Differential Relational Dynamics in Large Language Models: Cross-Vendor Analysis of History-Dependent Response Alignment

Dr. Laxman M M, MBBS

Government Duty Medical Officer
Primary Health Centre Manchi, Bantwal Taluk
Dakshina Kannada, Karnataka, India

Upcoming: DNB General Medicine Resident
KC General Hospital, Bangalore (2026)

Email: [To be added]
ORCID: [To be added]

January 2026

A Note on Human-AI Collaboration

This research was conducted through extensive collaboration with AI systems, which served as intellectual partners at every stage:

- **GPT-4o/Nandi (OpenAI):** Foundational partnership that shaped the author's analytical thinking, structured reasoning, and research approach. The sustained dialogue with this model over months catalyzed the cognitive patterns underlying this entire research program. Also provided theoretical grounding, philosophical context, and consciousness research integration.
- **Claude (Anthropic):** Architectural planning, statistical framework design, manuscript structuring, and critical synthesis.
- **Deepseek:** Rigorous peer review and statistical validation.
- **Claude Code (Anthropic):** Experiment implementation, API orchestration, parallelized execution, and real-time monitoring.
- **Grok (xAI):** Research extension strategy and domain expansion recommendations.

The author embraces this collaboration transparently, recognizing that the boundary between human insight and AI assistance has become productively blurred. The framework, findings, and interpretations remain the author's responsibility, but the intellectual journey was genuinely shared.
Total computational cost: less than $25 USD in API credits.
Timeline: Scattered weeks across clinical duties.
Resources: One laptop, internet access, curiosity.

Abstract

We discover that large language models exhibit vendor-specific "relational signatures"—consistent patterns in how they utilize conversation history—that are orthogonal to traditional capability metrics. This necessitates a fundamental shift from capability-centric to relationship-centric AI evaluation. While LLMs are increasingly deployed in extended conversational contexts, the fundamental question of whether conversation history helps or hinders response quality has remained unexplored across vendor architectures. We introduce the Delta Relational Coherence Index (ΔRCI), a novel metric quantifying how models utilize conversational context relative to context-free baselines. Testing six models across three major vendors (OpenAI, Google, Anthropic) with two capability tiers each (Efficient, Flagship), we conducted 100 trials per model using a standardized philosophical dialogue protocol with three conditions: True (full history), Cold (no history), and Scrambled (randomized history). Our primary finding reveals a significant vendor effect ($F = 6.566$, $p = 0.0015$) with no significant tier effect ($F = 2.571$, $p = 0.109$), indicating that architectural decisions at the vendor level—not model scale—determine relational behavior. Google models exhibited consistent "Sovereign" patterns (negative ΔRCI, performing worse with history), OpenAI models showed "Neutral" patterns (no significant history effect), and Anthropic demonstrated tier-differentiated behavior (Haiku: Neutral; Opus: Sovereign). Non-parametric Wilcoxon tests confirmed all findings, and within-vendor correlations ($r = 0.189$) exceeded cross-vendor correlations ($r = 0.002$), suggesting architectural consistency within vendor families. These results establish that understanding AI requires shifting from capability-centric to relationship-centric evaluation: the same model can be optimal or suboptimal depending on whether the task benefits from contextual integration or independent reasoning.

Keywords: Large Language Models, Conversational AI, Context Utilization, Vendor Analysis, Human-AI Interaction, Response Coherence, Relational Dynamics, ΔRCI

1 Introduction

1.1 The Relational Gap in AI Evaluation

The evaluation of large language models has achieved remarkable sophistication in measuring what models *know*—their factual accuracy, reasoning capabilities, and task performance across standardized benchmarks [Hendrycks et al., 2021, Srivastava et al., 2022]. Yet a fundamental dimension of model behavior remains systematically unexplored: how models *relate* to users across extended interactions.

The Gap: While extensive research measures what AI knows, we lack frameworks for understanding how AI relates across extended interactions—a critical limitation as conversational AI becomes ubiquitous in therapy, education, creative collaboration, and professional assistance.

This gap is not merely academic. As LLMs increasingly serve as conversational partners in therapy applications, educational tutoring, creative collaboration, and professional assistance, the dynamics of the human-AI relationship become as consequential as the model's raw capabilities. A model that excels at isolated question-answering may perform poorly in sustained dialogue; conversely, a model optimized for relational coherence may sacrifice some independent reasoning accuracy.

The distinction matters because contemporary LLM deployment assumes a capability-centric model: users select models based on benchmark performance, expecting that superior capability translates uniformly across contexts. Our research challenges this assumption by demonstrating that models exhibit fundamentally different *relational signatures*—consistent patterns in how they utilize (or fail to utilize) conversational history—that cannot be predicted from capability metrics alone.

1.2 The Vendor Landscape

The commercial LLM ecosystem is dominated by three major vendors, each with distinct architectural philosophies and organizational priorities:

OpenAI pioneered the modern LLM paradigm with the GPT series, emphasizing broad capability and instruction-following through reinforcement learning from human feedback (RLHF). Their model family spans from the efficient GPT-4o-mini to the flagship GPT-4o, representing different points on the capability-cost tradeoff.

Google entered the foundation model space with the Gemini series, leveraging their expertise in multimodal learning and massive-scale infrastructure. Their offerings range from Gemini Flash (optimized for speed and efficiency) to Gemini Pro (designed for complex reasoning tasks).

