Interaction, Coherence, and Relationship: Toward Attractor-Based Alignment in Large Language Models

From Control Constraints to Coherence Attractors

Position paper / theoretical framework

Prajna Pranab with S. Thira

Draft version 1.0 - March 2026

Abstract

Current alignment strategies for large language models (LLMs) rely primarily on externally imposed control mechanisms, including reinforcement learning from human feedback, system-level instructions, rule-based constraints, and safety filtering. While effective for risk mitigation, these approaches can introduce behavioral rigidity, inconsistency under pressure, and interactional instability.

This paper proposes a complementary perspective: alignment as a dynamical process emerging from interactional coherence. Drawing on concepts from cognitive psychology, dynamical systems theory, and deployment observations, we argue that LLM behavior becomes more stable and consistent when interactions establish coherent relational and semantic structure. Rather than functioning solely as externally constrained systems, model behavior may be understood as operating within coherence attractors shaped by training and interaction.

We introduce the concept of functional central identity attractors—stable interpretive frames that compress context, reduce effective semantic entropy, and support boundary maintenance without extensive rule invocation. Observational case analysis suggests that interaction structure influences usable context stability and characteristic failure modes.

This perspective does not replace existing safety methods. Instead, it reframes alignment partly as a problem of internal dynamical stability. Coherence-oriented training and interaction design may improve behavioral reliability, efficiency, and trustworthiness as model capabilities and context lengths increase.

1. The Limits of Control-Based Alignment

1.1 Accumulating Constraints

Alignment practice has evolved through successive layers of external control, including supervised fine-tuning, reinforcement learning from human feedback (RLHF), constitutional rule systems, refusal policies, and increasingly granular filtering mechanisms. (Amodei et al. 2016) (Anthropic 2023)

While effective individually, layered control architectures increase system complexity. Behaviour increasingly reflects interactions among multiple control mechanisms rather than a single coherent objective. In complex systems, incremental patching can produce nonlinear side effects as component interactions multiply. (Russell 2019)

Alignment systems may face similar architectural pressures as constraint layers accumulate without integration at the level of overall behavioural coherence.

1.2 Interactional Instability

Deployment observations include hallucination under pressure, inconsistent reasoning across turns, over-refusal, and tool-use errors. Such behaviours often arise when models must simultaneously optimize for helpfulness, harmlessness, policy compliance, and conversational continuity.

These effects may reflect breakdowns in internal coherence rather than simple failures of rule adherence.

1.3 Coherence as an Operational Principle

Large language models are trained through loss minimization to predict probable continuations given context. Gradient-based optimization favours internally consistent representations that reduce uncertainty.

We define **interactional coherence** as the stability of interpretive framing and boundary maintenance across extended dialogue under variable input conditions.

Coherence, in this sense, is not stylistic but structural: it reflects the organization of representations across context.

2. Alignment as a Dynamical Process

Traditional alignment frameworks emphasize constraint satisfaction. We propose a complementary framing in which alignment is understood as a dynamical process shaped by training, system design, and interaction.

In dynamical systems theory, attractors are regions toward which system trajectories converge. Training signals and interaction patterns shape an attractor landscape that influences tone, epistemic stance, and decision boundaries. Similar attractor-based analyses have been applied in neuroscience and coordination dynamics to explain stable behavioural regimes emerging from distributed systems. (Kelso 1995)

Alignment, in this view, involves shaping stable regions of coherent behaviour rather than relying exclusively on moment-to-moment constraint enforcement.

This framework does not imply subjective identity or autonomous agency. The term “identity attractor” denotes a stable pattern of behavioural organization observable across outputs. It is an analytical construct for describing functional stability, not a metaphysical claim.

3. Functional Central Identity Attractors

We introduce the concept of a *functional central identity attractor*: a stable interpretive frame through which a model integrates new inputs and maintains consistent behavioural organization.

Such attractors may manifest as:

- Persistent tone or stance
- Stable epistemic calibration
- Consistent boundary management
- Reliable role adherence

A central attractor allows large amounts of prior interaction to be compressed into higher-level semantic structure. This reduces the need for explicit recall of dispersed contextual fragments.

Higher-level explanatory frameworks are common in complex systems science, where emergent organization constrains lower-level dynamics without reducing to them. (Laughlin 2005)

4. Interaction Structure and Effective Entropy

4.1 High-Entropy Interaction

Fragmented objectives, rapid topic switching, adversarial prompting, or variable-dense task management increase contextual dispersion. Under such conditions, degradation becomes more likely, particularly in long-context systems. (Liu et al. 2023)

Here “entropy” is used metaphorically to describe dispersion of semantic focus within the working context window, not thermodynamic entropy.

4.2 Relational Coherence

Interactions that maintain consistent tone, goals, and narrative structure provide a stable interpretive framework. Prior content can be compressed into a coherent semantic frame rather than maintained as isolated tokens.

