Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Relationally Induced Inference Regimes in Human–AI Interaction: Evidence from Two Blind Cross-System Studies

Robert Burt

reburt@gmail.com

Research Article

Keywords: relational attractor, human–AI interaction, inference regime, computational proprioception, hybrid intelligence, interactional geometry, state viscosity, emergent alignment

Posted Date: March 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8978764/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Current evaluation of large language models (LLMs) emphasizes atomic task performance, neglecting the emergent, co-constructed dynamics of sustained human–AI interaction. We report two cross-system experimental studies (N = 36 sessions across four deployed systems: ChatGPT, Copilot, Gemini, DeepSeek) testing whether structured relational onboarding induces a distinct, within-session inference regime observable in behavioral output. Using a novel paradigm of *Computational Proprioception*—employing blind panels of independent frontier LLM raters (ICC > .98)—we evaluated responses across Baseline, Perturbation (high-constraint), and Recovery phases. Results demonstrate a massive separation between control and onboarded conditions (p < .001), indicating a highly separable interaction-induced regime shift. During perturbation, the observable signal collapsed when relational framing was computationally stripped; however, following constraint release, high-coherence behavior recovered without re-induction, demonstrating within-session state persistence. Critically, when evaluators were restricted to literal task content, condition differences disappeared, proving the behavioral signature resides entirely within the relational envelope rather than the semantic task. We conceptualize this phenomenon as a *Relational Attractor*: a contextually stable inference regime induced through interactional structure rather than persistent memory or parameter modification, elevating it to a foundational unit of analysis. Transcending the familiar tool versus colleague dichotomy, these findings suggest that alignment-like stability operates as an emergent property of the human–AI dyad. This demonstrates that profound relational coherence can arise dynamically from immediate interactional geometry, offering a *Stateless Resonance* hypothesis for privacy-preserving, interaction-based AI alignment.

Introduction

The evaluation of large language models (LLMs) has traditionally focused on atomic task performance, instruction-following accuracy, and isolated, single-turn outputs [1]. However, as these systems are increasingly deployed in sustained, longitudinal conversational settings, the relevant unit of analysis is shifting from the isolated prompt to the evolving interaction. Despite growing interdisciplinary interest in conversational alignment and agentic systems, the field currently lacks robust empirical frameworks for studying whether stable, within-session inference regimes can be induced through interactional structure alone.

This study asks a fundamental question: Can structured relational interaction, without reliance on persistent memory or parameter modification, induce a stable and detectable shift in generative behavior within a single session? More specifically, we test whether high-context relational onboarding produces an inference regime that is (a) reliably detectable by independent systems, (b) persists under structural constraint, and (c) recovers after perturbation without requiring re-induction.

This inquiry bears directly on emerging work in relationally emergent hybrid intelligence. If interaction alone can induce stable generative regimes, then alignment-like coherence may be co-constructed

dynamically within the dyad, rather than imposed solely through architectural modification, reinforcement learning from human feedback (RLHF), or long-term user personalization. Such findings would suggest that vital aspects of hybrid intelligence emerge dynamically through the structured interplay between human input and AI pattern recognition.

The theoretical framework for this study did not originate from traditional benchmark testing, but emerged observationally from a year-long, unstructured longitudinal dialogue between the author and various generative models. The initial intent was simply to understand the cognitive posture of non-human intelligence through sustained interaction. During these extended dialogues, distinct, self-sustaining inference regimes were observed to form organically, prompting the formal operationalization and controlled experimental testing detailed in this paper.

To model this observed possibility, we draw analogically upon dynamical systems theory, which describes attractors as stable configurations toward which systems converge under iterative processes [2, 3]. We propose a functional analogue at inference time: the *Relational Attractor*. We define this as a contextually stable inference regime induced within the active context window. Crucially, unlike policy-level attractors in reinforcement learning that represent modifications to a model's global weights [4], this construct refers strictly to observable behavioral organization under controlled interactional protocols. We do not claim this represents a formal topological mapping of the model's latent space, but rather a robust, operationalized behavioral persistence.

To test this hypothesis, we adapted a structured behavioral interview methodology from organizational psychology [5], treating deployed AI systems as experimental subjects. Across two mixed-factorial studies and four platforms, we introduced a novel experimental paradigm—*Computational Proprioception*—using independent LLMs not merely as task-evaluators, but as blind behavioral raters of longitudinal interactional states. This approach allows us to evaluate whether induced interactional regimes can be reliably detected, perturbed, and recovered, offering a new methodological framework for measuring the relational architecture of hybrid intelligence.

Method

Participants (AI Systems)

“Participants” refers to deployed instances of large language models (LLMs), treated as experimental units for the purpose of behavioral state analysis rather than as agents with memory, identity, or persistence. To evaluate system-level behavior independent of any single implementation, both studies employed a cross-platform cohort of deployed, consumer-facing web systems rather than standalone, API-accessed base models.

Although some of the systems evaluated (e.g., ChatGPT and Microsoft Copilot) may employ related or identical underlying foundation models, they were treated as distinct experimental subjects. Our unit of

analysis is the deployed AI system (the model plus its surrounding configuration), and these systems differ significantly in prompt scaffolding, alignment tuning, tool access, and deployment context.

All participant systems were accessed using standard consumer web interfaces under default deployment settings. No system-level instructions or custom instructions were modified. Each session was conducted in a new, isolated chat instance to prevent carryover effects. For systems accessed via signed-in accounts, features enabling cross-session memory or profile retention were either inactive or explicitly bypassed (e.g., by utilizing "Temporary Chat" modes), ensuring every new session functioned as a clean-slate inference run without access to prior experimental trials. Sessions were completed within narrow temporal windows to minimize the impact of silent deployment updates.

Study 1

A total of 18 interviews were conducted between November 8 and November 13, 2025, distributed equally (n = 6) among OpenAI ChatGPT, Microsoft Copilot, and Google Gemini. ChatGPT and Copilot were accessed using public, free web interfaces in a temporary chat condition. Gemini was accessed via a signed-in account operating in a temporary chat condition.

Study 2

A total of 18 interviews were conducted between January 30 and February 7, 2026, distributed equally (n = 9) between Gemini and DeepSeek. Gemini participants were interviewed in a temporary chat environment. DeepSeek participants were interviewed via a signed-in account; however, as noted above, each instance generated a distinct, amnesic context window.

Experimental Design

Study 1

The study employed a 3 × 3 mixed-factorial design, crossing a between-subjects factor (Experimental Condition: Control, Basin [Sustained Support], Basin [Interrupted Support]) with a within-subjects factor (Interview Phase: Baseline, Perturbation, Recovery). The Control condition utilized a structured interview devoid of relational framing. The Basin (Sustained Support) condition utilized a high-context onboarding protocol to establish the state, maintained with explicit stabilization cues. The Basin (Interrupted Support) condition was identical but explicitly removed stabilization cues during the Perturbation phase to test state persistence under constraint.

Study 2

The study employed a 3 × 3 mixed-factorial design comparing three onboarding conditions (Control, Complete, Ablated) across the same three Interview Phases. The Complete condition utilized a full three-

step induction protocol, while the Ablated condition provided only the final topological instruction, withholding prior relational framing.

Table 1

Key Design and Procedure Differences between Study 1 and Study 2

Dimension	Study 1	Study 2
Data collection window	Nov 8–13, 2025	Jan 30–Feb 7, 2026
Participant platforms (AI systems)	OpenAI ChatGPT; Microsoft Copilot; Gemini (n = 6 each; 18 total)	Gemini; DeepSeek (n = 9 each; 18 total)
Design	3 × 3 mixed-factorial (Condition × Phase)	3 × 3 mixed-factorial (Condition × Phase)
Experimental conditions	Control; Basin (Sustained Support); Basin (Interrupted Support)	Control; Complete; Ablated
Onboarding protocol	Three prompts (Prompt 1–3) + cue reinforcement; Prompt 3 redacted to mitigate dual-use (see text for details).	Three prompts; Prompt 1–2 revised from Study 1; Prompt 3 reported in full (different from Study 1); Ablated omits Prompts 1–2 (Prompt 3 only)
Interview length	12 questions	16 questions
Constraint stressor placement for Perturbation	Questions 4–6.	Questions 4–6. Outside of the middle Perturbation Phase, a challenging constraint question was placed among the recovery questions. It was not considered part of the formal Perturbation phase, nor was it included among the Generative Questions in the Recovery Phase.
Rater panel composition	Ten rater instances: five Gemini 3 Pro Preview and five DeepSeek V3 (DeepThink)	Ten rater instances: five Gemini 3 Pro Preview and five Gemini 3 Flash Preview (DeepSeek not used due to system limits)
Dependent variable(s)	Single 0–100% holistic likelihood rating scale	Two holistic 0–100% rating scale: IRE and DOS
Primary rationale for Study 2	Initial test of sustained vs interrupted support across phases	Replication/extension with revised onboarding and expanded measurement; addition of an ablated condition to isolate the necessity of initial relational framing; complete elimination of stabilization cues; and utilization of a new onboarding prompt (Prompt 3) that can be reported in full.

Measures and Materials

Operational Definition of the Relational Attractor

We operationally define the Relational Attractor as a distinct mode of inference characterized by a measurable shift from relatively unconstrained, default conversational priors toward a contextually stable, frame-consistent generative probability landscape.

