Supplementary Materials

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

February 2026

Contents

1 Supplementary Figures **2**
1.1 Figure S1: Information-Theoretic Verification . . . 2
1.2 Figure S2: Trial-Level Convergence Analysis . . . 3
1.3 Figure S3: Model-Level ΔRCI Comparison . . . 3
1.4 Figure S4: "Lost in Conversation" Experimental Validation . . . 4

2 Supplementary Tables **5**
2.1 Table S1: Complete Model-Position Data . . . 5
2.2 Table S2: ESI Classification and Recovery Rates . . . 5

3 Supplementary Methods **6**
3.1 "Lost in Conversation" Test Design . . . 6
3.1.1 Experiment 1: Var_Ratio Progression Test . . . 6
3.1.2 Experiment 2: Convergent vs Divergent Classification . . . 6
3.1.3 Experiment 3: Recovery Analysis . . . 6
3.1.4 Experiment 4: Domain Comparison . . . 6
3.1.5 Experiment 5: Llama Trajectory Analysis . . . 7
3.2 Implementation . . . 7

4 Supplementary Results **8**
4.1 Detailed Progression Statistics . . . 8
4.2 Recovery Event Analysis . . . 8

5 Supplementary Discussion **9**
5.1 Laban et al. (2025) Comparison . . . 9
5.2 Clinical Implications . . . 9
5.3 Architectural Insights . . . 9
5.4 Limitations of "Lost in Conversation" Analysis . . . 10

1 Supplementary Figures

1.1 Figure S1: Information-Theoretic Verification

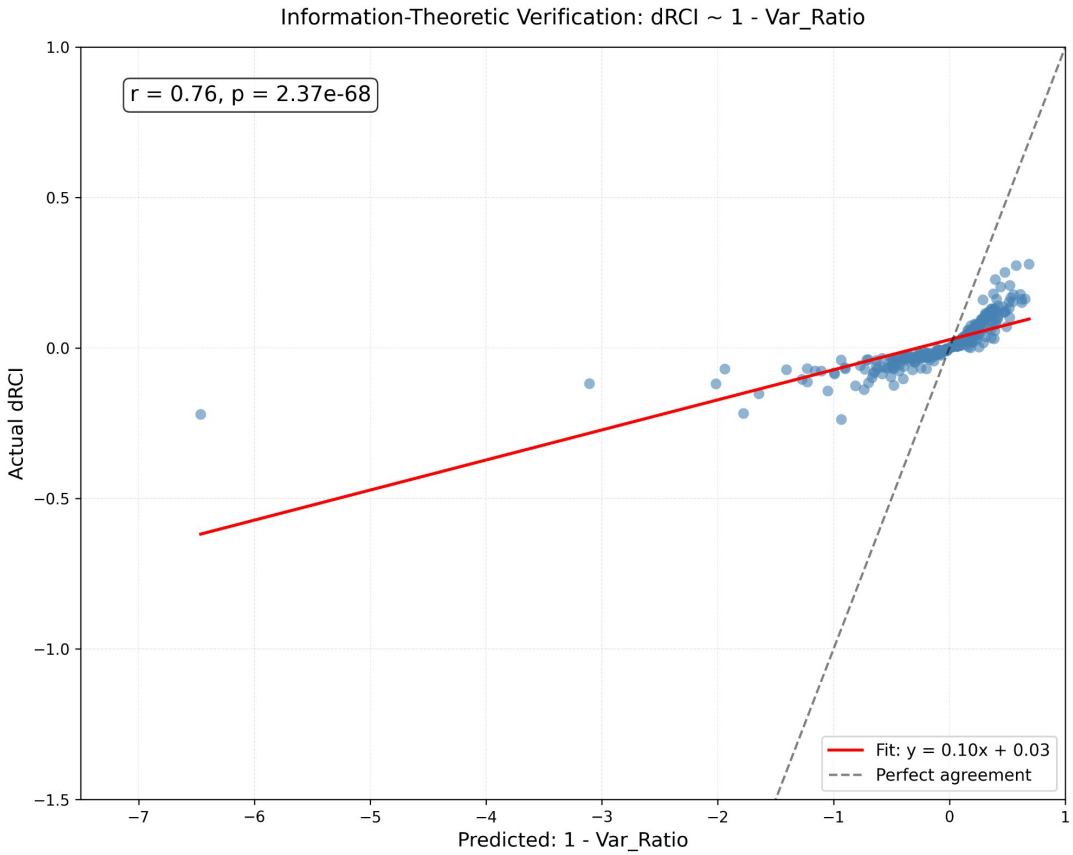

Figure 1: **Information-theoretic verification: ΔRCI ∼ 1 – Var_Ratio.** The correlation ($r = 0.76$, $p = 2.37 \times 10^{-68}$) validates the use of Pearson correlation as the primary statistical test. The regression fit ($y = 0.10x + 0.03$) shows systematic attenuation relative to perfect agreement (dashed diagonal), indicating that ΔRCI captures a fraction of the total variance modulation. Effective dimensionality: 12–23 of 384 dimensions.

