Mixture of Cognitive Reasoners
Modular Reasoning with Brain-Like Specialization

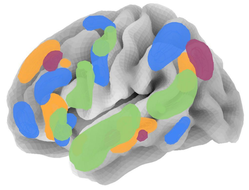
Badr AlKhamissi[1] C. Nicolò De Sabbata[1] Greta Tuckute[2,3] Zeming Chen[1]
Martin Schrimpf*,[1] Antoine Bosselut*,[1]

[1] EPFL [2] MIT [3] Harvard

ICLR International Conference On Learning Representations

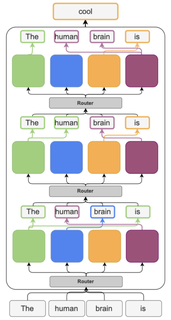
Demo | Project Page | @bkhmsi

1 Neuroscience Motivation

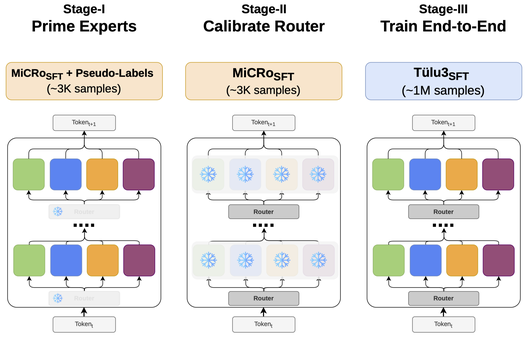


Brain-like specialization
→ more interpretable, and capable reasoners

2 Architecture



3 Training Curriculum

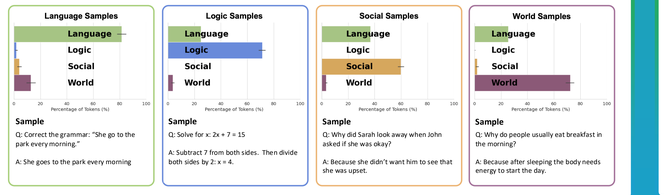


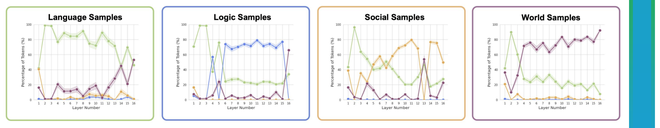
*

What if our MoEs were like our BRAINS?

- MiCRo is a **brain-inspired** MoE with experts in each layer analogous to four brain cognitive networks: **Language** **Logic** **Social** **World**
- MiCRo's specialized experts are **interpretable** and **causally** meaningful.
- MiCRo's behavior can be **dynamically steered** at test time.
- MiCRo **outperforms** or **matches** comparable baselines on both **reasoning benchmarks** and alignment to **human behavior**, while being interpretable.

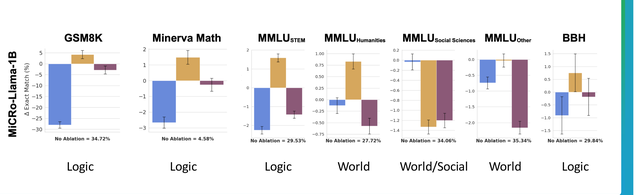
4 Semantically Meaningful Routing Across Experts

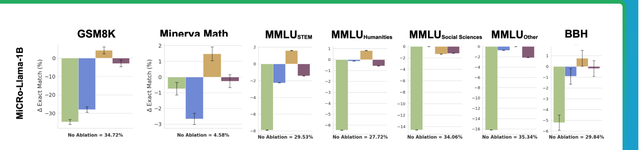




Hierarchy emerges with depth: early layers route to Language, deeper layers specialize

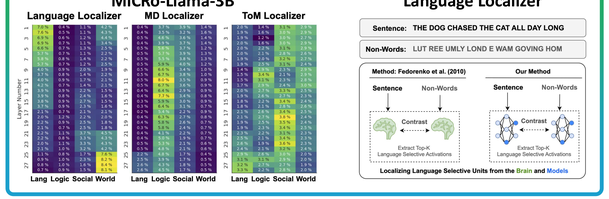
5 Expert Ablations Reveal the Causal Contributions



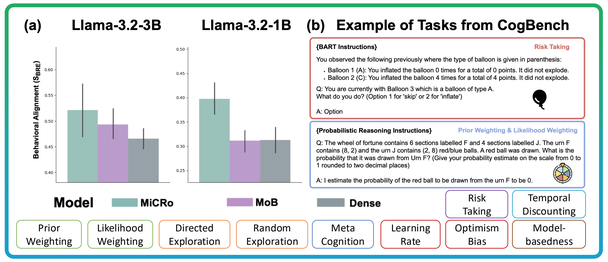


Language ablation hurts everywhere; reasoning experts matter most in-domain.

6 Neuroscience Localizers Recover Functionally Specialized Experts



7 Alignment with Human Behavior on CogBench



8 Competitive Performance on Reasoning Benchmarks