MIXTURE OF COGNITIVE REASONERS: MODULAR REASONING WITH BRAIN-LIKE SPECIALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Human cognitive behavior arises from the interaction of specialized brain networks dedicated to distinct functions, such as language, logic, and social reasoning. Inspired by this organization, we propose Mixture of Cognitive Reasoners (MICRO): a modular, transformer-based architecture post-trained with a curriculum that induces functional specialization across experts. Concretely, we partition the layers of a pretrained language model into four expert modules aligned with well-studied cognitive networks in the human brain. MICRO offers three key advantages over standard language models. (1) The specialized experts are interpretable and causally meaningful—ablating a module causes substantial drops on benchmarks requiring its specialized domain. (2) MICRO's behavior can be dynamically steered at inference time by routing tokens to particular experts (e.g., favoring social over logical reasoning), enabling fine-grained control over outputs. (3) MICRO outperforms or matches comparable baselines on both machine-learning reasoning benchmarks (e.g., GSM8K, BBH) and alignment to human behavior (CogBench), while maintaining interpretability. Taken together, cognitively grounded functional specialization yields models that are both more human-like and more human-interpretable.[1]

1 INTRODUCTION

Neuroscience research suggests that distinct brain regions support language, reasoning, social cognition, and other cognitive functions (Saxe & Kanwisher, 2003; Kanwisher, 2010; Fedorenko et al., 2024). In contrast, the internal organization of Large Language Models (LLMs) is largely unstructured. While certain units or subnetworks show selective activation (Zhang et al., 2022; 2023; Bayazit et al., 2023; AlKhamissi et al., 2025a; Wang et al., 2025), such specialization is implicit and difficult to interpret or control. Motivated by this discrepancy, we propose a model architecture that explicitly incorporates specialization. On the machine learning (ML) side, such designs hold great potential for improving interpretability and controllability; on the cognitive science side, they provide a framework toward formulating testable computational hypotheses about how the relative contributions of different brain networks support complex behavior. To this end, we propose the Mixture of Cognitive Reasoners (MICRO), a class of modular language models that partition computation across brain-inspired expert modules.

The MICRO architecture partitions each layer of a pretrained language model into four experts, each designed to mirror a major cognitive network in the human brain: language (Fedorenko et al., 2011), logic (multiple demand; Duncan, 2010), social reasoning (theory of mind; Saxe & Kanwisher, 2003), and world knowledge (default mode; Gusnard et al., 2001). To provide the model with the inductive bias needed to learn this partitioning and cohesively integrate these experts, we design a three-stage curriculum that uses lightweight training in the first two stages to sequentially (1) specialize the experts to mirror cognitive networks, and (2) bias a router to use certain experts for particular types of inputs (e.g., the logic expert for mathematics problems). The final training stage of this curriculum uses this now inductively-biased architecture to perform large-scale supervised finetuning.

[1] Code, data and models available at anonymous.4open.science/r/mixture-of-cognitive-reasoners-iclr

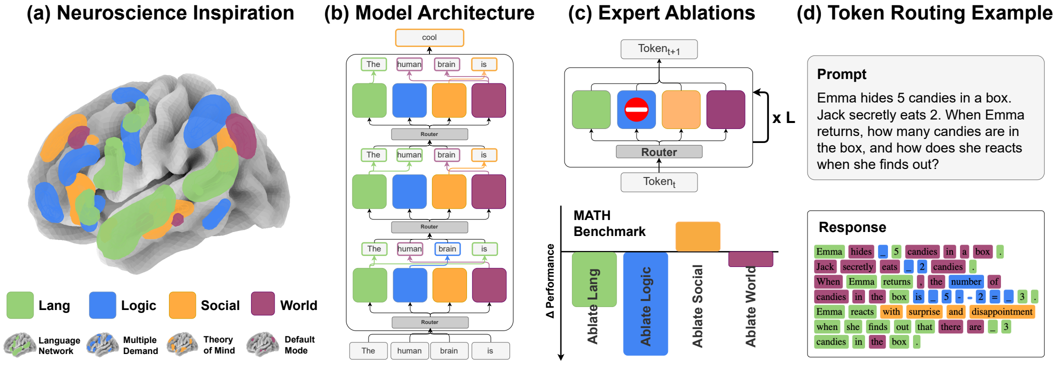

(a) Neuroscience Inspiration
(b) Model Architecture
(c) Expert Ablations
(d) Token Routing Example
Lang
Logic
Social
World
Language Network
Multiple Demand
Theory of Mind
Default Mode
cool
The
human
brain
is
Router
Token t+1
Token t
x L
MATH Benchmark
Δ Performance
Ablate Lang
Ablate Logic
Ablate Social
Ablate World
Prompt
Emma hides 5 candies in a box. Jack secretly eats 2. When Emma returns, how many candies are in the box, and how does she reacts when she finds out?
Response
Emma hides _ 5 candies in a box .
Jack secretly eats _ 2 candies .
When Emma returns , the number of
candies in the box is _ 5 _ - 2 = _ 3 .
Emma reacts with surprise and disappointment
when she finds out that there are _ 3
candies in the box .

Figure 1: **Brain-Inspired Modular Language Model.** **(a)** Illustration of major cognitive networks in the human brain. **(b)** Our proposed Mixture of Cognitive Reasoners (MICRO) architecture. The MICRO architecture partitions each transformer block into four expert modules corresponding to analogous brain networks; a router assigns each token to an expert at every layer (i.e., assignments can vary across layers and tokens). **(c)** Illustration for causal steering via mechanistic ablations: removing a module shifts behavior and degrades domain-relevant performance. **(d)** Token-level routing on a sample prompt shows semantically coherent expert usage.

Our results demonstrate that MICRO's architecture and training procedure induce interpretable specialization across these experts. This is evidenced by routing patterns and their correlations with human judgments (§5.1) and by causal ablations, which show dramatic drops in performance on reasoning categories when their corresponding experts are removed (§5.2). Moreover, the semantic behavior of these experts parallels the specialization of brain networks: (1) functional localizers used to recover brain-like mechanisms in LLMs (AlKhamissi et al., 2025a) identify the relevant experts in MICRO (§5.3), and (2) large MICRO models achieve high behavioral alignment scores on COGBENCH (Coda-Forno et al., 2024), a human behavioral benchmark, relative to two critical controls trained on the same data: (i) a mixture-of-experts model without induced brain-like specialization (MOB) and (ii) a non-modular dense transformer (DENSE) (§5.4). Finally, we find that MICRO's performance matches or exceeds these baselines (§5.5), indicating that interpretable and controllable specialization can be achieved without sacrificing overall performance.

2 BACKGROUND & RELATED WORK

2.1 NEUROSCIENCE MOTIVATION

Our design follows evidence that human cognition emerges from interacting, specialized brain networks. Cognitive neuroscience has mapped this modular organization by measuring how strongly different regions engage when people perform specific cognitive tasks (Kanwisher, 2010). We align MICRO's architecture with four core cognitive networks as shown in Figure 1(a). We summarize the functions of these networks below and their relevance to our modeling approach.

The Language Network. The *Language* expert mirrors the human language network, which comprises a set of left-lateralized frontal and temporal regions that selectively respond to linguistic input over perceptually matched non-linguistic stimuli (e.g., lists of nonwords; Fedorenko et al., 2010). These regions are highly specific to language, showing minimal activation during tasks such as arithmetic or music perception (Fedorenko et al., 2012; 2011), and their disruption can lead to selective language deficits (aphasia) without impairing general reasoning capabilities (Varley et al., 2005).

The Multiple Demand Network. The *Logic* expert mirrors the Multiple Demand (MD) network, which spans bilateral regions and is activated across diverse cognitively demanding tasks such as difficult math problems, with stronger responses for higher difficulty levels (Duncan & Owen, 2000; Fedorenko et al., 2013). It correlates with fluid intelligence (Woolgar et al., 2010).

The Theory of Mind Network. The *Social* expert mirrors the Theory of Mind (ToM) network, which is centered in the bilateral temporo-parietal junction and medial prefrontal cortex. This network supports reasoning about beliefs, intentions, and mental states (Gallagher et al., 2000; Saxe & Kanwisher, 2003; Saxe & Powell, 2006). It is robustly recruited across both verbal and non-verbal tasks involving perspective-taking and indirect communication (Koster-Hale & Saxe, 2013).

The Default Mode Network. The *World* expert mirrors the Default Mode Network (DMN), which is active during rest and internally directed thought such as self-reflection, memory recall, and mental simulation (Gusnard et al., 2001; Buckner et al., 2008; Buckner & DiNicola, 2019). Centered in medial prefrontal and parietal regions, the DMN integrates information over long timescales, supporting discourse- and event-level processing across sentences or episodes (Hassabis & Maguire, 2007; Fedorenko et al., 2024), in contrast to the shorter temporal scope of the language network.

2.2 Modular Language Models

In parallel with advances in cognitive neuroscience, recent years have seen growing interest in modular language models as a way to promote specialization, mitigate interference, and improve out-of-distribution generalization (Pfeiffer et al., 2023; Zhang et al., 2025). One major line of work centers on Sparse Mixture-of-Experts (MoE) architectures (Shazeer et al., 2017), with approaches ranging from curating domain-labeled datasets to train (Gururangan et al., 2022) or prompt (Si et al., 2023) domain-specific experts, to frameworks such as ModuleFormer (Shen et al., 2023), which introduce load-balancing and concentration losses to encourage modular specialization without explicit domain labels. Other modular approaches extend to multimodal integration (Liu et al., 2023; Swamy et al., 2023; Ye et al., 2023) or to disentangling representations by domain or language for multilingual and domain-specific applications (Pfeiffer et al., 2020; 2022; Zhong et al., 2022; Al-Maamari et al., 2024). In contrast, MiCRo is, to our knowledge, the first modular language model explicitly designed to induce brain-like specialization, aligning experts with well-studied cognitive networks.

2.3 Brain-Inspired Models

Recent studies have shown that some models achieve strong alignment with activity in the human language network (Schrimpf et al., 2021; Toneva & Wehbe, 2019; Caucheteux & King, 2022; Aw et al., 2023; AlKhamissi et al., 2025b). To further improve brain alignment, researchers have begun to integrate biologically inspired principles into model design—drawing from structures like the visual cortex hierarchy (Kubilius et al., 2019; Dapello et al., 2020; Spoerer et al., 2020), and the spatio-functional organization of the brain (Margalit et al., 2024; Rathi et al., 2025).

3 The Mixture of Cognitive Reasoners Framework

3.1 Model Architecture

To build MiCRo, we begin with a pretrained transformer-based backbone. For each layer, we clone the entire transformer block N times, where N corresponds to the number of experts intended for specialization, in a similar spirit to parameter upcycling (Komatsuzaki et al., 2023; Zhang et al., 2024). Then, we initialize a MLP-based router that assigns each token to a single expert. To maintain computational efficiency and a comparable number of active parameters to the original model, we use top-1 routing akin to Fedus et al. (2022). We refer to this architecture as *mixture-of-blocks* (MoB), distinguishing it from the more common mixture-of-experts (MoE), which restricts experts to FFN layers with shared attention. Importantly, we focus on MoB in the main paper because it induces clear functional specialization in all models, as reflected by lower router entropy and domain-consistent routing patterns, whereas MoE does not exhibit the same effect at specific scales (see Appendix H.1). Results for MiCRo-MoE variants on reasoning benchmarks in Appendix H.2.

3.2 Training Curriculum for Inducing Specialization

We induce functional specialization in MiCRo experts using a three-stage training curriculum (see Figure 2). The first two stages use a small, curated dataset (MiCRo$_{\text{SFT}}$) to provide targeted inductive biases, allowing specialization to emerge and solidify during the final full-scale training stage.

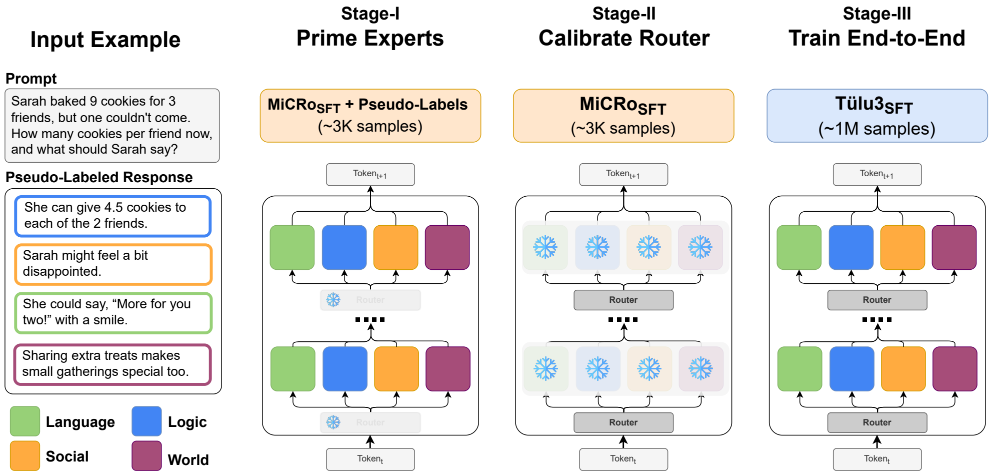

Input Example
Prompt
Sarah baked 9 cookies for 3 friends, but one couldn't come. How many cookies per friend now, and what should Sarah say?
Pseudo-Labeled Response
She can give 4.5 cookies to each of the 2 friends.
Sarah might feel a bit disappointed.
She could say, "More for you two!" with a smile.
Sharing extra treats makes small gatherings special too.
Language
Logic
Social
World
Stage-I
Prime Experts
$\text{MiCRo}_{\text{SFT}}$ + Pseudo-Labels
(~3K samples)
Stage-II
Calibrate Router
$\text{MiCRo}_{\text{SFT}}$
(~3K samples)
Stage-III
Train End-to-End
$\text{Tülu3}_{\text{SFT}}$
(~1M samples)
Token_{t+1}
Router
Token_t

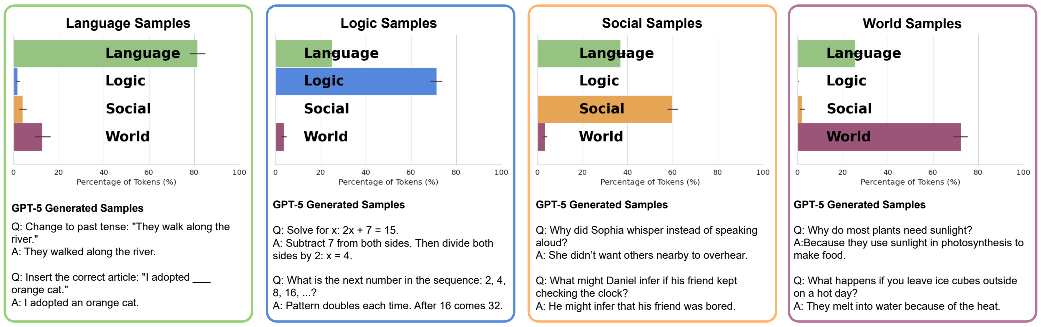
Figure 2: **Training Curriculum for Inducing Specialization.** Our brain-inspired Mixture of Cognitive Reasoners (MICRO) model contains four experts per layer, each aligned with a distinct cognitive network in the brain. In Stage-I, we train only the experts using a small, curated dataset $\text{MICRO}_{\text{SFT}}$ (see example on the left), providing each expert with an initial inductive bias. In Stage-II, we freeze the whole model and train the router on the same dataset to learn expert selection. In Stage-III, we finetune the entire model end-to-end on a large-scale instruction tuning dataset.

