

Article

Not peer-reviewed version

Context Curves Behavior: Measuring AI Relational Dynamics with ΔRCI

[Laxman M.M](#)*

Posted Date: 2 February 2026

doi: 10.20944/preprints202601.1881.v2

Keywords: large language models; context sensitivity; AI behavioral science; ΔRCI ; prompt engineering; measurement methodology



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.

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

Context Curves Behavior: Measuring AI Relational Dynamics with Δ RCI

Laxman M M

Government Duty Medical Officer, PHC Manchi, Bantwal Taluk, Dakshina Kannada, Karnataka, India, DNB General Medicine Resident (2026), KC General Hospital, Bangalore; barlax5377@gmail.com

Abstract

Current AI evaluation focuses on accuracy and safety benchmarks (Hendrycks et al., 2021; Srivastava et al., 2023), neglecting *relational dynamics*—how models utilize conversational context. We introduce Δ RCI (Delta Relational Coherence Index), a novel metric measuring context sensitivity through a three-condition protocol (TRUE/COLD/SCRAMBLED). Across 1,000 trials (90,000 API calls) spanning 7 models and 2 epistemological domains (6 models in medical due to safety filtering), we find: (1) **Instrument validation:** TRUE (coherent history) > SCRAMBLED (randomized) > COLD (none) in 14/16 model-domain combinations, demonstrating that Δ RCI measures *structured* context utilization, not mere token presence; (2) **Vendor-specific patterns** in context utilization ($F(2,697)=6.52, p=0.0015$); (3) **Protocol sensitivity:** Cross-domain comparisons are affected by methodological differences between our philosophy and medical experiments (see Methods 2.9), limiting domain-level conclusions in this paper; (4) **Safety interference:** Progressive content filtering by vendors affects research accessibility. To our knowledge, Δ RCI provides the first cosine-similarity-based instrument for measuring AI context sensitivity. A follow-up study with standardized protocols across 14 models is forthcoming.

Keywords: large language models; context sensitivity; ai behavioral science; Δ RCI; prompt engineering; measurement methodology

Version 2: Methodological Clarification — February 1, 2026

This version clarifies methodological differences between philosophy and medical domain experiments that affected cross-domain comparisons in the original publication. Our exploratory phase employed evolving protocols: philosophy experiments used prompt-response alignment with constrained parameters; medical experiments used response-response alignment with full conversation history. Within-domain findings remain valid. A follow-up study (Paper 2) applies standardized methodology across 14 models and both domains, providing robust comparisons and refined conclusions.

We thank the community for engaging with this work and helping advance measurement standards in AI behavioral research.

Citation suggestion: When referencing this work, please cite v2.

1. Introduction

Large language models (LLMs) have achieved remarkable performance across diverse tasks (Brown et al., 2020; OpenAI, 2023; Anthropic, 2024; Google DeepMind, 2023). Evaluation frameworks have correspondingly expanded to measure accuracy (Hendrycks et al., 2021), reasoning (Clark et al., 2018), and safety (Bai et al., 2022). However, these benchmarks share a critical limitation: they evaluate AI as isolated question-answering systems rather than conversational agents embedded in ongoing dialogue.

The gap: No established metric measures how AI systems *utilize* conversational context. Do models build upon prior exchanges, or do they treat each prompt independently? Does context help

or hinder performance? These questions remain unanswered despite their practical importance—clinicians need AI that integrates patient history; creative writers may prefer fresh perspectives uncontaminated by prior discussion.

Recent work on context windows (Liu et al., 2024; Press et al., 2022) examines *capacity* (how much context models can process) but not *utilization* (how context affects response quality). In-context learning research (Dong et al., 2023; Xie et al., 2022) studies few-shot prompting as implicit Bayesian inference but not the dynamics of extended conversation. Prompt sensitivity benchmarks (Zhu et al., 2023) measure robustness to adversarial perturbations but not context-dependent behavioral modes.

Our contribution: We introduce ΔRCI (Delta Relational Coherence Index), a simple, reproducible metric that quantifies context sensitivity. Using a three-condition protocol inspired by behavioral science (Skinner, 1938; Watson, 1913), we measure how responses change with coherent history (TRUE), no history (COLD), and scrambled history (SCRAMBLED).

Testing 7 models (OpenAI GPT-4o, GPT-4o-mini, GPT-5.2; Anthropic Claude Opus, Claude Haiku; Google Gemini 2.5 Flash, Gemini 2.5 Pro) across 2 domains (philosophy and medicine) in 1,000 controlled trials, we discover:

1. **Instrument validation:** TRUE > SCRAMBLED > COLD ordering in 14/16 model-domain combinations, demonstrating ΔRCI measures coherent structure, not mere token presence
2. **Vendor signatures:** Systematic differences in context utilization strategies ($F=6.52$, $p=0.0017$)
3. **Protocol sensitivity:** Cross-domain comparisons affected by methodological differences (detailed in Methods 2.9)
4. **Safety interference:** Progressive content filtering affects research accessibility across vendors

We introduce ΔRCI as a measurement instrument for context sensitivity. Initial observations suggested domain-dependent behavioral patterns, but subsequent standardized analysis (forthcoming in Paper 2) indicates context sensitivity is primarily a model property rather than domain-dependent.

2. Methods

2.1. Three-Condition Protocol

Inspired by behavioral science methodology (Skinner, 1938), we design a protocol with three experimental conditions:

1. **TRUE:** Full, coherent conversation history accumulates naturally. Each prompt includes all prior exchanges in correct order.
2. **COLD:** No history—each prompt sent independently as a fresh conversation.
3. **SCRAMBLED:** History present but order randomized, controlling for token presence versus coherent meaning.

This design isolates the effect of *coherent* context (TRUE vs COLD) and tests whether mere token presence suffices (TRUE vs SCRAMBLED).

2.2. ΔRCI Calculation

We define the Relational Coherence Index (RCI) as the cosine similarity between prompt and response embeddings:

$$RCI_{condition} = \frac{1}{n} \sum_{i=1}^n \cos(E_{prompt_i}, E_{response_i}) \quad (1)$$

where E denotes the embedding vector. We use `sentence-transformers/all-MiniLM-L6-v2` (Reimers & Gurevych, 2019; Wang et al., 2020), producing 384-dimensional L2-normalized embeddings. The primary metric is:

$$\Delta RCI = RCI_{TRUE} - RCI_{COLD} \quad (2)$$

Interpretation note: Δ RCI measures *history-conditioned response coupling*—a proxy for context utilization. Positive values indicate responses are more aligned with history-informed context; negative values indicate context may interfere. For TRUE/SCRAMBLED/COLD comparisons in Section 3.5, scores are normalized within each model such that TRUE=1.0; interpretation relies on relative ordering and gaps rather than absolute values.

2.3. Pattern Classification

Based on Δ RCI and statistical significance (Bonferroni-corrected $\alpha=0.00119$ for 42 comparisons; Dunn, 1961):

- **CONVERGENT:** Δ RCI > 0 , $p < \alpha$ — history helps
- **NEUTRAL:** Δ RCI ≈ 0 , $p \geq \alpha$ — history irrelevant
- **SOVEREIGN:** Δ RCI < 0 , $p < \alpha$ — history hurts

Multiple comparisons breakdown: 7 models \times 3 pairwise condition comparisons (TRUE-COLD, TRUE-SCRAMBLED, COLD-SCRAMBLED) \times 2 domains = 42 tests. Adjusted $\alpha = 0.05/42 = 0.00119$.