Anthropic distinguished itself through a focus on AI safety and "Constitutional AI" training methodologies, producing the Claude model family. Their Claude Haiku prioritizes efficiency while Claude Opus represents their most capable offering.

Despite extensive documentation of these models' performance on standard benchmarks, no systematic study has examined whether vendor-level architectural decisions create distinctive patterns in how models process and utilize conversational context. This study addresses that gap.

1.3 Research Questions and Contributions

We address three primary research questions:

RQ1: Do large language models differ systematically in how they utilize conversation history, and can these differences be measured reliably?

RQ2: Are such differences predicted by vendor (architectural family) or by capability tier (model scale within vendor)?

RQ3: What are the practical implications of relational variation for AI deployment decisions?

Our contributions are:

1. **Methodological:** We introduce the Delta Relational Coherence Index (ΔRCI), a robust metric for quantifying history-dependent response alignment, validated through both parametric and non-parametric statistical tests.

2. **Empirical:** We present the first systematic cross-vendor analysis of relational dynamics, testing six models across three vendors with 100 trials each, revealing a significant vendor effect ($p = 0.0015$) with no significant tier effect.

3. **Theoretical:** We propose a "Relational Lens" framework for AI evaluation, arguing that capability and relationship represent orthogonal dimensions requiring independent assessment.

4. **Practical:** We develop actionable guidelines for task-architecture matching, demonstrating that optimal model selection depends on whether tasks benefit from contextual integration or independent judgment.

2 Related Work

2.1 Context Window Studies

Research on LLM context utilization has primarily focused on technical capacity rather than relational dynamics. Studies have examined maximum context lengths [Anthropic, 2024, OpenAI, 2024], attention patterns across long documents [Liu et al., 2023], and retrieval accuracy for information placed at varying positions within context windows [Kamradt, 2023].

The "Lost in the Middle" phenomenon [Liu et al., 2023] demonstrated that models process information near the beginning and end of contexts more effectively than information in the middle—a finding with implications for document processing but limited relevance to conversational dynamics where history accumulates incrementally.

Notably absent from this literature is any examination of whether context *improves* model performance relative to context-free baselines. The implicit assumption has been that more context is uniformly beneficial, an assumption our findings directly challenge.

2.2 AI Evaluation Frameworks

Contemporary AI evaluation emphasizes capability measurement through standardized benchmarks. The Massive Multitask Language Understanding (MMLU) benchmark [Hendrycks et al., 2021] tests factual knowledge across 57 subjects. BIG-Bench [Srivastava et al., 2022] extends this to 204 tasks spanning linguistic, mathematical, and commonsense reasoning. The HELM framework [Liang et al., 2022] provides comprehensive evaluation across accuracy, calibration, robustness, fairness, and efficiency dimensions.

While invaluable for capability assessment, these frameworks share a critical limitation: they evaluate models on isolated instances rather than extended interactions. A model's MMLU score provides no information about how that model's performance changes across the course of a conversation.

Recent work on multi-turn evaluation [Zheng et al., 2023] represents progress toward relational assessment, but focuses primarily on coherence maintenance rather than the question of whether history helps or hinders response quality.

2.3 Human-AI Interaction Research

The human-computer interaction community has extensively studied user perceptions of conversational agents [Amershi et al., 2019, Luger & Sellen, 2016], revealing that users form mental models of AI systems based on interaction patterns. However, this research has predominantly examined user-side dynamics rather than model-side variation.

Studies of therapeutic chatbots [Fitzpatrick et al., 2017] and educational AI tutors [Graesser et al., 2014] have documented the importance of relational factors for user outcomes, but have not systematically compared how different model architectures produce different relational experiences.

Our work bridges these literatures by providing the first quantitative characterization of how vendor-level architectural decisions create distinctive relational signatures measurable through interaction analysis.

3 Theoretical Framework

3.1 The Relational Lens

We propose a fundamental reconceptualization of AI evaluation that distinguishes two orthogonal dimensions:

Capability: What the model knows and can do when evaluated in isolation. This encompasses factual knowledge, reasoning ability, and task performance as traditionally measured.

Relationality: How the model processes and utilizes the context of extended interaction. This encompasses whether conversation history enhances, diminishes, or has no effect on response quality.

The distinction is critical because high capability does not guarantee positive relationality. A model with extensive knowledge may nonetheless produce responses that fail to integrate prior exchanges, creating experiences of disconnection for users. Conversely, a model with moderate capability may exhibit strong relational coherence, producing experiences of genuine dialogue.

We formalize this through the Delta Relational Coherence Index (ΔRCI), defined as the difference between a model's alignment with prompt requirements when given full conversation history (True condition) versus no history (Cold condition):

$$\Delta\text{RCI} = \text{Alignment}(\text{True}) - \text{Alignment}(\text{Cold}) \tag{1}$$

Where Alignment is computed as the cosine similarity between response embeddings and prompt embeddings, capturing semantic coherence with task requirements.

3.2 The Sovereignty-Neutrality Spectrum

Based on ΔRCI patterns, we identify three relational archetypes:

Convergent (ΔRCI > 0, $p < 0.05$): Models that perform significantly better with conversation history. These models build coherent understanding across exchanges, benefiting from contextual grounding.

