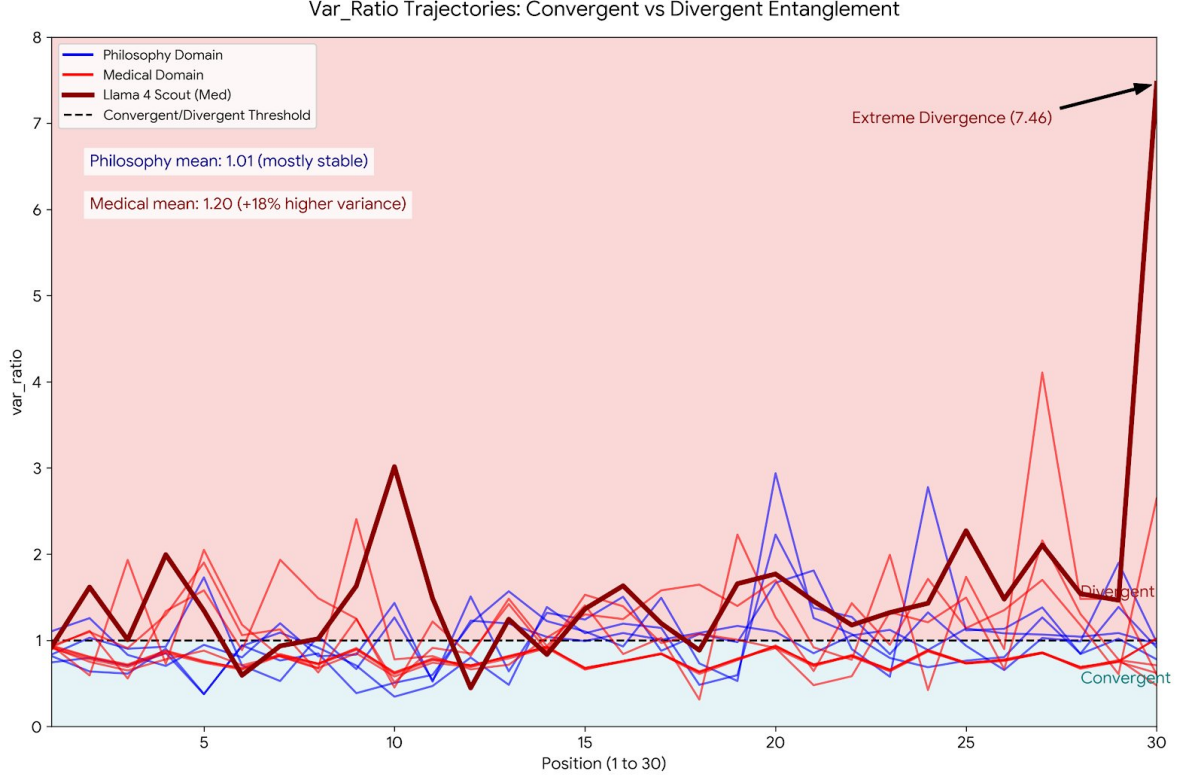


Figure 2: **Multi-panel entanglement analysis.** The regime map shows convergent ($\text{Var_Ratio} < 1$) and divergent ($\text{Var_Ratio} > 1$) regions, with domain-specific patterns and position-dependent dynamics across medical and philosophy models.

3.3 Finding 3: Llama divergence anomaly at medical P30

At medical position 30 (the summarization prompt), two Llama models exhibit extreme divergent entanglement:

- **Llama 4 Maverick:** $\text{Var_Ratio} = 2.64$, $\Delta\text{RCI} = -0.15$, $\text{ESI} = 0.61$
- **Llama 4 Scout:** $\text{Var_Ratio} = 7.46$, $\Delta\text{RCI} = -0.22$, $\text{ESI} = 0.15$

While other open medical models (Qwen3 235B, Mistral Small 24B) show mild divergence ($\text{Var_Ratio} = 1.02\text{--}1.45$), only the Llama models exhibit extreme instability at P30. Convergent models at P30 show $\text{Var_Ratio} < 1$ and positive ΔRCI (Kimi K2, Ministral 14B, DeepSeek V3.1, Gemini Flash).