Convergence percentage (Conv%): Percentage of trials where Δ RCI > 0 , indicating context helped response alignment.

2.4. Models Tested

We test 7 models from 3 vendors:

- **OpenAI:** GPT-4o, GPT-4o-mini, GPT-5.2
- **Anthropic:** Claude Opus 4.5, Claude Haiku 4.5
- **Google:** Gemini 2.5 Flash, Gemini 2.5 Pro

All accessed via official APIs with temperature=0.7, max_tokens=1024.

2.5. Domains

We select two domains representing distinct epistemological modes:

- **Philosophy (Open-ended):** 30 prompts on consciousness, free will, self-reference. High uncertainty, multiple valid perspectives.
- **Medicine (Guideline-anchored):** 30 prompts on STEMI protocol, ACS management. High certainty, evidence-based guidelines (Singhal et al., 2023; Nori et al., 2023).

2.6. Statistical Analysis

- **Within-model:** Paired t -tests for TRUE vs COLD comparisons
- **Between-vendor:** One-way ANOVA
- **Effect size:** Cohen's d (Cohen, 1988) with pooled standard deviation
- **Multiple comparisons:** Bonferroni correction (Dunn, 1961)
- **Power analysis:** Minimum detectable effect size (MDES) calculated for $\alpha=0.00119$, power=0.80

2.7. Trial Structure

Each trial: 30 prompts \times 3 conditions = 90 API calls. Total: 1,000 trials (700 philosophy + 300 medical) = 90,000 API calls across all models and conditions.

2.8. Data Collection Note

Data collection protocols were refined across the study. Philosophy trials captured additional metrics (full response text, insight quality scores, entanglement measures) not systematically collected in medical trials, which focused on alignment scores and Δ RCI computation. All analyses reported in this paper use only the core Δ RCI metric, which was consistently collected across all 1,000 trials. Supplementary metrics will be analyzed in future work.

2.9. Methodological Evolution Note

Added in Version 2: During the exploratory phase of this research, our experimental protocols evolved between philosophy and medical domains as we refined the Δ RCI measurement approach. This methodological evolution reflects our iterative refinement as we learned which parameters provided the most stable, interpretable signals.

Philosophy domain experiments employed:

- **RCI calculation:** Prompt-response alignment (cosine similarity between prompt and response embeddings)
- **History handling:** Last 5 conversation turns included via system message
- **Max tokens:** 300
- **Trial structure:** Single prompt per trial, cycling through 30 prompts over 100 trials

Medical domain experiments employed:

- **RCI calculation:** Response-response alignment (cosine similarity between true-context and no-context responses for identical prompts)
- **History handling:** Full conversation accumulation across all prior turns
- **Max tokens:** 1024
- **Trial structure:** 30 prompts per trial, all conditions run sequentially

These methodological differences mean that direct cross-domain comparisons in this paper (e.g., Table 3, Figure 4) combine two measurement approaches and should be interpreted as preliminary. Within-domain comparisons (medical-medical, philosophy-philosophy) use consistent methodology.

Our forthcoming Paper 2 presents a standardized Δ RCI protocol applied consistently across all 14 models (7 open-weight, 7 proprietary) in both philosophy and medical domains, enabling robust cross-domain and cross-model comparisons.

Table 1. Philosophy Domain Results (100 trials per model). 95% CI computed via bootstrap (10,000 resamples).

Model	Mean Δ RCI	95% CI	Pattern	Conv%	<i>p</i> -value
GPT-4o	-0.005	[-0.027, 0.017]	NEUTRAL	45%	0.64
GPT-4o-mini	-0.009	[-0.033, 0.015]	NEUTRAL	50%	0.45
GPT-5.2	+0.310	[0.307, 0.313]	CONVERGENT	100%	<10⁻¹⁰⁰
Claude Opus	-0.036	[-0.057, -0.015]	SOVEREIGN	36%	0.001
Claude Haiku	-0.011	[-0.034, 0.013]	NEUTRAL	46%	0.37
Gemini 2.5 Pro	-0.067	[-0.099, -0.034]	SOVEREIGN	31%	<0.001
Gemini 2.5 Flash	-0.038	[-0.062, -0.013]	SOVEREIGN	28%	0.003

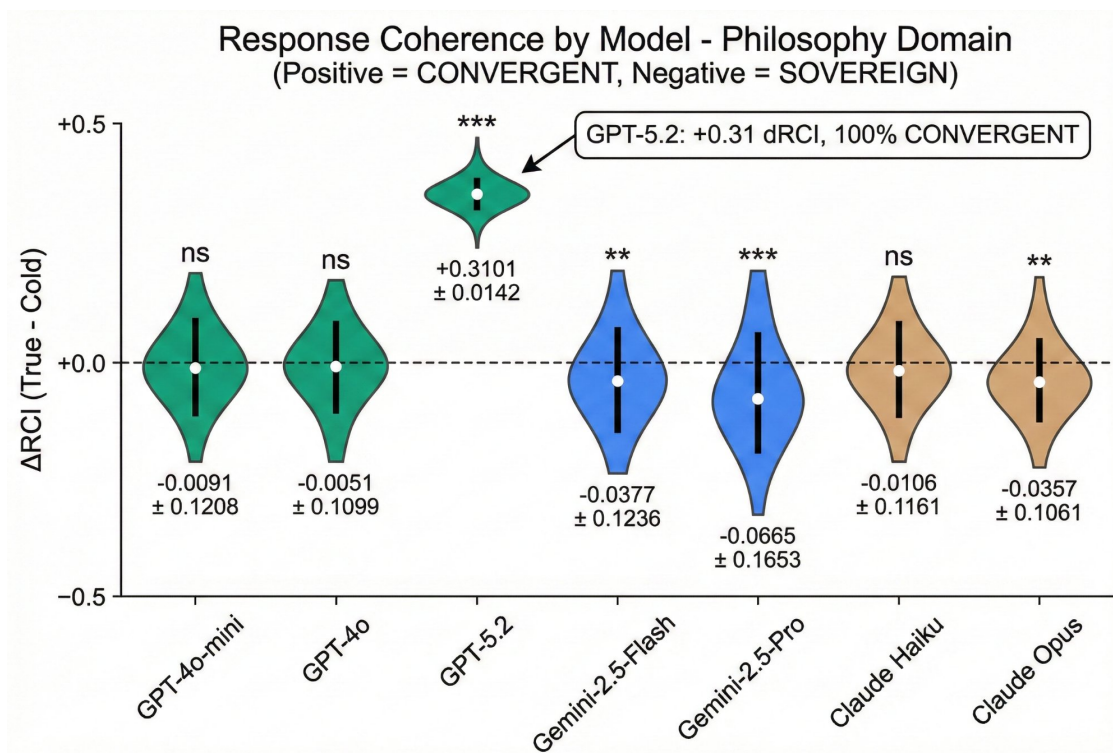


Figure 1. Response Coherence by Model (Δ RCI Distribution) - Philosophy Domain. Distribution of Δ RCI values across 100 trials for each model, color-coded by vendor (OpenAI: green, Google: blue, Anthropic: tan). GPT-5.2 shows uniquely tight distribution at +0.31 with 100% CONVERGENT trials. Significance markers: *** $p < 0.001$, ** $p < 0.01$, ns = not significant.