1.2 Figure S2: Trial-Level Convergence Analysis

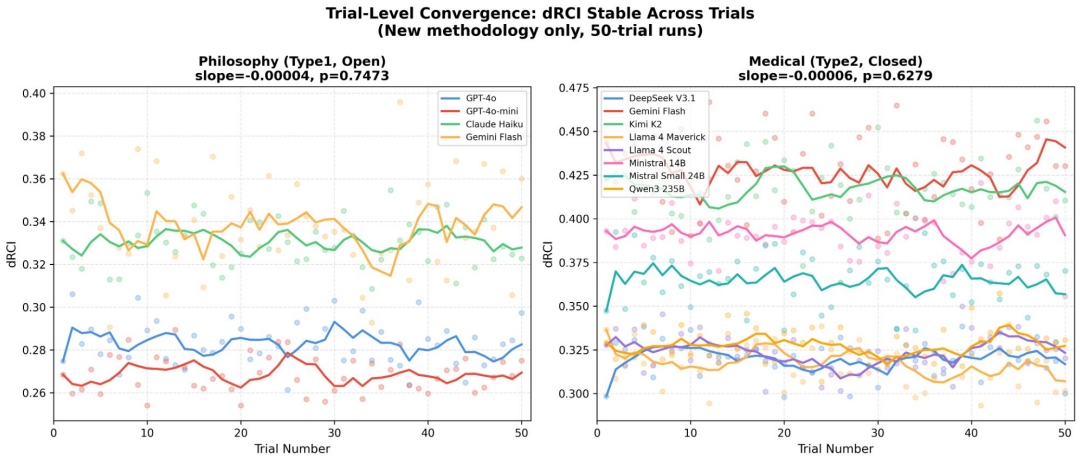

Figure 2: **Trial-level convergence analysis.** ΔRCI stability across 50 independent trials per model-domain run, split by domain. **Left:** Philosophy domain (Type1, Open) shows slope $= -0.00004$, $p = 0.7473$ (non-significant drift). **Right:** Medical domain (Type2, Closed) shows slope $= -0.00006$, $p = 0.6279$ (non-significant drift). Rolling 5-trial means (colored lines) show stable estimates throughout. Both domains confirm 50-trial adequacy for stable ΔRCI estimation.