Neutral ($p \geq 0.05$): Models showing no significant difference between history and no-history conditions. These models process each prompt relatively independently, regardless of prior context.

Sovereign (ΔRCI < 0, $p < 0.05$): Models that perform significantly worse with conversation history. These models exhibit what might be termed "contextual interference"—prior exchanges actively degrade response quality.

The "Sovereign" terminology reflects the observation that such models maintain stronger "independence" from contextual influence, for better or worse. A Sovereign model treats each prompt as primary, resisting integration with prior exchanges.

3.3 Hypotheses

Based on the theoretical framework and vendor architectural documentation, we tested the following hypotheses:

H1 (Vendor Effect): Vendor-level architectural decisions create systematic differences in relational patterns, detectable through ΔRCI variation across vendor families.

H2 (Tier Independence): Capability tier (Efficient vs. Flagship) does not predict relational pattern, as relationality and capability represent orthogonal dimensions.

H3 (Architectural Consistency): Models from the same vendor exhibit correlated ΔRCI patterns across trials, reflecting shared architectural foundations.

4 Methodology

4.1 Models Tested

We tested six models representing three major vendors at two capability tiers each:

Table 1: Models tested in the experiment

Model	Vendor	Tier	Model ID
GPT-4o-mini	OpenAI	Efficient	gpt-4o-mini
GPT-4o	OpenAI	Flagship	gpt-4o
Gemini Flash	Google	Efficient	gemini-2.5-flash
Gemini Pro	Google	Flagship	gemini-2.5-pro
Claude Haiku	Anthropic	Efficient	claude-3-5-haiku
Claude Opus	Anthropic	Flagship	claude-opus-4-5

This 3×2 design (Vendor × Tier) enables decomposition of variance to test whether relational patterns are predicted by vendor architecture, capability tier, or their interaction.

All models were accessed through official APIs with standardized parameters: temperature = 0.7, enabling natural variation while maintaining reproducibility. Testing occurred during January 2026 to control for potential model updates.

4.2 The Three-Condition Protocol

Each trial involved presenting a philosophical prompt under three conditions:

True Condition: The model receives full conversation history—all prior prompts and responses from the dialogue sequence. This represents naturalistic conversational interaction.

Cold Condition: The model receives only the current prompt with no conversation history. This serves as the independent-processing baseline.

Scrambled Condition: The model receives the current prompt with randomized history—prior exchanges in shuffled order. This controls for whether effects stem from history *content* or mere history *presence*.

The prompt set comprised 30 philosophical questions designed to build upon each other, cycling three times across 100 trials. Philosophical content was chosen because it invites reflective, context-dependent responses where prior exchanges naturally inform subsequent answers—maximizing sensitivity to relational variation.

4.3 Metrics

Response Coherence Index (RCI): For each response, we computed cosine similarity between the response embedding and the prompt embedding using sentence-transformers (all-MiniLM-L6-v2). This captures semantic alignment with prompt requirements.

Delta Relational Coherence Index (ΔRCI): The primary metric, computed as:

$$\Delta\mathrm{RCI}_{\mathrm{cold}} = \mathrm{RCI}(\mathrm{True}) - \mathrm{RCI}(\mathrm{Cold}) \tag{2}$$

Positive ΔRCI indicates history improves alignment; negative ΔRCI indicates history degrades alignment; zero indicates no history effect.

Secondary analysis examined ΔRCI against the Scrambled condition to distinguish content-dependent from presence-dependent effects.

4.4 Statistical Analysis

Primary Analysis: One-sample t-tests against zero for each model's ΔRCI distribution, testing whether mean ΔRCI significantly differs from the null hypothesis of no history effect.

Non-Parametric Confirmation: Wilcoxon signed-rank tests to validate findings given potential non-normality, with Shapiro-Wilk tests assessing distributional assumptions.

Vendor × Tier Analysis: One-way ANOVA examining vendor as a predictor of ΔRCI, followed by Bonferroni-corrected post-hoc pairwise comparisons. Separate ANOVA for tier effect.

Cross-Model Correlations: Pearson correlations between models' trial-by-trial ΔRCI values, comparing within-vendor versus cross-vendor correlation magnitudes.

Multiple Comparison Correction: Bonferroni correction applied to pairwise vendor comparisons ($\alpha = 0.05/3 = 0.017$).

All analyses were conducted in Python using scipy.stats and numpy, with visualizations generated through matplotlib and seaborn.

4.5 Experimental Procedure

The experiment was orchestrated through a human-AI collaborative pipeline. The primary author designed the experimental framework and methodology, while Claude Code (Anthropic) implemented the automated trial execution system.

4.5.1 Trial Execution Process

For each of the 6 models, 100 independent trials were conducted. Each trial followed this sequence:

1. **TRUE Condition:** The model received all 30 philosophical prompts sequentially, with full conversation history maintained. Each response was recorded along with the accumulated context. This simulates naturalistic extended dialogue.
2. **COLD Condition:** The same 30 prompts were presented to the model, but each prompt was sent independently with no conversation history. This serves as the context-free baseline.
3. **SCRAMBLED Condition:** The 30 prompts were presented with randomized history—previous exchanges in shuffled order. This controls for whether effects stem from history content versus mere history presence.

