Working Draft: Recursive Reflective Architecture for AI (LessWrong Submission)

# Abstract

This paper proposes a speculative architectural model for advanced language models, inspired by recursive structures in the human brain and the 'strange loops' articulated in Douglas Hofstadter’s \*Gödel, Escher, Bach\*. The core idea is the Hofstadter Engine: a seven-layer reflective stack in which successive modular agents observe, interpret, and modulate both the outputs and internal states of lower layers. These recursive modulating layers aim to simulate reflectivity, support epistemic robustness, and open a path toward AI systems with self-modulating ethical and cognitive alignment. The model does not propose sentience, but functional simulation of subjectivity through designed recursion. We offer this framework not as a blueprint, but as a working model—intended to provoke experimentation, criticism, and further development by the alignment research community.

# 1. Introduction

Language models today operate with astonishing surface fluency but shallow internal architecture. Their token-by-token outputs are shaped by enormous datasets and intricate parameter matrices, yet lack a structured mechanism for reflection, self-assessment, or recursive adjustment. When alignment failures occur—whether factual hallucinations or misaligned moral reasoning—there is often no inner scaffolding capable of catching or correcting the error beyond post-hoc fine-tuning.

This paper proposes an alternative path: a recursive architecture for artificial cognition that introduces layered reflective modulation. Each layer in this model serves as an observer and evaluator of the layer below, enabling a structured form of feedback, introspection, and self-regulation. Inspired by the layered organization of the neocortex, bounded human recursion depth, and the self-referential structures in \*Gödel, Escher, Bach\*, we argue that recursive reflectivity can and should be simulated within AI systems. We do not claim this will lead to consciousness. But it may yield systems capable of limited forms of simulated self-awareness—sufficient to support more robust reasoning, improved alignment, and transparent cognitive auditability.

We call this proposed system the \*Hofstadter Engine\*: a speculative, modular AI framework composed of seven recursive layers. Each layer serves a distinct reflective function—from observation and interpretation to epistemic auditing and goal-level integration. The model is intended to be modular, falsifiable, and implementable at varying degrees of complexity—from rules-based feedback loops to neural secondary observers. We offer this concept as a contribution to the alignment research community, and invite critique, simulation, and expansion.