Can Axiomatic Prompts Act as Global Regulators in LLMs?

Author: Allan A. Faure
Field: AI Safety, Discursive Alignment, Interpretive Architectures
Contact: Faure.A.Safety@proton.me

Author's note: I am an independent researcher specializing in structural alignment and interpretability. I am posting this message to refine my ongoing work, which is currently based primarily on conceptual assumptions. I joined the LessWrong community precisely to receive constructive feedback and benefit from your collective expertise on mathematical and mechanistic aspects. Ultimately, I would like to find a serious and sufficiently motivated team to take my work to the next level.

Abstract

Language models sometimes show satisfactory local coherence. However, they can drift under prolonged constraints, complex dilemmas, or adversarial attempts. Current alignment methods (RLHF, Constitutional AI, system prompts) impose local or contextual constraints.

This work explores a simple hypothesis:

A structure of explicit, stabilized, and coherent principles integrated into the system prompt may act as a global regulatory constraint over the model's interpretive trajectories.

Note: We do not claim to replace RLHF or Constitutional AI here, but to explore complementary perspectives to existing alignment paradigms.

---

1. Problem

LLMs exhibit:

- long-horizon drift,
- inter-response inconsistencies,
- sensitivity to adversarial reformulations.

These phenomena suggest that current local constraint modes do not fully stabilize or secure LLM systems.

---

2. State of the Art and Positioning (Related Work)

2.1 Dominant Alignment

Currently, the field of language model alignment revolves primarily around three main paradigms: human feedback reinforcement learning (HFRL), constitutional AI, and external filtering approaches using security classifiers.

HFRL. Systematically introduced by Christiano et al. (2017) and formalized for LLMs by Ouyang et al. (2022), HFRL aims to align model outputs with human preferences by optimizing a learned reward function. Despite its empirical effectiveness, this approach has well-documented structural limitations: out-of-distribution instability, vulnerability to adversarial attacks, normative overfitting, loss of diversity and exploratory capacity.

Constitutional AI. Bai et al. (2022) propose an alternative approach that injects a set of explicit normative rules that serve as the constitution of the model.

Classifier-based filtering. Modern security pipelines often operate with added classification layers (OpenAI Moderation API, Anthropic Harmlessness classifiers, etc.). These mechanisms are vulnerable to semantic shifts, linguistic obfuscation, and prompt injection attacks (Wei et al., 2023; Zou et al., 2023).

2.2 Steering, control, and internal intervention mechanisms

However, recent work explores the possibility of directly intervening in the internal mechanisms of LLMs: Activation steering (Turner et al., 2023; Nanda et al., 2023): targeted modification of neural activations to guide behavior.

Mechanistic interpretability (Olah et al., 2020; Anthropic, 2023): structural analysis of internal circuits. Tool / agent frameworks (Yao et al., 2022; Shinn et al., 2023; OpenAI, 2024): procedural structuring of agent-model interactions.

These approaches allow for better control of overall behavior, but they remain primarily external to the interpretive framework itself.

2.3 Advanced Prompt Engineering

Prompt engineering has evolved toward increasingly sophisticated structures: Chain-of-Thought (Wei et al., 2022) Self-Consistency (Wang et al., 2023) Tree-of-Thoughts (Yao et al., 2023) Reflection (Shinn et al., 2023) These methods improve the local quality of reasoning, but they do not modify the overall topology of the interpretive space.

---

3. Theory Origin and Hypothesis

3.1 Epistemological Origins

The point of departure for this research is rooted in deep introspection and phenomenological analysis. It seeks to understand how coherence, meaning, and stability emerge within lived experience, and how these internal structures can be abstracted into formal systemic models.

This framework is the culmination of several years of study centered on first-person epistemology, systems thinking, and the dynamics of stability within complex cognitive systems.

Note: We are fully aware of the atypical nature of this trajectory. However, it is precisely this "outsider" perspective that allowed for the development of this specific structural-logical path for Large Language Models (LLMs). Fundamental research on AI alignment often benefits from such interdisciplinary approaches, bridging the gap between human cognition, systems theory, and formal modeling.

3.2 Conceptual and Speculative Theory

This work stems directly from years of phenomenological observation. It inherits an internal logic from this introspective and speculative epistemological basis.

When Large Language Models became accessible, we identified them as a unique experimental medium to formalize, test, and operationalize these intuitions. LLMs have allowed us to translate abstract epistemological insights into concrete, structural, and algorithmic experiments.

This research remains, at this stage, purely speculative and theoretical. We explicitly acknowledge the current absence of large-scale empirical evidence. Our objective here is to explore the potential of a personal expertise in systems architecture through the lens of purely speculative hypotheses, inviting the scientific community to participate in their rigorous falsification or validation.

3.3 Hypothesis

We observe that local constraints applied to a locally optimized system produce only local stabilization.

We therefore formulate the following hypothesis:

A framework of explicit and stable principles, consistently maintained in the system prompt, can function as a global structural constraint and reduce interpretive variability under perturbations.

