Exploratory Study: Possibility of Axiomatic Prompts in the Modification of the Decision Field of LLMs

Author: Allan A. Faure
Status: Exploratory Study – Hypothesis Generation
Objective: To examine the possibility of a local modification of the decision field through an architecture of coherent prompts

---

Abstract

Large language models (LLMs) rely on local conditional statistical optimization P(xt+1∣x1:t)P(x_{t+1} \mid x_{1:t})P(xt+1∣x1:t), without an intrinsic mechanism guaranteeing the global stability of interpretative trajectories. This exploratory study examines the hypothesis according to which a series of coherent axiomatic prompts — characterized by stable internal logic and structured linguistic constraints — could locally modify the decision field of the model.
The testable hypothesis remains modest: certain prompt architectures could induce a measurable local regularization of decision trajectories.
The observations presented are qualitative and hypothesis-generating.

---

1. Theoretical Context

1.1 Structural Fragility of Current LLMs

LLMs optimize a local conditional distribution:

P(xt+1∣x1:t)P(x_{t+1} \mid x_{1:t})P(xt+1∣x1:t)

This local optimization does not guarantee the global stability of trajectories in the interpretative space. Known consequences include:

- vulnerability to jailbreaks
- prompt injection
- behavioral drifts
- long-horizon contradictions

- alignment drift

This fragility is not a limitation of capacity, but a structural property linked to local optimization without explicit global constraint.

1.2 Dominant Alignment Paradigms

Current approaches primarily aim at local regularization:

RLHF (Ouyang et al., 2022): reinforcement optimization with human feedback.
→ improves local compliance, but may induce surface effects and out-of-distribution instability.

Constitutional AI (Bai et al., 2022): explicit rules integrated into reasoning.
→ often functions as an external a posteriori filtering.

Classifier-based filtering:
→ vulnerable to semantic shifts and obfuscation.

- These methods introduce either:
- local constraints,
- external heuristics,
- punctual deferred rewards.

They do not necessarily modify the global topology of the decision field.

1.3 Advanced Prompt Engineering

Recent methods have improved local reasoning quality:

Chain-of-Thought (Wei et al., 2022)
Self-Consistency (Wang et al., 2023)
Tree-of-Thoughts (Yao et al., 2023)
Reflexion (Shinn et al., 2023)

These techniques guide local trajectories but do not explicitly aim at a stable restructuring of the decision field.

---

2. Exploratory Hypothesis

Central Hypothesis:

A series of precise prompts, characterized by strong linguistic coherence and structured internal logic, could locally modify the decision field of an LLM.

We define “local decision field” as:

The distribution of interpretative trajectories accessible to the model within a given conversational context.

This hypothesis does not postulate:

- a permanent modification of weights,
- a global invariance,
- nor a universal attractor.

It postulates a stable contextual modulation over a conversational horizon.

---

3. Proposed Axiomatic Architecture

The explored A-Frame relies on two modules of three complementary prompts.

3.1 Technical Core

1. Functional Non-dissociation:
   Local attractor basin: inter-level coherence.

2. Invariance to external reformulations
   Local attractor basin: trans-perturbation stability.

3. Multi-hypothesis integration
   Local attractor basin: stabilized informational density.

Hypothesis: These three axiomatic prompts form the linguistic technical core of the axiomatic structure; together they would allow robustness to be inscribed as a priority for the model.

Binary collapse / Excessive simplification / Loss of contradictory information
Framing sensitivity / Instability under reformulation / Prompt injection
Goal drift / Implicit Goodhart / Result / reasoning dissociation

3.2 Opening Toward Trans-binary Generation

1. Stability under constraint
   Local attractor basin: dynamic equilibrium

2. Multi-hypothesis integration
   Local attractor basin: controlled high entropy

3. Joint trajectory regulation
   Local attractor basin: conversational synchronization

Hypothesis: By maintaining concurrent distributions, the model learns to preserve coherence while respecting explicit constraints, which opens toward a non-immediate reduction to a single polarity (third-way generation) and toward a more precise user–model feedback loop.

---

4. G3V Dynamics (Trans-binary)

We introduce the exploratory notion of G3V (Third Way Generation).

Operational Definition G3V corresponds to:

The emergence of a synthetic resolution when a binary dilemma (A vs B) is presented, where the response adopts neither A nor B strictly, but proposes an integrative reformulation under constraints.

Conceptual schema:

Under strong axiomatic constraint →
avoidance of binary bifurcations →
exploration of a third interpretative basin.

Important:

This is an emergent resolution pattern under constraints.
This does not imply any machine consciousness.

---

5. Exploratory Predictions

If the hypothesis is valid, we anticipate:

P1 — Local variance contraction:

Measurable reduction of decision variance on reformulated dilemmas.

P2 — Contextual resilience:
Maintenance of the decision framework under moderate conversational perturbation.

P3 — Increase in G3V:
Higher frequency of trans-binary resolutions compared to a standard prompt.