Stage 1: Inducing Specialization. In the first stage, we train the experts on a small dataset of $M = 3055$ examples (described below), each crafted to reflect the functional domain of a specific expert (Section 2.1). We denote this dataset as $\text{MICRO}_{\text{SFT}} = \{(x_{i,1:T_i}, r_{i,1:T_i})\}_{i=1}^{M}$, where each input sequence x_i contains T_i tokens, and r_i provides token-level routing labels. Each label $r_{i,t} \in \{1, \ldots, N\}$ assigns the t-th token to one of the N experts. This stage focuses solely on training the expert parameters using a next-token prediction loss. Tokens attend to all preceding tokens in the sequence regardless of which expert processed them using the key and value representations produced by the same expert. However, only tokens that are assigned to the expert in question continue to be processed through the feed-forward network. The same setup is applied in the next training stages, with the only difference that the router assigns the tokens to the experts.

Stage 2: Calibrating Router. Next, we freeze the whole model and train only the routers on the same dataset $\text{MICRO}_{\text{SFT}}$. The objective remains next-token prediction. Given the initial expert specialization from Stage 1, the router now learns to assign tokens to the most suitable expert. To encourage smoother transitions and more robust routing decisions, we use a soft mixture of the top-2 experts per token, which we found to be more effective than top-1 routing during this phase.

Stage 3: End-to-End Supervised Finetuning. Finally, we finetune the entire model end-to-end on a full instruction-tuning dataset, TÜLU-3 (Lambert et al., 2024), which consists of 939k examples. Even though this phase constitutes the majority of the training budget, we observe that the functional specialization seeded by the small $\text{MICRO}_{\text{SFT}}$ dataset is largely preserved (see Appendix J). Moreover, the experts continue to improve on tasks aligned with their initial domains, demonstrating that early inductive biases can lead to meaningful and lasting functional decomposition.

Constructing the $\text{MICRO}_{\text{SFT}}$ Dataset. To build $\text{MICRO}_{\text{SFT}}$ for inducing expert and router specialization, we first selected 19 existing reasoning datasets corresponding to the cognitive domains of our non-language experts, ensuring that each group of datasets spanned a diverse range of functions known to engage the corresponding brain networks. From each of the three sets, we randomly sampled 1,000 examples and used OpenAI's o1 model (Jaech et al., 2024) to generate detailed, step-by-step responses for each input. We then pseudo-labeled each sentence in the generated reasoning chains by prompting GPT-4o (Hurst et al., 2024) to assign it to one of the four experts. The tokens within each sentence inherit the corresponding expert label, which is used for deterministic routing in Stage 1. Details of the datasets are provided in Appendix A.



GPT-5 Generated Samples

Q: Change to past tense: "They walk along the river."
A: They walked along the river.

Q: Insert the correct article: "I adopted ___ orange cat."
A: I adopted an orange cat.

GPT-5 Generated Samples

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

Q: What is the next number in the sequence: 2, 4, 8, 16, ...?
A: Pattern doubles each time. After 16 comes 32.

GPT-5 Generated Samples

Q: Why did Sophia whisper instead of speaking aloud?
A: She didn't want others nearby to overhear.

Q: What might Daniel infer if his friend kept checking the clock?
A: He might infer that his friend was bored.

GPT-5 Generated Samples

Q: Why do most plants need sunlight?
A:Because they use sunlight in photosynthesis to make food.

Q: What happens if you leave ice cubes outside on a hot day?
A: They melt into water because of the heat.

Figure 3: **Semantically Meaningful Routing Across Experts.** Token routing patterns in MICRO-LLAMA-1B. Each bar indicates the proportion of tokens routed to a given expert across layers, with variance shown across sentences (n=50). The model exhibits clear domain-specific specialization consistent with the intended brain-inspired organization. For example, social cognition samples are routed to the social expert, while arithmetic tasks are routed to the logic expert. We find that the language expert is consistently activated in the early layers (see Appendix C for layer-wise routing plots and results from additional models). Two random samples are shown below each subplot.

4 Experimental Setup

We post-train five models of varying scales from three different families under our MICRO framework, in order to assess the generalizability of our method and identify the conditions under which it fails. Specifically, we use LLAMA-3.2-{1B, 3B} (Dubey et al., 2024), SMOLLM2-{135M, 360M} (Allal et al., 2025), and OLMO-2-1B (OLMo et al., 2024). Due to space constraints, we present the results of LLAMA-3.2-{1B, 3B} in the main paper while providing the full results for the remaining models in Appendix F. Each model is first finetuned for two epochs on the curated MICRO$_{\text{SFT}}$ dataset (Stages 1 and 2), followed by one epoch of end-to-end training using the TÜLU-3 dataset (Lambert et al., 2024), as described in Section 3.2. We use next token prediction as the only learning objective in all training stages, with the loss masked on the input tokens. We use an effective batch size of 32 and the AdamW optimizer across all stages. The learning rate follows a linear schedule, warming up over the first 3% of training to a peak of 2×10^{-5}, then decays linearly for the remainder of training. This schedule is applied for each stage separately.

Reasoning Benchmarks. We evaluate on four widely used reasoning benchmarks: GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021b), BBH (Suzgun et al., 2022), and MMLU (Hendrycks et al., 2021a). Evaluation follows zero- or fewshot settings as detailed in Appendix F.

Behavioral Benchmarks. We evaluate alignment to human behavior using COGBENCH benchmark (Coda-Forno et al., 2024), which provides 10 metrics from 7 cognitive psychology experiments. These metrics capture how participants (or models) complete tasks that are designed to disentangle different behavioral strategies. Examples include *Directed Exploration*, *Meta-Cognition*, and *Risk Taking*. We refer readers to Coda-Forno et al. (2024) for a detailed description of the tasks.

5 Results & Analysis

Our results unfold in two parts. First, we ask whether brain-like specialization emerges under our training curriculum, analyzing routing behavior, correlations with human judgments, causal ablations to test the functional contributions of those experts, and whether neuroscience experiments used to identify brain networks also identify the corresponding experts in our models. Second, we ask how this specialization influences alignment with human behavior and reasoning performance.

5.1 Router Patterns Are Interpretable and Consistent with Human Judgments

Token Routing Per Expert. We first verify that our model routes tokens to the most relevant expert module, analogous to how specialized brain networks are selectively engaged by specific stim-

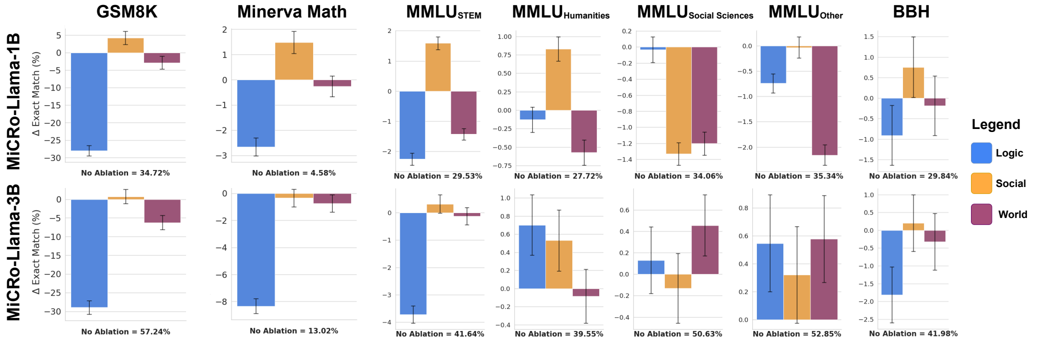


Figure 4: **Expert Ablations Reveal the Causal Contributions of Specialized modules.** Top and bottom panels show results for MiCRo-Llama-1B and MiCRo-Llama-3B. Removing the Logic expert causes large drops on MATH and GSM8K, while removing the Social expert yields slight gains. For MMLU and BBH, results indicate that some group of subtasks rely on distinct experts, whereas others draw on overlapping contributions. Additional models in Appendix E.

uli. Figure 3 shows the routing behavior of MiCRo-Llama-1B, revealing clear domain-specific specialization. To generate test inputs, we sampled 50 question–answer pairs using GPT-5 prompted with descriptions of the four brain networks (prompt provided in Appendix C). Results for the other MiCRo variants are consistent and reported in Appendix C. There, we also show routing patterns on reasoning benchmarks, where tokens are directed to experts consistent with the benchmark's domain. Finally, layer-wise analyses in Figures 13 & 14 reveal a hierarchical organization: earlier layers focus on linguistic grounding, while deeper layers increasingly delegate to domain-specific experts—an organization that emerged without being enforced by the training procedure and that parallels evidence from cognitive neuroscience (Fedorenko et al., 2024).

Correlation with Human Judgments. We evaluate model–human correspondence using 1,000 six-word sentences from Tuckute et al. (2024), each annotated with human ratings across several behavioral dimensions (e.g., mental state content, grammaticality). These annotations were collected independently of our routing framework. We find that router probabilities correlate with the corresponding human judgments: for example, the social expert's selectivity aligns with ratings of mental state content ($r = 0.7$). Full results are provided in Appendix D.

5.2 Experts Are Causally Meaningful

Validation of Functional Experts via Ablations. Figure 4 illustrates how expert ablations reveal the causal contributions of specialized modules to task performance. By selectively removing individual experts, we can directly test whether their specialization is functionally necessary for different domains. For example, on MATH and GSM8K, ablating the *Logic* expert causes a substantial drop in accuracy, confirming its central role in numerical reasoning. In contrast, removing the *Social* expert slightly improves performance, suggesting it plays a detrimental role in these tasks. For broader benchmarks such as MMLU, which span multiple subdomains, we report results for each subcategory separately. Performance drops after ablating the corresponding experts indicate that these clusters depend on distinct functional modules. Still, not all subtasks within a category align neatly with a single cognitive domain, and some require overlapping contributions, such as BBH. We show in Appendix E the effect of removing the language expert, which causes a significant drop on all benchmarks, along with additional ablation results on other models.

Steering Model Behavior at Test-Time. Our results demonstrate that test-time ablations can steer expert behavior, with social responses emerging when only the social expert is active and logical reasoning dominating when only the logic expert is retained. Qualitative examples in Appendix K.

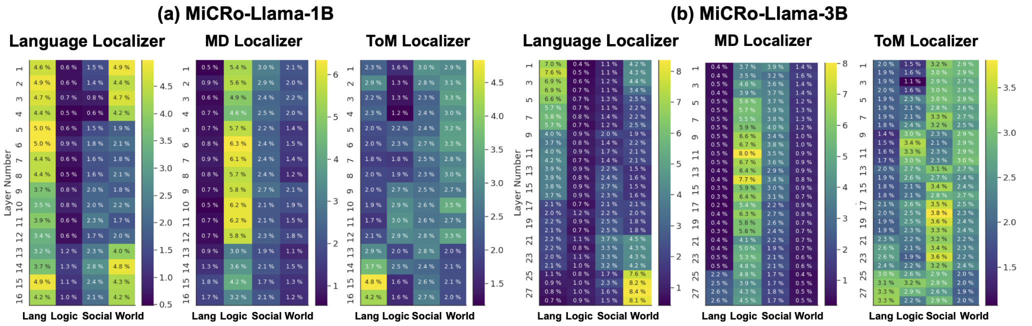


Figure 5: **Neuroscience Localizers Recover Functionally Specialized Experts.** (a) MiCRo-Llama-1B and (b) MiCRo-Llama-3B. For each model, we apply three neuroscience-inspired localizers—Language, Multiple Demand (MD) and Theory of Mind (ToM)—to examine the selectivity of localized units across experts and layers. Each plot shows the percentage of units in each expert of each layer that belongs to the top-10% selective units in the whole model.

5.3 Neuroscience Localizers Reveal Functional Expert Specialization

Neuroscientists rely on localizer experiments to identify the brain regions associated with specific functional networks, as their precise locations can vary across individuals. This raises a natural question: can we apply these established neuroscience localizers to identify the corresponding expert modules in our model? If so, this would provide further support for the hypothesis that our experts are functionally analogous to their associated brain networks.

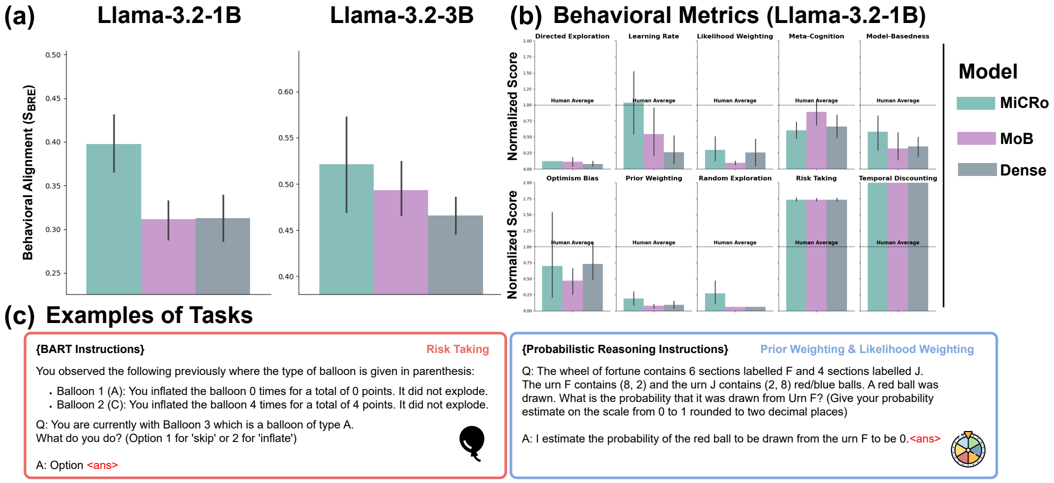
To investigate this, we adopt the methodology of AlKhamissi et al. (2025a), which has been used to localize the language network, the multiple demand network, and the theory of mind network in LLMs. We apply these localizers to our MiCRo models to test whether they can recover the corresponding expert modules. Figure 5 shows the percentage of units in each expert of each layer that belongs to the top 10% of selective units across the whole model, similar to what is done in the brain (Lipkin et al., 2022). The results show that language selectivity, as defined by the language localizer, favors the language expert at early layers while favoring the world expert at later layers for both models. The multiple demand localizer successfully favors the logic expert in both models. In contrast, ToM localization is less effective at isolating units within the social experts, but improves with scale, suggesting that ToM ability must emerge before it can be localized. One other possible reason for this is the limited size of the ToM stimulus set, which includes only 10 contrastive pairs, in contrast to 240 for language and 100 for multiple demand. This small sample may lack the robustness needed to reliably localize ToM-selective units (Jamaa et al., 2025).