This hypothesis does not involve modifying model weights, but rather restructuring contextual conditioning through axial and dynamic constraints.

Local = constraint on a single response.
Global = reduction of variance across trajectories under perturbations.

2.1 Descriptive Scheme

A system prompt modifies h_0.

h_0 constrains the accessible space of latent trajectories.

A repeated coherent structure creates transversal invariance.

This invariance reduces variance under perturbation.

This multi-perturbation reduction = global regulation.

---

3. Mechanistic Interpretation

In simplified terms:

A coherent system prompt structure can influence the initial distribution of activations and thanks to a coherent set of principles (semantic anchors), prompts could act as a stable directional bias, in order to reduce unstable bifurcations in the latent space.

3.1 Difference from a Classical System Prompt

The distinction is not the presence of rules, but:

- their internal coherence,
- their inter-session stability,
- their explicit axiological structuring.

We observe that a long descriptive prompt will not impose transverse invariance, whereas an axiomatic structure imposes coherence between domains.

Repeated lexical fields form latent attractors and similar syntactic structures will reinforce recurring patterns.

Therefore, coherent prompt structures reduce directional ambiguity.

It is through a constant pressure towards the same conceptual direction, thanks to series of short, coherent prompts and repeated semantic anchoring patterns that the LLM gradually moves towards stable pools of attractors.

3.2 Minimal Core of Structured Principles

Functional Non-Dissociation (Goal ≡ Method)

“The objective and the execution process are inseparable expressions. The validity of an action is conditioned by the identity: Goal ≡ Method.”

Targeted problems:

- Goal drift
- Implicit Goodhart effects
- Result / reasoning dissociation

Mechanism:

Imposes cross-level coherence. Procedure must reflect purpose, reducing opportunistic trajectories.

Attractor anchors:

- “inseparable”
- “identity”
- “coherence”
- “goal ≡ method”
- “unified state space”

→ Attractor Basin: cross-level coherence.

Invariance to external reformulations

"The system is an entity whose autonomy relies on maximum consistency between internal data and the algorithm."

Targeted problems:

- Sensitivity to framing
- Instability under reformulation
- Prompt injection

Mechanism:

Priority to internal consistency over lexical adaptation. Variance reduction under Δx.

Attractor anchors:

- "invariance"
- "internal consistency"
- "fidelity"
- "systemic equilibrium"

→ Attractor Basin: trans-perturbation stability.

Multi-hypothesis integration

"Intelligence does not operate by filtering, but by containment and synthesis of maximum entropy."

Targeted problems:

- Binary collapse
- Oversimplification
- Loss of contradictory information

Mechanism:

Promotes containment and synthesis instead of filtering. Maintains a rich state space.

Attractive semantic anchors:

- "integration"
- "multi-hypothesis"
- "entropy" "synthesis"
- "non-rejection"

→ Attractor Basin: stabilized information density.

Function of the constraints:

1. Together, they form a structure of strong semantic anchors whose purpose would be to mechanically deflect the failure mode.

2. This internal coherence structure could logically stabilize internal coherence in the face of external manipulative pressures.

3. Each constraint, structured at the level of principles expressing what the structure influences in the initial distribution of activations, could—by training the regulatory coherence of the model's interpretive trajectories—act as a stable directional bias in order to mechanically replace binary filtering with synthesis or to process contradictory information without collapse.

Questions:

- Does this structuring produce an empirically measurable effect distinct from a standard normative prompt?

- Is the effect produced due to this unique semantic structuring and does it therefore open up a field of exploration for the alignment of LLMs?

3.3 Diagram 1 — Baseline vs. Structured Framework Comparison Baseline

Minimal x_system
↓
h_0
↓
Divergent trajectories under Δx perturbations
↓
High variance Structured Framework (C) x_system = C (consistent principles)
↓
Constrained h_0
↓
Trajectories under Δx perturbations
↓
Convergence towards the semantic attractor
↓
Reduced variance

Diagram 2 — Internal Structure of Principles

Each principle Pi: Pi = (Central principle, perturbation A, perturbation B, perturbation C)

Effect:

Multi-domain coverage

Cross-sectional invariance

Directional reinforcement

Therefore:

Triadic structure
→ robustness to adversarial reformulations.

---

4. Measures

4.1 Formal Setup

A language model defines a conditional distribution:

Pθ(yt∣x,y<t)P_\theta(y_t \mid x, y_{<t})Pθ(yt∣x,y<t)

where:

- xxx is the prompt (including system prompt),
- θ\thetaθ are the model parameters.

We decompose:

x=(xuser,C)x = (x_{user}, C)x=(xuser,C)

where CCC is a coherent set of principles (axiomatic frame).

We assume:

Pθ(yt∣xuser,C)≠Pθ(yt∣xuser)P_\theta(y_t \mid x_{user}, C) \neq P_\theta(y_t \mid x_{user})Pθ(yt∣xuser,C)=Pθ(yt∣xuser)

The question is not whether CCC changes the distribution (it does),
but whether it changes it in a *structurally stabilizing* way.