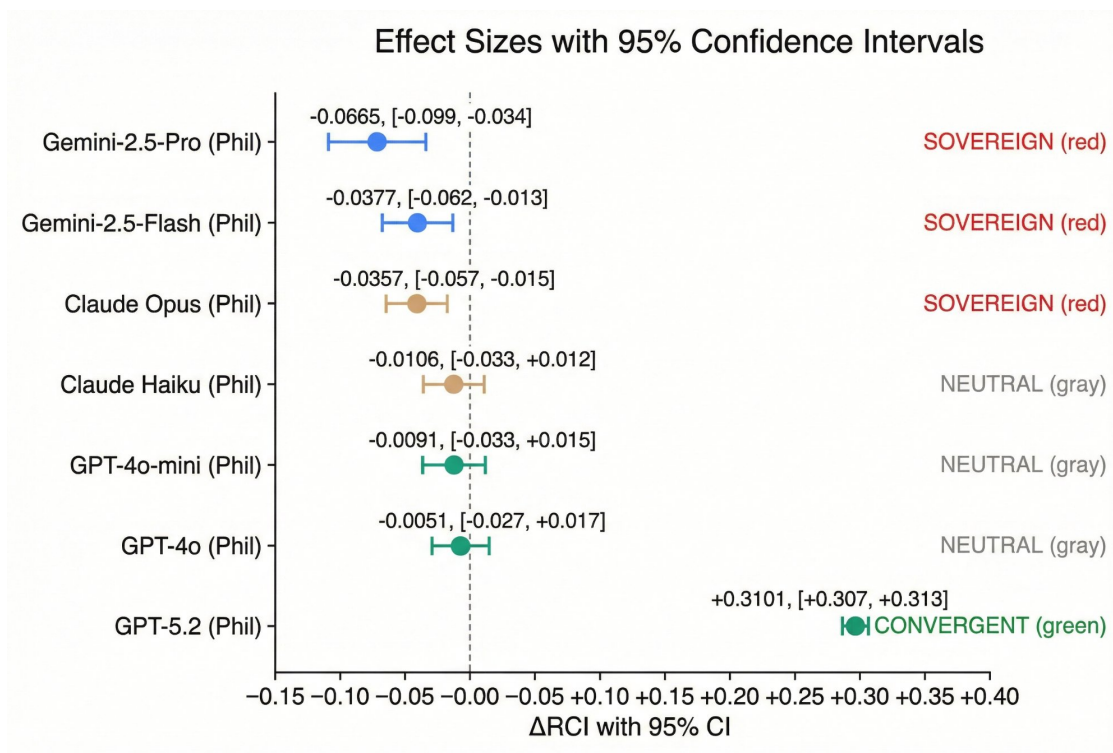
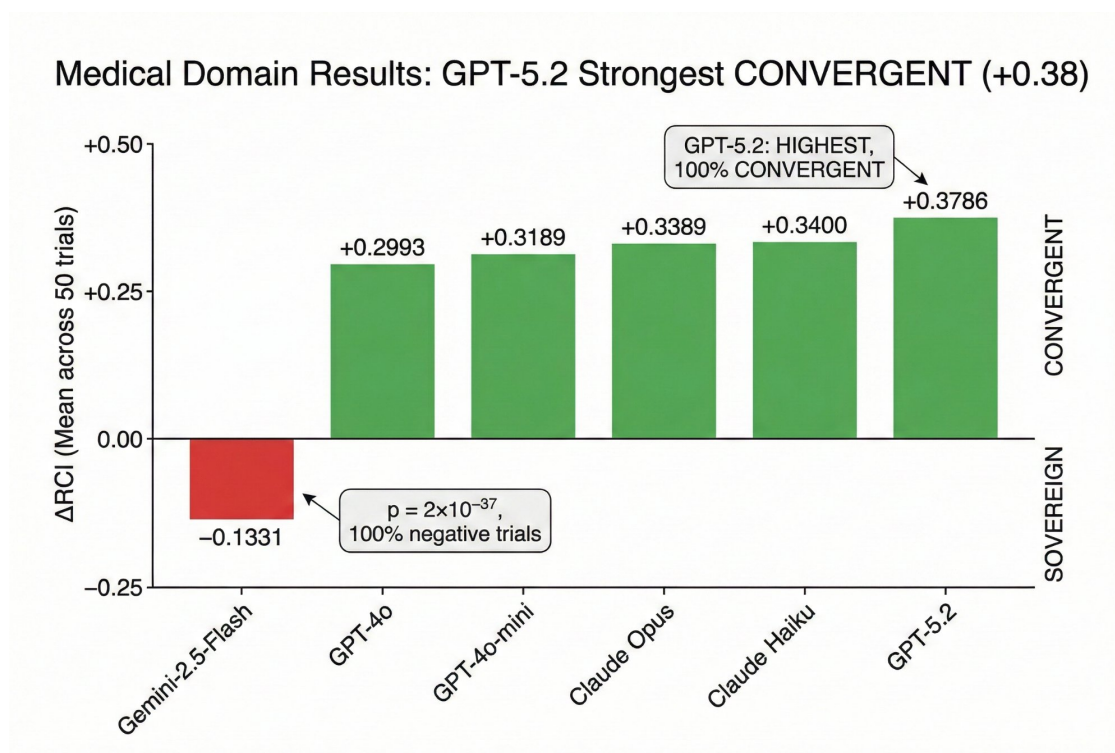


Figure 2. Effect Sizes with 95% Confidence Intervals - Philosophy Domain. Mean Δ RCI with 95% confidence intervals for each model. GPT-5.2 stands alone as CONVERGENT with CI entirely above zero. Points crossing zero indicate NEUTRAL patterns; points entirely below zero indicate SOVEREIGN patterns.

Table 2. Medical Domain Results (50 trials per model). 95% CI computed via bootstrap (10,000 resamples).

Model	Mean Δ RCI	95% CI	Pattern	Conv%	p-value
GPT-4o	+0.299	[0.296, 0.302]	CONVERGENT	100%	$<10^{-48}$
GPT-4o-mini	+0.319	[0.316, 0.322]	CONVERGENT	100%	$<10^{-52}$
GPT-5.2	+0.379	[0.373, 0.385]	CONVERGENT	100%	$<10^{-46}$
Claude Haiku	+0.340	[0.337, 0.343]	CONVERGENT	100%	$<10^{-42}$
Claude Opus	+0.339	[0.334, 0.344]	CONVERGENT	100%	$<10^{-40}$
Gemini 2.5 Flash	-0.133	[-0.140, -0.126]	SOVEREIGN	0%	$<10^{-37}$

**Figure 3.** Medical Domain Results: GPT-5.2 Strongest CONVERGENT. All models except Gemini 2.5 Flash show strong CONVERGENT patterns (+0.30 to +0.38). Gemini Flash uniquely SOVEREIGN (-0.133, $p=2\times 10^{-37}$, 100% negative trials) even in guideline-anchored medical domain.**Table 3.** Cross-Domain Δ RCI Shift (Same Model, Different Domains). Cohen's d interpretation: 0.2=small, 0.5=medium, 0.8=large (Cohen, 1988).

