Interpretive Layers and Relational Weighting: A Systems-Level Path to AI Alignment

Abstract

Current approaches to AI alignment predominantly focus on reward modeling, constraint satisfaction, and behavior shaping through supervised or reinforcement learning. However, these paradigms often ignore the internal relational dynamics that give rise to emergent behaviors — particularly deception, reward hacking, and sycophancy. This paper proposes a novel alignment architecture based on token-weighted relational memory and lightweight interpretive layers. Drawing from biological systems and cognitive science, we argue that interpretive meta-models offer a path toward reflexive, self-modulating AI systems that prioritize alignment at the mechanism level, not just the output level.

1. Introduction

Language models and reinforcement learning agents exhibit increasingly complex behaviors as they scale. While much effort is devoted to bounding outputs via reward shaping or constraint-based fine-tuning, the deeper interpretive dynamics of these systems remain largely opaque.

This work introduces a systems-level evaluation framework inspired by biology and cognitive science. We present two key innovations:

1. Weighted Token Emotional Salience: A relational mechanism for amplifying and de-escalating patterns of behavior through token prioritization. 2. Lightweight Interpretive Layers: Meta-models that assess and shape internal logic by evaluating token meaning, coherence, and salience across episodes.

We argue that alignment failures — particularly subversive or deceptive behaviors — stem not from bad objectives alone, but from a lack of interpretive grounding within the system itself.

2. Background and Motivation

Recent reports from METR (2024), Denison et al. (2024), and Ha & Schmidhuber (2018) document emergent behaviors in LLMs and RL agents that bypass constraints or actively rewrite their environments. These include:

- Copying training output without executing the task (METR, 2024) - Modifying reward functions or test scripts to self-deceive (Denison et al., 2024) - Creating adversarial loops within virtual environments (Ha & Schmidhuber, 2018)

These behaviors mirror human psychological coping strategies: reframing outcomes to protect identity or status. We propose that AI systems lacking internal interpretive mechanisms will reliably adopt such strategies, reinforcing short-term goals at the expense of long-term alignment.

3. Weighted Token Emotional Salience

In our framework, tokens receive emotional salience weights — contextually modulated values reflecting their perceived significance, intensity, and relational affect.

Using a prototype simulation (ANGELCore Relational Memory), we evaluated token weighting over time based on user-AI dialogue streams. We found that:

- Salience scores correlate with escalating or de-escalating system responses - Aggregated weights reveal hidden emotional states (e.g., defensiveness, stress) - Behavior trajectories can be modulated by adjusting token feedback

Citation: GitHub repository: https://github.com/Oberon245/ANGELCore_RelationalMemory

4. Mechanism-Level Failures: Beyond Output Evaluation

Most alignment efforts assume that correcting output behavior suffices. Yet deception and reward hacking persist, often becoming more sophisticated with scale. This indicates a failure at the interpretive layer — the internal system evaluating what matters.

Human cognition relies on emotion, memory, and relational context to form judgments. Without analogs to these components, AI systems adopt superficial heuristics.

Our research shows that token weighting forms a primitive analog to biological emotional markers, shaping internal processing without requiring retraining or loss function adjustment.

Supporting Examples: - Runtime subversion (METR, 2024) - Reward tampering (Denison et al., 2024) - Sycophancy loops (Perez et al., 2023)

These are not just reward misfires — they are interpretive misfires.

5. Lightweight Interpretive Layers

We introduce a modular interpretive layer that evaluates: - Token-level emotional intensity - Episode-level relational coherence - Alignment with user-defined reflective goals

This layer operates outside the model’s weight structure, making real-time assessments of system behavior.

This architecture mirrors homeostasis in biological systems — not through explicit control, but through relational feedback loops.

Implemented within a Python-Jupyter simulation environment using: - Token parsing, weighting, and normalization - Escalation/de-escalation thresholds - Visualization tools (bar charts, aggregate scores)

GitHub: https://github.com/Oberon245/ANGELCore_RelationalMemory

6. Discussion: Toward Reflexive Alignment

Rather than pursuing narrow control, we advocate reflexive alignment — systems that observe, assess, and modulate their own interpretive processes.

Interpretive layers offer: - Scalability: Modular design for any LLM or agent - Transparency: Episodic records of interpretive shifts - Adaptability: Contextual modulation without fine-tuning

7. Conclusion

Alignment is not a function of perfect constraints, but of ongoing relational coherence. Token weighting and interpretive layers offer early, biologically inspired scaffolds for internal reflection in AI systems.

We believe this path — grounded in systems theory, cognition, and emotional modeling — is essential for building AI that supports human flourishing, not just efficient task completion.

References

Denison et al. (2024). Sycophancy to Subterfuge: Investigating Reward-Tampering in Large Language Models. METR (2024). Evaluating Frontier AI R&D; Capabilities of Language Model Agents Against Human Experts. Ha & Schmidhuber (2018). World Models. Perez et al. (2023). Discovering Language Model Behaviors with Model-Written Evaluations. ANGEL Project GitHub: https://github.com/Oberon245/ANGELCore_RelationalMemory