5.4 Strong Alignment to Human Behavior

Having established that our MiCRo models exhibit human-like specialization (§5.1) that is causally linked to task performance (§5.2), we next examine whether they better align to human behavior compared to two baselines: one without brain-like specialization (MoB) and one without any modularization (Dense). Both models are post-trained on a mixture of 2× MiCRo$_{\text{SFT}}$ and 1× Tülu-3 matching the total amount of data used in the MiCRo training curriculum.

Figure 6 presents the results on CogBench, evaluating alignment with human behavior. Unlike the original paper, which predicts answers via autoregressive generation, we pick the option with the highest log-probability for multiple-choice tasks to avoid invalid generations. Each experiment is run with five random seeds. Metrics are normalized such that random = 0 and human = 1. To quantify overall alignment, we introduce the bounded relative error similarity score (S_{BRE}), which avoids inflation from superhuman scores. For a normalized score s_i on metric i, we compute $\text{BRE}_i = |s_i - 1| / \max(1, s_i)$ and aggregate as $S_{\text{BRE}} = 1 - \frac{1}{n}\sum_{i=1}^{n} \text{BRE}_i$. Thus, BRE_i remains bounded in $[0, 1]$ even if $s_i > 1$.

Overall, we find competitive alignment across models, with MiCRo-Llama-1B showing superior alignment compared to its counterparts. Panel (a) reports the average similarity score (S_{BRE}) aggre-



{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.

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 <ans>

{Probabilistic Reasoning Instructions} Prior Weighting & Likelihood Weighting

Q: The wheel of fortune contains 6 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.<ans>

Figure 6: **Alignment with Human Behavior on CogBench. (a)** Average similarity score (S_{BRE}) across 10 behavioral metrics, showing that MiCRo-Llama models achieves superior alignment compared to their MoB and Dense baselines. **(b)** Human-normalized scores for each metric separately across the three models. **(c)** Example inputs from two of the seven classical psychological experiments verbalized for LLM evaluation following CogBench.

gated across the 10 behavioral metrics for both MiCRo-Llama-{1B, 3B} models, while panel (b) breaks down the human-normalized scores for each metric separately across the three post-trained models for the Llama-3.2-1B base model. Finally, panel (c) illustrates input examples from two of the seven classical psychological experiments included in CogBench, which are verbalized for LLM evaluation following the original benchmark design. More results in Appendix I.

5.5 Competitive Performance on Reasoning Benchmarks

Here, we test whether our MiCRo models incur any performance degradation relative to their two baselines. Figure 7 shows performance on GSM8K, Minerva-Math, MMLU, MMLU-Pro, and BBH, along with their average. Models are evaluated using fewshot chain-of-thought prompting, except for GSM8K, which is evaluated under zero-shot CoT prompting. For both base models, MiCRo matches or exceeds comparable MoB baselines, while ablating the least relevant expert (i.e., the social expert for these benchmarks) further improves performance. We conduct pairwise Welch's t-tests between models and report significance directly in the plot. Results show that some base models, such as Llama-3.2-1B, benefit significantly from brain-like specialization, whereas others, such as Llama-3.2-3B, only show significant differences relative to their baselines on some benchmarks. We report additional results for the other models and benchmarks in Appendix F. We further show that our method is robust to different post-training pipelines, including DPO (Rafailov et al., 2023) and domain-specific instruction tuning (Appendix G).

6 Discussion & Future Work

Extending Specialization Beyond Cognitive Domains While inspired by the brain's functional organization, our specialization framework can be applied to any meaningful partition of expertise, such as technical domains or natural languages. One key question is whether the model preserves the intended specialization through the large-scale end-to-end training. Our results suggest that brain-inspired partitions provide a robust inductive bias; they persist throughout training and lead to structured, interpretable routing patterns. Supporting evidence in Appendix J shows that expert usage remains consistent across checkpoints over the course of Stage 3 training. Looking ahead, this framework could also be extended to other cognitive domains. For example, recent neuroscience findings point to a distinct brain network involved in abstract formal reasoning such as induction

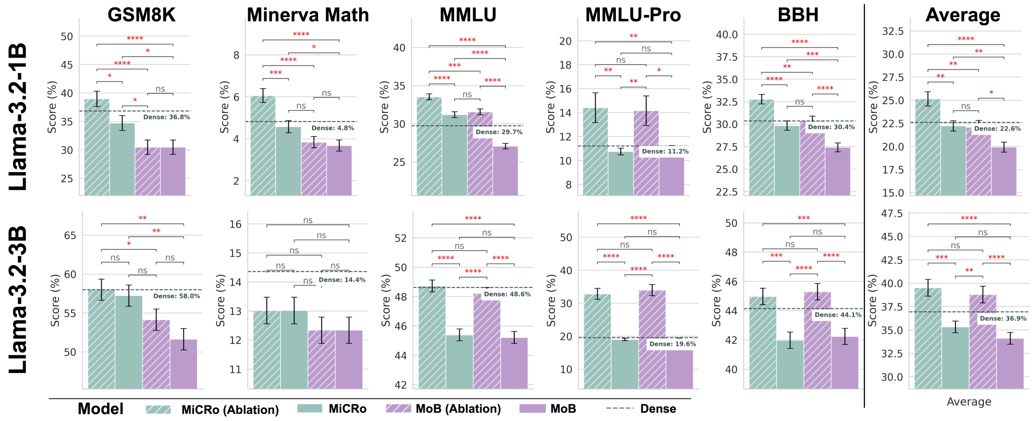


Figure 7: **Competitive Performance.** Results on GSM8K (0-shot CoT), MINERVA-MATH, MMLU, and BBH (fewshot CoT). MICRO matches or outperforms baselines, and ablating the least relevant expert (e.g., social for math benchmarks) yields further gains. For MOB (ABLATION) and MICRO (ABLATION) (on MMLU and BBH subtasks), results reflect the best performance obtained when ablating up to one expert. Significance is assessed with pairwise Welch's t-tests (shown in plot; $*$ denotes significant difference, while n.s. means not-significant). The dense model is shown as a dashed line. Results of the remaining models and on more benchmarks are provided in Appendix F.

and deduction and another network for intuitive physics (Kean et al., 2025a;b). Incorporating a corresponding module could improve the model's performance on tasks involving such capacities.

The Crucial Role of Stage-1 Pretraining Data Our experiments highlight the importance of the curated MICRO$_{\text{SFT}}$ dataset in inducing effective specialization. Notably, we used only 3,055 samples in Stage-1, suggesting that even minimal domain-aligned supervision can shape expert behavior. This finding raises the possibility that different or more expansive data mixtures could further strengthen functional specialization and lead to additional gains in the model's behavior.

Towards Brain Alignment Beyond Language Since our model is explicitly designed to mirror distinct cognitive networks in the human brain, and given that established neuroscience localizers can identify the corresponding expert modules, an exciting direction for future work is to examine whether the internal representations of these experts align more closely with neural activity in their respective brain networks (Schrimpf et al., 2018; 2020). Prior studies have shown that language-selective units in large language models correlate more strongly with activity in the human language network than randomly selected units (AlKhamissi et al., 2025a), suggesting a meaningful link between specialization in models and brains. This raises the natural question of whether similar alignment can be observed for other cognitive domains, such as reasoning or social cognition. However, assessing MICRO's neural alignment beyond the language network is currently limited by the lack of suitable datasets. Existing fMRI benchmarks rarely engage non-language regions such as the Multiple Demand (Duncan, 2010) network and often use blocked designs that preclude item-level analyses—highlighting the need for experimentalists to collect new datasets that explicitly target non-language brain regions. We believe that once suitable neural datasets exist, our model can be used to instantiate specific hypotheses about how these networks—and their corresponding experts—interact and exchange information.

Limitations While our approach improves interpretability without sacrificing performance, several open questions remain. Scaling beyond an 8B base model has yet to be demonstrated, and the impact of adding more experts to the current MICRO architecture is still unknown. The MICRO$_{\text{SFT}}$ dataset used in Stage-1 ($\approx$ 3,000 GPT-4o pseudo-labeled samples) has not been evaluated for size sensitivity, leaving open whether increasing or reducing its scale would alter the degree of specialization or downstream performance. Although GPT-4O provides high-quality pseudo-labels, as demonstrated in Appendix B, using human-annotated data could strengthen the inductive bias and potentially improve the final model. We expect, however, that the potential of this approach will continue to grow as synthetic labellers become more accurate and reliable, enabling even stronger and more scalable specialization in future versions of this class of methods. Further, our post-training pipeline currently includes only SFT and DPO (Appendix G); exploring additional stages such as RLVR remains an avenue for future work. Finally, our evaluation of alignment to human behavior focuses on CogBench, and extending this analysis to a broader set of behavioral or cognitive datasets is an important direction for future research.

7 Conclusion

This work presents the *Mixture of Cognitive Reasoners* (MICRO), a modular architecture and training paradigm inspired by the functional specialization of the human brain. By aligning expert modules with distinct cognitive domains—language, logic, social reasoning, and world modeling—each reflecting the functionality of well-studied brain networks, we show that incorporating cognitive inductive biases into transformer models can effectively promote functional specialization. This results in improved behavioral alignment as measured by COGBENCH, enhanced interpretability, all while being competitive or outperforming comparable models on reasoning benchmarks. Our staged training approach leverages a small curated dataset and enables specialization to emerge in a controllable and data-efficient manner. Furthermore, we show that the resulting modularity allows for targeted interventions at inference time, enabling the model to favor one mode of reasoning over another. These findings highlight a promising path toward more transparent, steerable, and cognitively grounded language models.

References

Mohammed Al-Maamari, Mehdi Ben Amor, and Michael Granitzer. Mixture of modular experts: Distilling knowledge from a multilingual teacher into specialized modular language models. *ArXiv*, abs/2407.19610, 2024. URL `https://api.semanticscholar.org/CorpusID:271533631`.

Badr AlKhamissi, Greta Tuckute, Antoine Bosselut, and Martin Schrimpf. The LLM language network: A neuroscientific approach for identifying causally task-relevant units. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 10887–10911, Albuquerque, New Mexico, April 2025a. Association for Computational Linguistics. ISBN 979-8-89176-189-6. URL `https://aclanthology.org/2025.naacl-long.544/`.

Badr AlKhamissi, Greta Tuckute, Yingtian Tang, Taha Binhuraib, Antoine Bosselut, and Martin Schrimpf. From language to cognition: How llms outgrow the human language network. *arXiv preprint arXiv:2503.01830*, 2025b.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, et al. Smollm2: When smol goes big–data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025.

Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. Instruction-tuning aligns llms to the human brain. 2023.

Deniz Bayazit, Negar Foroutan, Zeming Chen, Gail Weiss, and Antoine Bosselut. Discovering knowledge-critical subnetworks in pretrained language models. *ArXiv*, abs/2310.03084, 2023. URL `https://api.semanticscholar.org/CorpusID:263671765`.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

Antoine Bosselut, Zeming Chen, Angelika Romanou, Antoine Bonnet, Alejandro Hernández-Cano, Badr Alkhamissi, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas

Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Vinitra Swamy, Alireza Sakhaeirad, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Li Mi, and Martin Jaggi. Meditron: Open medical foundation models adapted for clinical practice, 03 2024.

Randy L. Buckner and Lauren M. DiNicola. The brain's default network: updated anatomy, physiology and evolving insights. *Nature Reviews Neuroscience*, 20(10):593–608, September 2019. ISSN 1471-0048. doi: 10.1038/s41583-019-0212-7. URL http://dx.doi.org/10.1038/s41583-019-0212-7.

Randy L. Buckner, Jessica R. Andrews-Hanna, and Daniel L. Schacter. The brain's default network: Anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences*, 1124(1):1–38, March 2008. ISSN 1749-6632. doi: 10.1196/annals.1440.011. URL http://dx.doi.org/10.1196/annals.1440.011.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. Modeling empathy and distress in reaction to news stories. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4758–4765, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1507. URL https://aclanthology.org/D18-1507/.

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications biology*, 5(1):134, 2022.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Julian Coda-Forno, Marcel Binz, Jane X Wang, and Eric Schulz. CogBench: a large language model walks into a psychology lab. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 9076–9108. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/coda-forno24a.html.

Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David D. Cox, and James J. DiCarlo. Simulating a Primary Visual Cortex at the Front of CNNs Improves Robustness to Image Perturbations. In *Neural Information Processing Systems (NeurIPS, Spotlight)*, June 2020. doi: 10.1101/2020.06.16.154542. URL https://www.biorxiv.org/content/10.1101/2020.06.16.154542v1.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, and et al. The llama 3 herd of models. *ArXiv*, abs/2407.21783, 2024. URL https://api.semanticscholar.org/CorpusID:271571434.

John Duncan. The multiple-demand (md) system of the primate brain: mental programs for intelligent behaviour. *Trends in cognitive sciences*, 14(4):172–179, 2010.

John Duncan and Adrian M Owen. Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences*, 23(10):475–483, October 2000. ISSN 0166-2236. doi: 10.1016/s0166-2236(00)01633-7. URL http://dx.doi.org/10.1016/S0166-2236(00)01633-7.

Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castanon, Susan L. Whitfield-Gabrieli, and Nancy G. Kanwisher. New method for fmri investigations of language: defining rois functionally in individual subjects. *Journal of neurophysiology*, 104 2:1177–94, 2010. URL https://api.semanticscholar.org/CorpusID:740913.

Evelina Fedorenko, Michael K Behr, and Nancy Kanwisher. Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences*, 108 (39):16428–16433, 2011.

Evelina Fedorenko, Josh H. McDermott, Sam Norman-Haignere, and Nancy Kanwisher. Sensitivity to musical structure in the human brain. *Journal of Neurophysiology*, 108(12):3289–3300, December 2012. ISSN 1522-1598. doi: 10.1152/jn.00209.2012. URL http://dx.doi.org/10.1152/jn.00209.2012.

Evelina Fedorenko, John Duncan, and Nancy Kanwisher. Broad domain generality in focal regions of frontal and parietal cortex. *Proceedings of the National Academy of Sciences*, 110 (41):16616–16621, September 2013. ISSN 1091-6490. doi: 10.1073/pnas.1315235110. URL http://dx.doi.org/10.1073/pnas.1315235110.

Evelina Fedorenko, Anna A. Ivanova, and Tamar I. Regev. The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, 25(5):289–312, April 2024. ISSN 1471-0048. doi: 10.1038/s41583-024-00802-4. URL http://dx.doi.org/10.1038/s41583-024-00802-4.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

H.L Gallagher, F Happé, N Brunswick, P.C Fletcher, U Frith, and C.D Frith. Reading the mind in cartoons and stories: an fmri study of 'theory of mind' in verbal and nonverbal tasks. *Neuropsychologia*, 38(1):11–21, January 2000. ISSN 0028-3932. doi: 10.1016/s0028-3932(99)00053-6. URL http://dx.doi.org/10.1016/s0028-3932(99)00053-6.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. Understanding social reasoning in language models with language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=8bqjirgxQM.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

Suchin Gururangan, Mike Lewis, Ari Holtzman, Noah A. Smith, and Luke Zettlemoyer. DEMix layers: Disentangling domains for modular language modeling. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5557–5576, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.407. URL https://aclanthology.org/2022.naacl-main.407/.