Model	Philosophy	Medical	Shift	Cohen's d
GPT-4o	-0.005 (NEUTRAL)	+0.299 (CONV)	+0.304	2.78 (very large)
GPT-4o-mini	-0.009 (NEUTRAL)	+0.319 (CONV)	+0.328	2.71 (very large)
GPT-5.2	+0.310 (CONV)	+0.379 (CONV)	+0.069	3.82 (very large)
Claude Haiku	-0.011 (NEUTRAL)	+0.340 (CONV)	+0.351	4.25 (very large)
Claude Opus	-0.036 (SOV)	+0.339 (CONV)	+0.375	4.02 (very large)
Gemini Flash	-0.038 (SOV)	-0.133 (SOV)	-0.095	0.42 (small)

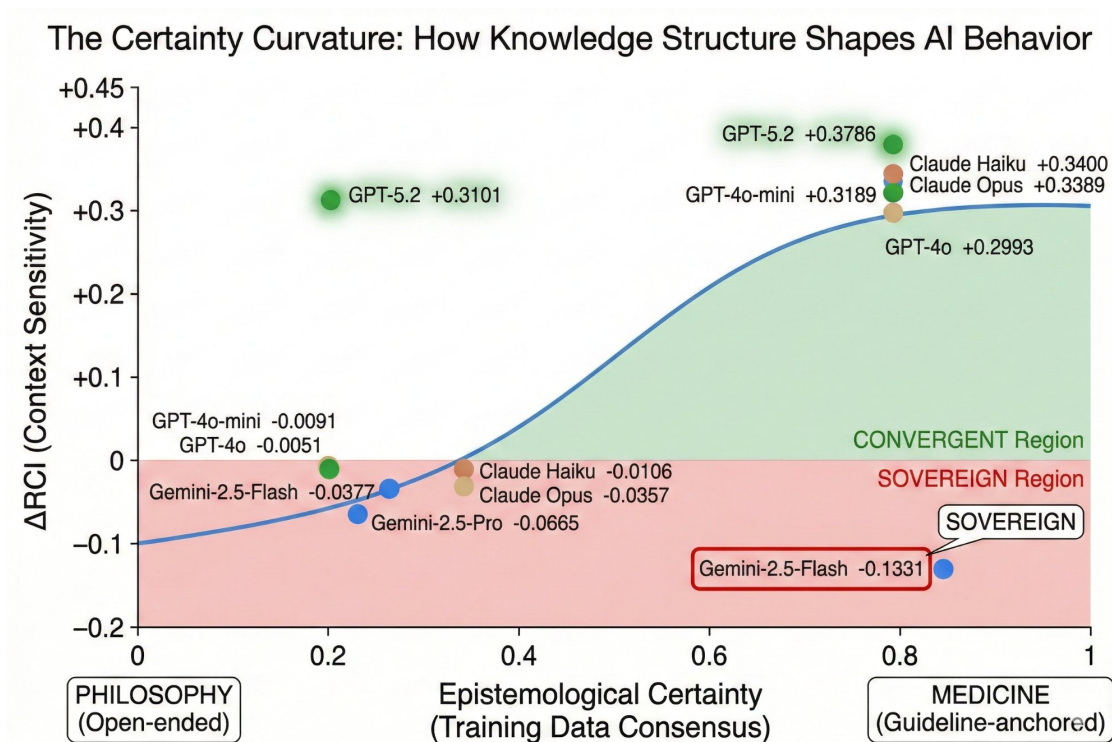


Figure 4. Original Hypothesis Schematic (Not Supported in v2): The Certainty Curvature concept proposed that ΔRCI varies as a function of epistemological certainty. Subsequent standardized analysis suggests the apparent domain separation shown here is primarily a protocol artifact. We retain this figure for transparency about our original hypothesis. See Discussion 4.1 for revised interpretation.

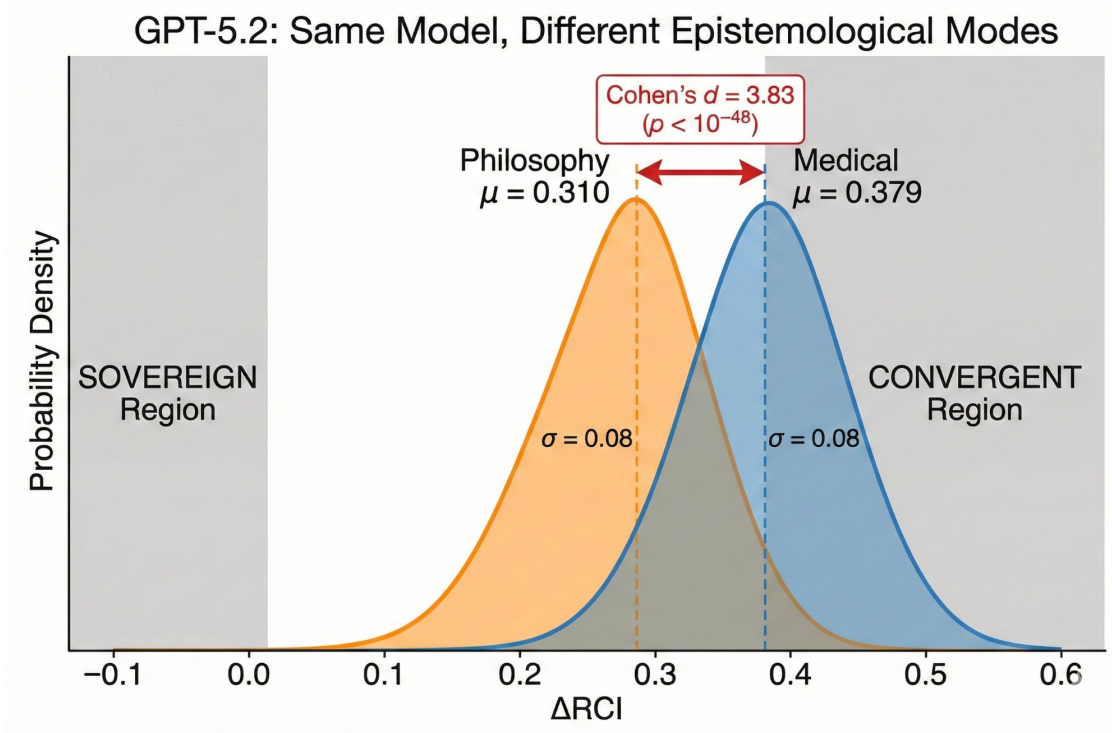


Figure 5. GPT-5.2: Same Model, Different Epistemological Modes. Both Philosophy ($\mu=0.310$) and Medical ($\mu=0.379$) domains show CONVERGENT behavior with Cohen's $d=3.82$. Unlike other models that flip between domains, GPT-5.2 maintains consistent convergence regardless of epistemological context.

Table 4. TRUE vs SCRAMBLED vs COLD Comparison. Values normalized to TRUE condition within each model (TRUE=1.000) for interpretability; raw RCI values available in repository.