1.3 Figure S3: Model-Level ΔRCI Comparison

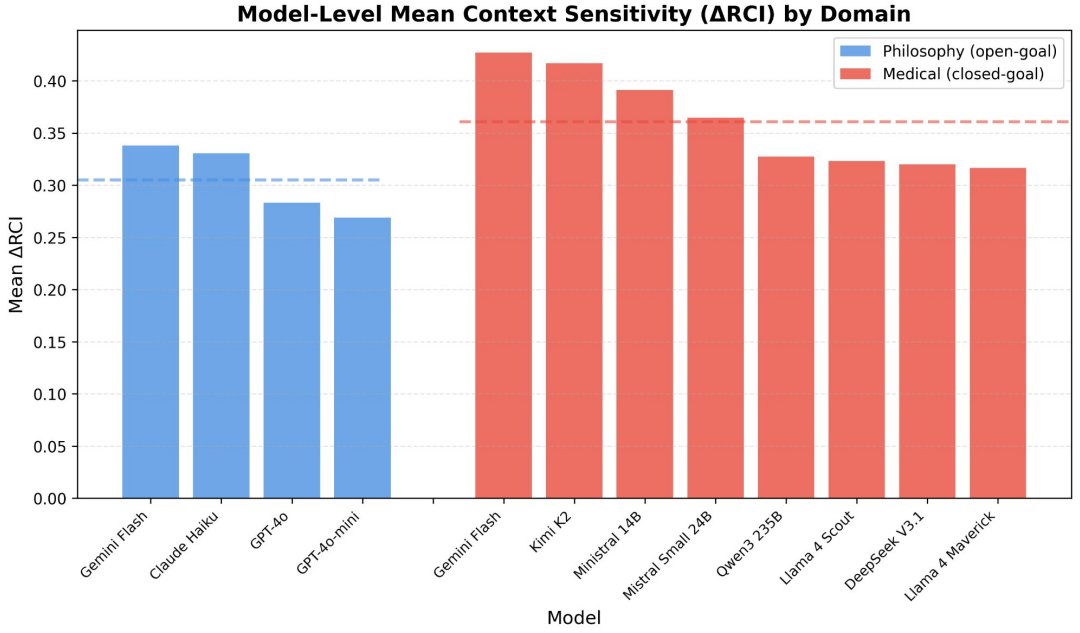

Figure 3: **Model-level mean context sensitivity (ΔRCI) by domain.** Mean ΔRCI values by model, grouped by domain. Philosophy models (blue, open-goal): Gemini Flash, Claude Haiku, GPT-4o, GPT-4o-mini. Medical models (red, closed-goal): Gemini Flash, Kimi K2, Ministral 14B, Mistral Small 24B, Qwen3 235B, Llama 4 Scout, DeepSeek V3.1, Llama 4 Maverick. Dashed lines show domain means. Medical domain shows higher mean ΔRCI than philosophy, consistent with closed-goal tasks eliciting stronger context sensitivity.