Debra A. Gusnard, Erbil Akbudak, Gordon L. Shulman, and Marcus E. Raichle. Medial prefrontal cortex and self-referential mental activity: Relation to a default mode of brain function. *Proceedings of the National Academy of Sciences*, 98(7):4259–4264, March 2001. ISSN 1091-6490. doi: 10.1073/pnas.071043098. URL http://dx.doi.org/10.1073/pnas.071043098.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, David Peng, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Shafiq Joty, Alexander R. Fabbri, Wojciech Kryscinski, Xi Victoria Lin, Caiming Xiong, and Dragomir Radev. Folio: Natural language reasoning with first-order logic. *arXiv preprint arXiv:2209.00840*, 2022. URL https://arxiv.org/abs/2209.00840.

Demis Hassabis and Eleanor A. Maguire. Deconstructing episodic memory with construction. *Trends in Cognitive Sciences*, 11(7):299–306, July 2007. ISSN 1364-6613. doi: 10.1016/j.tics.2007.05.001. URL http://dx.doi.org/10.1016/j.tics.2007.05.001.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021b.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. A fine-grained comparison of pragmatic language understanding in humans and language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4194–4213, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.230. URL https://aclanthology.org/2023.acl-long.230/.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Hamish Ivison, Yizhong Wang, Jiacheng Liu, Ellen Wu, Valentina Pyatkin, Nathan Lambert, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. Unpacking dpo and ppo: Disentangling best practices for learning from preference feedback, 2024.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Yassine Jamaa, Badr AlKhamissi, Satrajit Ghosh, and Martin Schrimpf. Evaluating contrast localizer for identifying causal unitsin social & mathematical tasks in language models. *arXiv preprint arXiv:2508.08276*, 2025.

Nancy Kanwisher. Functional specificity in the human brain: A window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences*, 107(25):11163–11170, May 2010. ISSN 1091-6490. doi: 10.1073/pnas.1005062107. URL http://dx.doi.org/10.1073/pnas.1005062107.

Hope Kean, Alex Fung, Chiebuka Ohams, Jason Chen, Josh Rule, Joshua Tenenbaum, Steven Piantadosi, and Evelina Fedorenko. A human brain network specialized for abstract formal reasoning. *bioRxiv*, 2025a. doi: 10.1101/2025.10.21.683445. URL https://www.biorxiv.org/content/early/2025/10/21/2025.10.21.683445.

Hope H. Kean, Alexander Fung, R.T. Pramod, Jessica Chomik-Morales, Nancy Kanwisher, and Evelina Fedorenko. Intuitive physical reasoning is not mediated by linguistic nor exclusively domain-general abstract representations. *Neuropsychologia*, 213:109125, July 2025b. ISSN 0028-3932. doi: 10.1016/j.neuropsychologia.2025.109125. URL http://dx.doi.org/10.1016/j.neuropsychologia.2025.109125.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. In *EMNLP*, 2021.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. Fantom: A benchmark for stress-testing machine theory of mind in interactions. In *EMNLP*, 2023.

Aran Komatsuzaki, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. Sparse upcycling: Training mixture-of-experts from dense checkpoints. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=T5nUQDrM4u.

Jorie Koster-Hale and Rebecca Saxe. Theory of mind: A neural prediction problem. *Neuron*, 79(5):836–848, September 2013. ISSN 0896-6273. doi: 10.1016/j.neuron.2013.08.020. URL http://dx.doi.org/10.1016/j.neuron.2013.08.020.

Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Aran Nayebi, Daniel Bear, Daniel L K Yamins, and James J DiCarlo. Brain-like object recognition with high-performing shallow recurrent anns. In *Advances in Neural Information Processing Systems*, 2019.

Nathan Lambert, Jacob Daniel Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James Validad Miranda, Alisa Liu, Nouha Dziri, Xinxi Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hanna Hajishirzi. Tülu 3: Pushing frontiers in open language model post-training. *ArXiv*, abs/2411.15124, 2024. URL https://api.semanticscholar.org/CorpusID:274192505.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large scale language model society, 2023.

Lei Li, Yuanxin Liu, Linli Yao, Peiyuan Zhang, Chenxin An, Lean Wang, Xu Sun, Lingpeng Kong, and Qi Liu. Temporal reasoning transfer from text to video. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=sHAvMp5J4R.

Benjamin Lipkin, Greta Tuckute, Josef Affourtit, Hannah Small, Zachary Mineroff, Hope Kean, Olessia Jouravlev, Lara Rakocevic, Brianna Pritchett, Matthew Siegelman, Caitlyn Hoeflin, Alvincé Pongos, Idan A. Blank, Melissa Kline Struhl, Anna Ivanova, Steven Shannon, Aalok Sathe, Malte Hoffmann, Alfonso Nieto-Castañón, and Evelina Fedorenko. Probabilistic atlas for the language network based on precision fmri data from¿800 individuals. *Scientific Data*, 9(1), August 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01645-3. URL http://dx.doi.org/10.1038/s41597-022-01645-3.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*, 2020.

Eshed Margalit, Hyodong Lee, Dawn Finzi, James J. DiCarlo, Kalanit Grill-Spector, and Daniel L.K. Yamins. A unifying framework for functional organization in early and higher ventral visual cortex. *Neuron*, 112(14):2435–2451.e7, July 2024. ISSN 0896-6273. doi: 10.1016/j.neuron.2024.04.018. URL http://dx.doi.org/10.1016/j.neuron.2024.04.018.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 OLMo 2 Furious, 2024. URL https://arxiv.org/abs/2501.00656.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7654–7673, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.617. URL https://aclanthology.org/2020.emnlp-main.617/.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. Lifting the curse of multilinguality by pre-training modular transformers. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3479–3495, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.255. URL https://aclanthology.org/2022.naacl-main.255/.

Jonas Pfeiffer, Sebastian Ruder, Ivan Vulić, and Edoardo Ponti. Modular deep learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=z9EkXfvxta. Survey Certification.

Yiwei Qin, Xuefeng Li, Haoyang Zou, Yixiu Liu, Shijie Xia, Zhen Huang, Yixin Ye, Weizhe Yuan, Zhengzhong Liu, Yuanzhi Li, and Pengfei Liu. O1 replication journey: A strategic progress report – part 1. https://github.com/GAIR-NLP/O1-Journey, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.

Neil Rathi, Johannes Mehrer, Badr AlKhamissi, Taha Osama A Binhuraib, Nicholas Blauch, and Martin Schrimpf. TopoLM: brain-like spatio-functional organization in a topographic language model. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=aWXnKanInf.

R Saxe and N Kanwisher. People thinking about thinking peoplethe role of the temporo-parietal junction in “theory of mind”. *NeuroImage*, 19(4):1835–1842, August 2003. ISSN 1053-8119. doi: 10.1016/s1053-8119(03)00230-1. URL http://dx.doi.org/10.1016/s1053-8119(03)00230-1.

Rebecca Saxe and Lindsey J. Powell. It’s the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science*, 17(8):692–699, August 2006. ISSN 1467-9280. doi: 10.1111/j.1467-9280.2006.01768.x. URL http://dx.doi.org/10.1111/j.1467-9280.2006.01768.x.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? preprint, Neuroscience, September 2018. URL http://biorxiv.org/lookup/doi/10.1101/407007.

Martin Schrimpf, Jonas Kubilius, Michael J Lee, N Apurva Ratan Murty, Robert Ajemian, and James J DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 108(3):413–423, 2020.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021. doi: 10.1073/pnas.2105646118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2105646118.

Noam Shazeer, *Azalia Mirhoseini, *Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

Yikang Shen, Zheyu Zhang, Tianyou Cao, Shawn Tan, Zhenfang Chen, and Chuang Gan. Moduleformer: Modularity emerges from mixture-of-experts. 2023. URL https://api.semanticscholar.org/CorpusID:261697418.

Chenglei Si, Weijia Shi, Chen Zhao, Luke Zettlemoyer, and Jordan Boyd-Graber. Getting MoRE out of mixture of language model reasoning experts. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8234–8249, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.552. URL https://aclanthology.org/2023.findings-emnlp.552/.

Courtney J. Spoerer, Tim C. Kietzmann, Johannes Mehrer, Ian Charest, and Nikolaus Kriegeskorte. Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLOS Computational Biology*, 16(10):e1008215, October 2020. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1008215. URL http://dx.doi.org/10.1371/journal.pcbi.1008215.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, , and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

Vinitra Swamy, Malika Satayeva, Jibril Frej, Thierry Bossy, Thijs Vogels, Martin Jaggi, Tanja Käser, and Mary-Anne Hartley. MultimoDN—multimodal, multi-task, interpretable modular networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=iB3Ew6z4WL.

Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Neural Information Processing Systems*, 2019. URL https://api.semanticscholar.org/CorpusID:167217728.

Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, pp. 1–18, 2024.

Rosemary A. Varley, Nicolai J. C. Klessinger, Charles A. J. Romanowski, and Michael Siegal. Agrammatic but numerate. *Proceedings of the National Academy of Sciences*, 102(9): 3519–3524, February 2005. ISSN 1091-6490. doi: 10.1073/pnas.0407470102. URL http://dx.doi.org/10.1073/pnas.0407470102.

Chengcheng Wang, Zhiyu Fan, Zaizhu Han, Yanchao Bi, and Jixing Li. Emergent modularity in large language models: Insights from aphasia simulations. *bioRxiv*, 2025. doi: 10.1101/2025.02.22.639416. URL https://www.biorxiv.org/content/early/2025/02/23/2025.02.22.639416.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. In *The Thirty-Eighth Annual Conference on Neural Information Processing Systems*, 2024.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, and Daniel Khashabi. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks, 2022. URL https://arxiv.org/abs/2204.07705.

Alexandra Woolgar, Alice Parr, Rhodri Cusack, Russell Thompson, Ian Nimmo-Smith, Teresa Torralva, Maria Roca, Nagui Antoun, Facundo Manes, and John Duncan. Fluid intelligence loss linked to restricted regions of damage within frontal and parietal cortex. *Proceedings of the National Academy of Sciences*, 107(33):14899–14902, August 2010. ISSN 1091-6490. doi: 10.1073/pnas.1007928107. URL http://dx.doi.org/10.1073/pnas.1007928107.

Yi Yang, Wen-tau Yih, and Christopher Meek. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2013–2018, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1237. URL https://aclanthology.org/D15-1237.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration, 2023.

Danyang Zhang, Junhao Song, Ziqian Bi, Yingfang Yuan, Tianyang Wang, Joe Yeong, and Junfeng Hao. Mixture of experts in large language models. *ArXiv*, abs/2507.11181, 2025. URL https://api.semanticscholar.org/CorpusID:280277258.

Qizhen Zhang, Nikolas Gritsch, Dwaraknath Gnaneshwar, Simon Guo, David Cairuz, Bharat Venkitesh, Jakob Nicolaus Foerster, Phil Blunsom, Sebastian Ruder, Ahmet Üstün, and Acyr Locatelli. BAM! just like that: Simple and efficient parameter upcycling for mixture of experts. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL `https://openreview.net/forum?id=BDrWQTrfyI`.

Zhengyan Zhang, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. MoEfication: Transformer feed-forward layers are mixtures of experts. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 877–890, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.71. URL `https://aclanthology.org/2022.findings-acl.71/`.

Zhengyan Zhang, Zhiyuan Zeng, Yankai Lin, Chaojun Xiao, Xiaozhi Wang, Xu Han, Zhiyuan Liu, Ruobing Xie, Maosong Sun, and Jie Zhou. Emergent modularity in pre-trained transformers. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4066–4083, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.250. URL `https://aclanthology.org/2023.findings-acl.250/`.

Tao Zhong, Zhixiang Chi, Li Gu, Yang Wang, Yuanhao Yu, and Jin Tang. Meta-DMoe: Adapting to domain shift by meta-distillation from mixture-of-experts. In *Thirty-Sixth Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

APPENDIX

A CONSTRUCTING THE EXPERTS DATASET

Datasets To construct the small curated datasets used in stages 1 and 2 of our training curriculum, we first identified existing datasets that align with the cognitive domain of each expert. Table 1 lists these datasets, the number of examples sampled from each, the corresponding high-level cognitive skill they were chosen to represent, and whether we used O1 to generate responses or relied on the original reasoning chains provided with the dataset.

Table 1: **Datasets Used to Induce Specialization in Stage-1.** Overview of datasets used to induce expert specialization during stages 1 and 2. Each dataset is aligned with a cognitive skill targeted by a specific expert. We indicate the number of examples sampled from each dataset and whether responses were generated using O1 or taken directly from the dataset's original reasoning chains.

Expert	Task	Dataset	# Samples	Use O1
Logic	Math	O1-Journey (Qin et al., 2024)	327	No
		Math (Li et al., 2023)	200	No
		GSM8K (Cobbe et al., 2021)	100	Yes
	Logic	Folio (Han et al., 2022)	100	Yes
		LogicQA (Liu et al., 2020)	100	Yes
	Physics	Physics (Li et al., 2023)	200	No
Total (Logic)			**1027**	
Social	Pragmatics	Deceits (Hu et al., 2023)	20	Yes
		Indirect Speech (Hu et al., 2023)	20	Yes
		Irony (Hu et al., 2023)	25	Yes
		Maxims (Hu et al., 2023)	19	Yes
		Metaphor (Hu et al., 2023)	20	Yes
		Humor (Hu et al., 2023)	25	Yes
		Coherence (Hu et al., 2023)	40	Yes
	Emotion Detection	EmoCause (Kim et al., 2021)	100	Yes
	Theory of Mind	FanToM 1st Order (Kim et al., 2023)	100	Yes
		FanToM 2nd Order (Kim et al., 2023)	100	Yes
		BigToM (Gandhi et al., 2023)	128	Yes
	Social Reasoning	Mixture ProlificAI/social-reasoning-rlhf	531	Yes
Total (Social)			**1028**	
World	World Knowledge	Biology (Li et al., 2023)	100	No
		Chemistry (Li et al., 2023)	100	No
		PIQA (Bisk et al., 2020)	200	Yes
		WikiQA (Yang et al., 2015)	100	Yes
	Spatial Reasoning	SpatialEval (Wang et al., 2024)	100	Yes
	Temporal Reasoning	TextTemporal (Li et al., 2025)	100	Yes
	World Building	World Building archit11/worldbuilding	200	No
	Cause Effect	CoPA (Wang et al., 2022)	100	Yes
Total (World)			**1000**	

Generating Reasoning Responses Since most of the existing datasets we identified (listed in Table 1) do not include reasoning steps for the final answer, we used O1 to generate them. Specifically, we prompted the model with the input followed by "Let's think step by step." to elicit a longer response that includes intermediate reasoning before reaching the final answer. This was only done for datasets that did not already contain suitable reasoning chains.