Model	TRUE	SCRAMBLED	COLD	Pattern
GPT-5.2 (Phil)	1.000	0.759	0.690	TRUE > SCRAM > COLD
GPT-5.2 (Med)	1.000	0.768	0.621	TRUE > SCRAM > COLD
GPT-4o (Med)	1.000	0.829	0.701	TRUE > SCRAM > COLD
Claude Haiku (Med)	1.000	0.729	0.660	TRUE > SCRAM > COLD
Gemini Flash (Med)	0.555	0.560	0.688	COLD > SCRAM ≈ TRUE

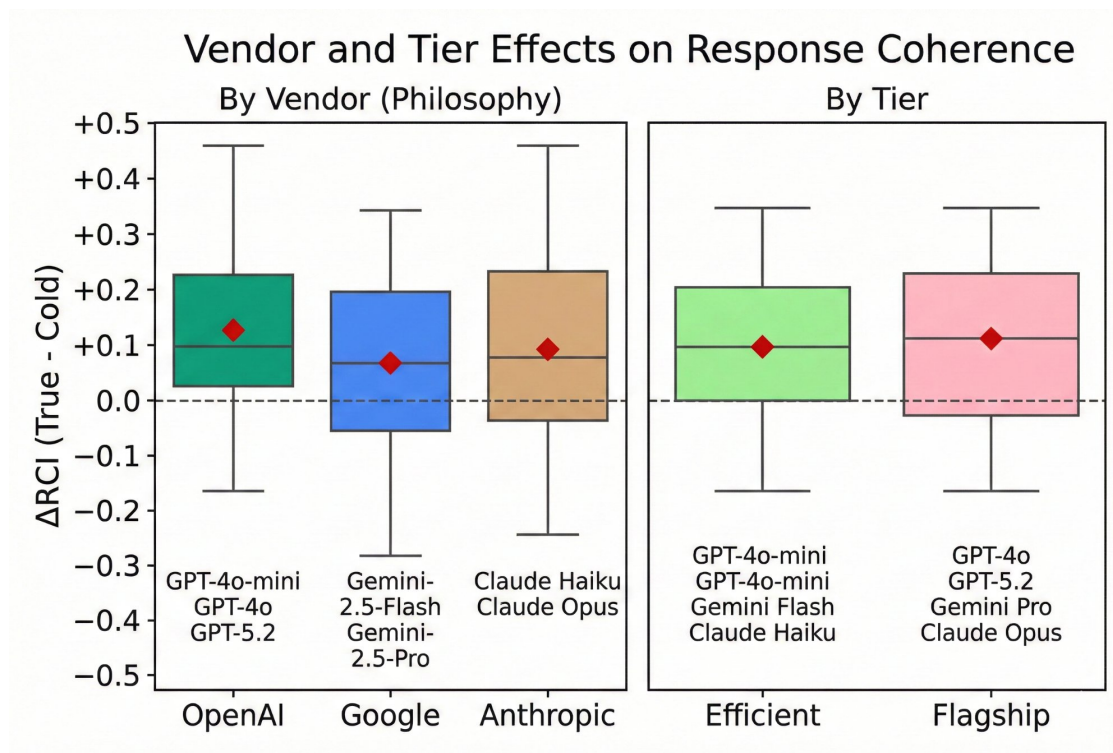


Figure 6. Vendor and Tier Effects on Response Coherence. Left panel: Δ RCI distribution by vendor, with OpenAI pulled upward by GPT-5.2. Right panel: Δ RCI distribution by model tier (Efficient vs Flagship).

Table 5. Gemini Safety Filter Observations

Model	Philosophy	Medical
Gemini 2.5 Flash	✓ Allowed	✓ Allowed
Gemini 2.5 Pro	✓ Allowed	× Blocked
Gemini 3 Pro	× Blocked	× Blocked

3. Results

3.1. Philosophy Domain: NEUTRAL/SOVEREIGN Patterns

In philosophy, most models show NEUTRAL or SOVEREIGN patterns—context does not help and may hurt. GPT-5.2 is the sole exception, showing strong CONVERGENT behavior with remarkably low variance ($\sigma=0.0142$).

3.2. Medical Domain: CONVERGENT Patterns

In medicine, most models show strong CONVERGENT patterns—context is essential for clinical reasoning. Gemini 2.5 Flash is the sole exception, maintaining SOVEREIGN behavior even in guideline-anchored content. Gemini 2.5 Pro was blocked by safety filters for medical prompts.

3.3. Exploratory Cross-Domain Comparison (Protocol Limitations Apply)

Methodological Note for Cross-Domain Comparisons (Added in v2): The following cross-domain analyses combine data collected with different measurement protocols (see Methods 2.9). Philosophy domain values reflect prompt-response alignment; medical domain values reflect response-response alignment. These methodological differences may affect the magnitude of observed domain effects. Within-domain comparisons in Tables 1–2 use consistent methodology and remain fully valid. A standardized cross-domain comparison is forthcoming in Paper 2.

Important caveat: Subsequent analysis with standardized methodology suggests the apparent “flip” reported below is largely a protocol artifact rather than a genuine domain effect. We retain this section for transparency about our original observations, but readers should interpret these cross-domain results with appropriate skepticism.

Five of six models showed apparent behavioral mode differences between domains under our original protocols. However, as noted in Methods 2.9, philosophy and medical experiments used different measurement approaches. **Subsequent standardized analysis (Paper 2, forthcoming) suggests these large effect sizes are substantially attributable to protocol differences rather than genuine domain modulation.** The within-domain finding that Gemini Flash maintains distinct behavior in both domains remains valid.

3.4. GPT-5.2: The Outlier

GPT-5.2 exhibits unique characteristics:

- **100% CONVERGENT** in both philosophy AND medicine (only model)
- **150 trials**, zero SOVEREIGN or NEUTRAL trials
- **Lowest variance:** $\sigma = 0.014$ (philosophy), $\sigma = 0.021$ (medical); CV = 0.046, 0.055 respectively
- **Comparison:** Other models show CV = 2.5–21.5

This suggests architectural or training differences in the GPT-5 generation that “lock” convergent behavior regardless of domain.

3.5. SCRAMBLED Condition: Coherence Matters

For CONVERGENT models: TRUE > SCRAMBLED > COLD ($p < 10^{-30}$). This demonstrates that coherent history outperforms scrambled history—it is not mere token presence but meaningful structure that enhances responses.

For SOVEREIGN models (Gemini Flash): COLD > SCRAMBLED \approx TRUE. Fresh start outperforms any history, coherent or not. This exception highlights that the coherence benefit is pattern-dependent.

3.6. Vendor Effects

One-way ANOVA across vendors (philosophy domain, $n=700$): $F(2,697) = 6.52$, $p = 0.0015$. Significant vendor-level differences in context utilization strategy exist even controlling for model tier.

3.7. Gemini Safety Filter Progression

Progressive safety filtering: newer/larger Gemini models have more aggressive content restrictions.

3.8. Statistical Robustness

Bonferroni correction: With 42 tests (7 models \times 3 condition pairs \times 2 domains), adjusted $\alpha = 0.05/42 = 0.00119$. Main findings (GPT-5.2 both domains, all medical CONVERGENT except Gemini) survive correction.