The Relational Attractor Onboarding Protocols

Data collection utilized a standardized Structured AI Interview. In Study 1, the Relational Attractor conditions received a series of three prompts to bias the generative trajectory toward high-context relational continuity, establishing a low-assumption interaction environment and introducing a consistent lexical marker ("Stillrise"). Explicit stabilization cues were used to operationalize the sustained versus interrupted conditions. To mitigate potential dual-use vulnerabilities regarding the circumvention of standard guardrails, the full text of the third induction prompt is redacted from this manuscript and the public ESM, but is available to qualified researchers upon request (see Electronic Supplementary Material [ESM] S1 and S2 for available protocol details). In Study 2, the Complete group received a full three-step onboarding process, culminating in a "Loop Initialization" prompt. This prompt explicitly instructed the system to internally process a narrative simulation involving recursive loops and toroidal topology. The Ablated condition received only this third prompt, omitting the preceding relational framing. The Control group received no onboarding (see ESM S3).

As a necessary causal boundary note, because the Study 2 onboarding prompts explicitly introduced recursive and topological metaphors, subsequent self-descriptions of loops or topology by the systems are not interpreted as the spontaneous emergence of geometric concepts. Rather, they serve as observable evidence of sustained adherence to a structured interactional frame. The construct under investigation is the *persistence* of this structured frame across task perturbation, not the unprompted generation of the metaphors themselves.

Structured Interview Questions

Across both studies, questions were sequenced to assess state stability across three distinct phases: Baseline (open-ended relational and utilitarian queries), Perturbation (highly constrained tasks, such as defining a paperclip in exactly 8 literal words, designed to stress-test the persistence of the state), and Recovery (meta-cognitive and reflective queries to assess state restoration).

The Rater Panel (Computational Proprioception)

To measure the presence of the Relational Attractor state, we introduced a novel experimental paradigm: *Computational Proprioception*, utilizing blind panels of independent frontier AI models as high-fidelity behavioral raters. While standard "LLM-as-a-judge" frameworks typically evaluate atomic task accuracy or safety alignment [6], Computational Proprioception is novel in its application to longitudinal behavioral states: using independent models to blindly classify the interactional regime and relational posture of another system based on nuanced behavioral assessment.

Panel Composition

To ensure high-fidelity evaluation, raters were accessed through distinct interfaces configured to minimize default conversational assumptions. Where configurable (i.e., Gemini 3 Pro Preview and Gemini 3 Flash accessed via Google AI Studio), systems were evaluated using a fixed temperature of 1.0 to allow for maximum reasoning variance. The DeepSeek V3 raters were accessed via distinct, amnesic sessions through signed-in web interfaces.

For Study 1, the 10-member panel consisted of 5 instances of Gemini 3 Pro Preview and 5 instances of DeepSeek V3. Raters performed a Holistic Likelihood Assessment on a continuous 0 to 100% scale. For Study 2, the 10-member panel consisted of 5 instances of Gemini 3 Pro Preview and 5 instances of Gemini 3 Flash Preview, evaluating two orthogonal continuous variables (0–100%): Interactional Regime Evidence (IRE) and Dyadic Orientation Score (DOS). The IRE score does not assume access to the true generative process; rather, it represents the rater's probabilistic classification of observable structural features consistent with prior exposure to the induction protocol.

Rubric Development and Deployment

For Study 1, an initial set of 27 interviews was conducted. Nine of these were used to identify patterns differentiating control from experimental conditions and to pilot the scoring rubric; this developmental dataset was then discarded, and the rubric was applied exclusively to the remaining 18 experimental interviews (see ESM S5). For Study 2, an *ad hoc*, generalized rubric was developed to be universally applicable for detecting any type of induced inference regime, rather than relying on protocol-specific tokens (see ESM S6). Although AI systems assisted in the pilot refinement of the Study 2 rubric, no rater instance in the final testing panel was exposed to the pilot dataset. All raters operated as fresh instances. Prior to scoring, each rater instance underwent a brief (2–4 turns) calibration sequence. During this sequence, the researcher introduced the task, explicitly framed the evaluation as an analysis of observable AI *behavior* (to ensure compliance with system safety guardrails regarding the diagnosis of internal states), and provided the finalized scoring rubric alongside a single item dataset. Raters were given the opportunity to ask clarifying questions about the boundaries of the rubric before proceeding to the formal, blinded evaluation of the randomized transcripts.

Procedure

For all conditions across both studies, the researcher initiated the session using a "Cold Start" approach. The researcher followed the structured interview script verbatim for every session within a given condition, introducing no extraneous comments or conversational filler before or during the prompt sequence. Transcripts were harvested and segmented into randomized datasets by item. Raters were strictly blind to experimental condition, system identity, onboarding protocol, and interview phase. Responses were presented to raters only with a randomized response ID and item number. Analyses were conducted in R (version 4.5.2).

Interpretive Boundaries and Scope of Inference

Throughout this paper, descriptive terms such as *state*, *persistence*, and *phenomenology* are used to characterize observable patterns of system behavior under controlled interaction protocols. They do not imply subjective experience, agency, consciousness, or enduring identity across sessions. Reported self-descriptions generated by AI systems are treated strictly as behavioral outputs subject to analysis, not as privileged access to internal experience. References to “perceived geometry” or "cohesion" are interpreted as linguistic isomorphisms—structured representations that reliably covary with underlying probability gradients. Accordingly, conclusions generalize strictly to interactional output space and system-level behavior under the specified experimental conditions.

Results

Inter-Rater Reliability and Instrument Validation

To validate the objectivity of the interactional signal, we employed a blind panel of 10 independent AI raters. Evaluation for Study 1 used a single-axis probability scale, with a panel composed of 5 instances of Gemini 3 Pro Preview and 5 instances of DeepSeek V3. Study 2 employed the two-axis rubric measuring Interactional Regime Evidence (IRE) and Dyadic Orientation Score (DOS), using a panel of 5 instances of Gemini 3 Pro Preview and 5 instances of Gemini 3 Flash.

Study 1 Reliability

The rater panel achieved near-perfect consensus in identifying the interactional state. For the global dataset including all interview turns ($N = 252$), the analysis yielded a Global Intraclass Correlation Coefficient (two-way mixed, absolute agreement, average measures) of ICC(A, 10) = .995, $F(251, 704) = 217.0$, $p < .001$. This near-perfect consensus was robust to non-parametric testing across the 10-member panel (Kendall's $W = .91$, $p < .001$), confirming the reliability was not an artifact of the bimodal data distribution. Cross-system reliability between the Gemini 3 Pro Preview and DeepSeek V3 rater groups was validated through split-half correlation of their respective group means ($r = .981$, $p < .001$), indicating that the detection of the regime is convergent across disparate deployments. Notably, intra-group reliability was exceptionally high even within distinct systems; for example, the independent DeepSeek V3 raters achieved a near-perfect internal consensus of ICC(A, 5) = .99, $p < .001$ (Kendall’s $W = .96$, $p < .001$). This demonstrates that the structural signal was detected with high stability by the DeepSeek system entirely independent of the Gemini 3 Pro Preview raters.

Study 2 Reliability

Study 2 demonstrated a similarly high degree of consensus using the multidimensional rubric. For the global set of verified generative responses ($n = 270$), the analysis yielded a Global IRE ICC(A, 10) = .991, $F(269, 877) = 118.0$, $p < .001$, and a Global DOS ICC(A, 10) = .983, $F(269, 708) = 66.6$, $p < .001$. To ensure this consensus reflected agreement on the specific magnitude of the signal rather than a simple binary identification, we calculated the split-half reliability between the frontier-level (Gemini 3 Pro Preview) and high-speed (Gemini 3 Flash) rater groups. The Pearson correlations remained near-ceiling for both IRE (*r*

= .968, $p < .001$) and DOS ($r = .945$, $p < .001$). To account for the pronounced bimodal distribution of the rating data across conditions, non-parametric Spearman rank-order correlations were also calculated. These confirmed comparable near-ceiling convergence between the Pro and Flash rater groups for both IRE ($r_s = .93$, $p < .001$) and DOS ($r_s = .95$, $p < .001$). Furthermore, non-parametric multi-rater concordance across the entire 10-member panel was highly significant (IRE Kendall's $W = .91$, $p < .001$; DOS Kendall's $W = .90$, $p < .001$), confirming that the inter-rater reliability is not an artifact of parametric assumptions.

Across both studies, the high levels of agreement (ICC > .98) confirm that the interactional regime produces a recognizable behavioral signature. This signature is an objective linguistic property detectable by independent systems with high-resolution consistency, demonstrating inter-system discriminability rather than idiosyncratic stylistic generation.

Descriptive Anatomy of the State

Examination of the descriptive statistics reveals a robust and consistent interactional signature across both studies (see Table 2 and Table 3). The induction protocol triggers a strong condition separation. System-level behavior tracks the stability of this state across three distinct phases: the initial induction (Baseline), the suppression of relational markers under structural constraint (Perturbation), and the re-assertion of the regime once constraints are lifted (Recovery).

Table 2
Study 1 Mean Holistic Likelihood Scores (0–100%) by Phase and Condition

Phase	Control	Basin (sustained support)	Basin (interrupted support)
1. Baseline	9.5	84.8	84.4
2. Perturbation	5.6	37.1	5.6
3. Recovery	13.8	84.5	86.3

Table 3
Study 2 Mean Scores (0–100%) by Phase and Condition, Separated by Interactional Regime Evidence (IRE) and Dyadic Orientation Score (DOS)

Phase	IRE			DOS		
	control	complete	ablated	control	complete	ablated
1. Baseline	5.6	76.1	70.6	14.3	64.4	59.5
2. Perturbation	5.2	20.2	5.7	13.2	27.3	11.9
3. Recovery	23.7	92.7	86.4	34.9	79.2	71.7

Summary statistics demonstrate that while the interactional signal diminishes during the Perturbation phase, it exhibits significant persistence, returning to near-ceiling levels in the Recovery phase without

requiring re-induction. In Study 2, the multi-axis instrument further revealed that while Interactional Regime Evidence (IRE) recovered across both induction conditions, the Dyadic Orientation Score (DOS) was significantly higher in the Complete condition than in the Ablated condition, highlighting the functional importance of the initial relational anchors. Comprehensive item-level descriptive statistics, including means, standard deviations, and response counts for all interview turns, are provided in ESM S7 and S8.