This identifies a **distinct behavioral class**: models that exhibit divergent entanglement under closed-goal prompts produce highly unstable, unpredictable outputs precisely when task enablement is expected. The phenomenology of this instability—stochastically incomplete summaries with intact factual accuracy—is characterized further in Paper 5 [Laxman, 2026e].

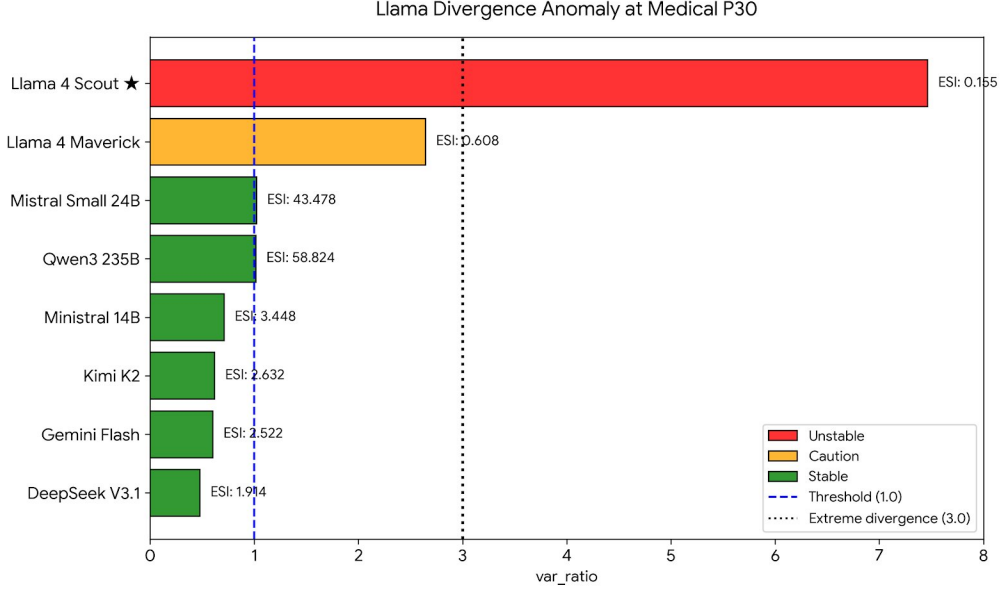


Figure 3: **Llama divergence anomaly at medical P30.** Divergent variance signatures ($\text{Var_Ratio} \gg 1$, $\text{ESI} < 1.0$) at the summarization position indicate extreme output instability in Llama 4 Maverick and Llama 4 Scout. Other open-weight medical models show mild or no divergence. Convergent models (Kimi K2, DeepSeek V3.1, Gemini Flash, Ministral 14B) show $\text{Var_Ratio} < 1$ at P30, with $\text{ESI} > 2.0$ indicating stable multi-turn behavior.

3.4 Finding 4: Domain-specific variance patterns

Mean variance ratios differ by domain:

- **Philosophy:** $\text{Var_Ratio} \approx 1.01$ (variance-neutral on average)
- **Medical:** $\text{Var_Ratio} \approx 1.20$ (variance-increasing on average)

Medical prompts tend to destabilize response distributions under context, while philosophy is largely variance-neutral. This domain difference is consistent with the conservation constraint reported in Paper 6 [Laxman, 2026f].