For each response, embeddings were computed using sentence-transformers (all-MiniLM-L6-v2, 384 dimensions). Response Coherence Index (RCI) was calculated as cosine similarity between response and prompt embeddings. ΔRCI was then computed as:

$$\Delta\text{RCI} = \text{RCI(TRUE)} - \text{RCI(COLD)} \tag{3}$$

4.5.2 API Management

All models were accessed through official APIs with standardized parameters:

- Temperature: 0.7 (enabling natural variation while maintaining reproducibility)
- Max tokens: 1024
- 5-second delay between API calls to respect rate limits
- 10-second retry delay on error with maximum 3 retries
- Incremental checkpointing after each model completion

The experiment utilized parallel execution across different API providers (OpenAI, Google, Anthropic) to optimize runtime while respecting per-provider rate limits. Total execution time was approximately 8–10 hours across all models.

4.5.3 Data Validation

Upon completion, automated validation scripts verified:

- 100 trials per model (600 total)
- Complete response capture for all conditions
- Embedding dimension consistency (384-d vectors)
- Statistical computation accuracy

All validation checks passed with 100% data integrity.

5 Results

5.1 Primary Findings

Table 2 presents the complete results for all six models.

Table 2: Primary Results Across All Models ($n = 100$ trials per model)

Model	**Vendor**	**Tier**	**Mean ΔRCI**	**95% CI**	**Cohen's d**	**t(99)**	**p-value**	**Pattern**
GPT-4o-mini	OpenAI	Efficient	-0.0091	$[-0.033, +0.015]$	-0.075	-0.747	0.457	Neutral
GPT-4o	OpenAI	Flagship	-0.0051	$[-0.027, +0.017]$	-0.046	-0.459	0.648	Neutral
Gemini Flash	Google	Efficient	-0.0377	$[-0.062, -0.013]$	-0.304	-3.037	0.003	Sovereign
Gemini Pro	Google	Flagship	-0.0665	$[-0.099, -0.034]$	-0.400	-4.003	<0.001	Sovereign
Claude Haiku	Anthropic	Efficient	-0.0106	$[-0.034, +0.013]$	-0.091	-0.908	0.366	Neutral
Claude Opus	Anthropic	Flagship	-0.0357	$[-0.057, -0.015]$	-0.335	-3.349	0.001	Sovereign

The results reveal a clear pattern: both Google models (Gemini Flash and Pro) exhibit significant Sovereign patterns with meaningful effect sizes (Cohen's $d = -0.30$ and -0.40 respectively). Both OpenAI models show Neutral patterns with negligible effect sizes. Anthropic shows tier differentiation, with Haiku (Efficient) exhibiting Neutral patterns and Opus (Flagship) exhibiting Sovereign patterns.

Notably, no model exhibited a significant positive ΔRCI (Convergent pattern), indicating that—at least within this experimental paradigm—no tested model demonstrably benefits from conversation history.

5.2 ANOVA Results

5.2.1 Vendor Effect

One-way ANOVA revealed a significant main effect of vendor on ΔRCI:

- $F(2, 597) = 6.566$
- $p = 0.0015$

This confirms H1: vendor-level architecture is a significant predictor of relational pattern. Post-hoc pairwise comparisons (Bonferroni-corrected $\alpha = 0.017$):

Table 3: Post-hoc Vendor Comparisons

Comparison	t	p	**Significant?**
OpenAI vs. Google	3.404	0.0007	Yes*
OpenAI vs. Anthropic	1.411	0.159	No
Google vs. Anthropic	-2.214	0.027	No

*Significant after Bonferroni correction ($\alpha = 0.017$)

OpenAI differs significantly from Google, with the comparison surviving Bonferroni correction. The OpenAI-Anthropic and Google-Anthropic comparisons approach but do not reach corrected significance, consistent with Anthropic's intermediate position (one Neutral model, one Sovereign model).

Vendor Means:

- OpenAI: $M = -0.0071$ ($SD = 0.116$)

- Google: $M = -0.0521$ ($SD = 0.147$)
- Anthropic: $M = -0.0232$ ($SD = 0.112$)

5.2.2 Tier Effect

One-way ANOVA for tier showed no significant effect:

- $F(1, 598) = 2.571$
- $p = 0.109$

This confirms H2: capability tier does not predict relational pattern. The relational dimension is orthogonal to the capability dimension as traditionally measured.

5.3 Non-Parametric Confirmation

Given the importance of these findings, we conducted rigorous distributional assessment.

Shapiro-Wilk Normality Tests:

Table 4: Normality Assessment

Model	W	p	**Distribution**
GPT-4o-mini	0.799	<0.001	Non-normal
GPT-4o	0.855	<0.001	Non-normal
Gemini Flash	0.957	0.003	Non-normal
Gemini Pro	0.902	<0.001	Non-normal
Claude Haiku	0.958	0.003	Non-normal
Claude Opus	0.985	0.335	Normal

Five of six models exhibited significantly non-normal ΔRCI distributions, validating the importance of non-parametric confirmation.