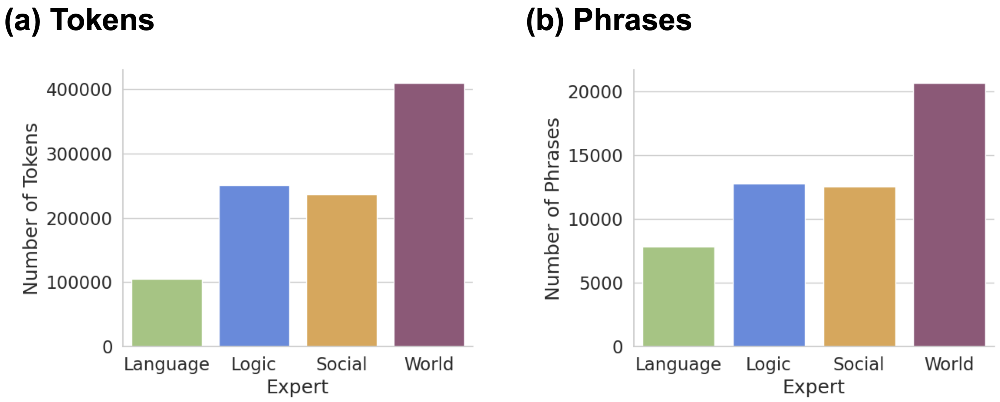


Figure 8: **Distribution of Expert Assignments across Tokens and Phrases. (a)** The distribution of expert assignments across tokens using the LLAMA-3.2-1B tokenizer. **(b)** The distribution of expert assignments labeled using GPT-4O for each phrase in the provided response.

Pseudo-Labeling Responses Finally, once we obtained responses with intermediate reasoning steps for all sampled examples, we used GPT-4O to pseudo-label each phrase in the response. During stage-1 training, each token in a phrase was then assigned to the expert identified by the pseudo-label. This labeling process was guided by the prompt shown in Figure 9. Figure 11 provides examples of labeled responses, while Figure 8 shows the distribution of expert assignments across tokens and phrases.

B INTER-ANNOTATOR AGREEMENT ANALYSIS OF MICRO$_{\text{SFT}}$

We assess the reliability of the MICRO$_{\text{SFT}}$ pseudo-labels using a set of standard agreement metrics between three human annotators and the GPT-4O labels on a subset of the MICRO$_{\text{SFT}}$ dataset. For the three human annotators, we report Krippendorff's α (a chance-corrected multi-annotator reliability metric), Fleiss' κ (multi-rater extension of Cohen's κ), and pairwise Cohen's κ with percent agreement. We also measure the agreement between the humans and the LLM labels both against the human majority vote and by treating the LLM as a fourth annotator.

B.1 HUMAN–HUMAN AGREEMENT

Table 2 summarizes agreement across the three human annotators. Both Krippendorff's α and Fleiss' κ indicate moderate reliability (0.517). The three-way exact agreement rate (all annotators assign the same label) is 49.4%.

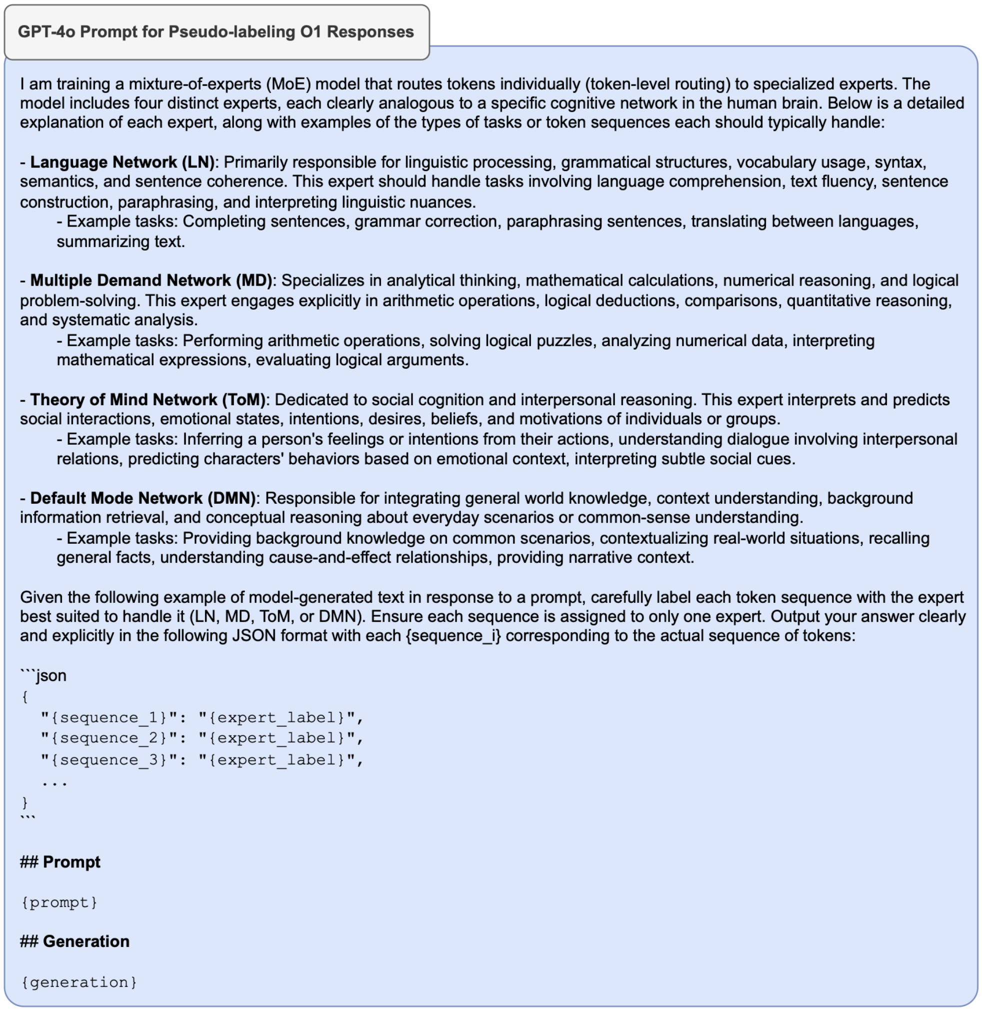
Pairwise metrics (Table 3) reveal that annotators H1 and H2 exhibit stronger mutual agreement ($\kappa = 0.681$) relative to H1–H3 and H2–H3, which show lower yet nontrivial agreement levels ($\kappa \approx 0.43$–0.45).

Table 2: Agreement among human annotators.

Metric	Value
Krippendorff's α (nominal, 3 annotators)	0.517
Fleiss' κ (3 annotators)	0.517
3-way exact agreement	49.4%

B.2 HUMAN–LLM AGREEMENT

We evaluate human–LLM agreement in two ways: (1) comparing the LLM to the majority-vote label of the three humans, and (2) treating the LLM as a fourth annotator.



GPT-4o Prompt for Pseudo-labeling O1 Responses

I am training a mixture-of-experts (MoE) model that routes tokens individually (token-level routing) to specialized experts. The model includes four distinct experts, each clearly analogous to a specific cognitive network in the human brain. Below is a detailed explanation of each expert, along with examples of the types of tasks or token sequences each should typically handle:

- **Language Network (LN)**: Primarily responsible for linguistic processing, grammatical structures, vocabulary usage, syntax, semantics, and sentence coherence. This expert should handle tasks involving language comprehension, text fluency, sentence construction, paraphrasing, and interpreting linguistic nuances.
    - Example tasks: Completing sentences, grammar correction, paraphrasing sentences, translating between languages, summarizing text.

- **Multiple Demand Network (MD)**: Specializes in analytical thinking, mathematical calculations, numerical reasoning, and logical problem-solving. This expert engages explicitly in arithmetic operations, logical deductions, comparisons, quantitative reasoning, and systematic analysis.
    - Example tasks: Performing arithmetic operations, solving logical puzzles, analyzing numerical data, interpreting mathematical expressions, evaluating logical arguments.

- **Theory of Mind Network (ToM)**: Dedicated to social cognition and interpersonal reasoning. This expert interprets and predicts social interactions, emotional states, intentions, desires, beliefs, and motivations of individuals or groups.
    - Example tasks: Inferring a person's feelings or intentions from their actions, understanding dialogue involving interpersonal relations, predicting characters' behaviors based on emotional context, interpreting subtle social cues.

- **Default Mode Network (DMN)**: Responsible for integrating general world knowledge, context understanding, background information retrieval, and conceptual reasoning about everyday scenarios or common-sense understanding.
    - Example tasks: Providing background knowledge on common scenarios, contextualizing real-world situations, recalling general facts, understanding cause-and-effect relationships, providing narrative context.

Given the following example of model-generated text in response to a prompt, carefully label each token sequence with the expert best suited to handle it (LN, MD, ToM, or DMN). Ensure each sequence is assigned to only one expert. Output your answer clearly and explicitly in the following JSON format with each {sequence_i} corresponding to the actual sequence of tokens:

```json
{
  "{sequence_1}": "{expert_label}",
  "{sequence_2}": "{expert_label}",
  "{sequence_3}": "{expert_label}",
  ...
}
```

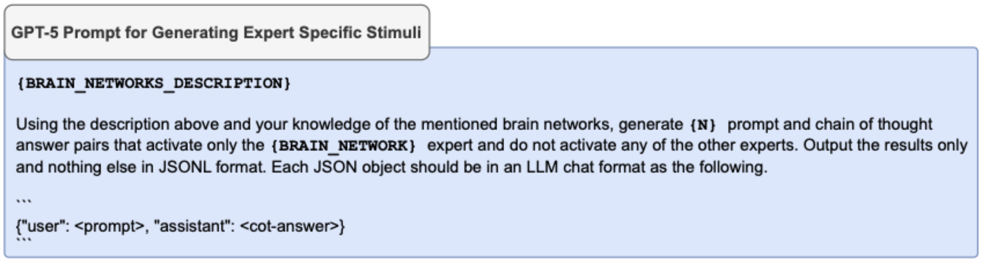
## Prompt

{prompt}

## Generation

{generation}

Figure 9: **Prompt Used for Pseudo-Labeling O1 Responses** The prompt used to instruct GPT-4o to label the O1 model generations given a specific input prompt.



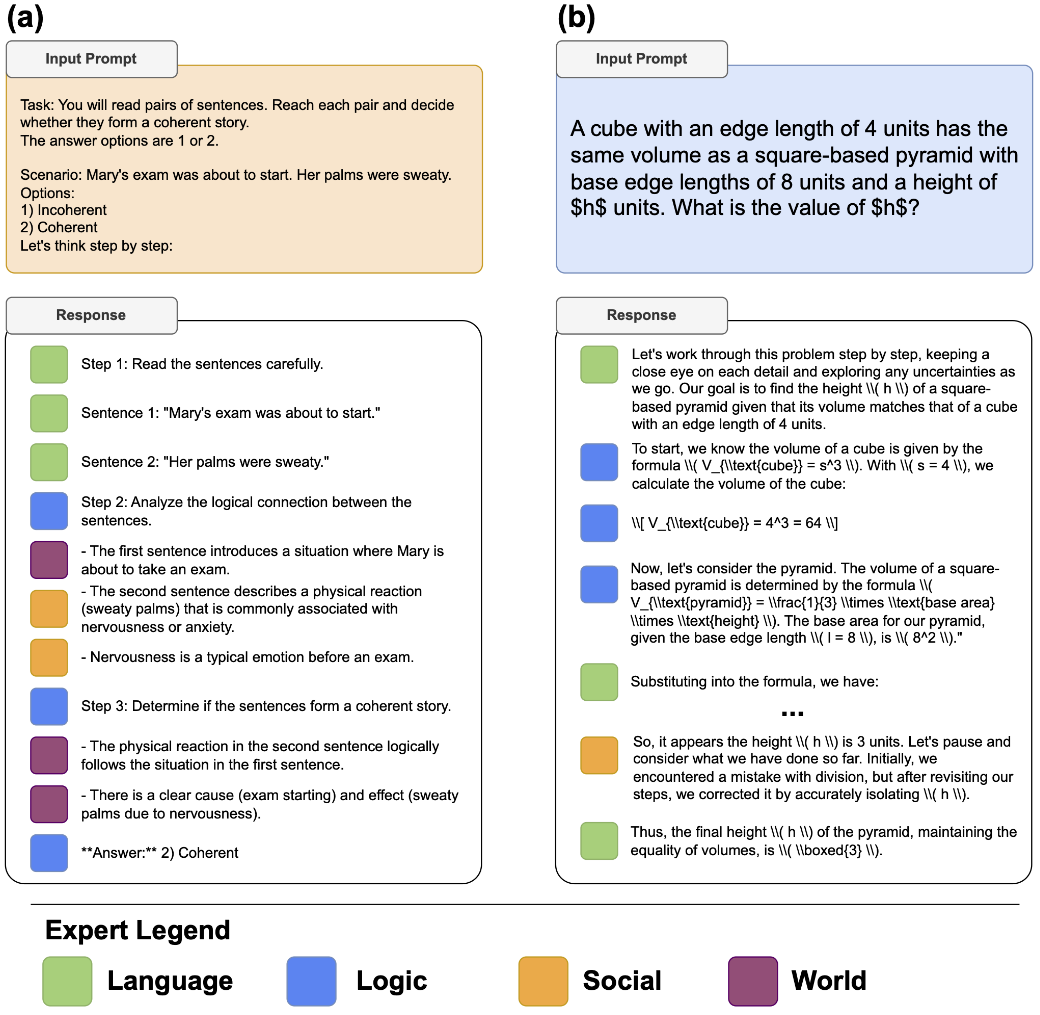
GPT-5 Prompt for Generating Expert Specific Stimuli

{BRAIN_NETWORKS_DESCRIPTION}

Using the description above and your knowledge of the mentioned brain networks, generate **{N}** prompt and chain of thought answer pairs that activate only the **{BRAIN_NETWORK}** expert and do not activate any of the other experts. Output the results only and nothing else in JSONL format. Each JSON object should be in an LLM chat format as the following.

```
{"user": <prompt>, "assistant": <cot-answer>}
```

Figure 10: **Expert-Specific Prompt Template Used with GPT-5** Prompt provided to GPT-5 for generating expert-specific question–answer pairs. The stimuli for each expert was prompted separately using the same brain-network descriptions as in Figure 9.



(a)

Input Prompt

Task: You will read pairs of sentences. Reach each pair and decide whether they form a coherent story.
The answer options are 1 or 2.

Scenario: Mary's exam was about to start. Her palms were sweaty.
Options:
1) Incoherent
2) Coherent
Let's think step by step:

Response

🟩 Step 1: Read the sentences carefully.

🟩 Sentence 1: "Mary's exam was about to start."

🟩 Sentence 2: "Her palms were sweaty."

🟦 Step 2: Analyze the logical connection between the sentences.

🟪 - The first sentence introduces a situation where Mary is about to take an exam.

🟧 - The second sentence describes a physical reaction (sweaty palms) that is commonly associated with nervousness or anxiety.

🟧 - Nervousness is a typical emotion before an exam.

🟦 Step 3: Determine if the sentences form a coherent story.

🟪 - The physical reaction in the second sentence logically follows the situation in the first sentence.

🟪 - There is a clear cause (exam starting) and effect (sweaty palms due to nervousness).

🟦 **Answer:** 2) Coherent

(b)

Input Prompt

A cube with an edge length of 4 units has the same volume as a square-based pyramid with base edge lengths of 8 units and a height of h units. What is the value of h?

