Mohamed bin Zayed University of Artificial Intelligence

ICML 2026
Mechanistic Interpretability Workshop

Dissecting Hierarchical Reasoning Models: A Mechanistic Study

Leo Rodrigues · Jian Kang — MBZUAI {leo.rodrigues, jian.kang}@mbzuai.ac.ae

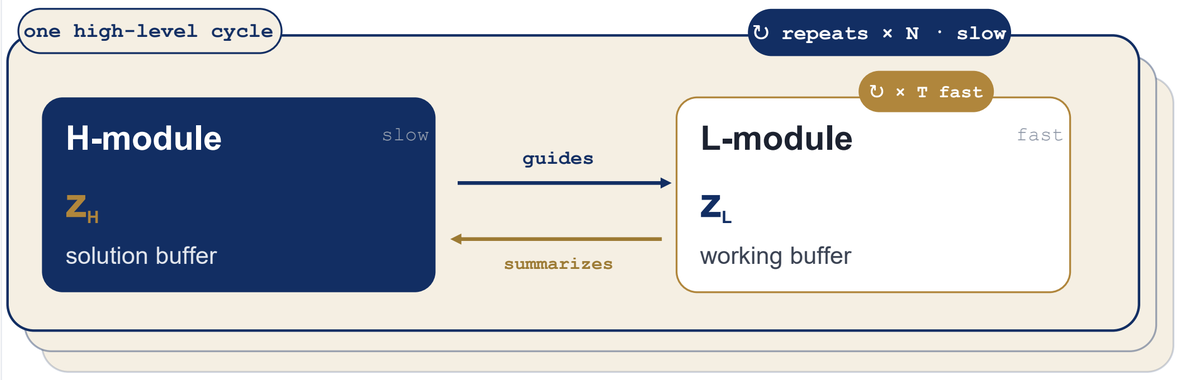
HRM refines a puzzle-specific solution in its latent state. Probes can read the constraints, but the constraints aren't localized to what we can decode.

1 TL;DR

- **Hierarchy has meaning.** High level activations as intermediate solution state; low level activations as fast-working buffer.
- **Probe readout ≠ causality.** Sudoku constraints are linearly decodable, but ablating these features has no causal impact on HRM performance.
- **Features aren't localized.** Top-50 SAE features are no more causal than 50 random ones; no small feature set captures HRM's computation.

2 Background & motivation

- **Why HRM?:** Small 27M-param; two-timescales (hierarchical); recurrent reasoning model that tops Sudoku, Maze and ARC-AGI
- **Probes & SAEs:** Tells what's *decodable*, not what's *causal;* Well studied on Transformers, but rarely on hierarchical architectures.
- **Sudoku as a testbed:** How do the high- and low-level modules coordinate in Sudoku? What does each module encode? Are decodable features the ones that the model uses?

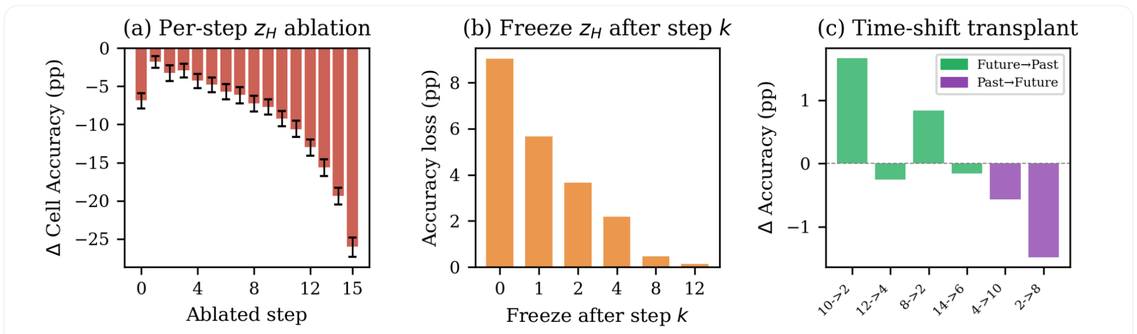


Each high-level cycle repeats × N; inside it, the L-module loops × T. zH carries the evolving solution; zL iterates fast detail.

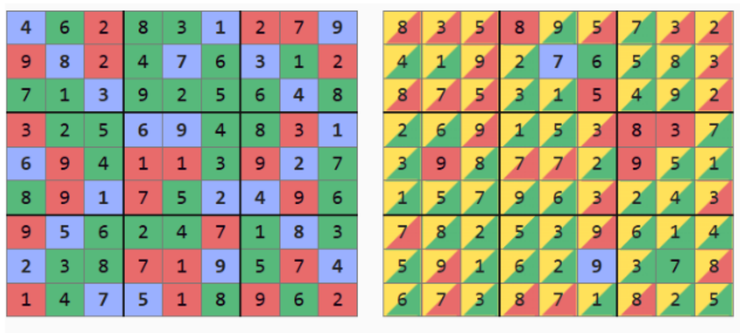
FINDING 1

3 HRM refines a solution in z_H

- **Causally essential, more so over time.** Zeroing z_H: −6.8% @ step 0 → −26.0% @ step 15. Zeroing z_L: ~0 per step.
- **Progressive, mostly done by step 8.** Freezing z_H loses −9.1% @ step 0; no effect from step 8 onward.