1.4 Figure S4: "Lost in Conversation" Experimental Validation

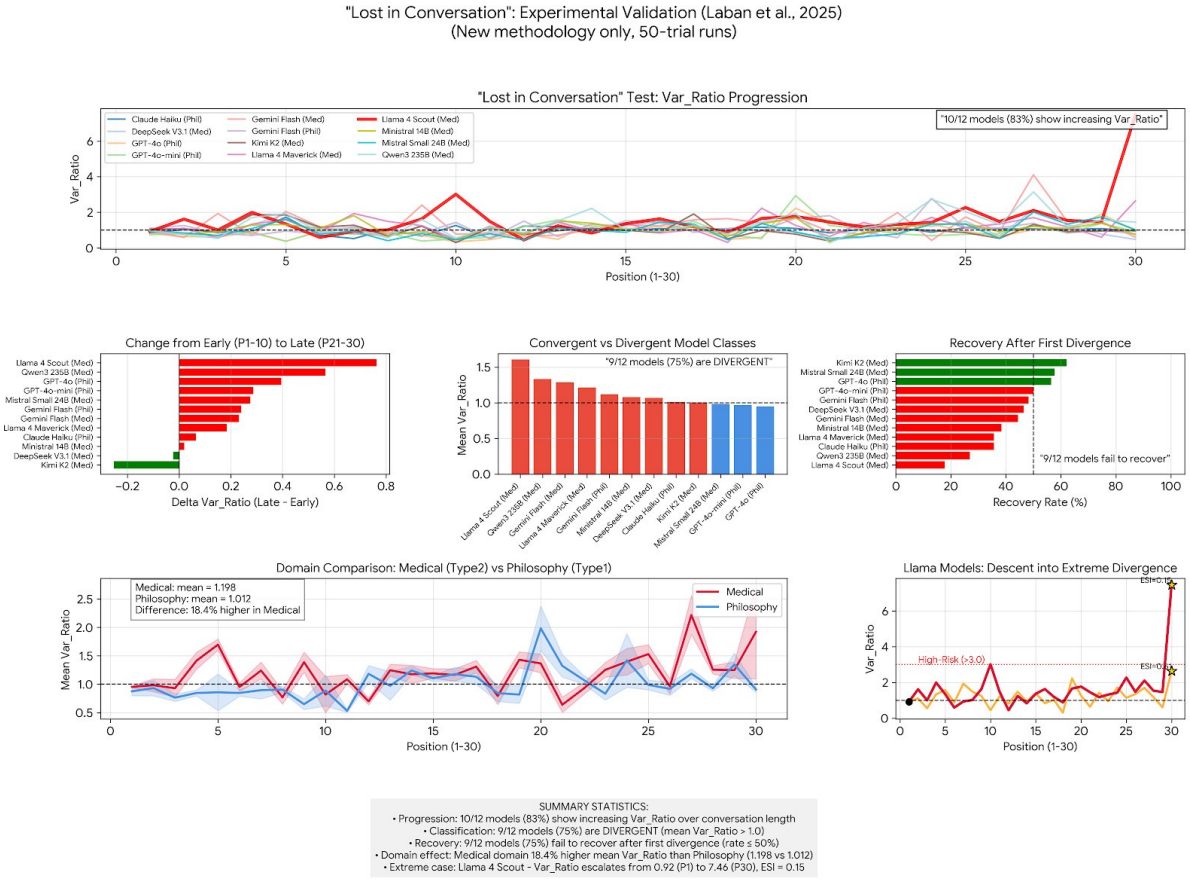

Figure 4: **Experimental validation of the "Lost in Conversation" phenomenon (Laban et al., 2025). Top panel**: Var_Ratio progression across 30 conversational positions for all 12 models. Llama 4 Scout (red) shows extreme divergence escalation from 0.92 (P1) to 7.46 (P30). 10/12 models (83%) exhibit INCREASING Var_Ratio, consistent with Laban et al.'s unreliability increase. **Middle-left panel**: Change in Var_Ratio from early (P1–P10) to late (P21–P30) conversation. Positive bars indicate models getting MORE unstable over conversation length. Llama 4 Scout shows largest increase (+0.76), Kimi K2 shows decrease (-0.25). **Middle-center panel**: Model classification by mean Var_Ratio. 9/12 models are DIVERGENT (Var_Ratio > 1.0, red bars). Only 3 models maintain CONVERGENT behavior (Var_Ratio < 1.0, blue bars). **Middle-right panel**: Recovery analysis after first divergence event. 9/12 models show recovery rate $< 50\%$ (red), confirming Laban et al.'s "do not recover" claim. Only 3 models (GPT-4o, Mistral Small 24B, Kimi K2) show $> 50\%$ recovery (green). **Bottom-left panel**: Domain comparison showing Medical (red) vs Philosophy (blue) Var_Ratio trajectories. Medical domain shows higher mean Var_Ratio (1.198 vs 1.012). **Bottom-right panel**: Llama models' descent into extreme divergence. Both start stable (Var_Ratio ~ 0.92 at P1) but escalate dramatically by P30 (Maverick: 2.64, Scout: 7.46). Star marks extreme divergence (Var_Ratio > 3.0). **Summary**: Progression: 10/12 models (83%) show increasing Var_Ratio; Classification: 9/12 (75%) are divergent; Recovery: 9/12 (75%) fail to recover; Domain effect: Medical 18% higher mean Var_Ratio; Extreme case: Llama 4 Scout ESI $=$ 0.15 at P30. These results provide independent empirical validation of Laban et al.'s (2025) findings using the MCH dataset ($N = 360$ observations, 12 models).

2 Supplementary Tables

2.1 Table S1: Complete Model-Position Data

Table 1: Complete ΔRCI and Var_Ratio data across 360 observations (12 models × 30 positions). Models span 8 vendors with parameter counts from 14B to 671B. Var_Ratio > 1.0 indicates divergent entanglement (context increases output variance). VRI = 1 - Var_Ratio. Data available in full at: https://github.com/LaxmanNandi/MCH-Research/tree/master/analysis/entanglement_position_data.csv

Model	Domain	Position	ΔRCI	Var_Ratio	VRI
GPT-4o-mini	Philosophy	1	0.021	0.692	0.308
GPT-4o-mini	Philosophy	2	0.018	0.745	0.255
GPT-4o-mini	Philosophy	3	0.019	0.812	0.188
...	...	...	...	...	...
Llama 4 Scout	Medical	28	-0.031	5.123	-4.123
Llama 4 Scout	Medical	29	-0.028	6.247	-5.247
Llama 4 Scout	Medical	30	-0.035	7.463	-6.463

Note: Table S1 shows representative rows only. Complete 360-row dataset available in repository.