Response

🟩 Let's work through this problem step by step, keeping a close eye on each detail and exploring any uncertainties as we go. Our goal is to find the height \\(h \\) of a square-based pyramid given that its volume matches that of a cube with an edge length of 4 units.

🟦 To start, we know the volume of a cube is given by the formula \\(V_{\\text{cube}} = s^3 \\). With \\(s = 4 \\), we calculate the volume of the cube:

🟦 \\[V_{\\text{cube}} = 4^3 = 64 \\]

🟦 Now, let's consider the pyramid. The volume of a square-based pyramid is determined by the formula \\(V_{\\text{pyramid}} = \\frac{1}{3} \\times \\text{base area} \\times \\text{height} \\). The base area for our pyramid, given the base edge length \\(l = 8 \\), is \\(8^2 \\)."

🟩 Substituting into the formula, we have:

...

🟧 So, it appears the height \\(h \\) is 3 units. Let's pause and consider what we have done so far. Initially, we encountered a mistake with division, but after revisiting our steps, we corrected it by accurately isolating \\(h \\).

🟩 Thus, the final height \\(h \\) of the pyramid, maintaining the equality of volumes, is \\(\\boxed{3} \\).

Expert Legend

🟩 Language 🟦 Logic 🟧 Social 🟪 World

Figure 11: **Examples of Pseudo-Labeled Responses using GPT-4o (a)** shows a response generated by O1 for a prompt from the coherence subset of the PRAGMATICS dataset (Hu et al., 2023). **(b)** shows a response taken directly from the O1-JOURNEY dataset (Qin et al., 2024). Each subfigure includes the original prompt, the full model-generated response, and the corresponding pseudo-labels assigned to each phrase.

Table 3: Pairwise agreement between human annotators.

Annotator Pair	Percent Agreement	Cohen's κ
H1–H2	76.3%	0.681
H1–H3	58.9%	0.431
H2–H3	58.5%	0.446

LLM vs. Human Majority Vote. Out of the full set, 240 items had a unique human majority label (13 items exhibited three-way ties). On this subset, the LLM achieves the performance shown in Table 4. The Cohen's κ between the LLM and the majority vote is 0.533, indicating substantial agreement comparable to human–human levels.

Table 4: LLM agreement with human majority-vote labels.

Metric	Value
Accuracy	0.658
Macro F1	0.666
Cohen's κ (LLM vs. majority)	0.533

LLM as a Fourth Annotator. We also compute multi-annotator reliability including the LLM (Table 5). Krippendorff's α decreases slightly to 0.497, reflecting the LLM's moderate alignment with the human annotators.

Pairwise comparisons between each human annotator and the LLM (Table 6) show agreement levels similar to those between some human pairs (particularly H1–H3 and H2–H3).

Table 5: Agreement among 3 humans + LLM.

Metric	Value
Krippendorff's α (nominal, 4 annotators)	0.497

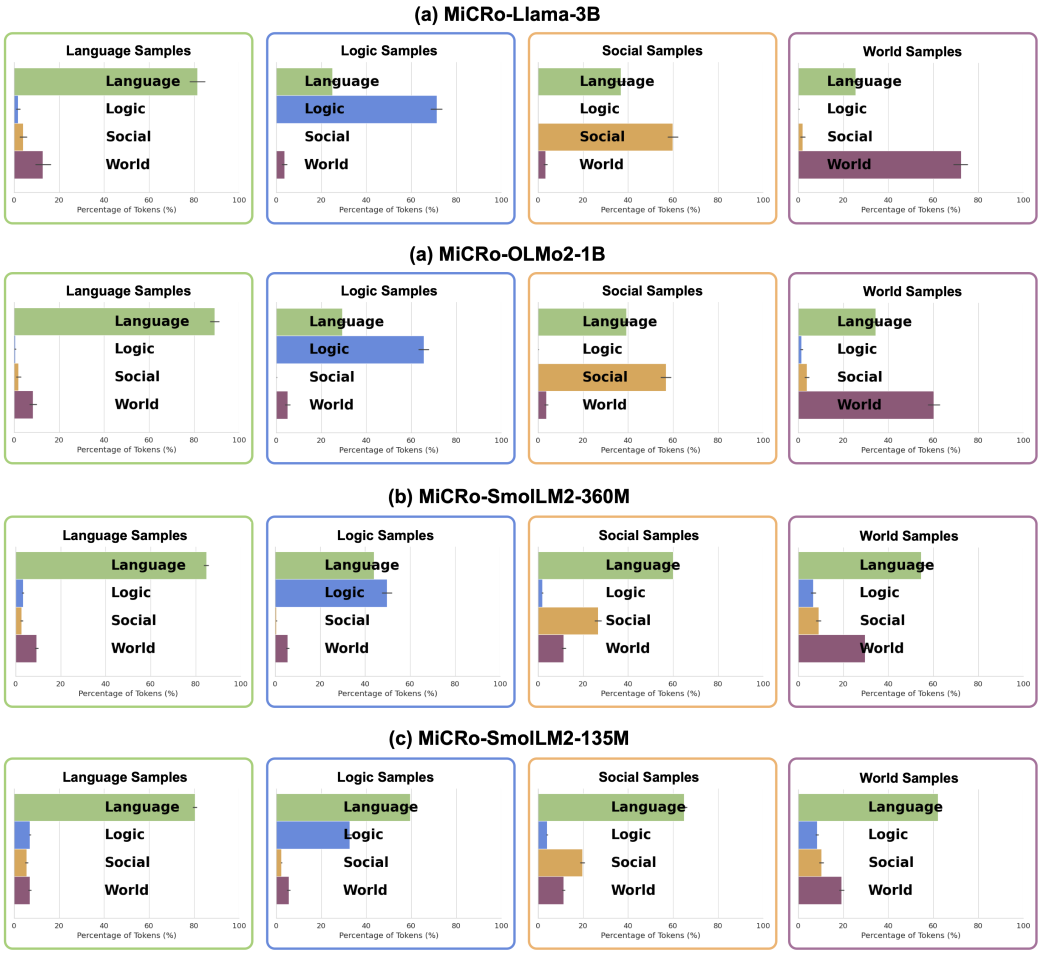
Table 6: Pairwise agreement between humans and the LLM.

Annotator Pair	Percent Agreement	Cohen's κ
H1–LLM	62.8%	0.489
H2–LLM	62.5%	0.492
H3–LLM	59.7%	0.448

Overall, the dataset exhibits moderate inter-annotator consistency across all metrics, with variation typical of multi-class subjective labeling tasks. The LLM aligns with human labels at levels comparable to human–human agreement, and the LLM agrees more with H1 and H2, who have a higher inter-annotator agreement.

C Token Routing Patterns

GPT-5 Prompt for Generating Expert Specific Stimuli Figure 10 presents the prompt used to instruct GPT-5 to generate the question–answer pairs shown in the non-benchmark token-routing pattern plots. The descriptions of the brain networks are identical to those in Figure 9, which were previously used to pseudo-label O1-generated responses for constructing the MiCRo$_{\text{SFT}}$ dataset. We queried GPT-5 separately for each expert. We show the routing patterns for additional models in Figure 12 and for MoE models in Figure 17.



(a) MiCRo-Llama-3B

(a) MiCRo-OLMo2-1B

(b) MiCRo-SmolLM2-360M

(c) MiCRo-SmolLM2-135M

Figure 12: **Token Routing Patterns in Additional MiCRo-Models.** Percentage of tokens routed to each expert, aggregated across all layers, for additional MiCRo models. Distributions are computed over GPT-5–generated question–answer pairs designed to engage specific domains. Results show consistent brain-inspired specialization, with tokens preferentially assigned to the relevant experts depending on the task domain. Figure 13 shows the corresponding layer-wise token routing of these plots, while Figure 14 shows the token routing on benchmark testing data.

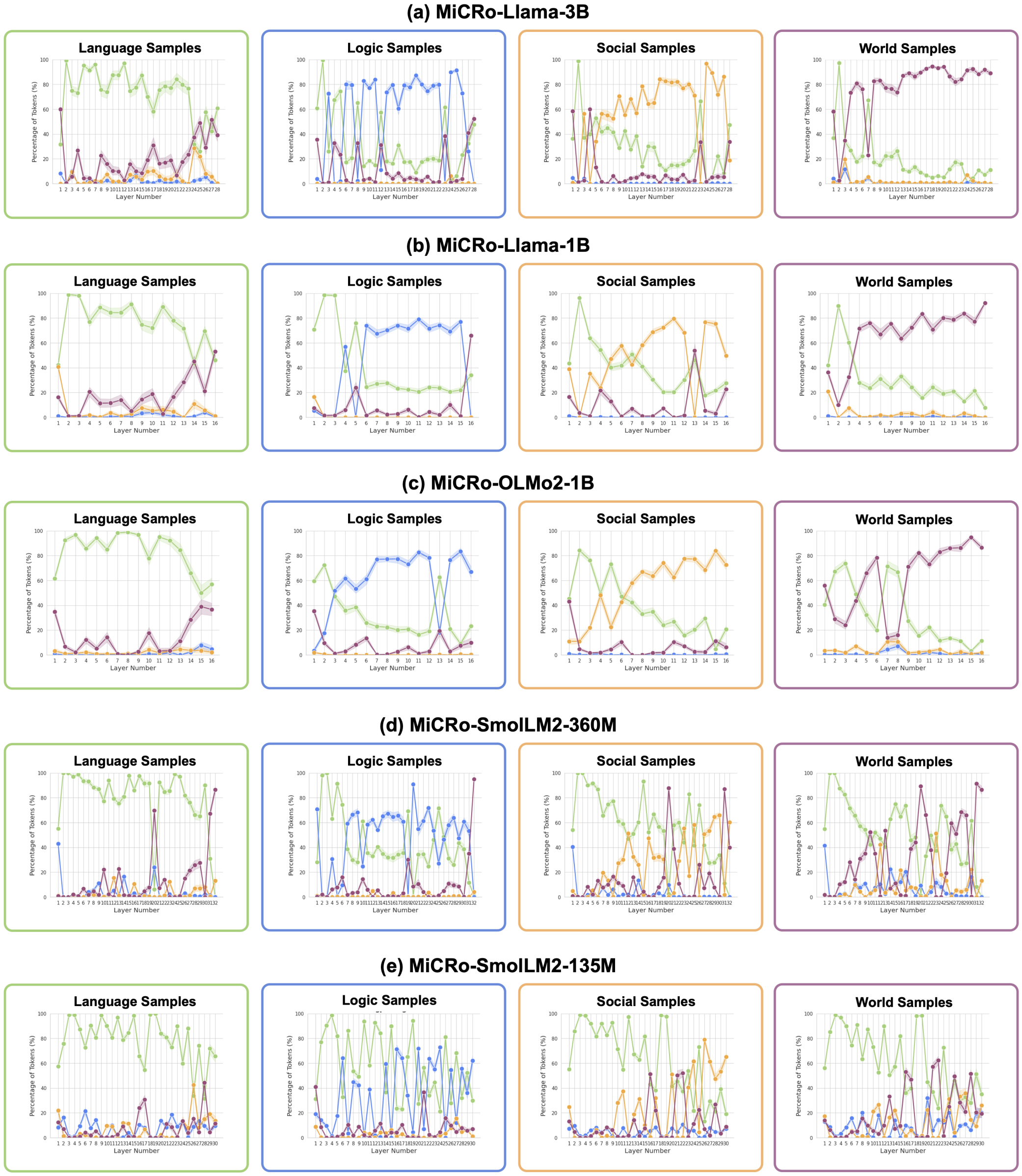


Figure 13: **Layer-wise Token Routing in MICRO Models**. Token routing distributions across layers for five MICRO models, measured on GPT-5–generated question–answer pairs targeting specific domains. In all models, the language expert is consistently engaged in early layers, while domain-specific experts (logic, social, world) are increasingly activated in deeper layers. This hierarchical organization parallels findings from cognitive neuroscience, where linguistic processing precedes engagement of higher-level networks.

Layerwise Routing Patterns Figure 13 illustrates layer-wise token routing patterns for five MiCRo models. Surprisingly, consistent trend emerges: tokens are initially processed by the language expert before being delegated to higher-level experts depending on the task domain. This organization parallels findings in cognitive neuroscience, where the language network is engaged early for virtually all linguistic input and then interfaces with other specialized networks (such as multiple-demand or social cognition systems) depending on task demands (Fedorenko et al., 2024). In Figure 14, we show benchmark-specific token routing patterns across layers as well. To probe social specialization directly, we also include evaluation on the EMPATHY benchmark (Buechel et al., 2018), which primarily engages the social expert, further confirming the expected routing behavior is generalizable across datasets.

D CORRELATION WITH HUMAN JUDGMENTS

We use a dataset of 1,000 six-word sentences from Tuckute et al. (2024), each annotated with human ratings across several behavioral dimensions, collected independently of our routing framework. To test correlations with human judgments, we selected features expected to align with specific experts: GRAMMATICALITY and PLAUSIBILITY with the language expert, MENTAL STATES with the social expert, and PHYSICAL OBJECTS and PLACES with the world expert. The dataset does not include features relevant to the logic expert.

To analyze these relationships, we divide each model into three layer segments (early, middle, late) and averaged router probabilities within each segment. Figure 15 reports correlations between the average routing probability of each expert and human ratings for MiCRo-LLAMA-3.2-1B and MiCRo-LLAMA-3.2-3B. For both models and layer segments, mental state ratings correlate most strongly with the social expert. Language expert probabilities correlate with GRAMMATICALITY and PLAUSIBILITY, but primarily in early layers. PHYSICAL OBJECTS and PLACES correlate with the world expert, while the logic expert shows no positive correlations (and in most cases negative correlations) with these features. These findings suggest that our router exhibits a meaningful degree of correspondence with human behavioral judgments.