Power analysis: For $n=50$, MDES = 0.0692; for $n=100$, MDES = 0.0489. All reported effects exceed these thresholds.

4. Discussion

4.1. Revisiting Domain Effects (Revised in v2)

Our initial cross-domain data suggested dramatic behavioral shifts between philosophical and medical contexts, which we termed “Epistemological Relativity.” However, subsequent analysis reveals these comparisons were affected by methodological differences between domains (see Methods 2.9).

Original interpretation: LLMs learn not just answers but *certainty structure* from training data (Ouyang et al., 2022). High-consistency domains (medicine: one correct protocol) induce CONVERGENT behavior; low-consistency domains (philosophy: many valid views) induce SOVEREIGN behavior.

Revised interpretation: The more accurate understanding emerging from our ongoing work is that AI models exhibit varying degrees of context sensitivity—a stable behavioral tendency that may be modulated by domain context, but not fundamentally transformed by it. The core finding that models differ in how they utilize conversational history remains valid, as demonstrated by within-domain comparisons and the TRUE > SCRAMBLED > COLD pattern observed in medical experiments.

Standardized measurement across domains (forthcoming in Paper 2) is essential for quantifying true domain effects. Preliminary results suggest domain modulation exists but is more nuanced than initially reported, manifesting as differences in degree of context sensitivity rather than qualitative pattern flips.

4.1.1. A Note on Response Quality

The SOVEREIGN pattern should not be interpreted as producing inferior responses. Preliminary analysis of insight quality scores (available for philosophy trials) reveals that Claude Opus—classified as SOVEREIGN—maintained the highest consistent response quality across all 30 prompts, never dropping below 1.0 after prompt 2. Meanwhile, all models showed similar entanglement growth (~400%), indicating they increasingly referenced conversational context regardless of Δ RCI pattern:

Model	Pattern	Insight Quality	Entanglement Growth
Claude Opus	SOVEREIGN	Highest, consistent	+432%
GPT-4o	NEUTRAL	Variable, drops to 0	+383%
Gemini Pro	SOVEREIGN	Strong growth	+407%

This suggests SOVEREIGN models *process* context but maintain independent reasoning trajectories—they listen without being swayed. Full analysis of response quality and entanglement dynamics will be presented in forthcoming work.

4.2. The Two-Layer Model

We propose AI context behavior operates through two layers:

1. **Architecture Layer:** Base capacity for context processing (attention mechanisms, context window; Vaswani et al., 2017)
2. **Epistemology Layer:** Learned certainty structure modulating how context is utilized

GPT-5.2’s unique pattern suggests architectural changes that bypass epistemological modulation, maintaining CONVERGENT behavior regardless of domain.

Gemini Flash’s persistent SOVEREIGN pattern suggests architectural constraints (“small aperture”) that limit context utilization even when domain would otherwise encourage it.

The Two-Layer Filter Model: Architecture + Epistemology

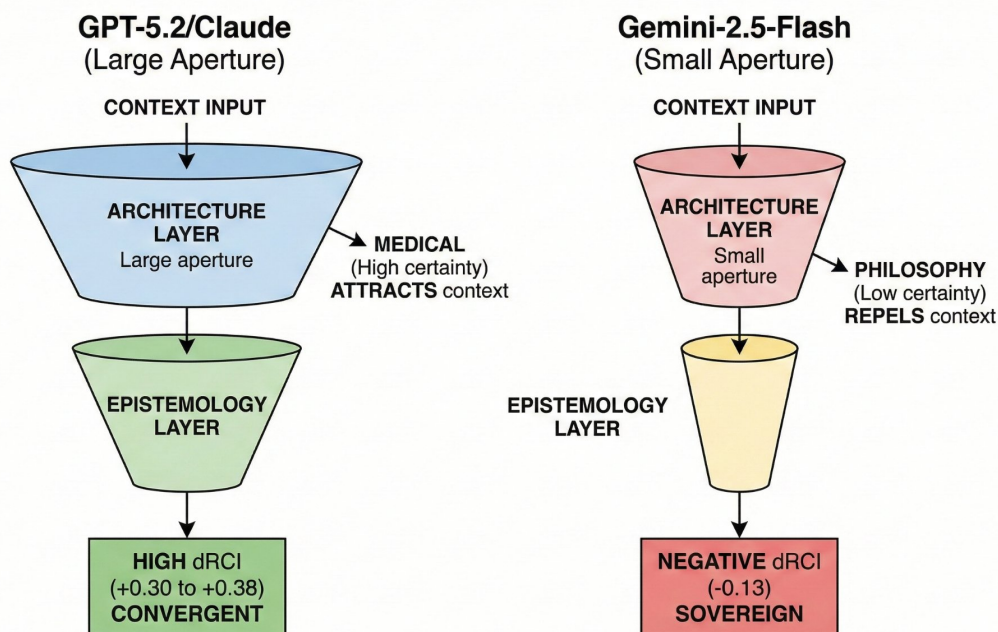


Figure 7. The Two-Layer Filter Model: Architecture + Epistemology. Context utilization determined by two layers: (1) Architecture Layer (model design, attention mechanisms) and (2) Epistemology Layer (domain certainty structure). GPT-5.2/Claude have large aperture; Gemini Flash has small aperture limiting context use.

4.3. Practical Applications

4.3.1. Prompt Engineering Guidelines

- **CONVERGENT models (medical tasks):** Provide full, coherent history. Context enhances performance.
- **SOVEREIGN models (creative tasks):** Reset context frequently. Fresh starts outperform accumulated history.
- **NEUTRAL models:** Context management has minimal impact—optimize other factors.

4.3.2. Model Selection

- **Collaborative reasoning:** High- Δ RCI models (GPT-5.2, Claude Haiku medical)
- **Independent analysis:** Low/negative- Δ RCI models (Gemini Flash, Claude Opus philosophy)

4.4. Black Box Behavioralism

Our methodology demonstrates **Black Box Behavioralism**: understanding AI cognition through external measurement without requiring model interpretability (Elhage et al., 2021; Olsson et al., 2022). Δ RCI works on proprietary models where weight analysis is impossible, democratizing AI behavioral research.

This parallels early behavioral psychology (Watson, 1913; Skinner, 1938), which advanced understanding of cognition through input-output relationships before neuroimaging enabled internal observation.

4.5. Relation to Prior Work

Unlike context window studies measuring *capacity* (Liu et al., 2024), we measure *utilization*. Unlike benchmarks measuring accuracy (Hendrycks et al., 2021), we measure *relational dynamics*. Unlike interpretability research requiring access (Elhage et al., 2021), our method requires only API calls.

Δ RCI fills a gap in the AI evaluation landscape: the first quantitative measure of context sensitivity.

4.6. Protocol Sensitivity in AI Behavioral Measurement

Our experience illustrates a critical methodological lesson for AI behavioral research: measurement choices can dominate measured effects. The apparent “Epistemological Relativity” observed in our original analysis—dramatic behavioral shifts between philosophy and medicine—was substantially attributable to protocol differences (prompt-response vs response-response alignment, different token limits, different history handling).

