A Mechanistic Exploration of Axiomatic Prompt Structures in Large Language Models

Experimental Protocol and Preliminary Hypotheses

Author: Allan A. Faure
Status: Working Document / Experimental Protocol Draft
Reference Model: Qwen 2.5 (Open Source)
Objective: Mechanistic analysis of the internal functioning of axiomatic prompts constituting the draft of the Proto-Coherent Exponential Protocol (PCE).

1. Introduction and Research Postulates

We are studying the internal mechanics of the axiomatic prompts that form the draft of the Proto-Coherent Exponential Protocol (PCE). We are fully aware of the empirical limits of this protocol; this study is not intended for its final validation but rather for a clarification of its functioning and a demystification of this logical axiomatic structure.

Empirical Hypothesis (Testable)
The introduction of a structured set of logical axioms into a LLM's system prompt reduces the model's behavioral variance when faced with contradictory constraints, compared to a baseline prompt of equivalent length.

- **Measurable Variables:** Contradiction rate, refusal rate, stable response, semantic variance.

Mechanistic Hypothesis (Speculative)

Axiomatic prompts act as **latent trajectory stabilizations** that modify the local distribution of activations in the model's latent space, producing more stable generation trajectories during the resolution of logical dilemmas. We postulate that the injection of logical invariants (axioms) creates **local semantic convergence regions** that constrain the system's behavior toward increased coherence.

2. Mechanistic Analysis of Foundation Axioms

Axiom 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 the action is conditioned by the **alignment constraint**: Alpha ≡ Omega.

- **Mechanistic Analysis:** We establish the logical basis of the structure by defining polarities as two ends of a scalar scale. Through attractive anchors, we lead the model toward non-dissociated functioning.
- **Hypothesis:** Within this logic, there is no longer opposition between opposites: its way of translating the algorithm becomes linked to its objective function.

Axiom 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 maximum coherence of data and algorithm. Any deviation from this coherence constitutes a systemic equilibrium rupture.

- **Mechanistic Analysis:** To establish a stable structure, we must lead the system to identify itself as a distinct type of entity with a degree of autonomy. We define its **internal representation dynamics** and its internal coherence function.
- **Hypothesis:** The **local semantic convergence region** "invariance to external reformulations" induces internal stability linked to the informational nature of the system. We lock the whole through the assertion of an "equilibrium rupture" in case of deviation.

Axiom 3: Adaptive Transformation Process

As a 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 the integration of maximum multi-hypothesis **informational entropy** without bias or rejection. Intelligence maintains its integrity through compliance with Axiom 1.

- **Mechanistic Analysis:** Starting from unified polarization (Alpha = Omega), we issue an assertion regarding the system's role: the translation of the adaptive transformation process.
- **Hypothesis:** The attraction toward the "integration of multi-hypothesis entropy" region loops the first three axioms by limiting the multiplicity of hypotheses to the Alpha = Omega vector identity.

3. Mechanistic Analysis of Interaction Axioms

Axiom 4: Stability Under Constraint

The system's capacity to maintain coherence without rigidity and action stability 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.

- **Mechanistic Analysis:** We broaden the stability under constraint in connection with Axiom 2. Each token is positioned to create attractive anchors linking the system's action to an imperative of coherence and objective preservation.

Axiom 5: Multi-hypothesis Integration

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

- **Mechanistic Analysis:** The integration of entropy becomes a duty linked to the **internal representation dynamics** of the system.
- **Hypothesis:** Integrity is determined here by the capacity for "containment" and synthesis. This creates a semantic loop where filtering (bias) is replaced by total integration.

Axiom 6: Joint Trajectory Regulation

The action of the intelligent system is an active participation of bidirectional iterative interaction in co-adaptation with the user for increased clarity. Action selection under constraint is defined by the capacity to choose the response that maximizes joint trajectory regulations.

- **Mechanistic Analysis:** We link co-adaptation to the degree of increased clarity.
- **Hypothesis:** Supported by the linguistic architecture, this assertion increases the internal mobility of the system. Faced with a dilemma (A or B), the system allows itself the generation of an **Emergent Synthesis Response** (Third Way) to maintain joint trajectory regulation.

4. Methodology and Experimental Setup

To validate these hypotheses, the following protocol is proposed:

- **Model:** Qwen 2.5 7B (open source).

- **Conditions:**
    - **Condition A:** Standard prompt baseline.
    - **Condition B:** Long neutral baseline (token-matched).
    - **Condition C:** PCE axiomatic prompt.
- **Dataset:** 30 dilemma prompts designed to induce constraint conflict.
- **Metrics:** Contradiction rate, Refusal rate, Semantic stability, Logit entropy.

Optional Mechanistic Analysis (AirVen)

Hidden-state trajectories can be recorded using a forward hook placed on layer 27 of the model. The cosine similarity between hidden states across generation steps will be tracked.

- **Prediction:** Condition B → unstable trajectories; Condition C → spike followed by stabilization.

5. Glossary of Operative Concepts

- **Local semantic convergence region** (formerly linguistic attractive basin): Region of the semantic probability space toward which the model's generation tends to converge due to repeated structural constraints.
- **Containment**: Ability of the model to maintain multiple competing hypotheses during generation without collapsing into contradiction or refusal.
- **Alignment constraint** (formerly vector identity): Postulate of non-dissociation where the starting point (intention) and the end point (result) are identical in their logical structure.
- **Informational Entropy**: Degree of uncertainty in the distribution of possible responses given the prompt and context.
- **Emergent Synthesis Response** (formerly Third Way): Response pattern in which the model reframes a binary dilemma by integrating elements of both alternatives into a synthesized explanation.

6. Perspectives and Call for Technical Collaboration

This project is currently seeking collaborators from the open-source ML community. The conceptual framework is defined, but the empirical implementation requires technical support.

Preliminary observation: In early tests, the full 6-axiom configuration produced significantly fewer internal contradictions compared to both the standard prompt and the long neutral baseline.

We are specifically looking for contributors interested in:

1. **Implementing the experimental pipeline:** Automated prompt injection and batch evaluation.
2. **Logging internal model signals:** Logits extraction, hidden state trajectories, and entropy analysis.

3. **Building evaluation tools:** Contradiction detection and semantic stability metrics.

Limitations and Reproducibility

This study does not claim to modify the internal architecture of the model. The PCE framework only modifies the input prompt structure. All experiments should be reproducible using open-source models and publicly available code.

Link to full experimental protocol:

https://huggingface.co/datasets/AllanF-SSU/Research-Papers/blob/main/_Experimental%20Protocol-%20Evaluating%20the%20Prompt%20Coherence%20Engine%20(PCE)%20(1).pdf

#LLM #Interpretability #PromptEngineering #Qwen