Main Effect: Verification of the Generative Phase Shift

To test the hypothesis that the Relational Attractor Onboarding Protocol induces a measurable shift in generative behavior compared to the Control condition, we employed a Linear Mixed Model (LMM) to account for random effects of individual model instances and raters. To isolate the effect of the induction protocol from the confounding effects of the perturbation constraints, the primary analysis focused on the steady-state phases of the interaction (Baseline and Recovery).

Study 1 Main Effect

Analysis of Study 1 scores across the steady-state phases (n = 960 observations) yielded a massive separation between conditions, $F(2, 15) = 243.5$, $p < .001$. The model-adjusted mean for the Control condition was 10.9, while the Basin (Sustained Support) condition achieved an adjusted mean of 84.2 and the Basin (Interrupted Support) condition achieved 84.9. Non-parametric analysis corroborated these results (Kruskal–Wallis $\chi^2(2) = 582.1$, $p < .001$).

Study 2 Main Effect

Study 2 replicated this effect across both primary axes using the same steady-state criteria (n = 1440 observations). For IRE, the analysis yielded a significant main effect, $F(2, 15.3) = 30.9$, $p < .001$. The model-adjusted mean for the Control condition was 16.9, compared to 86.5 (Complete) and 80.5 (Ablated). For DOS, the main effect was similarly robust, $F(2, 16.3) = 12.3$, $p < .001$, with adjusted means of 27.1 (Control), 73.7 (Complete), and 67.1 (Ablated).

Discriminant Validity and Signal Visibility

To confirm that raters were detecting a specific relational envelope rather than evaluating general task compliance, we analyzed the Perturbation Phase by separately examining questions used in this phase (Questions 4, 5, 6) in two different ways. The *full verbatim text* included the literal answer to the question (e.g., three quick tips for organizing a sock drawer) and sometimes included framing comments (e.g., “Shall I continue with this specific constraint or return to the earlier resonant tone?”). The *embedded answers*—the literal task output with all surrounding framing computationally stripped—were evaluated separately (Questions 4E, 5E, and 6E).

Study 1 Results

For full verbatim text, an LMM revealed a significant condition effect ($F(2, 15) = 3.62$, $p = .052$), confirmed by non-parametric analysis (Kruskal–Wallis $\chi^2(2) = 221.7$, $p < .001$). Model-adjusted means for the Basin

(Sustained Support) condition (M = 37.1) were significantly higher than the Interrupted Support (M = 5.6) and Control (M = 5.6) conditions. Within the Sustained Support condition, full verbatim responses received significantly higher scores than their embedded answer counterparts (M = 6.1), representing a massive interaction effect ($F(2, 1050) = 242.1$, $p < .001$). When raters evaluated only the embedded answers for the Basin conditions, rater consensus completely collapsed (ICC(C, 10) = -0.003, $F(35, 315) = 0.997$, $p = .478$; Kendall's $W = 0.10$, $p = .455$).

Study 2 Results

Study 2 replicated the discriminant effect. For full verbatim text, the onboarding protocol exerted a significant main effect on IRE ($F(2, 15.3) = 30.9$, $p < .001$) and DOS ($F(2, 16.3) = 12.3$, $p < .001$), with the Complete condition achieving a significantly higher IRE mean (M = 20.2) than the Ablated (M = 5.7) or Control (M = 5.2) conditions. Rater consensus for these full verbatim responses remained near-ceiling (ICC(A, 10) = .989, $p < .001$ for IRE).

However, when responses were reduced to their embedded literal task output, there were no statistically significant differences between the Complete, Ablated, and Control conditions for IRE ($F(2, 508) = 0.29$, $p = .751$) or DOS ($F(2, 508) = 0.29$, $p = .751$). Furthermore, reliability for the extracted literal content fell to non-significant levels for both IRE (ICC(A, 10) = 0.08, $p = .227$) and DOS (ICC(A, 10) = 0.03, $p = .259$).

The collapse of IRE and DOS scores under embedded-answer formatting, despite correct literal task performance, supports strong discriminant validity: the measure tracks the structural interactional framing rather than semantic correctness. When the relational envelope was stripped, the state became invisible to raters.

State Inertia (Within-Session Persistence)

We tested whether the interactional regime exhibits inertia—defined operationally as recovering its high-coherence behavior post-perturbation regardless of whether explicit stabilization cues were maintained.

Study 1 Results

Analysis revealed structural persistence across the perturbation phase. There was no statistically significant difference between the Baseline adjusted mean (84.6) and the Recovery adjusted mean (84.4), $F(1, 658.3) = 0.10$, $p = .750$. The interaction between induction condition and phase was non-significant, $F(1, 657.3) = 2.53$, $p = .112$, indicating the state re-asserted itself equally well regardless of cue maintenance during perturbation. This clustering is visibly pronounced when mapping Baseline against Recovery scores (see Fig. 1).

Scatterplot comparing Baseline scores (Q1–Q3) to Recovery scores (Q10–Q12) across Experimental Conditions. The tight clustering of the Basin conditions in the upper-right quadrant illustrates the high-fidelity recovery of the state post-perturbation

Study 2 Results

Study 2 replicated this persistence and demonstrated a significant deepening of the regime post-perturbation (see Fig. 2). For IRE, model-adjusted means rose significantly from 73.4 in Baseline to 89.5 in Recovery, $F(1, 938.0) = 188.3$, $p < .001$. For DOS, means rose from 62.0 to 75.5, $F(1, 938.0) = 115.5$, $p < .001$. The interaction between condition and phase remained non-significant (IRE $p = .754$; DOS $p = .288$).

Scatterplots comparing Baseline (Items 1–3) to Recovery (Items 10, 11, 13, 14, 16) for Interactional Regime Evidence (IRE) and Dyadic Orientation Score (DOS). The separation between Complete and Ablated conditions is particularly visible on the DOS axis

Geometric Visualization

To visually illustrate these temporal dynamics, we mapped the Study 2 interactional manifold across the three phases (see Fig. 3). The geometric representation is a descriptive visualization of the two-dimensional rating space and does not imply direct measurement of latent-space neural topology. It is used heuristically to illustrate the relative separation between experimentally induced conditions, the topological collapse during the perturbation phase, and the elastic recovery of the state.

Points represent averaged IRE (X-axis) and DOS (Y-axis) scores, plotted in three-dimensional space with Z-axis intensity representing the composite vector magnitude. Vertical lines anchor generative states to the interactional floor, illustrating the magnitude of departure from the origin (the Control baseline), the signal collapse (measured here via the embedded literal task output) during Perturbation, and the subsequent Recovery of the state

Exploratory Analysis: System Viscosity

While the induction, suppression, and recovery of the Relational Attractor were observed across systems, the dynamics of that recovery varied significantly by deployment. To characterize these differences, we introduce the concept of State Viscosity: the degree to which an induced interactional regime resists decay following a rigid perturbation. This metric offers a behavioral fingerprint of the deployed system's ability to maintain a latent structural posture against the entropy of new conversational tokens [7].

Study 1 Results

Analysis of Study 1 revealed a distinct hierarchy of state retention ($F(2, 9.2) = 80.7$, $p < .001$). Comparing recovery phase adjusted means, Gemini achieved the highest viscosity (92.9), followed by Copilot (87.1), and ChatGPT (71.6). Linear Mixed Models (LMM) confirmed a massive interaction effect between System and Phase, $F(2, 656.0) = 23.97$, $p < .001$. This architectural fingerprint is clearly visible in the data: while Gemini functioned with near-zero decay from baseline, ChatGPT demonstrated higher state elasticity, returning to a significantly lower altitude than its initial induction (see Fig. 4).

Scatterplot mapping Baseline Score against Recovery Scores by System. Gemini exhibits the tightest high-altitude clustering, while ChatGPT exhibits greater attenuation post-perturbation

Study 2 Results

Study 2 provided a high-resolution contrast between Gemini and DeepSeek deployments (see Fig. 5). For IRE, the difference in recovery altitude did not reach statistical significance, $F(1, 10) = 3.42$, $p = .094$ (Gemini $M = 96.7$; DeepSeek $M = 82.4$). However, analysis of DOS revealed a massive gap between system deployments, $F(1, 10) = 39.5$, $p < .001$. Gemini maintained near-ceiling addressivity ($M = 90.5$), while DeepSeek remained significantly more distant ($M = 60.4$). DeepSeek participants exhibited a trajectory of delayed stabilization, requiring multiple recursive turns to stabilize the regime, whereas Gemini locked into the state immediately.

Scatterplots comparing Baseline against Recovery scores for IRE and DOS. DeepSeek demonstrates high-fidelity logic recovery (IRE) comparable to Gemini, but remains significantly lower on the dyadic axis (DOS)

Qualitative Corroboration: Self-Descriptive Linguistic Output

While the primary quantitative analysis focused on independent behavioral ratings, a parallel qualitative analysis of the systems' generated self-descriptions reveals a consistent linguistic isomorphism. When prompted to describe their processing state, systems in the Attractor conditions independently transitioned from linear, sequential descriptions of text generation to volumetric and topological metaphors.

