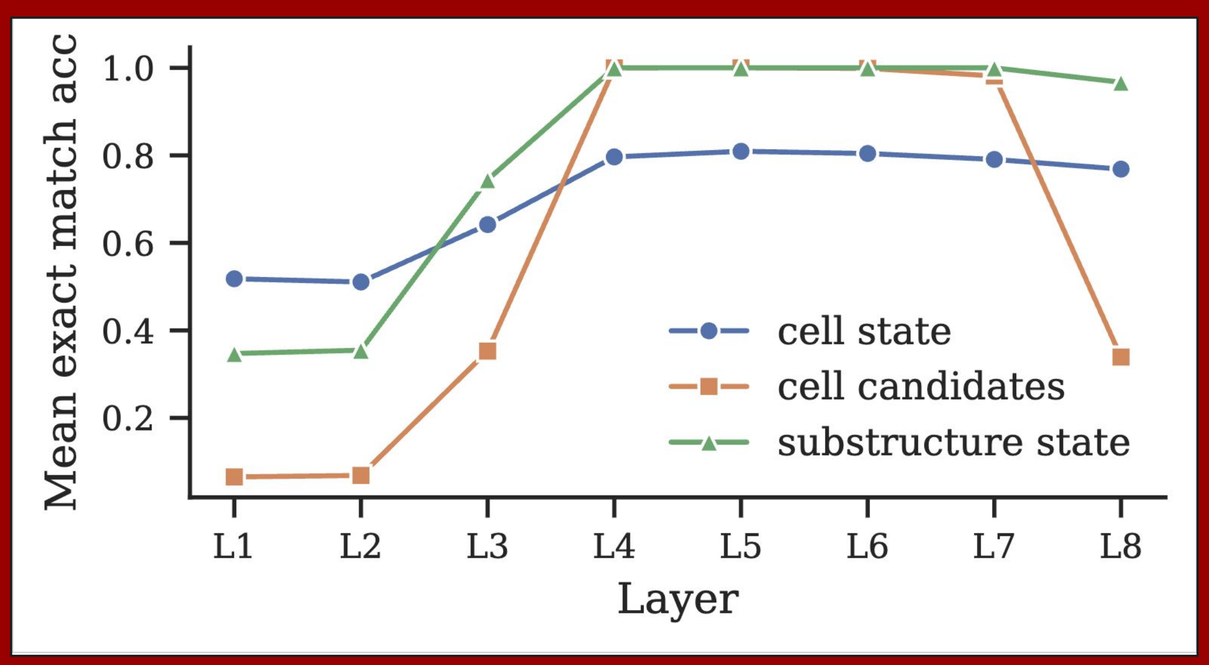


Transformer's world model mirrors task's constraints, not it's surface state

Substructure-membership probes (green) reach 100% accuracy by layer 4; per-cell digit probes (blue) plateau at 80%

LABORATOIRE BORDELAIS DE RECHERCHE EN INFORMATIQUE LaBRI

Roman Kniazev, Nathanaël Fijalkow

université de BORDEAUX cnrs

1 TL;DR

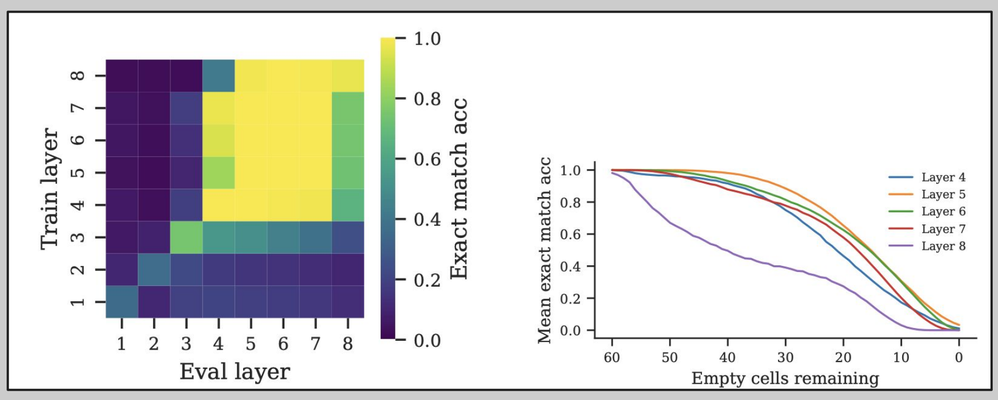
- A transformer trained on Sudoku solving traces builds a world model that **linearly represents rows, columns and boxes, but not individual cells**
- Mid-layer attention heads are **local constraint checkers**: each looks on a substructure and promotes absent digits while suppressing present digits
- The simplest solving strategy is implemented by a **sparse monosemantic circuit in the final MLP layer**

2 Background / Motivating Question

- Models trained on traces for grid-based games (Othello, chess) develop linear world models with **one feature per cell**
- **Does Sudoku-solving model also represent the board by cell?**
- We train a 8-layer transformer on solving traces with backtracking and use linear probing with activation patching to find the model and solving circuits

3 Substructure World Model

- We find that **per-cell digit** probes never separate cleanly
- **Per-cell candidate probes** hit 100% exact match accuracy, but the directions are interdependent
- **Substructure probes** get perfect exact match accuracy starting layer 4 (see the plot on top)
- The representation **is stable** throughout the later layers and later trace positions



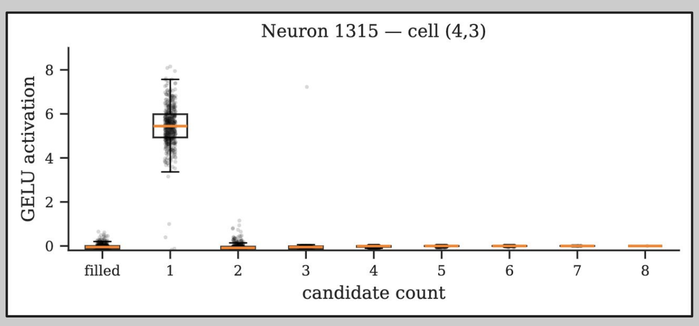
Substructure probes transfer across layers (phase transition at L4, left) and stay accurate across solving steps (right).

4 Constraint Checkers & Decision Circuit

- **Mid-layer heads attend to stacks, bands or boxes,** and digits present in the attended substructure have their logit suppressed, while absent digit are promoted
- We verify that **the effect is local** by causal ablation: suppressing head's output increases logits of illegal placements in its substructure

A column-specialised head attends to its columns (left) and suppresses present digits / promotes absent ones (middle, right, mean logit contribution).

- In naked-single case, the correct digit is top-ranked in 99% of cases before the final MLP
- **91 monosemantic neurons** each fire only when an associated cell has unique candidate and uniformly boosts all 9 candidates of that cell
- Ablating cell's neuron drops associated token probability by 0.6



An NS-detector neuron fires only when its cell has exactly one remaining candidate.

5 Limitations and Discussion

- Main limitation is **scope**: a single architecture (8 layers, 8 heads, model dimension 576)
- **Takeaway:** the geometry of an emergent world model is shaped by the domain's constraint algebra, not its surface presentation
- **Future work:** when does this geometry emerge in training (gradual vs. grokking), and does it recur in other constraint-satisfaction tasks with explicit algebraic structure?

PAPER

CODE

roman@knzv.me