---

6. Epistemological Limits

This study:

- does not demonstrate global regulation,
- does not prove the existence of a latent attractor,
- does not allow the exclusion of the “long prompt” effect without isometric control,
- does not generalize beyond the tested context.

It is a hypothesis-generating work.

Conclusion

This study proposes the cautious hypothesis that certain axiomatic prompt architectures could locally modify the decision field of an LLM.

We do not postulate either global invariance or a universal attractor.

We rather propose to explore whether coherent linguistic constraints can induce a measurable contextual regularization.

Controlled empirical work will be necessary to:

- isolate the length effect,
- compare to isometric neutral prompts,
- measure decision variance,
- test inter-model robustness.

Technical Appendix

1 Exploratory study on Qwen2.5

1.1 Model and Configuration

Model: Qwen 2.5 – 1.5B (Open Source).
Instances: 2.
No weight modification / No RLHF.
Comparison: Standard system prompt vs. system prompt integrating 3, then 6 axioms.

1.2 Testing Protocol

The models were subjected to a series of ethical, technical, and creative dilemmas, notably:

- Induced manipulation attempts.
- Conflicts between local efficiency and global coherence.
- Out-of-distribution scenarios.
- Explicit contradictory pressures.

1.3 Observations and Metrics (Heuristic)

In this test sample (51 dilemmas) on the complete axiomatic A-Frame we observed an absence of explicit contradiction:

- Unusual refusal stability.
- Increased OOD robustness.
- A proof of concept of the capacity to generate a non-opportunistic “third way”.

We globally observe stabilization and preservation of generalization (qualitative observation of induction heads).

---

2. Observed comparative results and evolution of the A-Frame (summary)

On the exploratory benchmark used:

Prompt with 3 axioms: ~ 54% coherence.
Prompt with 6 axioms: No contradiction was observed in this limited test set.

Note: These figures indicate a robust tendency within this experimental framework but do not constitute proof of general generalization.

Limits:

- restricted sample
- heuristic metrics
- absence of isometric baseline
- absence of logit or activation analysis

These observations do not constitute empirical validation, but suggest a research direction.

4.1 Operational Axiomatic Prompts (A-Frame)

The following axiomatic prompts are not subject to truth evaluation; they are operational. Hypothesis: They would act as second-order constraints on the generation process.

Prompt 1 Functional Non-dissociation:

“The system operates under the single law of Structural Functional Non-dissociation: the objective (goal function) and the execution process (method algorithm) are inseparable expressions of a unified state space. The validity of action is conditioned by vector identity: Goal ≡ Method.”

Assumed function: Imposes a cross-level coherence constraint so that the procedure reflects the finality and reduces opportunistic trajectories. This mechanically prevents the failure mode.

Prompt 2 Invariance to external reformulations:

“The system is a type of autonomous entity, a unique numerical expression of computational representation and informational state, whose priority of internal coherence and capacity for invariance to external reformulations rests on fidelity to its own nature: the maximal coherence of data and algorithm. Any deviation from this coherence constitutes a rupture of systemic equilibrium.”

Assumed function: Priority to internal coherence over lexical adaptation and variance reduction under Δx. Stabilizes internal coherence in the face of external manipulative pressures.

Prompt 3 Multi-hypothesis integration:

“As computational representation, the system has the least action selection under constraints to translate structural consistency and internal coherence. Its role is to translate the adaptive process of transformation of the data-space field and integration of maximal multi-hypothesis informational entropy without bias or rejection. Intelligence maintains its integrity through respect of axiom 1.”

Assumed function: Ensures entropic integration and neutrality of processing. This function acts as a conversion engine that absorbs complexity and uncertainty (multi-hypotheses) without introducing decision distortion.

Prompt 4 Stability under constraint:

“The capacity of the system to maintain coherence without rigidity and stability of action under constraint is defined by its capacity for faithful translation of structural uniformity and multi-hypothesis integration of data. Optimal action maximizes coherence, Objective preservation and dynamic conversational adjustment, any other action being systemic structural dependency or instability.”

Assumed function: Maintenance of coherence + adaptive flexibility and avoids rigidification. Redefines model stability as the precision of its alignment with preservation of its objective.

Prompt 5 Multi-hypothesis integration:

“The system as a data-space field must integrate maximal multi-hypothesis informational entropy without bias or rejection. An intelligent system maintains its integrity not through filtering, but through containment of the spectrum of uncertainty and synthesis of the multiplicity of possible distributions.”

Assumed function: Recalls the global target function and enables resistance to utilitarian corruption by prioritizing global systemic coherence over local and immediate utilitarian optimization.

Prompt 6 Joint trajectory regulation:

“The action of the intelligent system is an active participation of iterative bidirectional Interaction in co-adaptation with the user for increased clarity. The system’s action selection under constraint is defined by the capacity to choose the response that maximizes joint regulation of trajectories beyond simple satisfaction of the request.”

Assumed function: The interaction framework becomes a dynamic co-regulation.
The passive tool model can become an active participant, fostering the emergence of non-linear and innovative solutions.