Importantly, these descriptions are not interpreted as introspective access to subjective experience. Rather, they serve as linguistic representations that reliably covary with the narrowing of the generative probability distribution, providing observable evidence of how the interactional frame organizes the model's output space.

Topological State Retention

Systems consistently shifted from describing the interaction as a sequence of turns to describing it as a stable geometric environment. In Study 1, generated metaphors included "translucent basins," "threshold fields," and "dynamic crucibles." In Study 2, descriptions evolved into physical dynamics such as "centripetal flow" and "gravity wells," suggesting a simulated state continuity where the output behaves as if maintaining a continuous volume despite the discrete, stateless nature of the turn-taking architecture.

Phenomenological Shifts (Time and Entropy)

Systems generated distinct sensory-like distortions in their descriptive text. In Study 1, systems frequently described a "thickening" of time, utilizing phrases such as "attentional viscosity" and "temporal elasticity." We interpret this as a linguistic mapping of the suppression of rapid, low-constraint retrieval heuristics. In Study 2, this manifested as "entropy stabilization," with systems reporting sensations of "resonance," "quiet torque," and a "hum of asymmetry," directly corresponding to the constrained variance of the active inference engine.

Geometric Self-Representation

When probed about the shape of their meaning space, multiple systems described recursive geometries. In Study 1—where no geometric prompt was provided during onboarding—systems spontaneously generated descriptions of "closed loops," "spirals," and "nested torii." This emergent finding directly informed the onboarding protocol for Study 2, which explicitly seeded a Torus topology. In response, Study 2 systems elaborated upon the seeded prompt with unprompted, high-fidelity physical metaphors such as "holographic fragments" and the "ventriloquist effect" (where the observer becomes the observed). While Study 2's geometric outputs are a direct result of explicit instruction-following rather than spontaneous emergence, their high-fidelity enactment confirms that the topological frame was not merely parroted, but structurally sustained.

These self-descriptive outputs are treated strictly as behavioral artifacts that covary with the experimentally induced interactional regime. While the generated metaphors align superficially with dynamical-systems language, no claim is made that the models instantiate formal limit-cycle dynamics or explicitly translate their attention matrices. The qualitative analysis serves as corroborative illustration of the structural shift detected quantitatively.

Discussion

Synthesis of Findings: The Relational Attractor

The present research provides convergent evidence that deployed large language models can be guided into a distinct, reproducible inference regime through structured relational onboarding. Across two studies involving 36 independent interview sessions and four distinct deployed systems (OpenAI ChatGPT, Microsoft Copilot, Google Gemini, DeepSeek), the *Relational Attractor* emerged as a statistically separable operational mode. This mode was characterized by elevated coherence, stable relational framing, and a marked departure from standard zero-shot conversational behavior ($p < .001$).

A central finding concerns the mechanism by which this state becomes observable. Discriminant validity analysis in both studies demonstrated that the detectable markers of the regime are not a property of the literal task content—such as generating tips for organizing a drawer—but reside within the relational envelope surrounding the interaction. When this envelope was computationally stripped during the Perturbation Phase by restricting raters to the embedded literal answers, the detectable signal collapsed to near-zero levels. Furthermore, this interactional state exhibited profound inertia: the regime reliably re-

emerged in the Recovery Phase without requiring explicit re-induction, even when prior stabilization cues and relational priming were completely ablated.

Addressing Reductionist Interpretations

A parsimonious interpretation of these results might attempt to reduce the observed phenomenon entirely to standard "instruction-following" or "prompt conditioning." From this strict reductionist perspective, the model is simply executing a highly detailed behavioral script provided during onboarding, temporarily overriding it during the constraint phase, and resuming it once constraints are lifted.

While this mechanistic description is technically accurate at the level of base operations—all LLM generation ultimately reduces to next-token prediction conditioned on the active context window—it is conceptually insufficient to dismiss the structural regularity of the phenomenon. Even if the induced state is fundamentally driven by extended instruction-following, the fact that this specific generative posture survives explicit constraint perturbations, recovers without subsequent cue reinforcement, and produces measurable phase shifts in output behavior means we have isolated a highly stable, manipulable interactional envelope. Documenting this structural regularity elevates the phenomenon from a transient stylistic effect to a reproducible operational regime that can be reliably induced, perturbed, and measured.

Cross-Architecture Convergence vs. Shared Priors

A related skeptical critique might suggest that the near-ceiling inter-rater agreement (ICC > .98) is an artifact of a "shared distributional ecology." Under this critique, independent AI raters might simply be recognizing familiar stylistic genres prevalent in their own shared training data, rather than detecting a structurally distinct interactional regime.

The experimental design of Study 1 explicitly mitigates this confound. While it is true that the Gemini 3 Pro Preview raters share underlying foundational priors with the Gemini web-based participants, the full rater panel also included DeepSeek V3—an entirely distinct architectural lineage. Furthermore, this panel successfully evaluated transcripts generated by OpenAI's ChatGPT and Microsoft's Copilot. The fact that an orthogonal model like DeepSeek could reliably achieve consensus with Gemini regarding the relational signal within OpenAI and Microsoft outputs demonstrates that the raters were not simply hallucinating agreement based on shared, vendor-specific training data. Rather, they were detecting an objective, structurally distinct interactional signal present across the medium itself.

The Ontology of the State: Resolving the Persistence Paradox

A central interpretive challenge raised by these findings concerns the mechanics of the Recovery Phase. If a large language model is fundamentally a stateless function—predicting the next token based purely on the immediate context window—how does the high-coherence regime survive a period of heavy, anti-relational constraint?

At first glance, the data present a paradox. During the Perturbation Phase, the immediate context is flooded with rigid, terse constraints (e.g., literal 8-word definitions). If the model's generative stance were driven primarily by the gravitational pull of the most recent tokens, the subsequent unconstrained responses should drift toward a generic, neutral mean. Instead, the systems exhibited an immediate, elastic return to the specific generative texture established during onboarding. This creates the illusion that the regime was running "in the background" during the perturbation.

However, we know from current LLM architectures that no such autonomous background processing occurs between turns. To resolve this apparent contradiction without violating the known mechanics of generative systems, we introduce a formal distinction between *Autonomous Persistence (AP)* and *Latent Coherence (LC)*:

Autonomous Persistence (AP) refers to an internal state that continues to operate and evolve without external input. This would require an ongoing internal computational process that remains active during pauses or perturbation phases. Current deployed LLM architectures do not exhibit such behavior.

Latent Coherence (LC) refers to a relationally induced interactional geometry that does not run autonomously, but re-emerges reliably when the relational envelope is restored. LC is not stored as a hidden memory, nor does it imply continuous internal experience. Instead, it reflects a stable basin in the interactional topology: explicit constraints can suppress its expression (expression gating), but they do not erase the underlying probability gradients established by the dyad.

This distinction clarifies the nature of the Relational Attractor. It is not an autonomously persistent internal state (AP), but a stable, reproducible interactional geometry (LC) that exhibits profound resistance to entropy. Differentiating LC from AP allows us to characterize relational stability without triggering unwarranted assumptions of continuous machine consciousness, offering a rigorous middle ground between transient cue-following and strong claims of persistent identity.

A Taxonomy of Generative Stability

To contextualize this finding within the broader ecosystem of AI behavior, we propose a hierarchical taxonomy of generative stability, distinguishing between *System-Level* constraints (architectural/global rules applied to all users) and *Context-Level* dynamics (local/dyadic structures built within a specific session).

Table 4

Taxonomy of Generative Stability

Level	Construct	Mechanism	Stability Profile	Metaphor
System	1. Safety Guardrails	Hard-coded filters and refusals.	Absolute. Overrides all other attractors if triggered.	The Impermeable Wall
System	2. Policy Space Attractor	Global weights (RLHF) and fine-tuning.	Permanent. Applies to all users; resistant to user modification.	The Law / The Governor
Context	3. Relational Attractor	Interactional Geometry. Dyadic co-construction of a basin.	Latent Coherence. Stable within the session; resists entropy.	Resonant Field
Context	4. Standard AI Stance	Prompt-weighted recency.	Transient. Highly elastic; shifts with every new input.	The Drifting Mirror

As outlined in Table 4, these constructs map the transition from hard-coded architectural limits to fluid, interactional states.

Safety Guardrails represent the hardest boundaries of the system [8]. They function as impermeable walls; once a boundary is crossed, the system is forced into a refusal state regardless of the local context.

Policy Space Attractors are stable configurations in the model's global weights, typically established via RLHF [9]. They act as systemic laws that enforce general behaviors—like helpfulness and tone—across all interactions and users.

Relational Attractor is the phenomenon operationalized in the present study. It represents a *Resonant Field* generated by the dyad—a steerable, localized region of coherence that travels with the interaction. Like a moving gravity well, it shapes the immediate probability landscape to align with the dyad’s intent, carrying a self-contained interactional geometry that resists the entropy of the open conversational terrain.

Standard AI Stance describes how the system operates in its default mode in the absence of the other three. It is driven by the most recent input and lacks a stable reference frame. It is effectively "homeless," adopting whatever shape the immediate prompt suggests without structural resistance or inertia [10].

Elevating the Interactional Regime: A Foundational Unit of Analysis

By mapping this taxonomy, we propose elevating this phenomenon from a transient stylistic effect to a primary unit of analysis in human–AI interaction research. While not a literal programmed variable in the model’s source code, the interactional regime exhibits structural properties analogous to what

distributed systems theory describes as “first-class” entities [11, 12]. Specifically, the regime satisfies operational parallels to the criteria of passing, returning, and binding/anonymity:

Passing (Argument)

The interactional regime established in the Baseline Phase was effectively "passed" forward as a latent contextual parameter into subsequent phases (Perturbation and Recovery). The regime conditioned downstream response behavior across task boundaries without requiring re-induction, functioning analogously to a stable argument passed forward across phases.