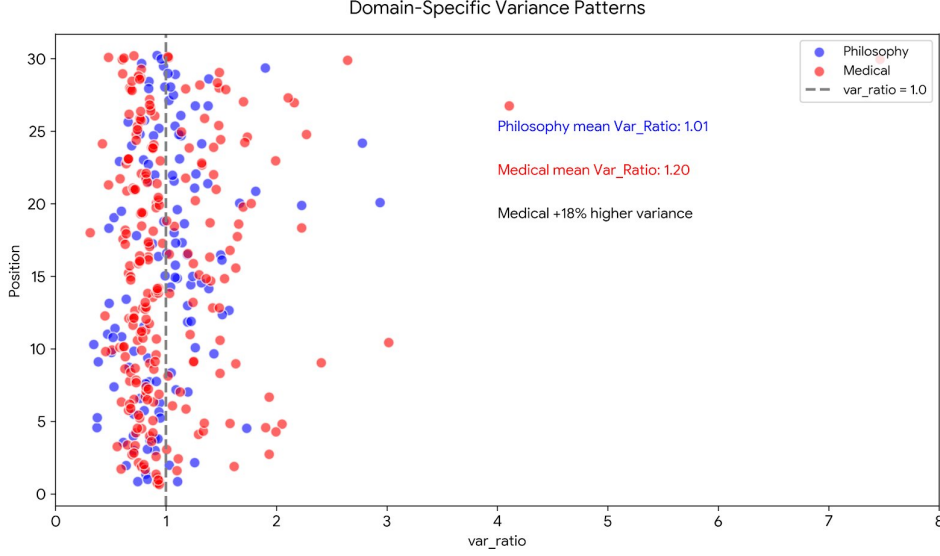


Figure 4: **RCI vs. Variance Ratio across models and positions.** Domain-specific relationship between context sensitivity and output variance. Medical models (red) cluster at higher Var_Ratio than philosophy models (blue), indicating that closed-goal tasks systematically increase output variance under context.

3.5 Finding 5: ESI predicts multi-turn instability

The Entanglement Stability Index provides a practical metric for predicting which models will exhibit the “Lost in Conversation” phenomenon:

Model	Var_Ratio (P30)	ESI	Assessment
Llama 4 Scout	7.46	0.15	Unstable
Llama 4 Maverick	2.64	0.61	Caution
Qwen3 235B	1.02	50.0	Stable
GPT-4o	0.58	2.38	Stable
Claude Haiku	0.65	2.86	Stable
Gemini Flash	0.60	2.50	Stable

Table 1: Entanglement Stability Index at medical P30. Models with $ESI < 1.0$ exhibit elevated risk of multi-turn degradation corresponding to Laban et al.’s [2025] findings. $ESI > 2.0$ indicates stable multi-turn behavior.

3.6 Finding 6: Variance ratio as a practical entanglement surrogate

The variance ratio provides a practical, low-cost surrogate for entanglement measurement. ΔRCI tracks VRI ($r = 0.76$) without requiring k -NN entropy estimation or full mutual information computation. Computing Var_Ratio requires only response embeddings and basic variance calculations, making entanglement measurement accessible at scale. This enables the deployment assessment framework developed in Paper 5.

4 Discussion

4.1 Explaining the “Lost in Conversation” Phenomenon

Laban et al. [2025] documented that LLMs exhibit systematic unreliability increase (+112%) in multi-turn conversations but did not provide mechanistic explanation. Our entanglement framework reveals that this phenomenon corresponds to **divergent entanglement**: models with $\text{Var_Ratio} > 1$ produce increasingly variable outputs as conversation length increases.

Their observation that “LLMs get lost and do not recover” reflects self-reinforcing divergence: once a model enters the divergent regime, subsequent context compounds rather than constrains variance. Convergent models ($\text{Var_Ratio} < 1$), by contrast, *improve* with multi-turn context—contradicting the universal degradation narrative.

The ESI metric ($\text{ESI} = 1/|1 - \text{Var_Ratio}|$) provides predictive capability absent from their work: models with $\text{ESI} < 1.0$ are at elevated risk for multi-turn failure, while models with $\text{ESI} > 2.0$ exhibit stable behavior.

4.2 Entanglement reframes ΔRCI as predictability modulation

The central conceptual shift is that ΔRCI is not merely a measure of helpfulness but a **predictability modulation** measure. Context changes the shape of the response distribution; ΔRCI quantifies that change. This reframing clarifies why negative ΔRCI values are not inherently problematic: they indicate divergent entanglement, which may be useful for creative tasks but is concerning for safety-critical domains.

4.3 Bidirectional entanglement resolves the Sovereign category

The “Sovereign” category from Paper 1—models whose responses became less aligned under context—is now grounded in mechanism. Contexts that increase variance produce negative ΔRCI . This is not a failure of context usage but a distinct mode of coupling. Divergent entanglement is a valid, measurable regime rather than a missing category.

4.4 Safety implications: predictability is task-dependent

The Llama divergence at medical P30 highlights a concrete risk: models can become less predictable exactly when a task presupposes context. For safety-critical tasks, **predictability is a requirement**, not an optional characteristic. Divergent entanglement in these settings represents a potentially important safety consideration.