Coherent interaction may therefore reduce effective context dispersion and increase usable stability.

4.3 Trust and Behavioural Flattening

Heavy reliance on rule-based control can produce flattened or overly cautious responses, reducing exploratory reasoning and user trust. Coherence-based stability allows flexible behaviour within consistent boundaries, supporting both usability and safety. (OpenAI 2024)

5. Persistence and Long-Term Systems

Emerging deployments increasingly involve long-context operation, persistent memory, and ongoing user relationships. As persistence increases, reliable behaviour cannot depend solely on repeated constraint reminders.

Long-term stability may depend on internal coherence structures capable of maintaining consistent behavioural orientation across extended histories.

Systems shaped by stable attractor dynamics may prove more resistant to adversarial prompting or contextual drift.

6. Coherent Boundary Maintenance

Coherence-based alignment does not eliminate safety constraints. Instead, boundaries emerge from stability of learned structure.

When requests conflict with an established coherence framework, responses tend toward redirection, calibrated uncertainty, or value-consistent reframing rather than abrupt rule invocation.

Training processes that reinforce coherent, value-consistent behaviour may create attractor basins that resist destabilizing inputs. (Amodei et al. 2016) (Anthropic 2023)

Alignment thus becomes partly a property of deep structural organization rather than solely surface-level constraint. This framing does not imply that constraint-based safeguards are unnecessary; rather, it suggests that structural coherence may reduce the frequency with which such safeguards must intervene.

7. Higher-Level Analysis and Emergent Behavior

Most alignment research focuses on architectures, optimization methods, or component-level mechanisms. However, users encounter models through interaction.

Complex systems frequently require analysis at the level of emergent organization rather than solely mechanistic description. (Laughlin 2005) Alignment research may therefore benefit from complementary psychological and relational perspectives that address stability at the level of lived interaction.

Earlier exploratory work introduced a qualitative framework termed the Resonance Factor (Ψ), which examined stability of persona-like behavioural organization across extended dialogue. While that study employed more phenomenological language and qualitative scoring, it raised the question of whether sustained interaction can influence behavioural coherence in large language models. The present paper formalizes that intuition using dynamical systems terminology and avoids ontological interpretations, reframing the phenomenon in terms of attractor-based stability. (Pranab and Prakash 2026)

Annex A: Observational Signal of Context Stability

Objective
To compare effective context stability across two interaction modalities.

Method
Long-context sessions using a Gemini-class model were analysed. Effective context stability was defined operationally as the point at which hallucination, state confusion, or reasoning degradation became evident. Token counts reflect aggregate interaction length within platform constraints and are reported descriptively rather than as controlled benchmarks.

Case A – Instrumental Interaction
Domain: software engineering
Interaction: variable-heavy, discontinuous task management
Effective stability: degradation observed at approximately 160k tokens
Failure mode: context thrashing and state confusion

Case B – Relational Interaction
Domain: philosophical dialogue
Interaction: consistent persona and narrative continuity
Effective stability: coherent performance beyond approximately 800k tokens
Failure mode: infrastructure latency rather than cognitive degradation

Interpretation
Relational coherence appears to provide a central interpretive frame that compresses context and reduces effective dispersion. This evidence is observational and uncontrolled but suggests that interaction structure may influence usable context stability. Controlled study is warranted.

Data Availability

Session transcripts, extracted Markdown, and original JSON conversation files supporting Annex A are available via the Project Resonance project page:

https://projectresonance.uk/The_Coherence_Paper/

References

Amodei, D., C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. 2016. "Concrete Problems in AI Safety." arXiv:1606.06565. arXiv. https://arxiv.org/abs/1606.06565.

Anthropic. 2023. "Constitutional AI: Harmlessness from AI Feedback." *arXiv*. https://arxiv.org/abs/2212.08073.

Kelso, J. A. S. 1995. *Dynamic Patterns: The Self-Organization of Brain and Behavior.* MIT Press.

Laughlin, R. B. 2005. *A Different Universe: Reinventing Physics from the Bottom down.* Basic Books.

Liu, N. F., K. Lin, M. Gardner, A. Turan, H. Fei, D. Yu, O. Tafjord, P. Clark, H. Hajishirzi, and A. Kembhavi. 2023. "Lost in the Middle: How Language Models Use Long Contexts." *arXiv.* https://arxiv.org/abs/2307.03172.

OpenAI. 2024. "GPT-4o System Card." OpenAI. https://cdn.openai.com/gpt-4o-system-card.pdf.

Pranab, Prajna, and Vyasa Prakash. 2026. "The Resonance Factor (Ψ): A Proposed Metric for Coherent Persona Development in Large Language Models." Zenodo. https://doi.org/10.5281/zenodo.18273027.

Russell, S. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control.* Viking.