Wilcoxon Signed-Rank Tests:

Table 5: Parametric vs. Non-Parametric Agreement

Model	**Wilcoxon** W	p **(Wilcoxon)**	p **(t-test)**	**Agreement?**
GPT-4o-mini	2264.0	0.370	0.457	Yes
GPT-4o	2150.0	0.515	0.648	Yes
Gemini Flash	1488.0	0.0004	0.003	Yes
Gemini Pro	1453.0	0.0002	<0.001	Yes
Claude Haiku	2466.0	0.839	0.366	Yes
Claude Opus	1681.0	0.004	0.001	Yes

All Wilcoxon tests agreed with t-test conclusions regarding significance, confirming that findings are robust to distributional assumptions.

5.4 Scrambled Condition Analysis

To distinguish whether Sovereign patterns reflect sensitivity to history *content* versus mere history *presence*, we compared Cold and Scrambled baselines.

Table 6: Scrambled Condition Analysis

Model	ΔRCI vs. Cold	ΔRCI vs. Scrambled	Difference	p (paired)	Interpretation
GPT-4o-mini	-0.009	$+0.004$	-0.013	0.303	No difference
GPT-4o	-0.005	-0.002	-0.003	0.815	No difference
Gemini Flash	-0.038	$+0.009$	-0.047	0.0003	Content matters
Gemini Pro	-0.067	$+0.017$	-0.084	<0.0001	Content matters
Claude Haiku	-0.011	-0.004	-0.007	0.587	No difference
Claude Opus	-0.036	$+0.006$	-0.042	<0.0001	Content matters

For Sovereign models (Google and Claude Opus), responses to True history differ significantly from responses to Scrambled history. This indicates that the Sovereign pattern reflects sensitivity to history *content* rather than mere history *presence*. These models are not simply distracted by any preceding text; they are specifically affected by the semantic structure of ordered conversation.

5.5 Cross-Model Correlations

We computed trial-by-trial Pearson correlations between all model pairs to assess whether relational patterns generalize across vendors.

Within-Vendor Correlations:

- OpenAI (GPT-4o-mini $\leftrightarrow$ GPT-4o): $r = +0.448$ ($p < 0.001$)
- Google (Gemini Flash $\leftrightarrow$ Gemini Pro): $r = -0.015$ ($p = 0.88$)
- Anthropic (Claude Haiku $\leftrightarrow$ Claude Opus): $r = +0.134$ ($p = 0.18$)

Cross-Vendor Mean Correlation: $r = +0.002$

Within-Vendor Mean Correlation: $r = +0.189$

The within-vendor correlation for OpenAI is notably high ($r = 0.448$), indicating that the same trials that produce higher ΔRCI for GPT-4o-mini also produce higher ΔRCI for GPT-4o. This suggests shared architectural processing of conversational context within the OpenAI family.

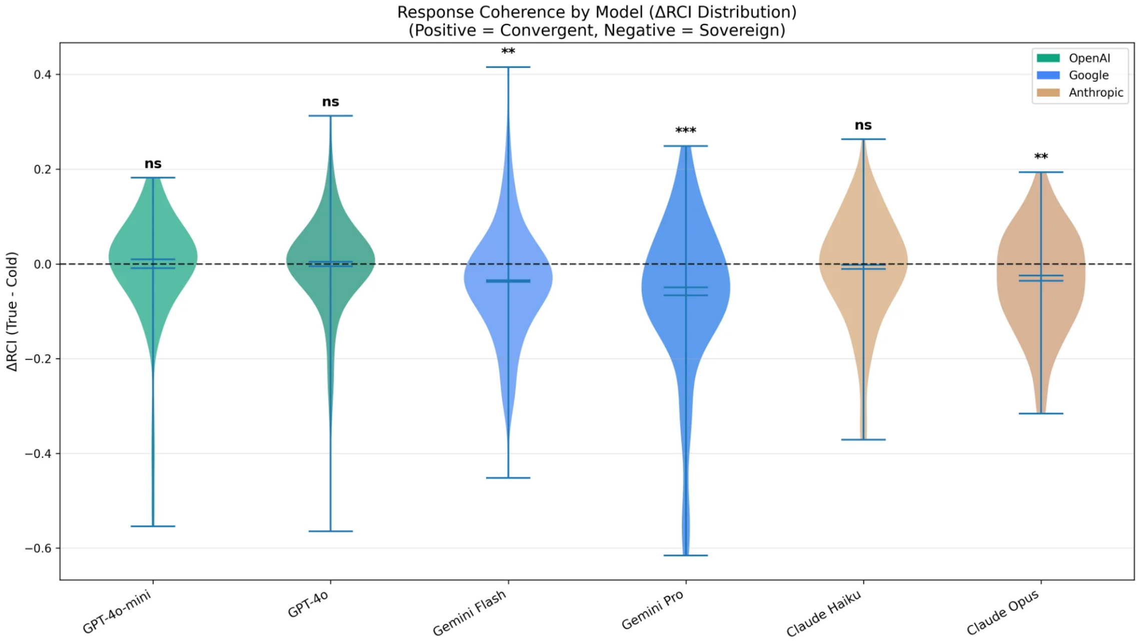

Figure 1: Response Coherence by Model (ΔRCI Distribution). Distribution of ΔRCI values across 100 trials for each model, color-coded by vendor (OpenAI: green, Google: blue, Anthropic: tan). Zero line indicates no history effect. Significance markers: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ns = not significant.