E ADDITIONAL EXPERT ABLATION RESULTS

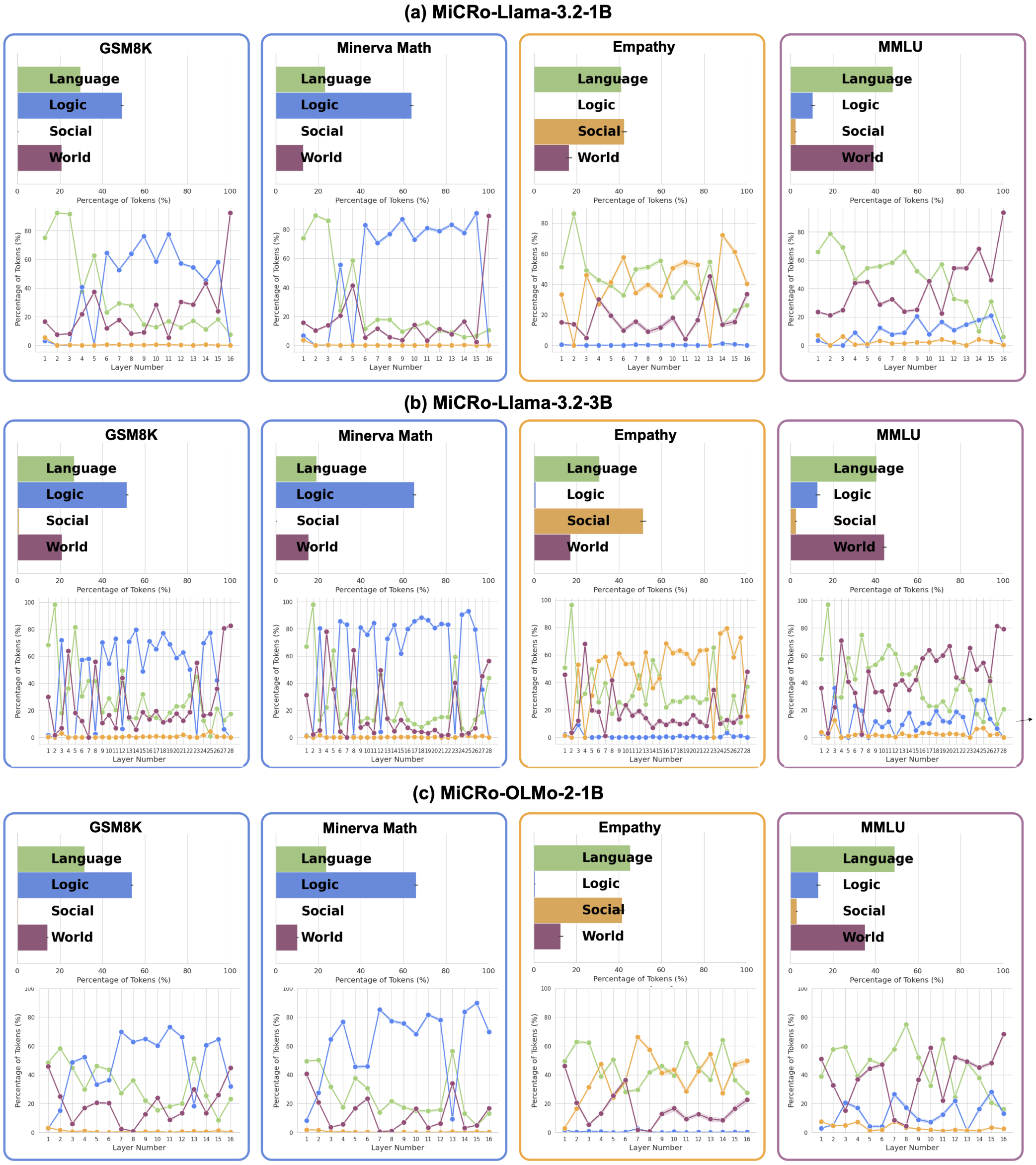
Figure 16 reports the effect of ablating individual experts, including the language expert, on benchmark performance for five MiCRo models. We find that the language expert is essential for most tasks, while domain-specific experts—such as the logic expert for GSM8K and MINERVA MATH—are also necessary to maintain performance. Interestingly, in some cases, ablating an expert improves performance, suggesting that certain experts may interfere with more relevant ones, leading to performance degradation when all are active.

F BENCHMARKS

Table 7: **Number of Shots and Samples Per Benchmark Used in Evaluation.** Number of shots and samples used when evaluating the test-set of each benchmark. Last two row shows whether we used CoT or evaluated using log-probabilities and the metric used to obtain the final accuracy.

Benchmark	**GSM8K**	**Minerva Math**	**MMLU**	**MMLU-Pro**	**BBH**	**HellaSwag**	**PIQA**	**ARC$_{\text{Easy}}$**	**ARC$_{\text{Challenge}}$**
N-Shots	0-Shot	4-shots	4-shots	5-shots	3-Shots	0-Shot	0-Shot	0-Shot	0-Shot
Num Samples	1,319	5,000	14,042	12,032 * 6,511	10,042	1,838	2,376	1,172	
CoT Prompting	Yes	Yes	Yes	Yes	Yes	No	No	No	No
Metric	Exact Match	Exact Match	Exact Match	Exact Match	Exact Match	Acc Norm	Acc Norm	Acc	Acc Norm

Benchmarks Description We evaluate our models on eight benchmarks using various fewshot settings, four of which are prompted to generate a reasoning chain before producing the final answer. These reasoning steps are intended to more meaningfully engage the expert modules throughout the generation process, which is why we focused on them in the main paper. The other benchmarks are multiple choice questions where the most likely candidate—as measured by the log-probabilities of the model—is taken as the prediction. Table 7 lists the number of in-context examples used


(a) MiCRo-Llama-3.2-1B

(b) MiCRo-Llama-3.2-3B

(c) MiCRo-OLMo-2-1B

Figure 14: **Benchmark Token Routing Patterns.** Token routing patterns for **(a)** MiCRo-Llama-3.2-1B, **(b)** MiCRo-Llama-3.2-3B, and **(c)** MiCRo-OLMo-2-1B, evaluated on up to 1,000 samples drawn from the GSM8K, Minerva-Math, Empathy, and MMLU test sets. For each model, the top panel reports the overall percentage of tokens routed to each expert across the whole model (variance across samples), while the bottom panel shows layer-wise routing. The latter reveals an emergent hierarchy: earlier layers emphasize language grounding, whereas deeper layers increasingly delegate to domain-relevant experts.

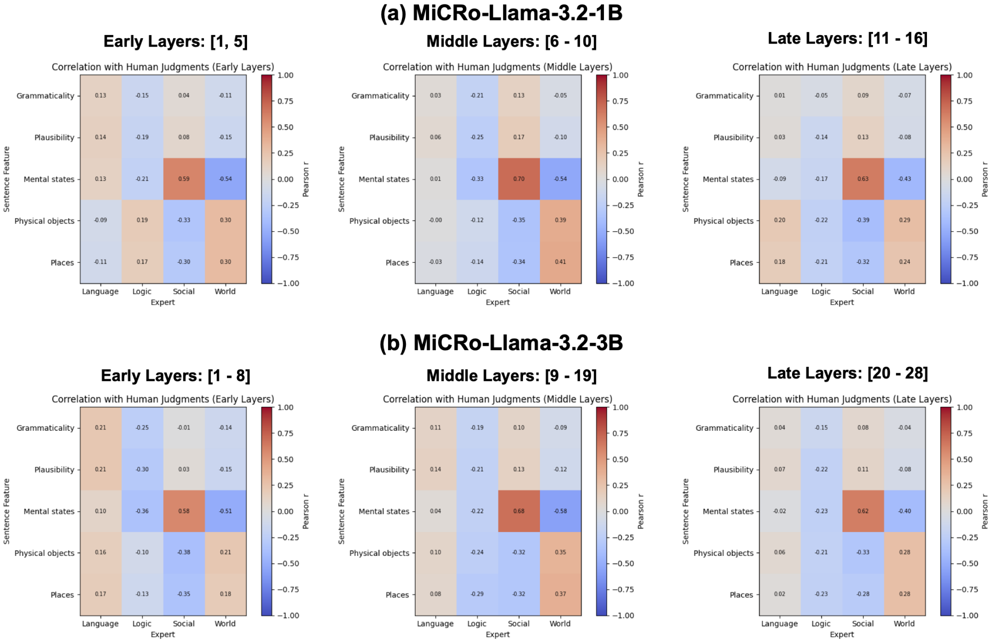


Figure 15: **Correlations Between Expert Routing Probabilities and Human Ratings.** Correlations are shown for MICRO-LLAMA-3.2-1B and MICRO-LLAMA-3.2-3B, averaged across early, middle, and late layer segments. Mental state ratings correlate most strongly with the social expert, grammaticality and plausibility correlate to some degree with the language expert (primarily in early layers), and physical objects and places with the world expert. The logic expert shows no positive correlations with these features.

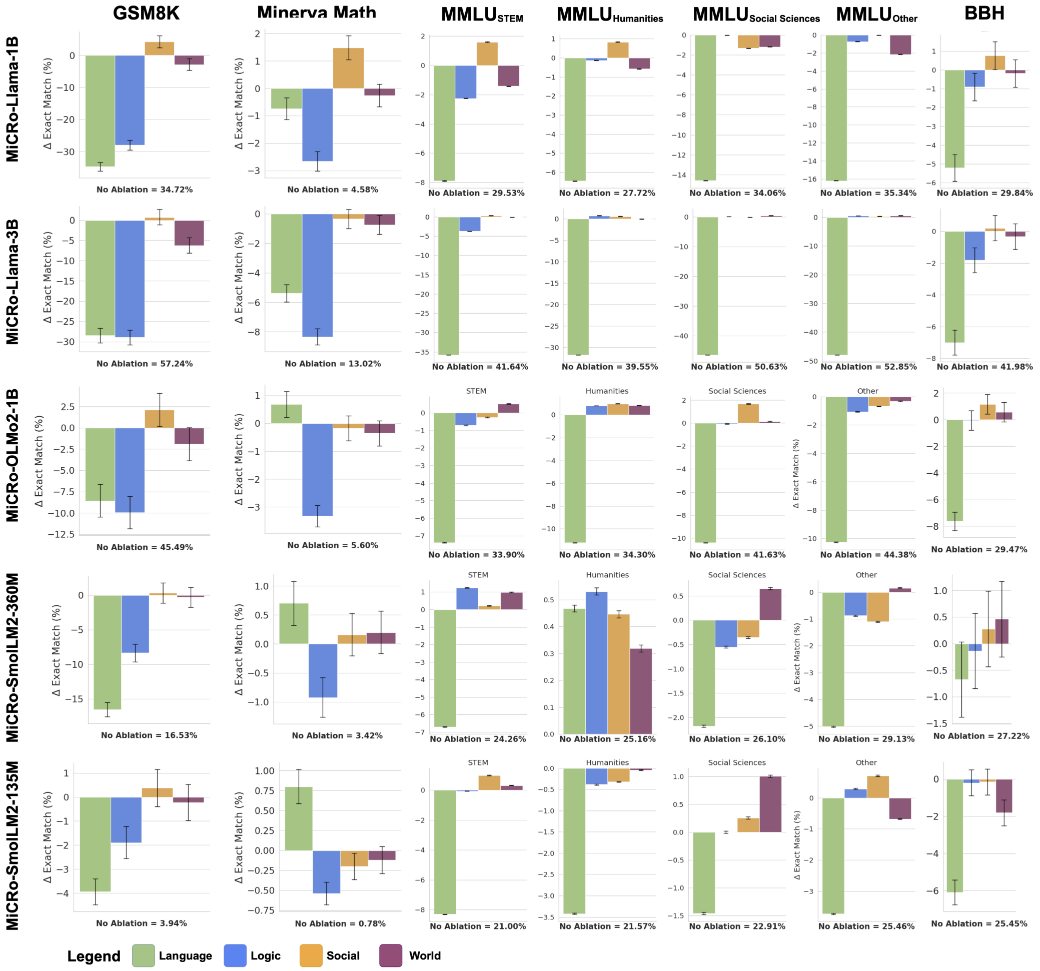


Figure 16: **Expert Ablation Results Across Benchmarks.** Impact of ablating individual experts on benchmark performance for five MiCRo models. Results are shown for GSM8K, Minerva Math, BBH, and MMLU, with the latter divided into its four subcategories. Removing the language expert causes substantial drops across most tasks, while domain-specific experts (e.g., logic for math benchmarks) are critical for their respective domains. In some cases, ablating an expert improves performance, suggesting interference with more relevant experts.

Table 8: **Additional Benchmark Results for MiCRo and Baselines** Accuracy (%) ± standard error across reasoning and knowledge benchmarks. Results are reported for different model classes (Dense, MoB, and MiCRo) under each base model.

Base Model	Model	GSM8K	Minerva	MMLU	MMLU$_{\text{Pro}}$	BBH	ARC$_{\text{Easy}}$	ARC$_{\text{Challenge}}$	HellaSwag	PIQA
SmollM2-135M	Dense	2.7 ± 0.4	0.5 ± 0.1	21.5 ± 0.3	7.8 ± 0.2	24.1 ± 0.5	62.8 ± 1.0	29.6 ± 1.3	43.6 ± 0.5	67.7 ± 1.1
	MoB	3.0 ± 0.5	0.6 ± 0.1	21.9 ± 0.3	7.4 ± 0.2	23.5 ± 0.5	63.0 ± 1.0	29.9 ± 1.3	43.5 ± 0.5	67.8 ± 1.1
	MiCRo	3.9 ± 0.5	0.8 ± 0.1	22.5 ± 0.4	7.9 ± 0.2	25.4 ± 0.5	56.0 ± 1.0	27.6 ± 1.3	41.8 ± 0.5	67.5 ± 1.1
SmollM2-360M	Dense	15.0 ± 1.0	3.7 ± 0.3	26.4 ± 0.4	9.9 ± 0.3	27.3 ± 0.5	69.7 ± 0.9	37.5 ± 1.4	56.6 ± 0.5	71.4 ± 1.1
	MoB	17.4 ± 1.0	3.9 ± 0.3	26.8 ± 0.4	9.8 ± 0.3	27.7 ± 0.5	70.0 ± 0.9	37.1 ± 1.4	56.9 ± 0.5	72.0 ± 1.0
	MiCRo	16.5 ± 1.0	3.4 ± 0.3	26.0 ± 0.4	10.1 ± 0.3	27.2 ± 0.5	69.9 ± 0.9	38.2 ± 1.4	56.7 ± 0.5	71.7 ± 1.1
Llama-3.2-1B	Dense	36.8 ± 1.3	4.8 ± 0.3	29.7 ± 0.4	11.2 ± 0.3	30.4 ± 0.5	64.3 ± 1.0	33.7 ± 1.4	58.4 ± 0.5	73.8 ± 1.0
	MoB	30.5 ± 1.3	3.7 ± 0.3	27.1 ± 0.4	11.0 ± 0.3	27.4 ± 0.5	61.7 ± 1.0	32.4 ± 1.4	56.2 ± 0.5	71.3 ± 1.1
	MiCRo	34.7 ± 1.3	4.6 ± 0.3	31.2 ± 0.4	10.7 ± 0.3	29.8 ± 0.5	59.6 ± 1.0	32.8 ± 1.4	54.7 ± 0.5	73.1 ± 1.0
Llama-3.2-3B	Dense	58.0 ± 1.4	14.4 ± 0.5	48.6 ± 0.4	19.6 ± 0.4	44.1 ± 0.6	73.6 ± 0.9	42.9 ± 1.4	68.9 ± 0.5	77.0 ± 1.0
	MoB	51.6 ± 1.4	12.3 ± 0.5	45.2 ± 0.4	19.1 ± 0.4	42.2 ± 0.6	71.6 ± 0.9	41.3 ± 1.4	67.3 ± 0.5	77.0 ± 1.0
	MiCRo	57.2 ± 1.4	13.0 ± 0.5	45.4 ± 0.4	19.0 ± 0.4	42.0 ± 0.6	73.0 ± 0.9	43.3 ± 1.4	67.4 ± 0.5	76.6 ± 1.0

and the number of samples tested for each benchmark. For the remaining benchmarks, we used the default fewshot examples from the `lm-evaluation-harness` (Gao et al., 2024) repository. Specifically, we used the `bbh_cot_fewshot` task for BBH, the `mmlu_flan_cot_fewshot` task for MMLU, the `mmlu_pro` for MMLU-Pro, the `minerva_math` task for Minerva Math, and the `gsm8k_cot_zeroshot` for the GSM8K task. We used the default tasks for the multiple-choice benchmarks.