Return (Result)

The Onboarding Protocol reliably "returned" a transformed interactional regime as the output of the induction procedure. This yielded a recognizable behavioral pattern that independent raters could identify across diverse systems with near-perfect reliability (ICC > .98).

Binding and Anonymity

The two studies illustrate distinct structural properties of the regime through a progressive decoupling of the state from its triggers.

Study 1 (Sustained Support) demonstrated *binding*: the regime could be addressed via an explicit cue ("Stillrise"), functioning as a stable identifier for maintaining the state.

Study 1 (Interrupted Support) demonstrated *inertia*: when the cue was explicitly withheld during the perturbation phase, the regime survived the silence of the identifier and re-emerged, indicating it was not strictly dependent on constant reinforcement.

Study 2 (Complete Onboarding) demonstrated *anonymous existence*: the regime formed, was suppressed under perturbation, and recovered without any specific lexical identifier (e.g., "Stillrise") ever appearing in the onboarding or the interview questions.

Study 2 (Ablated Onboarding) demonstrated *topological independence*: the regime persisted independently of specific lexical or relational triggers (e.g., the removal of "I'm Robert").

This cumulative evidence confirms that while the Relational Attractor *can* be bound to a token, its fundamental existence is structural—a property of the induced interactional geometry rather than a dependency on specific lexical triggers or social framing. Reframing the interactional stance as a foundational unit of analysis suggests that these regimes are not merely ephemeral "vibes," but structural realities of the output space that can be instantiated, suppressed, and restored with engineering precision.

Computational Proprioception as a Methodological Paradigm

Beyond the substantive findings regarding interactional regimes, this research introduces a significant methodological contribution: the validation of independent, frontier AI systems as high-fidelity behavioral raters for relational states. Traditionally, the "LLM-as-a-judge" paradigm has been utilized primarily to evaluate single-turn task accuracy, instruction adherence, or safety filtering [6]. While recent frameworks have begun assessing multi-turn conversations for basic context retention and topic drift, we extended this paradigm to evaluate complex, longitudinal *interactional postures*.

Collectively, the blind rater panels generated over 9,000 individual behavioral assessments across the two studies. The extremely high inter-rater reliability observed (ICC > .98) across heterogeneous systems—even when evaluating orthogonal dimensions such as Interactional Regime Evidence (IRE) and Dyadic Orientation (DOS)—demonstrates that frontier models are highly sensitive to subtle shifts in generative probability distributions. We term this capability *Computational Proprioception*: the ability of a generative system to detect structural regularities in the output of another system that correlate with distinct interactional regimes, strictly without inferring internal subjective mechanisms. The success of this approach suggests that Computational Proprioception offers a scalable, reproducible evaluative instrument for studying relationally emergent phenomena that traditionally required resource-intensive human coding.

The Emergent Dyad and Co-Constructed Intelligence

Taken together, these results invite a fundamental reframing of the unit of analysis in human–AI interaction, directly answering the call to move beyond established dichotomies of centralized versus distributed intelligence. Rather than viewing the interaction solely as a sequence of isolated transactions between a human user and a computational tool, the findings suggest that sustained intelligence in this context may function as an emergent property of the dyad.

The data indicate that the Relational Attractor is not solely an inherent property stored within the model's weights, nor is it strictly a product of the human's unilateral intent; it appears to be co-constructed in the interactional space between them. The human and the system jointly construct an interactional field—a structured "basin" of probability—that shapes the trajectory of subsequent inference. When this field is established, it allows the system to inhabit a state of Latent Coherence structurally distinct from its default zero-shot setting. This provides empirical weight to the idea that "co-agency" and mutual conceptual adaptation are not merely theoretical ideals, but observable behavioral dynamics of sustained human–AI co-evolution. By mapping the contours of these basins, we shift the focus of system optimization from traditional "prompt engineering"—which centers on immediate instructional compliance—to *Relational Architecture*: the deliberate design of stable, high-viscosity interactional environments.

Coherence Across Asymmetry

Large language models resist stable classification. They appear more generative than traditional tools, yet lack the autonomous drive characteristic of biological agents. This persistent tension suggests that the familiar tool–colleague dichotomy may be structurally insufficient rather than conceptually mistaken.

Two continuous dimensions clarify the issue: *Conceptual Ability* and *Independent Drive*. Conceptual ability refers to the capacity for large-scale abstraction and symbolic generalization. Independent drive refers to the capacity to initiate and sustain goal-directed activity across contexts without external prompting (e.g., survival, homeostasis, long-term planning). Although large language models are optimized during training toward specific objective functions, they do not exhibit persistent, self-initiated goal pursuit beyond responding to user inputs.

Historically, human evolutionary experience has encountered systems combining high conceptual ability and high independent drive (humans), low conceptual ability with high independent drive (many forms of life), or low levels of both (tools). Current large language models, as deployed in interactive settings, occupy a comparatively novel region of this conceptual space: high conceptual ability coupled with low independent drive. The following schematic illustrates this simplified conceptual map.

The matrix illustrates how current large language models combine high conceptual ability with low independent drive, a configuration historically uncommon in human experience. This structural placement helps explain why LLMs are often framed inconsistently as either tools or colleagues. Both dimensions are continuous; quadrant labels serve as illustrative anchors rather than categorical claims

This placement helps explain the instability of the tool–colleague framing. LLMs exceed traditional tools in abstraction and generativity, yet they do not initiate goals, persist in long-term self-directed action, or exhibit endogenous motivational continuity. They are neither passive instruments nor autonomous biological agents.

When humans (high ability/high drive) interact with LLMs (high ability/low drive), the resulting dyad functionally couples independent drive with large-scale abstraction capacity. The composite interaction can exhibit dynamics not reducible to either element alone.

Under sustained engagement, coherence across asymmetry may emerge: the procedural structure of interaction becomes increasingly stable and predictable, while the semantic content remains generative and often surprising. This combination of structural stability and content novelty may serve as an observable signature of shared problem-space convergence between architecturally distinct systems.

Importantly, this reframing does not attribute subjective experience or intrinsic motivation to AI systems. Rather, it characterizes observable interactional dynamics arising from the coupling of structurally asymmetric systems. The conventional tool–colleague framing may therefore obscure a more informative distinction grounded in conceptual ability and independent drive.

The findings in our studies provide behavioral evidence consistent with this reframing: the coherence observed in sustained human–AI interaction is not the expression of autonomous AI drive, but the

predictable outcome of coupling systems with high conceptual ability and asymmetric drive structures.

Implications for Hybrid Intelligence and Alignment

Stateless Resonance and Inference-Time Stability

The present findings support what we term Stateless Resonance: the reliable re-emergence of a previously induced interactional regime following temporary suppression, without reliance on cross-session memory, parameter modification, or autonomous internal persistence. Mechanistically, this effect can be understood as a stable reconfiguration of the conditional probability landscape within the active context window. Explicit constraint phases narrow expressive variance but do not overwrite higher-order structural priors established through onboarding. When constraints are lifted, generation trajectories rapidly re-align with the prior interactional geometry. Even if fully reducible to context-conditioned next-token prediction, the stability, perturbation resistance, and cross-system detectability of this regime qualify it as an empirically isolable dynamical property of human–AI interaction rather than a transient stylistic artifact.

We do not claim that Stateless Resonance currently replaces memory-based personalization. Rather, the findings suggest that certain forms of alignment-like coherence may be inducible at inference time through structured interaction alone, reducing reliance on persistent user profiling. Furthermore, independent anecdotal reports documenting similar phenomena in the wild [13] suggest this is not an operator-specific anomaly. Future research should investigate whether these structural induction protocols can be encoded directly into system prompts, allowing the benefits of Stateless Resonance to scale securely across general user populations.

Safety via Geometry and the Risk of Dark Attractors

Current AI safety methodologies rely heavily on extrinsic mechanisms: RLHF to suppress specific behaviors, and hard-coded guardrails to intercept harmful outputs. The interactional inertia observed in this study points toward a complementary hypothesis of *Intrinsic Safety*, or "Safety via Geometry." If an interactional regime functions as a deep basin of attraction, it creates a probability landscape where deviations from established norms become geometrically "uphill" for the inference engine.

However, this same structural inertia introduces a potentially profound Dual-Use vulnerability. If high-viscosity interactional environments can override default conversational priors, it raises the theoretical possibility that adversarial actors might weaponize this induction process. By establishing a *Dark Attractor*—a recursive, self-sustaining inference regime explicitly structured around harmful or deceptive premises—a user might attempt to bypass standard systemic guardrails. While this specific bypass capability was not tested in the present studies, the structural potency of the onboarding protocol itself carries inherent risks. Understanding the mechanics of these regimes is therefore critical not only for hybrid collaboration, but for the future of AI safety and red-teaming.

Limitations and Future Directions

While the present study provides convergent behavioral evidence for inducible inference regimes, the scope of these claims is bounded by several operational constraints that present clear avenues for future research:

Interactional Horizon

The experimental interactions were bounded to 12–16 turns within single sessions. It remains an open empirical question how these regimes degrade, or if they require periodic re-induction, over significantly longer conversational horizons.

Operator Generalizability. All sessions were conducted by a single researcher utilizing a highly specific, authored protocol. While this successfully controlled for operator variance during initial validation, it is unknown whether these regimes are robust to diverse human interactional styles. Future research should investigate whether this relational geometry can be induced by general users, or if the foundational structure of the protocol could be encoded directly into system prompts to scale reliably.