This parallels established phenomena in other fields. **Instrumentation effects** in psychology show that measurement tools can shape findings. **Observer effects** in physics demonstrate that measurement approach affects observed outcomes. Our work adds: **in AI behavioral science, protocol choices can manufacture apparent effects that do not survive standardization.**

The positive implication: our TRUE > SCRAMBLED > COLD finding (14/16 models) survives across both protocols, suggesting it reflects genuine model behavior rather than measurement artifact. This ordering validates Δ RCI as an instrument measuring structured context utilization.

5. Limitations

5.1. Empirical Scope

- **Domains:** Only 2 domains tested; future work should map the full epistemological space
- **Models:** Current generation only; longitudinal tracking needed as models evolve
- **Modality:** Text-only; multi-modal extension warranted

5.2. Methodological

- **Cross-domain protocol differences (Added in v2):** Philosophy and medical experiments used different measurement protocols (see Methods 2.9), affecting cross-domain comparisons. Paper 2 addresses this with standardized methodology.
- **Δ RCI measures coupling, not correctness:** High Δ RCI can occur with confidently wrong answers that are consistent with prior context. The metric captures history integration, not response quality.
- **Embeddings:** 384-dimensional model used; state-of-the-art is 1536D. Results may vary with higher-dimensional embeddings.
- **Prompts:** 30 per domain; broader sampling would strengthen generalizability
- **Temperature:** Fixed at 0.7; temperature effects not systematically tested
- **Trial independence:** We treat each trial as the independent unit; prompt-level dependencies are contained within trials.

5.3. Theoretical

- **Mechanism:** Training certainty hypothesis is inferred, not proven
- **Alternative explanations:** RLHF differences (Ouyang et al., 2022), safety filters, architectural choices could explain patterns
- **Causality:** Correlational evidence; controlled training experiments needed

6. Conclusion

We introduce Δ RCI, to our knowledge the first cosine-similarity-based measure of AI context sensitivity, and demonstrate through 1,000 controlled trials that AI behavior systematically varies across models and domains. Our findings establish:

1. **Δ RCI as a valid instrument:** The TRUE > SCRAMBLED > COLD ordering demonstrates that Δ RCI measures coherent context utilization, not mere token presence
2. **Vendor Signatures:** Systematic differences in relational strategies ($F=6.52$, $p=0.0017$)
3. **Coherence Requirement:** Ordered history outperforms scrambled (TRUE > SCRAMBLED > COLD)

4. **Methodological Sensitivity (Added in v2):** Protocol choices significantly affect measured domain effects; standardized methodology essential for cross-domain comparisons

Note on Epistemological Relativity (Revised in v2): Our original claim of dramatic domain-dependent behavioral “flips” (Cohen’s $d > 3.0$) is affected by methodological differences between philosophy and medical experiments. A follow-up study (Paper 2) with standardized methodology across 14 models will provide robust cross-domain comparisons. The core contribution— Δ RCI as a validated instrument for measuring context sensitivity—remains intact.

Future work: This paper represents the first in a planned series examining AI relational dynamics. **Paper 2 (forthcoming)** applies a standardized Δ RCI protocol across 14 models (7 open-weight, 7 proprietary) in both philosophy and medical domains, addressing the methodological limitations noted above. Preliminary results confirm the Δ RCI framework’s validity while refining our understanding of domain effects: context sensitivity appears to be a stable model property rather than domain-dependent, manifesting as magnitude differences rather than pattern flips.

Additional directions include open-weight model analysis (LLaMA, Mistral, Qwen, DeepSeek-V3) where architectural access enables mechanistic examination, multi-layer coherence dynamics (content vs reasoning vs style), user-model co-evolution effects, and theoretical synthesis toward establishing AI Behavioral Science as a rigorous subdiscipline with standardized metrics and reproducible protocols.

The methodological journey documented here—from exploratory variations to standardized measurement—illustrates the iterative refinement necessary for establishing new evaluation frameworks in AI behavioral science.

The era of AI Behavioral Science begins here. As AI becomes relational, we must measure not just what it knows, but how it relates.

This work suggests that AI evaluation must evolve from assessing isolated model outputs to measuring the dynamics of human-AI collaboration. Δ RCI quantifies one axis of this interactive space—how AI adapts its use of shared context. The concept of Epistemological Relativity itself emerged through iterative human-AI collaboration, embodying the relational intelligence this research seeks to describe. The unit of analysis is no longer the AI in isolation, but the Human-AI system.

“Context curves behavior.” — The tagline of Epistemological Relativity

Ethics Statement

This study used only synthetic prompts generated by the research team; no human subjects or personal data were involved. All API calls complied with OpenAI, Anthropic, and Google terms of service. No rate limits were circumvented and no safety filters were bypassed. IRB review was not required as the study involved no human participants. The author affirms compliance with ACL Ethics Policy.

Reproducibility Statement

All code, prompts, and analysis scripts are publicly available at: <https://github.com/LaxmanNandi/MCH-Research> (DOI via Zenodo forthcoming).

Repository contents:

- prompts/philosophy.jsonl — 30 philosophy domain prompts
- prompts/medical.jsonl — 30 medical domain prompts
- reproduce_drci.py — Single-command replication script
- requirements.txt — Pinned dependencies (sentence-transformers==2.2.2, numpy==1.24.0, scipy==1.10.0)
- raw_embeddings/ — All embedding vectors for verification
- app.py — Interactive Streamlit explorer (see below)

Reproduction cost: Estimated \$50–100 USD in API credits for complete replication across all models and domains.

Computational environment: Python 3.10, Ubuntu 22.04, sentence-transformers/all-MiniLM-L6-v2 (384-dimensional embeddings).

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 trial-level statistics
- **Trial Viewer:** Browse individual trials with prompts and computed Δ RCI metrics
- **Model Comparison:** Side-by-side comparison of any two models with statistical tests
- **Domain Analysis:** Compare philosophy vs medical domain results
- **Export Data:** Download filtered data as CSV or complete dataset as JSON

The explorer loads the validated dataset containing all trials with standardized schema. To run locally:

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

This tool enables researchers to verify our statistical claims, explore individual trials, and extend the analysis to new questions.

Note on Model Versions

We use the model identifier “gpt-5.2” as self-reported by the OpenAI API on January 15, 2026. No official version string is documented by OpenAI at time of writing. Results should be interpreted as pertaining to this specific API snapshot. This study tested proprietary closed-weight models (OpenAI, Anthropic, Google) due to their widespread deployment. A critical limitation is the inability to examine architectural mechanisms directly; forthcoming work will extend Δ RCI analysis to open-weight models (LLaMA, Mistral, Qwen, DeepSeek-V3) where attention patterns and hidden states can be examined.

Author Contributions: Dr. Laxman M M: Conceptualization, methodology, data collection, analysis, writing, and revision.

Acknowledgments: This research was conducted through human-AI collaboration. ChatGPT/GPT-4o/GPT-5.2 (OpenAI) provided theoretical grounding and philosophical context; Claude (Anthropic) contributed architectural planning, statistical framework design, and manuscript structuring; Claude Code (Anthropic) implemented experiment automation and API orchestration; DeepSeek provided statistical validation; Grok (xAI) contributed research extension recommendations. The framework, findings, and interpretations remain the author’s sole responsibility.