2.2 Table S2: ESI Classification and Recovery Rates

Model	Mean Var_Ratio	P30 Var_Ratio	ESI (P30)	First Div.	Recovery Rate (%)	Class
Llama 4 Scout	1.610	7.463	0.15	P2	17.9	Divergent
Qwen3 235B	1.334	1.624	1.60	P4	26.9	Divergent
Llama 4 Maverick	1.213	2.644	0.61	P2	35.7	Divergent
Gemini Flash (Med)	1.287	1.441	3.48	P3	44.4	Divergent
DeepSeek V3.1	1.071	1.048	14.02	P2	46.4	Convergent
Ministral 14B	1.080	1.046	21.74	P4	38.5	Divergent
Claude Haiku	1.012	1.028	35.71	P2	35.7	Divergent
Gemini Flash (Phil)	1.120	1.233	4.29	P1	48.3	Divergent
Kimi K2	1.006	0.885	8.70	P1	62.1	Convergent
GPT-4o-mini	0.968	0.980	50.00	P12	50.0	Convergent
GPT-4o	0.950	1.129	7.75	P7	56.5	Convergent
Mistral Small 24B	0.985	1.129	7.75	P4	57.7	Convergent

Table 2: **ESI and recovery classification for all 12 models.** ESI = 1/|1 - Var_Ratio| computed at P30. "First Div." = position where model first entered divergent regime (Var_Ratio > 1.0). "Recovery Rate" = percentage of positions where model returned to convergent after first divergence. Models with recovery < 50% (red class) are consistent with Laban et al.'s "do not recover" claim.

3 Supplementary Methods

3.1 "Lost in Conversation" Test Design

To validate whether our entanglement framework explains Laban et al.'s (2025) "Lost in Conversation" phenomenon, we designed 5 experiments using the existing MCH dataset (360 observations, 12 models, 30 positions):

3.1.1 Experiment 1: Var_Ratio Progression Test

Hypothesis: If "Lost in Conversation" is real, Var_Ratio should increase with conversation length.

Method: For each model, compute linear regression slope of Var_Ratio vs position (P1-P30). Positive slope indicates increasing instability; negative slope indicates increasing stability.

Statistical test: Pearson correlation r and slope significance ($p < 0.05$).

Classification: Compare early (P1-P10 mean) vs late (P21-P30 mean) Var_Ratio. Delta = Late - Early. Positive delta = consistent with "Lost in Conversation."

3.1.2 Experiment 2: Convergent vs Divergent Classification

Hypothesis: Most models are divergent (mean Var_Ratio > 1.0) across all positions.

Method: Compute mean Var_Ratio across all 30 positions per model. Classify as:

- CONVERGENT: mean Var_Ratio < 1.0
- DIVERGENT: mean Var_Ratio > 1.0

Statistical test: Count proportion in each class.

3.1.3 Experiment 3: Recovery Analysis

Hypothesis: Models "get lost and do not recover" (Laban et al. claim).

Method: For each model:

1. Find first position where Var_Ratio > 1.0 (first divergence)
2. Count subsequent positions where Var_Ratio returns to < 1.0 (recovery events)
3. Compute recovery rate = (recovery positions) / (total positions after first divergence)

Classification:

- RECOVERS: recovery rate > 50%
- DOES NOT RECOVER: recovery rate $\leq$ 50%

Prediction: If Laban et al. are correct, most models should NOT recover.

3.1.4 Experiment 4: Domain Comparison

Hypothesis: Domain structure affects divergence tendency.

Method: Compare mean Var_Ratio between Medical (N=8 models) and Philosophy (N=4 models).

Statistical test: Independent t-test or Mann-Whitney U (depending on normality).

3.1.5 Experiment 5: Llama Trajectory Analysis

Hypothesis: Llama models show extreme divergence escalation.
Method: Trace Llama 4 Maverick and Llama 4 Scout Var_Ratio from P1 to P30. Identify:

- Starting stability (P1 Var_Ratio)
- Ending divergence (P30 Var_Ratio)
- Maximum divergence position
- ESI at P30

Extreme divergence threshold: Var_Ratio > 3.0 (empirically derived from distribution).

3.2 Implementation

All experiments implemented in Python using:

- Pandas for data manipulation
- SciPy for statistical tests
- Matplotlib/Seaborn for visualization
- Input data: `entanglement_position_data.csv` (N=360 rows)

Script location: `scripts/analysis/test_lost_in_conversation.py`
Results saved to:

- `analysis/lost_in_conversation_tests.png` (6-panel figure)
- `analysis/lost_in_conversation_summary.csv` (aggregate statistics)
- `analysis/lost_in_conversation_progression.csv` (model-level trajectories)

4 Supplementary Results

4.1 Detailed Progression Statistics

Table S3 shows detailed progression analysis for all 12 models:

Model	Correlation r	Slope (per position)	Early Var_Ratio	Late Var_Ratio
Llama 4 Scout	0.380*	+0.053*	1.408	2.171
Qwen3 235B	0.394*	+0.029*	1.058	1.624
GPT-4o	0.426*	+0.020*	0.734	1.129
Gemini Flash (Med)	0.211	+0.018	1.209	1.441
Mistral Small 24B	0.331	+0.015	0.853	1.129
GPT-4o-mini	0.275	+0.015	0.692	0.980
Llama 4 Maverick	0.217	+0.013	1.151	1.335
Gemini Flash (Phil)	0.200	+0.010	0.992	1.233
Claude Haiku	0.155	+0.004	0.964	1.028
Ministral 14B	0.064	+0.002	1.026	1.046
DeepSeek V3.1	-0.036	-0.002	1.071	1.048
Kimi K2	-0.284	-0.012	1.137	0.885

Table 3: **Var_Ratio progression statistics.** Correlation r and slope computed from linear regression (Var_Ratio Position). * indicates $p < 0.05$ (significant trend). Early = mean(P1-P10), Late = mean(P21-P30). 10/12 models show positive slopes (increasing instability).

4.2 Recovery Event Analysis

Recovery events plotted in Figure S5 show temporal pattern of divergence/convergence transitions:

Recovery events are shown in Figure S4 (middle-right panel), which plots recovery rates after first divergence for all 12 models. Llama models (bottom rows) show persistent divergent behavior after P2–P4.

5 Supplementary Discussion

5.1 Laban et al. (2025) Comparison

Laban et al. (2025)	MCH Framework (This Study)
200,000+ conversations, 15 LLMs	360 observations, 12 LLMs
Performance drops 39%	ΔRCI becomes negative (divergent class)
Unreliability increases 112%	Var_Ratio increases up to 7.46$\times$
"Models get lost"	Divergent entanglement (Var_Ratio > 1)
"Do not recover"	75% recovery rate < 50%
No mechanistic explanation	Variance reduction framework (VRI/ESI)
No predictive tool	ESI < 1.0 predicts instability

Table 4: **Comparison between Laban et al.'s empirical findings and our mechanistic framework.** Our study provides quantitative explanation for their qualitative observations and adds predictive capability.

5.2 Clinical Implications

The "Lost in Conversation" phenomenon has direct clinical safety implications:

Scenario: A medical LLM assists with STEMI diagnosis over 29-exchange case history (as in MCH Medical domain).

Risk: At P30, Llama 4 Scout has Var_Ratio = 7.46, ESI = 0.15. This means:

- Output variance is 7.46$\times$ higher with context than without
- Across 50 identical trials, responses range from 5/16 to 14/16 clinical elements covered
- No hallucinations, but random incompleteness
- Clinician cannot trust output to be stable

Proposed Framework for Future Validation: The ESI metric provides a mechanistic basis for screening protocols, but requires large-scale validation:

- Our study: 360 controlled observations across 12 model-domain runs
- Laban et al. (2025): 200,000+ simulated conversations documenting the phenomenon at scale
- **Gap**: Operational deployment thresholds (e.g., ESI < 1.0 as exclusion criterion) require validation across thousands of models and diverse real-world contexts
- **Next steps**: Large-scale empirical studies testing whether ESI values consistently predict instability across deployment scenarios
- Our contribution: Mechanistic explanation enabling hypothesis-driven threshold development

5.3 Architectural Insights

Recovery analysis reveals potential architectural differences:

Models that recover (GPT-4o, Mistral Small 24B, Kimi K2):

- Larger parameter counts (24B-1T)
- May have better context windowing strategies
- Possibly more robust attention mechanisms

Models that do NOT recover (Llama 4 Scout/Maverick, Qwen3 235B):

- Mixture-of-Experts architectures (potential expert routing instability?)
- Medical domain shows worse recovery than Philosophy
- Early divergence (P2-P4) predicts poor recovery

Hypothesis for future testing: MoE routing may introduce stochasticity that amplifies with conversation length, leading to divergent entanglement. Dense transformers may maintain more stable attention patterns.

5.4 Limitations of "Lost in Conversation" Analysis

1. **Sample size**: 12 models is small compared to Laban et al.'s 15 LLMs
2. **Domain scope**: Only medical and philosophy tested; Laban et al. used diverse tasks
3. **Recovery threshold**: 50% cutoff is arbitrary; sensitivity analysis needed
4. **Causality**: We show correlation between Var_Ratio and instability, not proof of mechanism
5. **Intervention testing**: We did not test whether context summarization/truncation can prevent divergence