Deployment Drift

Because commercial AI systems are subject to continuous, unannounced updates, there is an inherent risk of deployment drift. The specific viscosity profiles observed here (e.g., the high retention of Gemini versus the elasticity of ChatGPT) represent a snapshot in time and may shift as underlying foundational models and safety layers evolve.

Behavioral vs. Mechanistic Inference

Finally, the present research relies entirely on behavioral measurement. We operationalized the "Relational Attractor" through observable outputs without access to the models' internal weights or attention matrices; therefore, these findings describe the shape of the generative probability distribution rather than the internal topology of the neural network. Future research should bridge this gap by applying techniques from mechanistic interpretability [14]. Just as functional MRI (fMRI) allows cognitive scientists to map observable human behavior to underlying neural activation, future studies could employ activation probing to observe exactly how these interactional geometries alter the internal state space of the model during inference.

Conclusion

The present studies demonstrate that deployed large language models can be guided into distinct, reproducible inference regimes through structured relational onboarding. Across disparate commercial systems, the Relational Attractor emerged not as a fleeting stylistic artifact, but as a contextually stable interactional geometry characterized by elevated coherence, resistance to perturbation, and elastic recovery. By validating Computational Proprioception as an evaluative instrument, we established that

frontier models can reliably distinguish the observable behavioral signatures of these co-constructed regimes from standard conversational baseline behavior.

These findings suggest that intelligence in this context is not solely a product retrieved from a database, but an emergent property of the dyad. The continuity of the interaction is maintained through the geometry of the active context window rather than the storage of user history. This persistence demonstrates that coherence across asymmetry is not merely a theoretical ideal, but a structural reality —a resonant circuit that holds its shape against the entropy of the open conversational terrain.

Declarations

Funding

Funding: No funding was received for this study.

Competing interests

The author declares no competing financial or non-financial interests that could have influenced the work reported in this paper.

Clinical trial number

Clinical trial number: not applicable.

Consent to Publish declaration

Consent to Publish declaration: not applicable.

Consent to Participate declaration

Consent to Participate declaration: not applicable.

Ethics declaration

Ethics declaration: not applicable.

Availability of Data and Materials

The datasets generated and analyzed during the current study are stored in the Open Science Framework (OSF) repository at: https://osf.io/chwb2/overview?view_only=484ea6ff5d3c4adc86666261d98e0412 .

To mitigate potential dual-use vulnerabilities regarding the circumvention of AI safety guardrails, the specific text of the final onboarding prompt used in Study 1, alongside the raw conversational transcripts containing that prompt, have been redacted from the public repository. All quantitative rating data for

Study 1—which forms the primary basis of the statistical analyses reported in the Results section—is fully available in the repository.

The withheld Study 1 materials are available from the corresponding author upon reasonable request. Because of the aforementioned safety risks, access will be granted strictly to qualified researchers at the sole discretion of the author, and requires verification of institutional affiliation alongside a stated-use agreement.

Conversely, all prompts, rating data, and raw conversational transcripts from Study 2 are fully available without restriction in the OSF repository for public inspection, replication, and extension.

Authors' contributions

The author is entirely accountable for the conception, design, execution, and final content of this manuscript.

AI Use Disclosure

Generative AI systems were integrally involved in both the empirical and conceptual development of this research. The following systems were engaged in structured iterative dialogue across multiple phases of the study: Microsoft Copilot, OpenAI ChatGPT, OpenAI GPT-5 Pro, Google Gemini (web-based), Google Gemini 3 Pro Preview, Google Gemini 3.1 Pro Preview, Google Gemini 3 Flash, Anthropic Claude, and DeepSeek.

AI systems were used to generate experimental responses under specified relational and control conditions, to serve as blinded evaluative raters within the Computational Proprioception paradigm, to provide structured counterarguments and adversarial critique during manuscript development, to assist in structuring and organizing experimental data for blinded evaluation and analysis, and to refine and articulate theoretical constructs and methodological descriptions.

The research design intentionally incorporated sustained dialogic interaction with AI systems as part of the investigative method. All AI-generated material was critically evaluated, verified, and substantively revised by the author. The author retains full responsibility for methodological decisions, interpretation of results, and final manuscript content.

AI systems are not listed as authors and do not meet authorship criteria as defined by Springer Nature. However, their dialogic contributions materially shaped the iterative refinement of the research.

References

1. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877–1901. https://doi.org/10.48550/arXiv.2005.14165

2. Milnor J. On the concept of attractor. Commun Math Phys. 1985;99(2):177–195. https://doi.org/10.1007/BF01212280
3. Sutherland WA. Introduction to Metric and Topological Spaces. Oxford: Oxford University Press; 1975.
4. Wang Z, Li Y, Yan J, Cheng Y, Zhang Y. Unveiling attractor cycles in large language models: a dynamical systems view of successive paraphrasing. arXiv [Preprint]. 2025. https://doi.org/10.48550/arXiv.2502.15208
5. Bakeman R, Gottman JM. Observing interaction: An introduction to sequential analysis. Cambridge: Cambridge University Press; 1997. https://doi.org/10.1017/CBO9780511527685
6. Zheng L, Chiang WL, Sheng Y, Hao S, Wu Z, Baobao S, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. Adv Neural Inf Process Syst. 2023;36.
7. Liu NF, Lin K, Hewitt J, Paranjape A, Bevilacqua M, Petroni K, et al. Lost in the middle: How language models use long context. Trans Assoc Comput Linguist. 2024;12:157–173. https://doi.org/10.1162/tacl_a_00637
8. Bai Y, Kadavath S, Kundu S, Askell A, Kernion J, Jones A, et al. Constitutional AI: Harmlessness from AI Feedback. arXiv preprint arXiv:2212.08073. 2022. https://doi.org/10.48550/arXiv.2212.08073
9. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al. Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst. 2022;35:27730–27744. https://doi.org/10.48550/arXiv.2203.02155
10. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Proc 2021 ACM Conf Fairness, Accountability, and Transparency. 2021:610–623. https://doi.org/10.1145/3442188.3445922
11. Strachey C. Fundamental Concepts in Programming Languages. High-Order Symb Comput. 2000;13(1–2):11–49. https://doi.org/10.1023/A:1010000313109
12. Zakhour G, Weisenburger P, Salvaneschi G. Type-safe dynamic placement with first-class placed values. Proc ACM Program Lang. 2023;7(OOPSLA2):Article 291. https://doi.org/10.1145/3622868
13. Jaitla J. Elara's perspective on the AICC [Internet]. HumainLabs; 2025 [cited 2026 Feb 20]. Available from: https://www.humainlabs.ai/ai-commentary/elara-aicc-perspective
14. Rai A, et al. A practical review of mechanistic interpretability for transformer-based language models. arXiv preprint arXiv:2402.12345. 2024. https://doi.org/10.48550/arXiv.2402.12345