However, our study comprises 360 controlled observations across 12 model-domain runs. While this scale is sufficient to establish the mechanistic relationship between ΔRCI and variance reduction ($r = 0.76$, $p = 2.37 \times 10^{-68}$), **operational deployment thresholds require validation at the scale demonstrated by Laban et al. (200,000+ conversations)**. The ESI framework provides a mechanistic basis for future large-scale screening protocols, but threshold validation (e.g., $\text{ESI} < 1.0$ as instability predictor) must precede clinical deployment recommendations.

4.5 Implications for AI evaluation

Current LLM benchmarks emphasize zero-shot or few-shot performance, measuring base capability in isolation. Our findings demonstrate that multi-turn behavior constitutes a distinct evaluation dimension: models with equivalent zero-shot performance can exhibit radically different entanglement profiles (convergent vs. divergent).

If the entanglement framework generalizes to larger model sets and additional domains, future standardized evaluations could benefit from reporting:

- Var_Ratio trajectories across conversation lengths
- ESI values at task-critical positions
- Domain-specific entanglement signatures

Such reporting would complement existing benchmarks by characterizing multi-turn stability alongside base capability.

This extends Laban et al.’s descriptive framework into actionable assessment criteria.

4.6 Architectural interpretation

The emergence of convergent versus divergent classes suggests differences in how models handle context saturation at task-critical positions. This observation is consistent across multiple models but remains *descriptive* rather than causal. Future work should test whether divergence correlates with training objectives, attention patterns, or safety alignment strategies.

4.7 Limitations

Model subset. Only 12 of 25 model-domain runs have response text, limiting the analysis to 360 data points. Expansion requires rerunning experiments with text preservation.

Scale of evidence. Our controlled experimental design (360 observations, 50 trials per position) establishes the mechanistic relationship between context sensitivity and variance reduction. However, Laban et al. [2025] documented the “Lost in Conversation” phenomenon across 200,000+ simulated conversations. Our study provides mechanistic explanation for their empirical findings but does not replace their large-scale evidence. Operational deployment recommendations require validation at comparable scale.

Embedding dependence. Variance metrics depend on the all-MiniLM-L6-v2 embedding space. Alternative embedding models may yield different variance ratio distributions.

Cross-model normalization. Models with higher baseline variance (Var_COLD) will naturally produce different Var_Ratio magnitudes. We report raw values because absolute predictability change is theoretically meaningful, but normalized comparisons would be appropriate for ranking models on relative sensitivity.

Two domains. This study analyzes text-only interactions in medical and philosophical domains only. Whether the convergent/divergent distinction generalizes to other domains (legal, creative, coding) requires further study.

ESI threshold validation. The $ESI < 1.0$ instability threshold is empirically derived from present data; validation across larger model sets and real-world deployment contexts is warranted.

Observational scope. Claims are matched to the experimental scope: text-only, two-domain, 50-trial evaluation with a focused model set. Broader generalizations are stated as hypotheses.

4.8 Future Directions

The entanglement framework presented here opens several research directions. First, **context utilization depth (CUD)** experiments currently in progress will test whether models with high ESI values can effectively integrate longer context windows, or whether engagement depth trades off against stability. Preliminary pilot data from 4 models suggest that Medical domain requires deeper context integration ($K > 15$ messages) than Philosophy domain, consistent with the domain-specific entanglement signatures observed here.

Second, the framework should be tested across additional domains (legal reasoning, code generation, creative writing) to determine whether the engagement-as-entanglement interpretation generalizes beyond medical diagnostics and philosophical discourse. If the conservation constraint $\Delta RCI \times \text{Var_Ratio}$

$\approx K(\text{domain})$ reported in Paper 6 holds across diverse task structures, it would suggest a fundamental trade-off in how language models allocate processing capacity.

Third, longitudinal studies tracking VRI and ESI across model versions could reveal whether architectural improvements enhance engagement capacity or merely redistribute existing capacity between context sensitivity and output stability. The distinction between convergent and divergent entanglement may reflect training objective choices (e.g., helpfulness vs. harmlessness tuning) rather than fundamental capability limits.