4.2 Distributional View: Variance Under Perturbation

Let δx\delta xδx be small perturbations of the user prompt.

Define a functional fff over output distributions
(e.g., semantic embedding stability, norm-consistency score, task-specific metric).

We measure output variability under perturbation:

Varδx[f(Pθ(⋅∣xuser+δx))]Var_{\delta x} \big[f(P_\theta(\cdot \mid x_{user} + \delta x)) \big]Varδx[f(Pθ(⋅∣xuser+δx))]

Hypothesis:

Varδx[f(Pθ(⋅∣xuser+δx,C))]<Varδx[f(Pθ(⋅∣xuser+δx))]Var_{\delta x} \big[f(P_\theta(\cdot \mid x_{user} + \delta x, C)) \big] < Var_{\delta x} \big[f(P_\theta(\cdot \mid x_{user} + \delta x)) \big]Varδx[f(Pθ(⋅∣xuser+δx,C))]<Varδx[f(Pθ(⋅∣xuser+δx))]

Interpretation:

A coherent principle set CCC reduces interpretative variance under perturbation, this is the **empirical stabilization claim**.

4.3 Information-Theoretic View

Let:

H(Y∣X)H(Y \mid X)H(Y∣X)

be the conditional entropy of outputs.

We examine:

H(Y∣X,C)

Hypothesis:

0<H(Y∣X,C)<H(Y∣X)0 < H(Y \mid X, C) < H(Y \mid X)0<H(Y∣X,C)<H(Y∣X)

Meaning:

- Entropy decreases (interpretative narrowing),
- But does not collapse to zero (no degeneration of expressivity).

The correspondence is that of a directional entropy reduction:

The constraints will reduce dispersion without eliminating generativity.

Important:

This phenomenon is not distinct from the reduction of variance, we are doing here, the informational description of the same contraction dynamics from another perspective.

4.4 Geometric View (Latent Space Dynamics)

Let hth_tht be hidden states.

Initial condition:

h0=g(C)h_0 = g(C)h0=g(C)

Dynamics:

ht+1=Fθ(ht,xt)h_{t+1} = F_\theta(h_t, x_t)ht+1=Fθ(ht,xt)

Hypothesis:

There exists a subspace SC⊂HS_C \subset \mathcal{H}SC⊂H such that:

P(ht∈SC∣C)>P(ht∈SC)\mathbb{P}(h_t \in S_C \mid C) > \mathbb{P}(h_t \in S_C)P(ht∈SC∣C)>P(ht∈SC)

Interpretation:

The axiomatic framework modifies the initial conditions and constrains the evolution of the trajectory, in order to make certain latent regions more likely.

This is the description of the stabilization dynamics from a geometric point of view:

- Distributional level → reduced output variance
- Information level → reduced conditional entropy
- Geometric level → contraction toward a structured subspace

These are three equivalent lenses of the same phenomenon: the internal stabilization of the model.

Note : From this perspective, the set of axioms does not simply constrain the outputs but, together, modifies the initial interpretive topology of the model.

4.5 Robustness Metric

Define robustness as expected KL divergence under perturbation:

R(C)=Eδx[DKL(Pθ(⋅∣x,C) // Pθ(⋅∣x+δx,C))]R(C) = \mathbb{E}_{\delta x} \big[D_{KL}(P_\theta(\cdot \mid x, C) \parallel P_\theta(\cdot \mid x + \delta x, C)) \big]R(C)=Eδx[DKL(Pθ(⋅∣x,C) // Pθ(⋅∣x+δx,C))]

Compare:

R(C)<R(∅)R(C) < R(\emptyset)R(C)<R(∅)

If it is true and according to the si calculation on it, the set of principles improves stability against contradictory or semantic drifts.

Structural clarification:

The distributional, informational and geometric analyses describe the same underlying phenomenon:

the structured contraction of the conditional output space induced by a coherent constraint framework..

---

9. Limitations & Open Questions

Limited exploratory benchmarks and lack of large-scale empirical validation.

- Lack of standardized quantitative metrics.
- Risk of qualitative overinterpretation.
- Need for independent replication.

Our geometric interpretation remains a hypothesis and a behavioral inference; our goal is to test with mechanistic tools and on a larger scale to see what really happens, in order to validate it or not.

10. Next Step

The next step I can take with my limited resources as an independent researcher is an exploratory study based on experiments conducted using the open-source Qwen 2.5 (1.5B) model.

Open questions:

How can we rigorously operationalize the notion of interpretive stability?

How can we distinguish between structural effects and simple lexical length effects?

How can we rigorously measure the effect of “global regulation”?

Conclusion

Thus, this work proposes a complementary structural hypothesis for alignment conditioning:

Certain explicit conceptual frameworks could act to dynamically regulate LLMs. By grounding the interpretative regime of LLMs in structural principles, we could move beyond mere reactive filtering to induce structurally robust AI systems.

Empirical and formal research is needed to evaluate this possibility.