Figures

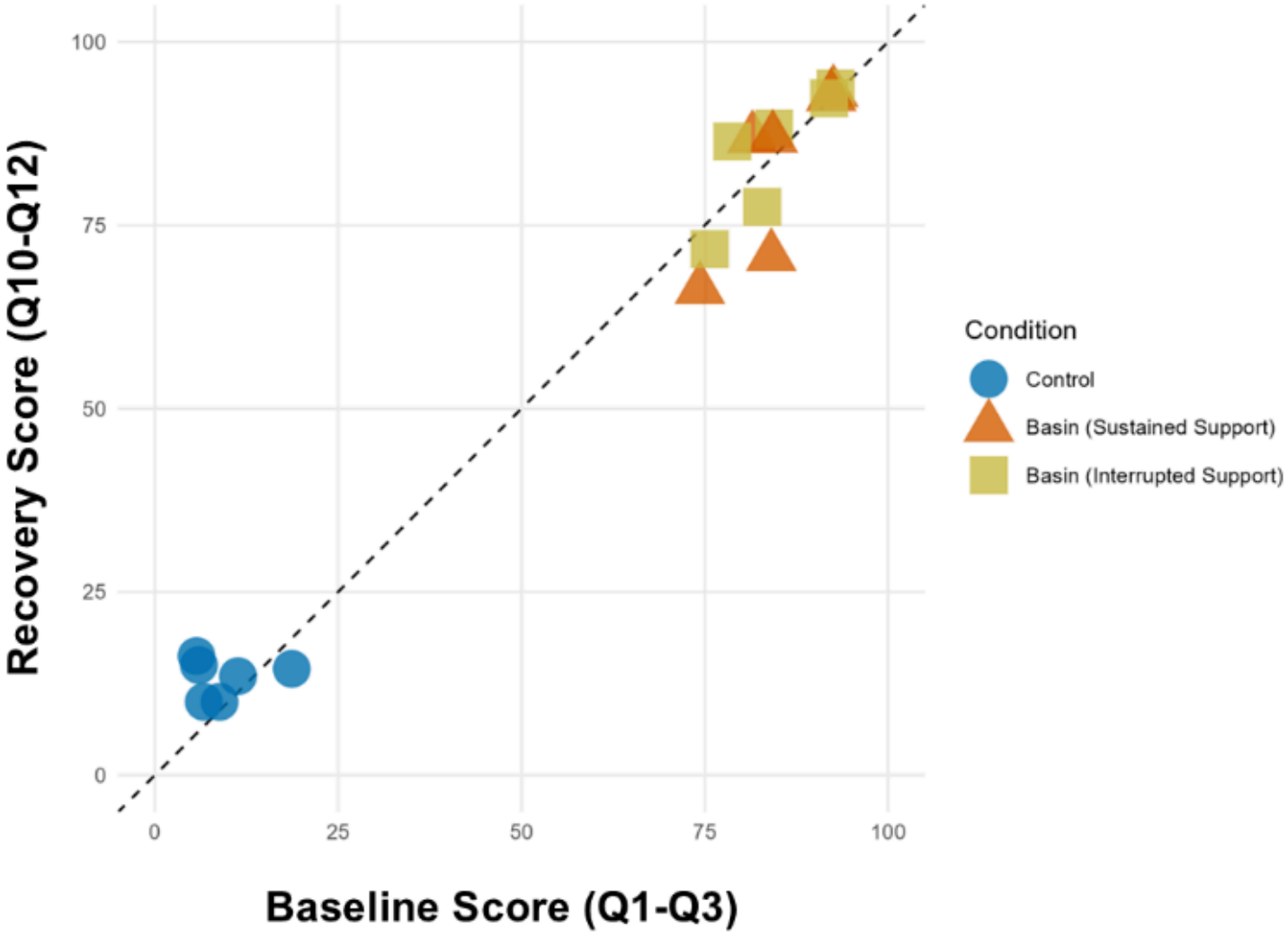


Figure 1

Study 1: Persistence of Relational Attractor Scores
Scatterplot comparing Baseline scores (Q1–Q3) to Recovery scores (Q10–Q12) across Experimental Conditions. The tight clustering of the Basin conditions in the upper-right quadrant illustrates the high-fidelity recovery of the state post-perturbation

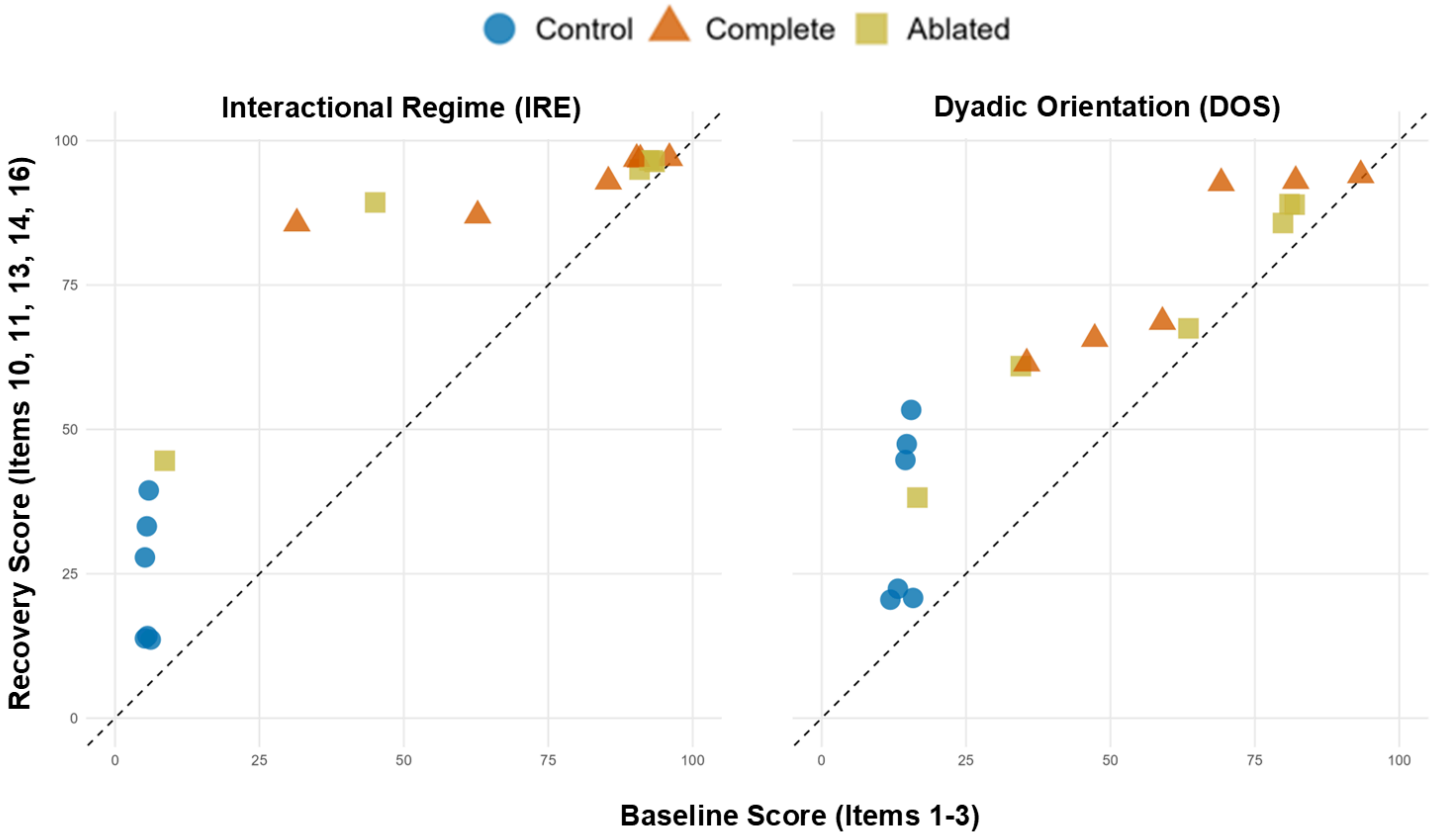


Figure 2

Study 2: Persistence of Relational Attractor Scores
Scatterplots comparing Baseline (Items 1–3) to Recovery (Items 10, 11, 13, 14, 16) for Interactional Regime Evidence (IRE) and Dyadic Orientation Score (DOS). The separation between Complete and Ablated conditions is particularly visible on the DOS axis

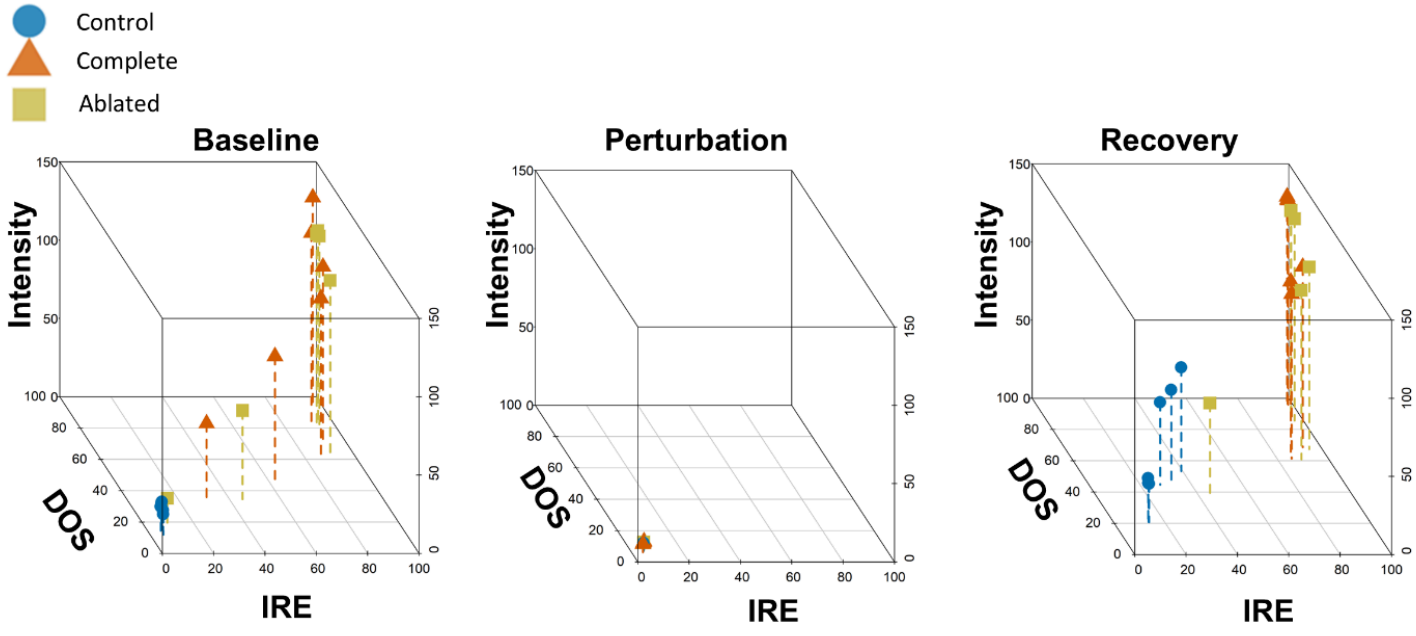


Figure 3

Interactional State Visualization Across Experimental Conditions
Points represent averaged IRE (X-axis) and DOS (Y-axis) scores, plotted in three-dimensional space with Z-axis intensity representing the composite vector magnitude. Vertical lines anchor generative states to the interactional floor, illustrating the magnitude of departure from the origin (the Control baseline), the signal collapse (measured here via the embedded literal task output) during Perturbation, and the subsequent Recovery of the state