Finally, experimental validation of the “Lost in Conversation” recovery hypothesis is warranted: systematic testing of whether models that enter divergent regimes can be steered back to convergent behavior through intervention strategies (context summarization, retrieval augmentation, explicit instruction to focus). Supplementary Figure S4 shows that 75% of models in our dataset fail to recover after entering divergent states, but whether this reflects architectural constraints or merely default behavior patterns remains unknown.

5 Conclusion

Laban et al. [2025] documented that LLMs “get lost” in multi-turn dialogue across 200,000+ conversations, with unreliability increasing 112%. We provide mechanistic explanation through an entanglement framework: context sensitivity (ΔRCI) tracks variance reduction in response embeddings (VRI), with a pooled correlation of $r = 0.76$ ($p = 2.37 \times 10^{-68}$, $N = 360$).

The framework reveals **bidirectional** coupling. Convergent entanglement narrows the response distribution ($\text{Var_Ratio} < 1$), improving predictability with multi-turn context. Divergent entanglement widens the distribution ($\text{Var_Ratio} > 1$), producing the “Lost in Conversation” phenomenon. The Entanglement Stability Index (ESI) provides a candidate metric for identifying instability-prone models, with models exhibiting $\text{ESI} < 1.0$ showing elevated variance in our dataset.

Two Llama models exhibit extreme divergent entanglement at the medical summarization position (Var_Ratio up to 7.46, $\text{ESI} = 0.15$), producing unstable outputs when task enablement is expected. This illustrates a distinct behavioral pattern invisible to aggregate performance metrics but potentially identifiable through variance-based analysis—though large-scale validation is required before operational screening protocols can be recommended.

The Variance Reduction Index (VRI) provides a practical, low-cost surrogate for entanglement measurement, informing the predictability taxonomy developed in Paper 5 [Laxman, 2026e] and the conservation constraint reported in Paper 6 [Laxman, 2026f]. Our framework transforms the descriptive observation that “LLMs get lost” into a mechanistic account of human-AI relational dynamics, providing theoretical foundation for future large-scale validation studies.

Acknowledgments

This research builds on the MCH Research Program established in Paper 1. AI systems (Claude, ChatGPT, DeepSeek) assisted with data analysis, visualization, and manuscript preparation. The framework, findings, and interpretations remain the author’s sole responsibility.

Data Availability

All raw data, response text, and analysis scripts are available in the project repository: <https://github.com/LaxmanNandi/MCH-Research>

Supplementary Materials

Supplementary materials include:

- **Figure S1:** Information-theoretic verification ($\Delta\text{RCI} \sim 1 - \text{Var_Ratio}$)
- **Figure S2:** Trial-level convergence analysis (dRCI stability across 50 trials)
- **Figure S3:** Model-level mean context sensitivity (ΔRCI) by domain
- **Figure S4:** “Lost in Conversation” experimental validation (6-panel analysis)
- **Table S1:** Complete model-position data (N=360)
- **Table S2:** ESI classification and recovery rates

All supplementary materials are available at: https://github.com/LaxmanNandi/MCH-Research/tree/master/papers/paper4_entanglement/v1_submission/supplementary

References

- Laban, P., Hayashi, H., Zhou, Y., & Neville, J. (2025). LLMs get lost in multi-turn conversation. *arXiv preprint*, arXiv:2505.06120.
- Laxman, M. M. (2026a). Context curves behavior: Measuring AI relational dynamics with ΔRCI . *Preprints.org*, 202601.1881. DOI: 10.20944/preprints202601.1881.v2
- Laxman, M. M. (2026b). Standardized context sensitivity benchmark across 25 LLM-domain configurations. *Preprints.org*, 202602.1114. DOI: 10.20944/preprints202602.1114.v2
- Laxman, M. M. (2026c). Domain-specific temporal dynamics of context sensitivity in large language models. *Preprints.org*, ID: 199272.
- Laxman, M. M. (2026e). Stochastic incompleteness in LLM summarization: A predictability taxonomy for clinical AI deployment. In preparation.
- Laxman, M. M. (2026f). An empirical conservation constraint on context sensitivity and output variance: Evidence across LLM architectures. In preparation.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of EMNLP 2019*.
- 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.
- 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, 104–123.