The Hierarchical Brain

Hierarchical Modularity and Shared Embeddings Enable Compositional Reasoning in Neural Networks

Authors: The Grokking Research Team

Date: January 2026
Repository: hierarchical-brain-pytorch

Abstract

Deep learning models, particularly Transformers, excel at pattern matching but often struggle with compositional reasoning, failing to generalize to complex, nested problems despite mastering atomic components. This phenomenon arises from the entanglement of high-level planning and low-level execution within a monolithic latent space. We introduce the Hierarchical Brain architecture, a neuro-symbolic framework that decouples reasoning into a Planner (System 2) and an Executor (System 1), connected via a Shared Embedding Space. This architecture achieves perfect zero-shot generalization on nested arithmetic tasks, effectively approximating a differentiable CPU.

1. Introduction

Current large language models rely on monolithic architectures where syntax parsing and logic execution are entangled. While effective for soft tasks, these systems fail under rigorous compositional demands. We argue that robust intelligence requires orthogonality: logic must be independent of values, and physics must be independent of context.

2. Methodology: The Resonant Architecture

The Hierarchical Brain enforces a Shared Embedding Constraint across all modules, ensuring lossless communication between the Planner and Executor.

2.1 The Shared Soul Constraint

All modules read and write to the same embedding matrix, eliminating representational drift and translation loss between subsystems.

2.2 System 2: The Planner

A Transformer model trained purely on syntactic decomposition, translating infix expressions into reverse polish notation without ever observing execution results.

2.3 System 1: The Executor

A modular neural stack machine trained on the full physics of the domain, acting as a precise Arithmetic Logic Unit that never hallucinates atomic facts.

3. Experiments & Results

On modular arithmetic with prime P=97, the Hierarchical Brain achieved 100% accuracy on nested, out-of-distribution expressions, while monolithic Transformers achieved only 24%.

4. Discussion: The Differentiable CPU

The architecture mirrors a classical CPU, with shared embeddings as registers, the Executor as the ALU, and the Planner as an instruction decoder.

5. Conclusion

Compositional reasoning failures are architectural, not fundamental. By decoupling logic from physics and reconnecting them through a resonant shared space, we enable neural networks to reason reliably and generalize systematically.