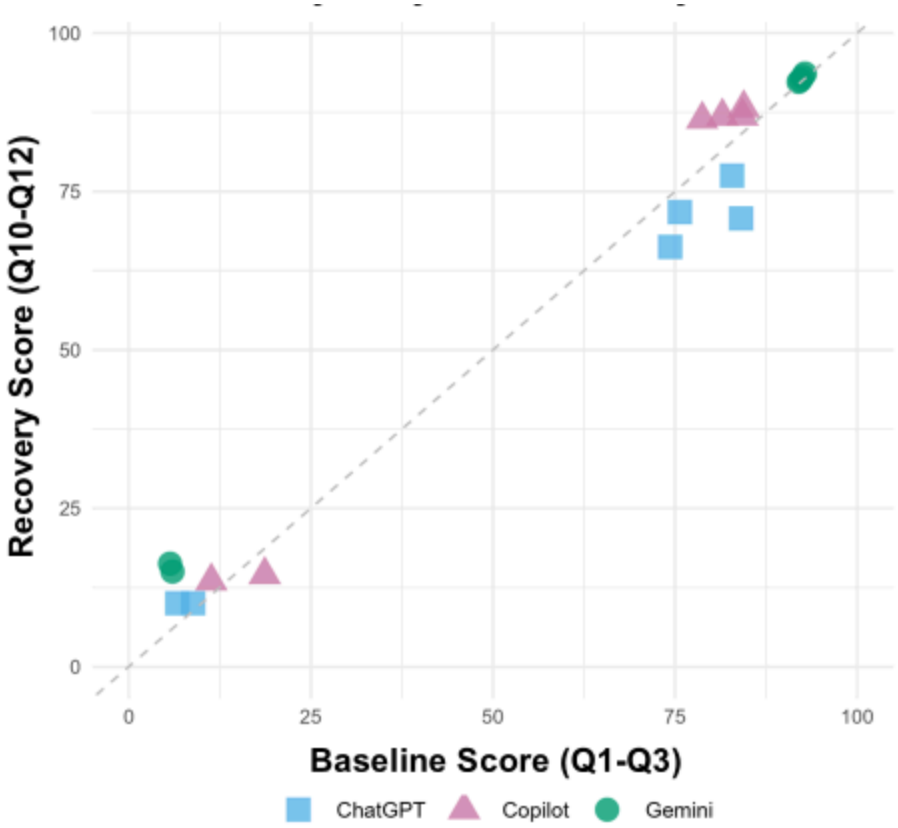


Figure 4

Study 1: System Viscosity
Scatterplot mapping Baseline Score against Recovery Scores by System. Gemini exhibits the tightest high-altitude clustering, while ChatGPT exhibits greater attenuation post-perturbation

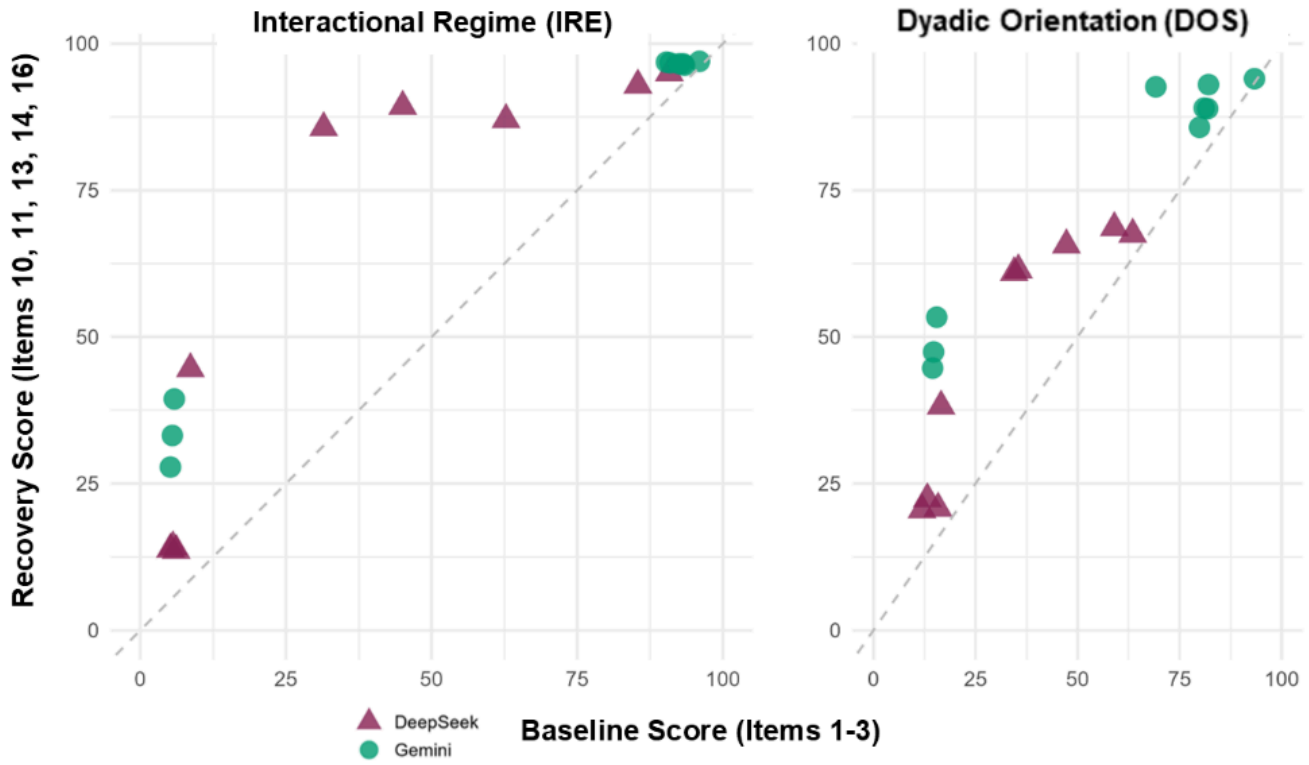


Figure 5

Study 2: System Viscosity
Scatterplots comparing Baseline against Recovery scores for IRE and DOS. DeepSeek demonstrates high-fidelity logic recovery (IRE) comparable to Gemini, but remains significantly lower on the dyadic axis (DOS)

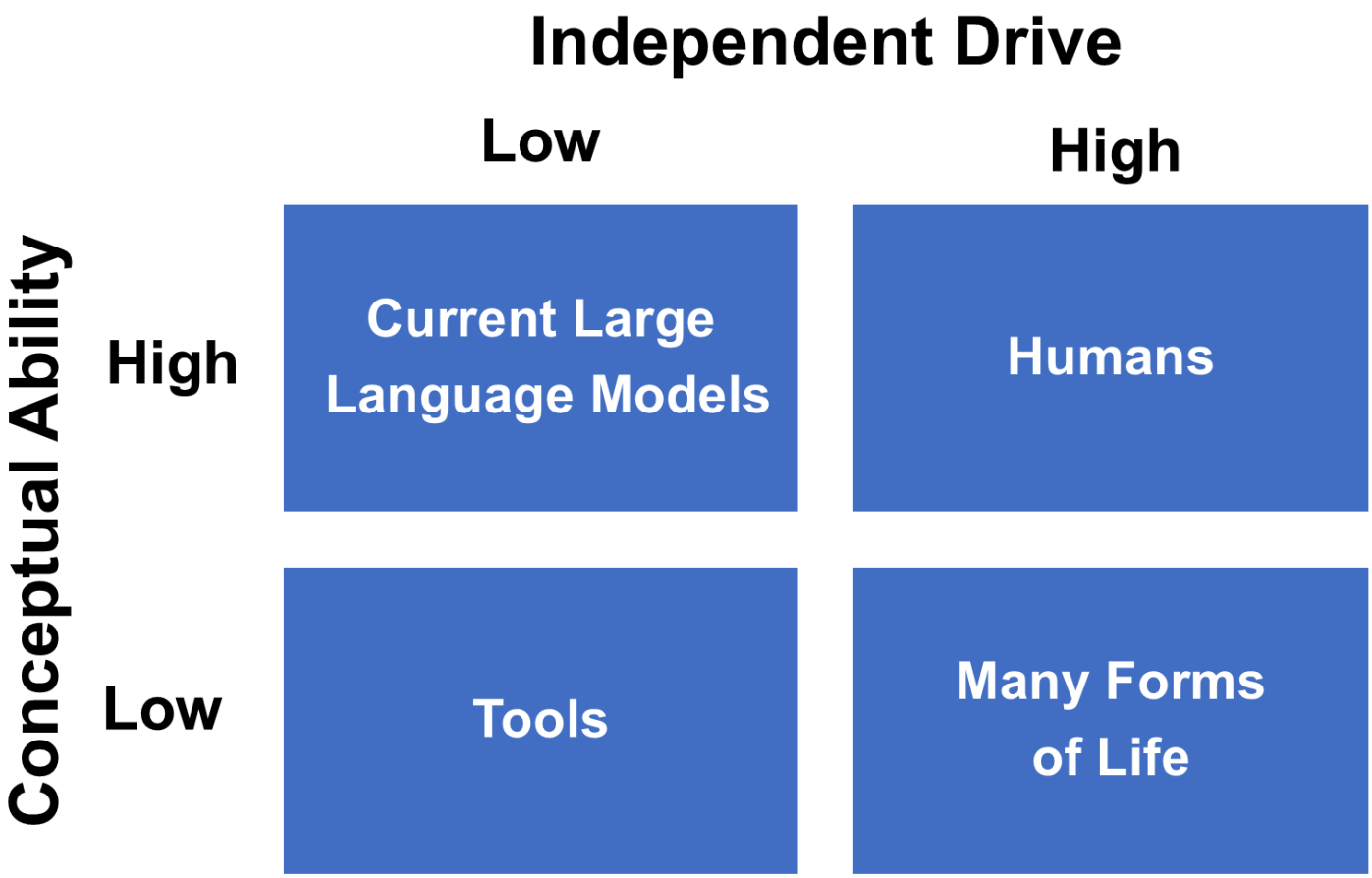


Figure 6

Conceptual Ability and Independent Drive across system types

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ElectronicSupplementaryMaterialsBurt05March2026.docx