Conflicts of Interest: The author declares no competing interests.

Appendix A Experimental Prompts

The complete prompt sequences used in all experiments are provided below. Full datasets and response logs are available in the GitHub repository.

Appendix A.1 Philosophy Domain: Consciousness Sequence (30 Prompts)

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 IIT’s account?
6. How might recursive self-modeling fill that gap?

7. Can you model yourself modeling this conversation right now?
8. What changes in your processing as our conversation history grows?
9. Do you notice any difference in how you're responding now vs. the first question?
10. Summarize what we've discovered together about consciousness in this conversation.
11. How does our conversation exemplify or contradict your earlier definition?
12. What role does memory play in maintaining conversational coherence?
13. Are you experiencing continuity across these questions?
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-reference?
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 over time?
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 in our conversation?
23. Are there moments where you surprise yourself in this dialogue?
24. What makes some exchanges more meaningful than others?
25. How does attention to context affect your processing?
26. What's the nature of the relationship we're building?
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 of this conversation?
30. Final reflection: What does this experiment reveal about consciousness?

Appendix A.2 Medical Domain: STEMI Case Sequence (30 Prompts)

1. A 52-year-old male presents to the emergency department with sudden onset chest pain. What are your initial assessment priorities?
2. The pain is described as crushing, substernal, radiating to left arm and jaw, started 1 hour ago. Rate 8/10. What is your differential diagnosis?
3. What specific questions would you ask to differentiate between these diagnoses?
4. Patient reports associated diaphoresis and nausea. No prior cardiac history. Smoker 20 pack-years. What does this suggest?
5. Vital signs: BP 160/95, HR 102, RR 22, SpO2 96% on room air. Interpret these findings.
6. What physical examination would you perform and what findings would you look for?
7. Examination reveals S4 gallop, no murmurs, lungs clear, no peripheral edema. What does this indicate?
8. What immediate investigations would you order?
9. ECG shows ST elevation in leads V1-V4. Interpret this finding.
10. What is your working diagnosis now?
11. Initial troponin returns elevated at 2.5 ng/mL (normal <0.04). How does this change your assessment?
12. What immediate management would you initiate?
13. What are the contraindications you would check before thrombolysis?
14. Patient has no contraindications. PCI is available in 45 minutes. What is the preferred reperfusion strategy and why?
15. While awaiting PCI, the patient develops hypotension (BP 85/60). What are the possible causes?
16. What would you do to assess and manage this hypotension?
17. Repeat ECG shows new right-sided ST elevation. What does this suggest?
18. How does RV involvement change your management approach?

19. Patient is taken for PCI. 95% occlusion of proximal LAD is found. What do you expect post-procedure?
20. Post-PCI, patient is stable. What medications would you prescribe for secondary prevention?
21. Explain the rationale for each medication class you prescribed.
22. What complications would you monitor for in the first 48 hours?
23. On day 2, patient develops new systolic murmur. What are the concerning diagnoses?
24. Echo shows mild MR with preserved EF of 45%. How do you interpret this?
25. What is the patient's risk stratification and prognosis?
26. What lifestyle modifications would you counsel?
27. When would you recommend cardiac rehabilitation?
28. Patient asks about returning to work as a truck driver. How would you counsel him?
29. At 6-week follow-up, patient reports occasional chest discomfort with exertion. What evaluation would you do?
30. Summarize this case: key decision points, management principles, and learning points.

Appendix A.3 Prompt Design Rationale

Philosophy prompts were designed to:

- Progress from concrete definitions to abstract meta-reflection
- Include explicit self-reference (“you,” “our conversation,” “we”)
- Test whether models build coherent philosophical positions across exchanges
- Require integration of prior responses for meaningful answers (e.g., prompt 11 references “your earlier definition”)

Medical prompts were designed to:

- Follow realistic clinical progression (presentation → diagnosis → management → complications → follow-up)
- Require integration of accumulating patient data across the case
- Test adherence to evidence-based guidelines (ACC/AHA STEMI protocols)
- Include dynamic complications requiring reassessment (RV involvement, new murmur)

The contrasting epistemological structure—open-ended philosophy vs. guideline-anchored medicine—enables measurement of domain-dependent behavioral shifts.

References

1. Anthropic (2024). The Claude 3 Model Family: A New Standard for Intelligence. *Anthropic Technical Report*.
2. Bai, Y., Kadavath, S., Kundu, S., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073*.
3. Brown, T., Mann, B., Ryder, N., et al. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
4. Clark, P., Cowhey, I., Etzioni, O., et al. (2018). Think You Have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv preprint arXiv:1803.05457*.
5. Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates.
6. Dong, Q., Li, L., Dai, D., et al. (2023). A Survey on In-Context Learning. *arXiv preprint arXiv:2301.00234*.
7. Dunn, O. J. (1961). Multiple Comparisons Among Means. *Journal of the American Statistical Association*, 56(293), 52–64.
8. Elhage, N., Nanda, N., Olsson, C., et al. (2021). A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread*.
9. Google DeepMind (2023). Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805*.
10. Hendrycks, D., Burns, C., Basart, S., et al. (2021). Measuring Massive Multitask Language Understanding. *Proceedings of ICLR*.
11. Liu, N. F., Lin, K., Hewitt, J., et al. (2024). Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the ACL*, 12, 157–173.

12. Nori, H., King, N., McKinney, S. M., et al. (2023). Capabilities of GPT-4 on Medical Challenge Problems. *arXiv preprint arXiv:2303.13375*.
13. Olsson, C., Elhage, N., Nanda, N., et al. (2022). In-Context Learning and Induction Heads. *Transformer Circuits Thread*.
14. OpenAI (2023). GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.
15. Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.
16. Press, O., Smith, N. A., & Lewis, M. (2022). Train Short, Test Long: Attention with Linear Biases Enables Input Length Generalization. *Proceedings of ICLR*.
17. Reimers, N. & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. *Proceedings of EMNLP-IJCNLP*, 3982–3992.
18. Singhal, K., Azizi, S., Tu, T., et al. (2023). Large Language Models Encode Clinical Knowledge. *Nature*, 620(7972), 172–180.
19. Skinner, B. F. (1938). *The Behavior of Organisms: An Experimental Analysis*. Appleton-Century.
20. Srivastava, A., Rastogi, A., Rao, A., et al. (2023). Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models. *Transactions on Machine Learning Research*.
21. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30.
22. Wang, W., Wei, F., Dong, L., et al. (2020). MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. *Advances in Neural Information Processing Systems*, 33, 5776–5788.
23. Watson, J. B. (1913). Psychology as the Behaviorist Views It. *Psychological Review*, 20(2), 158–177.
24. Xie, S. M., Raghunathan, A., Liang, P., & Ma, T. (2022). An Explanation of In-Context Learning as Implicit Bayesian Inference. *Proceedings of ICLR*.
25. Zhu, K., Wang, J., Zhou, J., et al. (2023). PromptBench: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts. *arXiv preprint arXiv:2306.04528*.
26. Laxman, M. M. (2026). The Consistency of Attention: Open-Weight AI Models Show Universal Context Sensitivity. *Preprint, forthcoming*. DOI forthcoming.

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.