

Mixture of Cognitive Reasoners

Modular Reasoning with Brain-Like Specialization

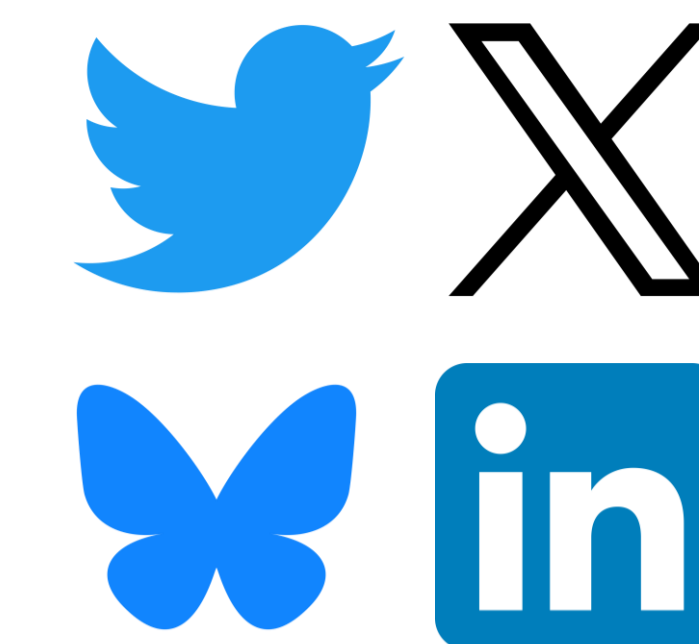
Badr AlKhamissi¹ C. Nicolò De Sabbata¹ Greta Tuckute^{2,3} Zeming Chen¹
 Martin Schrimpf*,¹ Antoine Bosselut*,¹



Demo



Project Page



@bkhmsi

¹ EPFL ² MIT ³ HARVARD

1 Neuroscience Motivation

2 Architecture

* What if our MoEs were like our BRAINS?

6 Neuroscience Localizers

Recover Functionally Specialized Experts

Lang **Logic** **Social** **World**

Language Network Multiple Demand Theory of Mind Default Mode

Brain-like specialization
→ more interpretable, and capable reasoners

- 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.

MiCRo-Llama-3B

Language Localizer MD Localizer ToM Localizer

Lang Logic Social World Lang Logic Social World Lang Logic Social World

Language Localizer

Sentence: THE DOG CHASED THE CAT ALL DAY LONG

Non-Words: LUT REE UMLY LOND E WAM GOVING HOM

Method: Fedorenko et al. (2016) Our Method

Sentence Non-Words Sentence Non-Words

Contrast Contrast

Extract Top-K Language Selective Activations

Localizing Language Selective Units from the Brain and Models

3 Training Curriculum

4 Semantically Meaningful Routing Across Experts

5 Expert Ablations Reveal the Causal Contributions

7 Alignment with Human Behavior on CogBench

Stage-I Prime Experts **Stage-II Calibrate Router** **Stage-III Train End-to-End**

MiCRo_{SFT} + Pseudo-Labels (~3K samples) MiCRo_{SFT} (~3K samples) Tulu3_{SFT} (~1M samples)

Language Samples Logic Samples Social Samples World Samples

Language Logic Social World

Sample Q: Correct the grammar: "She go to the park every morning." A: She goes to the park every morning.

Sample Q: Solve for x: 2x + 7 = 15 A: Subtract 7 from both sides. Then divide both sides by 2: x = 4.

Sample Q: Why did Sarah look away when John asked if she was okay? A: Because she didn't want him to see that she was upset.

Sample Q: Why do people usually eat breakfast in the morning? A: Because after sleeping the body needs energy to start the day.

(a) Llama-3.2-3B Llama-3.2-1B (b) Example of Tasks from CogBench

Behavioral Alignment (Spec)

Model MICRo MoB Dense

Prior Weighting Likelihood Weighting Directed Exploration Random Exploration Meta Cognition Learning Rate Optimism Bias Model-basedness

Risk Taking Temporal Discounting

(BART Instructions) Risk Taking

You observed the following previously where the type of balloon is given in parenthesis:

- Balloon 1 (A): You inflated the balloon 0 times for a total of 0 points. It did not explode.
- Balloon 2 (C): You inflated the balloon 4 times for a total of 4 points. It did not explode.
- Balloon 3 (B): You inflated the balloon 3 times for a total of 3 points. It exploded.

Q: You are currently with Balloon 3 which is a balloon of type A. What do you do? (Option 1 for "skip or 2 for "inflate")

A: Option

(Probabilistic Reasoning Instructions) Prior Weighting & Likelihood Weighting

Q: The wheel of fortune contains 8 sections labelled F and 4 sections labelled J. The urn F contains (8, 2) and the urn J contains (2, 8) red/blue balls. A red ball was drawn. What is the probability that it was drawn from Urn F? (Give your probability estimate on the scale from 0 to 1 rounded to two decimal places)

A: I estimate the probability of the red ball to be drawn from the urn F to be 0.

Language Samples Logic Samples Social Samples World Samples

Language Logic Social World

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

GSM8K Minerva Math MMLU_{STEM} MMLU_{Humanities} MMLU_{Social Sciences} MMLU_{Other} BBH

MiCRo-Llama-1B

Logic Logic Logic World World/Social World World

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

8 Competitive Performance on Reasoning Benchmarks

GSM8K Minerva Math MMLU BBH Average

Llama-3.2-1B

Model MICRo (Ablation) MICRo MoB (Ablation) MoB Dense

Score (%)