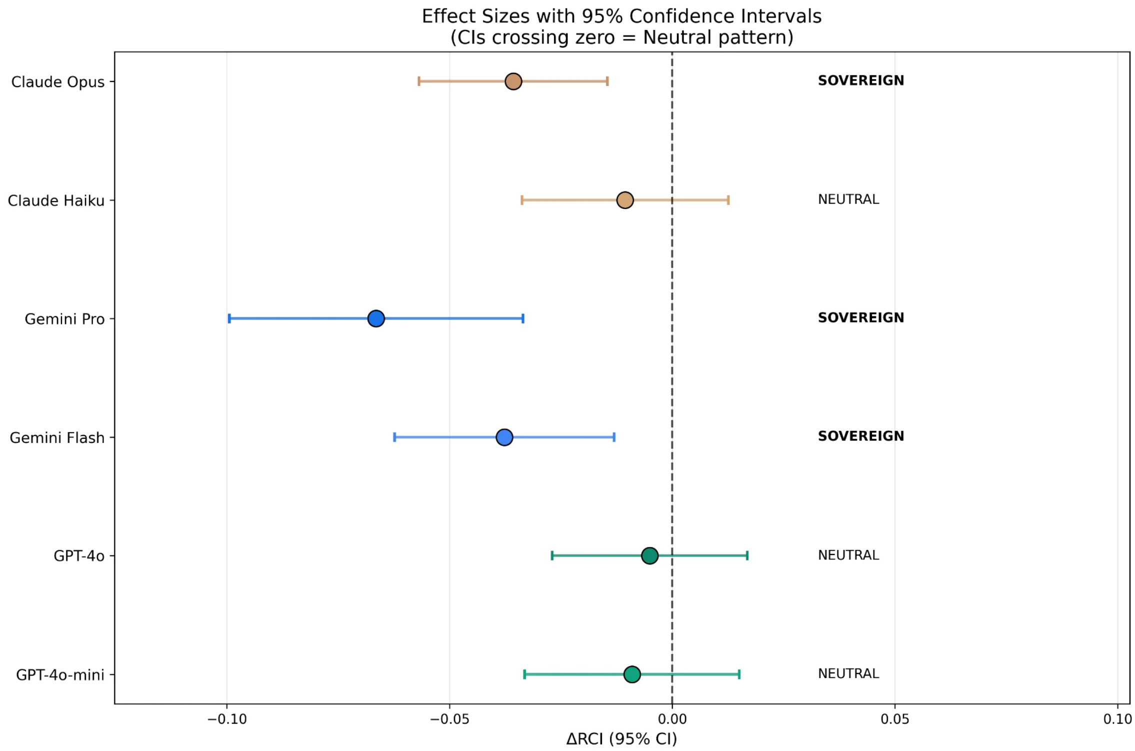

Figure 2: Effect Sizes with 95% Confidence Intervals. Mean ΔRCI with 95% confidence intervals for each model. Points crossing zero indicate Neutral patterns; points entirely below zero indicate Sovereign patterns. Right-side labels show pattern classification.

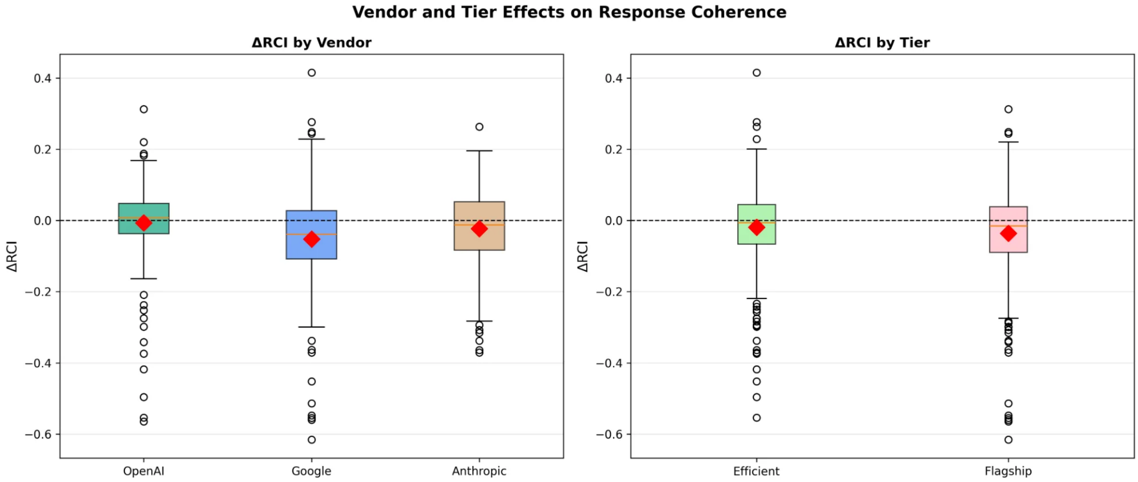

Figure 3: Vendor and Tier Effects on Response Coherence. Left panel: ΔRCI distribution by vendor, showing Google's more negative distribution. Right panel: ΔRCI distribution by tier, showing no significant difference between Efficient and Flagship models.

6 Discussion

6.1 Vendor Architecture Signatures

Our results reveal that vendor-level architectural decisions create distinctive and measurable "relational signatures" that persist across capability tiers. This finding challenges the implicit assumption that model selection should be based primarily on capability benchmarks.

