Hierarchical Reasoning Model

Guan Wang[1,†], Jin Li[1], Yuhao Sun[1], Xing Chen[1], Changling Liu[1],
Yue Wu[1], Meng Lu[1,†], Sen Song[2,†], Yasin Abbasi Yadkori[1,†]

[1]Sapient Intelligence, Singapore

Abstract

Reasoning, the process of devising and executing complex goal-oriented action sequences, remains a critical challenge in AI. Current large language models (LLMs) primarily employ Chain-of-Thought (CoT) techniques, which suffer from brittle task decomposition, extensive data requirements, and high latency. Inspired by the hierarchical and multi-timescale processing in the human brain, we propose the Hierarchical Reasoning Model (HRM), a novel recurrent architecture that attains significant computational depth while maintaining both training stability and efficiency. HRM executes sequential reasoning tasks in a single forward pass without explicit supervision of the intermediate process, through two interdependent recurrent modules: a high-level module responsible for slow, abstract planning, and a low-level module handling rapid, detailed computations. With only 27 million parameters, HRM achieves exceptional performance on complex reasoning tasks using only 1000 training samples. The model operates without pre-training or CoT data, yet achieves nearly perfect performance on challenging tasks including complex Sudoku puzzles and optimal path finding in large mazes. Furthermore, HRM outperforms much larger models with significantly longer context windows on the Abstraction and Reasoning Corpus (ARC), a key benchmark for measuring artificial general intelligence capabilities. These results underscore HRM's potential as a transformative advancement toward universal computation and general-purpose reasoning systems.

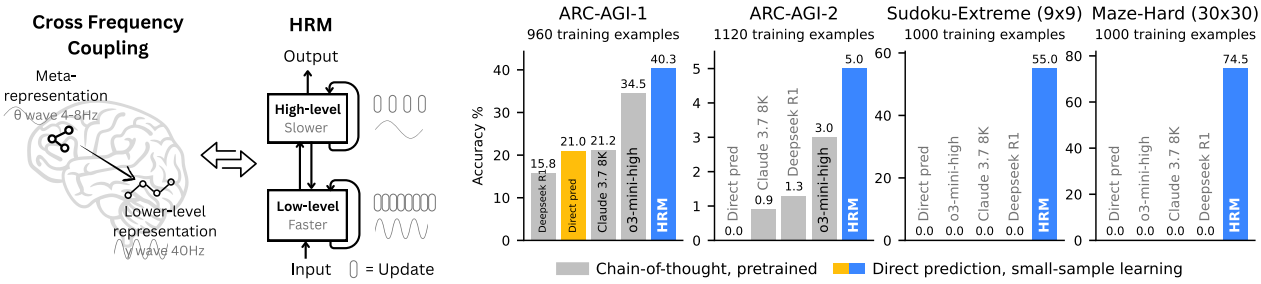

Figure 1: **Left:** HRM is inspired by hierarchical processing and temporal separation in the brain. It has two recurrent networks operating at different timescales to collaboratively solve tasks. **Right:** With only about 1000 training examples, the HRM (~27M parameters) surpasses state-of-the-art CoT models on inductive benchmarks (ARC-AGI) and challenging symbolic tree-search puzzles (*Sudoku-Extreme*, *Maze-Hard*) where CoT models failed completely. The HRM was randomly initialized, and it solved the tasks directly from inputs without chain of thoughts.

[2]Tsinghua University † Corresponding author. Contact: research@sapient.inc.
Code available at: github.com/sapientinc/HRM

1 Introduction

Deep learning, as its name suggests, emerged from the idea of stacking more layers to achieve increased representation power and improved performance[1,2]. However, despite the remarkable success of large language models, their core architecture is paradoxically shallow[3]. This imposes a fundamental constraint on their most sought-after capability: reasoning. The fixed depth of standard Transformers places them in computational complexity classes such as AC^0 or TC^0[4], preventing them from solving problems that require polynomial time[5,6]. LLMs are not Turing-complete and thus they cannot, at least in a purely end-to-end manner, execute complex algorithmic reasoning that is necessary for deliberate planning or symbolic manipulation tasks[7,8]. For example, our results on the Sudoku task show that increasing Transformer model depth *can* improve performance,[1] but performance remains far from optimal even with very deep models (see Figure 2), which supports the conjectured limitations of the LLM scaling paradigm[9].

The LLMs literature has relied largely on Chain-of-Thought (CoT) prompting for reasoning[10]. CoT externalizes reasoning into token-level language by breaking down complex tasks into simpler intermediate steps, sequentially generating text using a shallow model[11]. However, CoT for reasoning is a crutch, not a satisfactory solution. It relies on brittle, human-defined decompositions where a single misstep or a misorder of the steps can derail the reasoning process entirely[12,13]. This dependency on explicit linguistic steps tethers reasoning to patterns at the token level. As a result, CoT reasoning often requires significant amount of training data and generates a large number of tokens for complex reasoning tasks, resulting in slow response times. A more efficient approach is needed to minimize these data requirements[14].

Towards this goal, we explore "latent reasoning", where the model conducts computations within its internal hidden state space[15,16]. This aligns with the understanding that language is a tool for human communication, not the substrate of thought itself[17]; the brain sustains lengthy, coherent chains of reasoning with remarkable efficiency in a latent space, without constant translation back to language. However, the power of latent reasoning is still fundamentally constrained by a model's *effective computational depth*. Naively stacking layers is notoriously difficult due to vanishing gradients, which plague training stability and effectiveness[1,18]. Recurrent architectures, a natural alternative for sequential tasks, often suffer from early convergence, rendering subsequent computational steps inert, and rely on the biologically implausible, computationally expensive and memory intensive Backpropagation Through Time (BPTT) for training[19].

The human brain provides a compelling blueprint for achieving the effective computational depth that contemporary artificial models lack. It organizes computation hierarchically across cortical regions operating at different timescales, enabling deep, multi-stage reasoning[20,21,22]. Recurrent feedback loops iteratively refine internal representations, allowing slow, higher-level areas to guide, and fast, lower-level circuits to execute—subordinate processing while preserving global coherence[23,24,25]. Notably, the brain achieves such depth without incurring the prohibitive credit-assignment costs that typically hamper recurrent networks from backpropagation through time[19,26].

Inspired by this hierarchical and multi-timescale biological architecture, we propose the Hierarchical Reasoning Model (HRM). HRM is designed to significantly increase the effective computational depth. It features two coupled recurrent modules: a high-level (H) module for abstract, deliberate reasoning, and a low-level (L) module for fast, detailed computations. This structure

[1] Simply increasing the model width does not improve performance here.