Cross-puzzle z_H patching

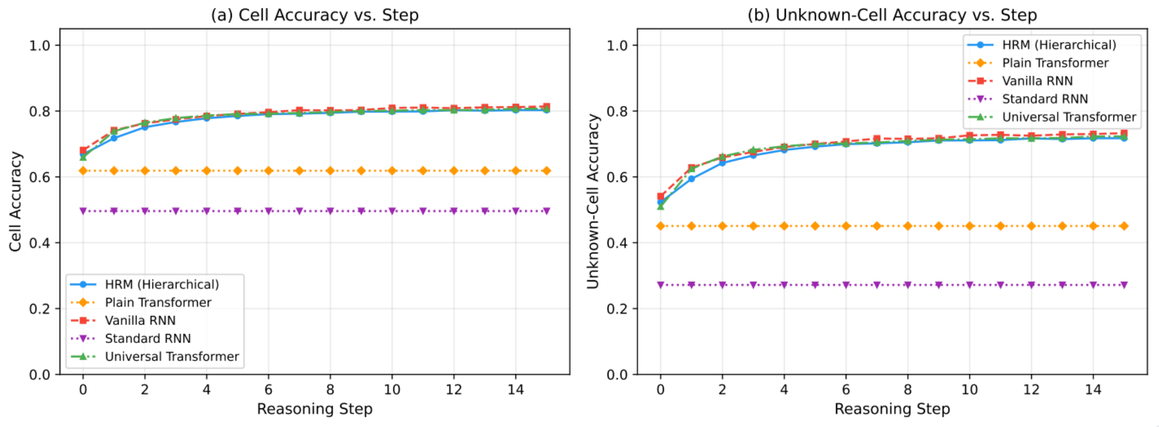


Clean run (left) vs. z_H patched from a donor at step 5 (right): Patching in a donor's high-level state rewrites the board to the donor's constraints(right), even overwriting given clues (blue).

BASELINE COMPARISON

Recurrence drives the refinement

Only recurrent models iteratively improve over ACT steps; the two-timescale hierarchy aids *interpretability*, not raw accuracy.

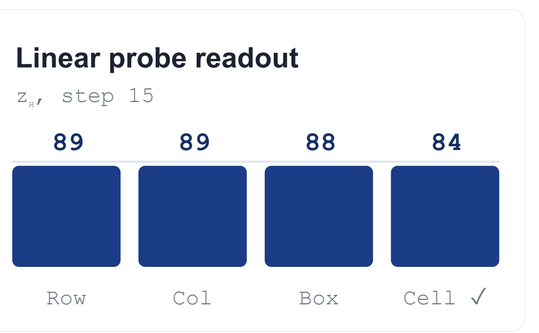


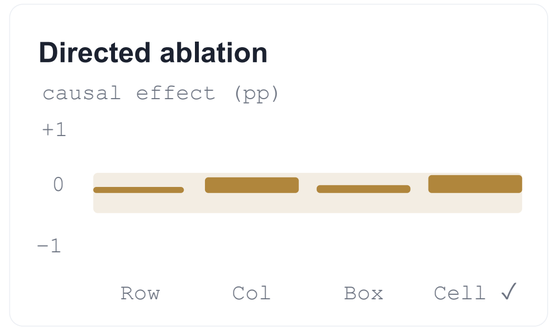
FINDING 2

4 Decodable is not always causal

Decodable ≠ causal: high readout, near-zero causal effect.

- **z_H is puzzle-specific.**
  - Swapping in another puzzle's z_H rewrites the board to satisfy the donor's constraints, collapsing accuracy by 53–62% (Sudoku grids, Finding 1).
- **High readout, zero causal effect.**
  - Constraints decode from z_H at 90% accuracy.
  - Yet ablating those probe directions shifts accuracy <1%, indistinguishable from random.
- **Computation is distributed.**
  - SAE features bite harder than probes, yet the top-50 are no more causal than 50 random ones.
  - Signal is spread across the full 512-d state.





Probe reads constraints at ~89%; ablating those directions moves accuracy <1% — inside the random-control band.

5 Discussion & Future Work

- **Hierarchy buys interpretability, not accuracy.** A flat recurrent transformer matches HRM; the two-timescale split aids analysis, not performance.
- **z_L stays underexplored.** A fast-working buffer with ~0 per-step causal effect, ablated only once; its multi-step role is open.
- **Narrow setting.** Sudoku-Extreme only, single SAE config (d=2048, λ=0.01); transfer to Maze and ARC-AGI is untested.

SCAN **Paper**

SCAN **Code**