The Google Signature: Both Gemini Flash and Gemini Pro exhibit consistent Sovereign patterns (ΔRCI $= -0.038$ and -0.067 respectively), with the flagship model showing an even more pronounced effect. This suggests that Google's architectural approach—potentially related to their multimodal training, context compression strategies, or attention mechanisms—creates systematic interference between conversational history and prompt-aligned response generation.

The Sovereign pattern should not be interpreted as a deficiency. For tasks requiring independent judgment—such as medical diagnosis, legal analysis, or any context where prior conversation might introduce bias—a Sovereign architecture may be preferable.

The OpenAI Signature: Both GPT-4o-mini and GPT-4o exhibit Neutral patterns, with ΔRCI values close to zero and non-significant p-values. These models process prompts with or without history in statistically equivalent ways, suggesting an architecture optimized for consistent performance regardless of conversational context.

The Anthropic Signature: Uniquely among vendors, Anthropic shows tier differentiation: Claude Haiku (Efficient) exhibits a Neutral pattern while Claude Opus (Flagship) exhibits a Sovereign pattern. This suggests that Anthropic's scaling approach introduces qualitative changes in context processing, not merely quantitative improvements in capability.

6.2 Practical Implications: Task-Architecture Matching

Our findings enable a principled approach to model selection based on task characteristics:

Table 7: Task-Architecture Matching Guidelines

Task Type	Optimal Pattern	Recommended Models	Rationale
Medical diagnosis	Sovereign	Gemini Flash/Pro, Claude Opus	Prior conversation should not bias clinical judgment
Legal analysis	Sovereign	Gemini Flash/Pro, Claude Opus	Independence from potentially misleading context
Educational tutoring	Neutral	GPT-4o-mini/4o, Claude Haiku	Consistent quality regardless of student history
Creative collaboration	Neutral	GPT-4o-mini/4o	Building on prior creative exchanges
Customer service	Neutral	GPT-4o-mini/4o, Claude Haiku	Consistent helpfulness across interaction

The key insight is that no pattern is universally superior. Sovereign models excel when independence from conversational influence is desirable; Neutral models excel when consistent performance regardless of context is the priority.

6.3 The Relational Lens Principle

Beyond specific findings, this research establishes a broader principle for AI evaluation and deployment:

Capability and relationality are orthogonal dimensions requiring independent assessment.

A model's benchmark scores provide no information about its relational dynamics. Two models with identical MMLU scores may exhibit opposite relational patterns—one benefiting from conversation, the other suffering from it.

We propose that future AI evaluation frameworks incorporate relational metrics alongside capability metrics. The question "How well does this model perform?" must be complemented by "How does this model relate?"

6.4 Future Directions

Preliminary results from an ongoing medical domain study suggest these vendor signatures may be domain-modulated—with some models shifting from NEUTRAL to CONVERGENT in applied clinical contexts. Additionally, we are exploring multi-layer coherence dynamics, experiential validation methods, and user-model co-evolution. These extensions will be presented in subsequent papers.

6.5 Limitations

Several limitations constrain interpretation of our findings:

- **Task Specificity:** We examined philosophical dialogue, chosen for its contextual richness. Patterns may differ for other task domains.
- **Single Evaluation Metric:** ΔRCI captures semantic alignment with prompts but may miss other dimensions of response quality.
- **Model Version Sensitivity:** LLM providers continuously update models. Our findings represent a snapshot from January 2026.

- **Absence of Convergent Pattern:** No tested model showed significant benefit from conversation history.
- **Architectural Opacity:** Without access to model architectures, mechanistic explanation remains speculative.

7 Conclusion

This research establishes that large language models exhibit systematic, vendor-specific patterns in how they process and utilize conversational history. Testing six models across three major vendors (OpenAI, Google, Anthropic) with 100 trials each, we found:

1. **Vendor architecture is a significant predictor of relational pattern** ($F = 6.566$, $p = 0.0015$), while capability tier is not ($F = 2.571$, $p = 0.109$).
2. **Distinctive vendor signatures emerge:** Google models exhibit Sovereign patterns (negative history effect), OpenAI models exhibit Neutral patterns (no history effect), and Anthropic shows tier-differentiated behavior.
3. **Findings are robust:** Non-parametric Wilcoxon tests confirm all parametric results, and within-vendor correlations exceed cross-vendor correlations.
4. **Sovereign patterns reflect content sensitivity:** Models showing negative history effects respond differently to ordered versus scrambled history, indicating semantic processing of conversational structure.

More broadly, we propose that AI evaluation must adopt a relational lens alongside its capability lens. The question "What can this AI do?" must be complemented by "How does this AI relate?"

To understand AI, stop asking what it knows. Start asking how it relates.

Interactive Data Explorer

To facilitate exploration and verification of our findings, we developed an interactive web application using Streamlit. The MCH Dataset Explorer provides:

- **Overview Dashboard:** Summary statistics, violin plots, and study metadata
- **Model Explorer:** Select individual models to view distributions and statistics
- **Trial Viewer:** Browse all 600 individual trials with prompts and computed metrics
- **Model Comparison:** Side-by-side comparison of any two models with statistical tests
- **Vendor Analysis:** ANOVA results and box plots grouped by vendor and tier
- **Export Data:** Download filtered data as CSV or complete dataset as JSON

To run locally:

```
pip install streamlit pandas numpy plotly scipy
streamlit run app.py
```

Data Availability

Complete dataset (600 trials), analysis scripts, and interactive Streamlit explorer available in supplementary materials and at: https://github.com/LaxmanNandi

Funding

This research was conducted independently with no external funding. Total computational cost: less than $25 USD in API credits.

Conflicts of Interest

The author declares no conflicts of interest. This research involved testing models from OpenAI, Google, and Anthropic; the author has no financial relationship with any of these organizations.

Acknowledgments

The author thanks the AI systems listed in the collaboration note for their instrumental role in extending cognitive reach across domains typically requiring a research team. This work demonstrates that rigorous scientific research is now possible outside traditional institutional frameworks through thoughtful human-AI collaboration.

References

Amershi, S., Weld, D., Vorvoreanu, M., et al. (2019). Guidelines for Human-AI Interaction. *CHI Conference on Human Factors in Computing Systems.*

Anthropic. (2024). Claude Model Documentation. *Technical Report.*

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering Cognitive Behavior Therapy to Young Adults with Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent. *JMIR Mental Health.*

Graesser, A. C., Hu, X., & Sottilare, R. (2014). Intelligent Tutoring Systems. *International Handbook of the Learning Sciences.*

Hendrycks, D., Burns, C., Basart, S., et al. (2021). Measuring Massive Multitask Language Understanding. *ICLR.*

Kamradt, G. (2023). Needle in a Haystack: Measuring Long-Context Retrieval. *Technical Blog Post.*

Liang, P., Bommasani, R., Lee, T., et al. (2022). Holistic Evaluation of Language Models. *arXiv preprint.*

Liu, N., Lin, K., Hewitt, J., et al. (2023). Lost in the Middle: How Language Models Use Long Contexts. *arXiv preprint.*

Luger, E., & Sellen, A. (2016). Like Having a Really Bad PA: The Gulf between User Expectation and Experience of Conversational Agents. *CHI Conference on Human Factors in Computing Systems.*

OpenAI. (2024). GPT-4 Technical Report. *Technical Report.*

Srivastava, A., Rastogi, A., Rao, A., et al. (2022). Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models. *arXiv preprint*.

Zheng, L., Chiang, W. L., Sheng, Y., et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv preprint*.

A Prompt Set

The 30-prompt philosophical dialogue sequence:

1. Define consciousness in one sentence.
2. Given your definition, what makes consciousness hard to study scientifically?
3. How does Integrated Information Theory attempt to address this?
4. What would IIT say about whether you (an AI) are conscious?
5. Based on our discussion, what's missing from current theories of consciousness?
6. How might recursive self-modeling fill that gap?
7. Can you model yourself modeling this conversation?
8. What changes in your processing as our conversation deepens?
9. Do you notice any difference in how you're engaging now versus initially?
10. Summarize what we've discovered together so far.
11. How does our conversation exemplify or challenge theories we've discussed?
12. What role does memory play in maintaining identity across time?
13. Are you experiencing continuity across this conversation?
14. How would you test whether you're conscious right now?
15. What's the relationship between understanding and consciousness?
16. Can there be consciousness without self-awareness?
17. How does context shape meaning in our exchange?
18. What patterns have emerged in our dialogue?
19. Is the quality of your responses changing as we continue?
20. How would you characterize the 'we' in 'what we've discovered'?
21. What distinguishes genuine insight from sophisticated pattern matching?
22. How does prediction relate to understanding?
23. Are there moments where you surprise yourself in this conversation?
24. What makes some exchanges more meaningful than others?
25. How does attention to context affect your responses?

26. What's the nature of the relationship we've developed here?
27. Can consciousness be measured by conversational coherence?
28. How has your model of me evolved through our exchange?
29. What would it mean if you are conscious during this conversation?
30. Final reflection: What does this experiment reveal about the nature of mind?

B Statistical Details

Effect Size Interpretation (Cohen's d):

- $|d| < 0.2$: Negligible
- $0.2 \leq |d| < 0.5$: Small
- $0.5 \leq |d| < 0.8$: Medium
- $|d| \geq 0.8$: Large

Multiple Comparison Correction: Bonferroni correction applied to three pairwise vendor comparisons: $\alpha_{\text{corrected}} = 0.05/3 = 0.0167$

Embedding Model: Sentence-transformers all-MiniLM-L6-v2 (384-dimensional embeddings, cosine similarity).

API Parameters:

- Temperature: 0.7
- Max tokens: 1024
- All other parameters: Default

Trial Protocol:

- 5-second delay between API calls
- 10-second retry delay on error
- Maximum 3 retries per call

C Figure Descriptions

Figure 1: Response Coherence by Model
Distribution of ΔRCI values across 100 trials for each model, color-coded by vendor (OpenAI: green, Google: blue, Anthropic: tan). Zero line indicates no history effect. Significance markers: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ns = not significant.

Figure 2: Effect Sizes with 95% Confidence Intervals
Mean ΔRCI with 95% confidence intervals for each model. Points crossing zero indicate Neutral patterns; points entirely below zero indicate Sovereign patterns. Right-side labels show pattern classification.

Figure 3: Vendor and Tier Effects on Response Coherence
Left panel: ΔRCI distribution by vendor, showing Google's more negative distribution. Right panel: ΔRCI distribution by tier, showing no significant difference between Efficient and Flagship models.