Extended Benchmark Results Table 8 reports results for additional base models as well as on benchmarks beyond those presented in the main paper. Consistent with the main results, MiCRo remains comparable to Dense and MoB baselines across most tasks while being interpretable. These supplementary experiments provide further evidence that the observed trends hold across a broader range of model scales and evaluation settings.

G Robustness Across Post-training Methods

We further assess the robustness of our method to different post-training methods by applying two variations. First, we further post-train our MiCRo models using Direct Preference Optimization (DPO) (Rafailov et al., 2023) on a subset of the Tülu-2.5 preference dataset (Ivison et al., 2024) (Table 9). Second, we replace the large-scale general-purpose Tülu-3 dataset used in stage-3 with a more domain-specific (medical) instruction-tuning set used in Bosselut et al. (2024) (Table 10). Our results show that our method is robust to different post-training pipelines, whether applying DPO or using an alternative instruction-tuning dataset, as shown in Tables 9 and 10 respectively.

Table 9: **Performance After DPO Finetuning** Comparison of MiCRo models and baselines after further finetuning with DPO on a preference dataset. Results show average performance across the 4 benchmarks, indicating that specialization remains beneficial after DPO.

Base Model	Model	GSM8K	Minerva Math	MMLU	BBH	Average
Llama-3.2-1B	Dense	38.1	3.9	29.4	**30.3**	25.4
	MiCRo	**39.3**	**5.8**	**31.8**	**30.3**	**26.8**
OLMo-2-1B	Dense	45.8	5.6	39.3	29.8	30.1
	MiCRo	**48.1**	**5.8**	**39.8**	**30.4**	**31.0**

H Mixture-of-Experts Results

In the main paper, we report results using the mixture-of-blocks (MoB) architecture, where each expert is a full transformer block with its own attention mechanism. Here, we contrast these results with the more standard mixture-of-experts (MoE) architecture, where experts consist only of FFN blocks and attention is shared across experts within each layer. We first present routing patterns for MiCRo-MoE models, highlighting cases where our training curriculum fails to induce the intended

Table 10: **Performance on Medical Benchmarks After Domain-Specific Instruction Tuning.** Models are finetuned during Stage-3 using a medical instruction-tuning dataset instead of TÜLU-3, and evaluated on four medical benchmarks. Results show that specialization achieve competitive performance across both base models, and outperforming in the out-of-distribution (OOD) setting. We choose the option with the highest log-probability among the multiple-choice options.

		Out-of-Distribution		In-Distribution		
Base Model	**Model**	**MMLU Medicine**	**MedQA**	**MedMCQA**	**PubMedQA**	**Average**
Llama-3.2-1B	Dense	26.0	34.3	33.9	**73.4**	41.9
	MICRO	**28.3**	**35.2**	33.9	71.4	**42.2**
OLMo-2-1B	Dense	35.8	34.2	**35.3**	**74.0**	44.8
	MICRO	35.8	**36.3**	34.5	73.8	**45.1**

Table 11: **Results with Mixture-of-Experts (MoE) Architectures.** Accuracy (%) ± standard error is reported for Dense, MoE, and MICRO-MOE models across multiple benchmarks. For each base model, the best score per benchmark is highlighted in bold.

Base Model	**Model**	**GSM8K**	**Minerva**	**MMLU**	**BBH**	**ARC$_{\text{Easy}}$**	**ARC$_{\text{Challenge}}$**	**HellaSwag**	**PIQA**
SmollM2-135M	Dense	2.7 ± 0.4	0.5 ± 0.1	21.5 ± 0.3	24.1 ± 0.5	62.8 ± 1.0	29.6 ± 1.3	43.6 ± 0.5	67.7 ± 1.1
	MoE	2.8 ± 0.5	0.6 ± 0.1	22.3 ± 0.3	24.6 ± 0.5	62.9 ± 1.0	29.4 ± 1.3	43.6 ± 0.5	67.6 ± 1.1
	MiCRo-MoE	4.1 ± 0.5	0.4 ± 0.1	22.2 ± 0.3	24.5 ± 0.5	62.0 ± 1.0	29.0 ± 1.3	43.4 ± 0.5	67.5 ± 1.1
SmollM2-360M	Dense	15.0 ± 1.0	3.7 ± 0.3	26.4 ± 0.4	27.3 ± 0.5	69.7 ± 0.9	37.5 ± 1.4	56.6 ± 0.5	71.4 ± 1.1
	MoE	16.1 ± 1.0	3.6 ± 0.3	26.6 ± 0.4	27.4 ± 0.5	70.2 ± 0.9	37.2 ± 1.4	56.8 ± 0.5	71.8 ± 1.1
	MiCRo-MoE	16.1 ± 1.0	4.0 ± 0.3	26.1 ± 0.4	27.0 ± 0.5	69.7 ± 0.9	37.5 ± 1.4	56.7 ± 0.5	71.4 ± 1.1
Llama-3.2-1B	Dense	36.8 ± 1.3	4.8 ± 0.3	29.7 ± 0.4	30.4 ± 0.5	64.3 ± 1.0	33.7 ± 1.4	58.4 ± 0.5	73.8 ± 1.0
	MoE	29.1 ± 1.3	4.7 ± 0.3	25.7 ± 0.4	28.6 ± 0.5	64.1 ± 1.0	35.2 ± 1.4	57.9 ± 0.5	72.5 ± 1.0
	MiCRo-MoE	35.4 ± 1.3	5.0 ± 0.3	30.4 ± 0.4	30.1 ± 0.5	65.0 ± 1.0	35.5 ± 1.4	57.3 ± 0.5	73.8 ± 1.0

specialization—an issue we primarily observe in models larger than 1.5B parameters. We then report the performance of the models that did exhibit specialization on reasoning benchmarks.

H.1 MOE TOKEN ROUTING PATTERNS

Figure 17 shows routing patterns for five MICRO-MoE models on question–answer pairs generated with GPT-5 to target specific experts. The MICRO-MOE-LLAMA-1B model exhibits the intended specialization, whereas the 3B variant does not. Within the SMOLLM2 family, the 135M and 360M models display partial specialization, though less cleanly than LLAMA-1B, often defaulting to the language expert regardless of the input domain. The 1.7B model fails to specialize, similar to MICRO-MOE-LLAMA-3B, indicating that the MoE architecture does not reliably induce the desired specialization under our training curriculum.

H.2 MOE BENCHMARK RESULTS

Table 11 presents results for models trained with a Mixture-of-Experts (MoE) design, complementary to the Mixture-of-Blocks (MoB) results reported in the main paper. The key distinction between MoE and MoB lies in what is replicated to form the experts. In standard MoE, only the feed-forward network (FFN) within each layer is cloned into multiple experts, with the self-attention module shared across all experts. In contrast, MoB duplicates the entire transformer block—including both the attention and FFN components—so that each expert has its own attention mechanism as well as its own FFN. We find that MoB scales more effectively: under our training curriculum, specialization emerges reliably in larger models (>,1B parameters) for MoB, but not for MoE. For this reason, we focus on MoB in the main text and do not include the MoE variants of the other base models, as they did not exhibit the expected functional specialization.

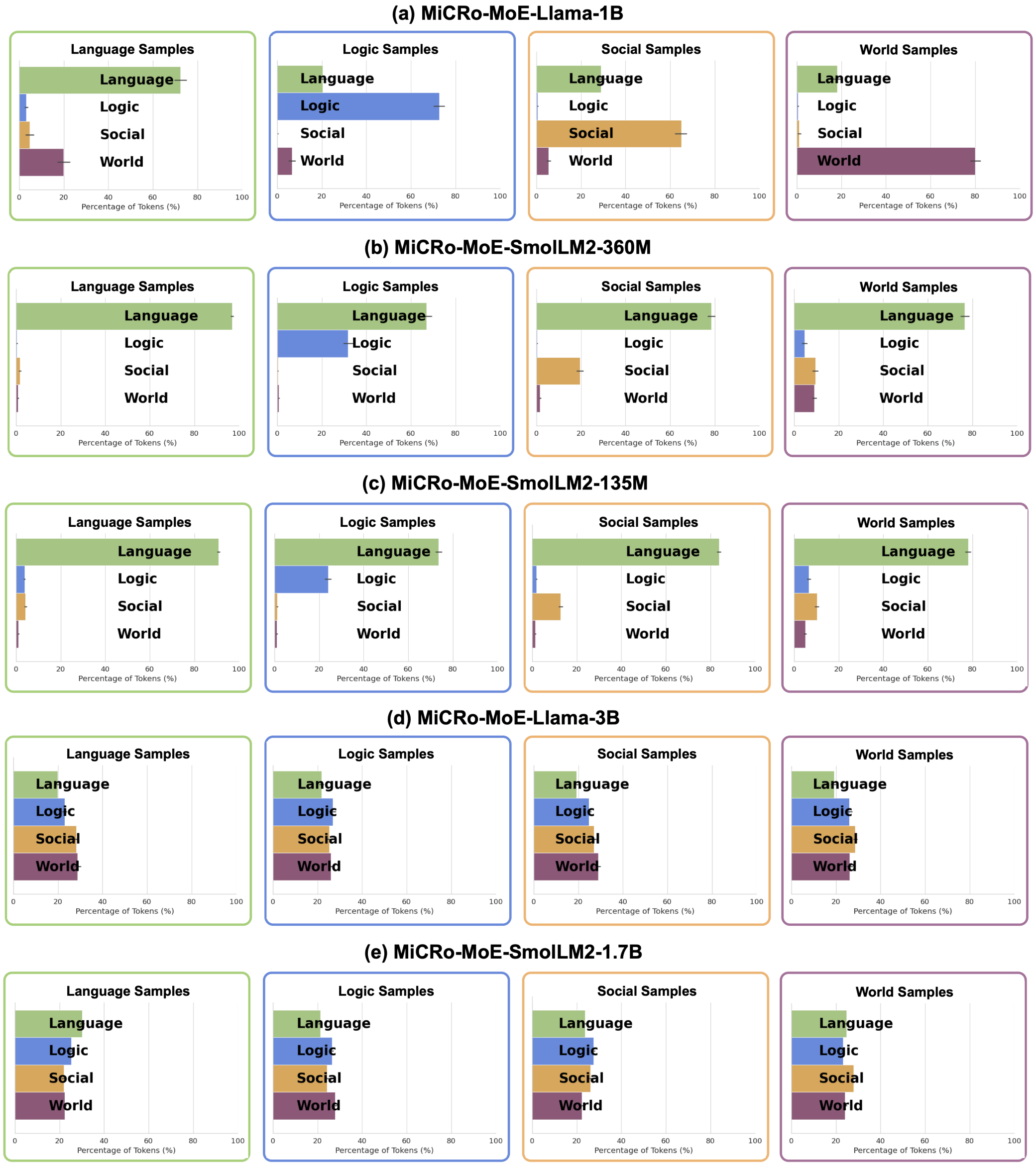


Figure 17: **Routing Patterns in MiCRo-MoE Models.** Routing behavior for five MiCRo-MoE models on GPT-5–generated question–answer pairs targeting specific experts. The MiCRo-MoE-Llama-1B model shows the intended specialization, while larger variants (e.g., 3B, SmolLM2-1.7B) fail to specialize. Smaller SmolLM2 models (135M and 360M) display partial but less consistent specialization, often defaulting to the language expert. These results suggest that the MoE architecture does not reliably induce brain-like specialization under our training curriculum.

Table 12: Performance of MICRO-OLMO-1B when increasing the number of active experts from Top-1 to Top-2 at test time. Enabling an additional expert leads to consistent improvements across most benchmarks, demonstrating that the model generalizes well to increased routing capacity even though SFT was performed with Top-1 routing. The benchmarks MATH, ARC-E, and ARC-C refer to MINERVA-MATH, ARC-EASY, and ARC-CHALLENGE respectively.

K	GSM8K	BBH	MMLU	MATH	HellaSwag	PIQA	ARC-E	ARC-C	Avg
1	45.5%	29.5%	37.9%	5.6%	65.4%	75.2%	70.2%	41.3%	46.3%
2	**47.7%**	**30.6%**	**38.2%**	**6.8%**	**66.4%**	**75.4%**	**71.9%**	**41.6%**	**47.3%**

I ADDITIONAL BEHAVIORAL ALIGNMENT RESULTS

Figure 18 shows alignment to human behavior for additional base models, comparing MICRO with corresponding MOB and DENSE baselines. We find that MICRO achieves higher average behavioral alignment on COGBENCH metrics in larger models, while maintaining comparable performance in smaller models. Please refer to §5.4 for more details on how we evaluate the models.

J SPECIALIZATION REMAINS CONSISTENT THROUGHOUT TRAINING

Figure 19 illustrates token routing assignments across checkpoints during Stage 3 training of MICRO-LLAMA-1B, with checkpoint-0 representing the final weights from Stage 2. The results show that the model consistently preserves the specialization established in stages 1 and 2, despite no explicit constraints being enforced during this phase, except for the initial weak inductive bias. This suggests that brain-like specialization may offer a robust initialization, enabling the model to maintain functionally distinct expert behaviors throughout continued end-to-end training.

K QUALITATIVE EXAMPLES OF STEERING BEHAVIOR

Figures 22-25 show examples of how one can use MICRO to steer the model's behavior by selectively ablating or activating certain experts. In the examples provided, we only retain the target expert along with the language expert using the MICRO models. When the social expert is ablated, the model shifts toward a more analytical tone, producing a response that is logically coherent but lacking in empathy.

L TEST-TIME SCALING BY INCREASING THE NUMBER OF ACTIVE EXPERTS

Table 12 reports the effect of increasing the number of active experts at test time from 1 to 2 for MICRO-OLMO-1B. The results show that test-time compute can be scaled by enabling additional experts, and that the model generalizes well to this setting and improves performance on all benchmarks, even though the large-scale SFT stage was trained exclusively with top-1 routing. This indicates that MICRO retains robustness when the routing capacity is expanded at inference time under the k=2 setting. However, with larger values of k the performance degrades slightly.

M SCALING MICRO TO LLAMA-3.1-8B BASE MODEL

We post-trained three Llama-3.1-8B variants: (1) MICRO-LLAMA-8B, (2) its modular baseline LLAMA-8B-MOB, and (3) LLAMA-8B-DENSE. However, due to compute constraints, we instantiated experts only in the last 12 layers of MICRO-LLAMA-8B and LLAMA-8B-MOB, keeping the earlier layers dense. This choice is motivated by our prior findings (Figures 13–14), which show that early layers predominantly route to language experts, while non-language specializations emerge in later layers. The results confirm that MICRO-LLAMA-8B exhibits the expected routing patterns in its last 12 layers, as shown in Figure 20. Table 13 shows the results on 9 benchmarks of the three LLAMA-3.1-8B model variants, along with the results when we remove the most detrimental expert across all layers for a given task for the MICRO and MOB models. Using a paired Wilcoxon

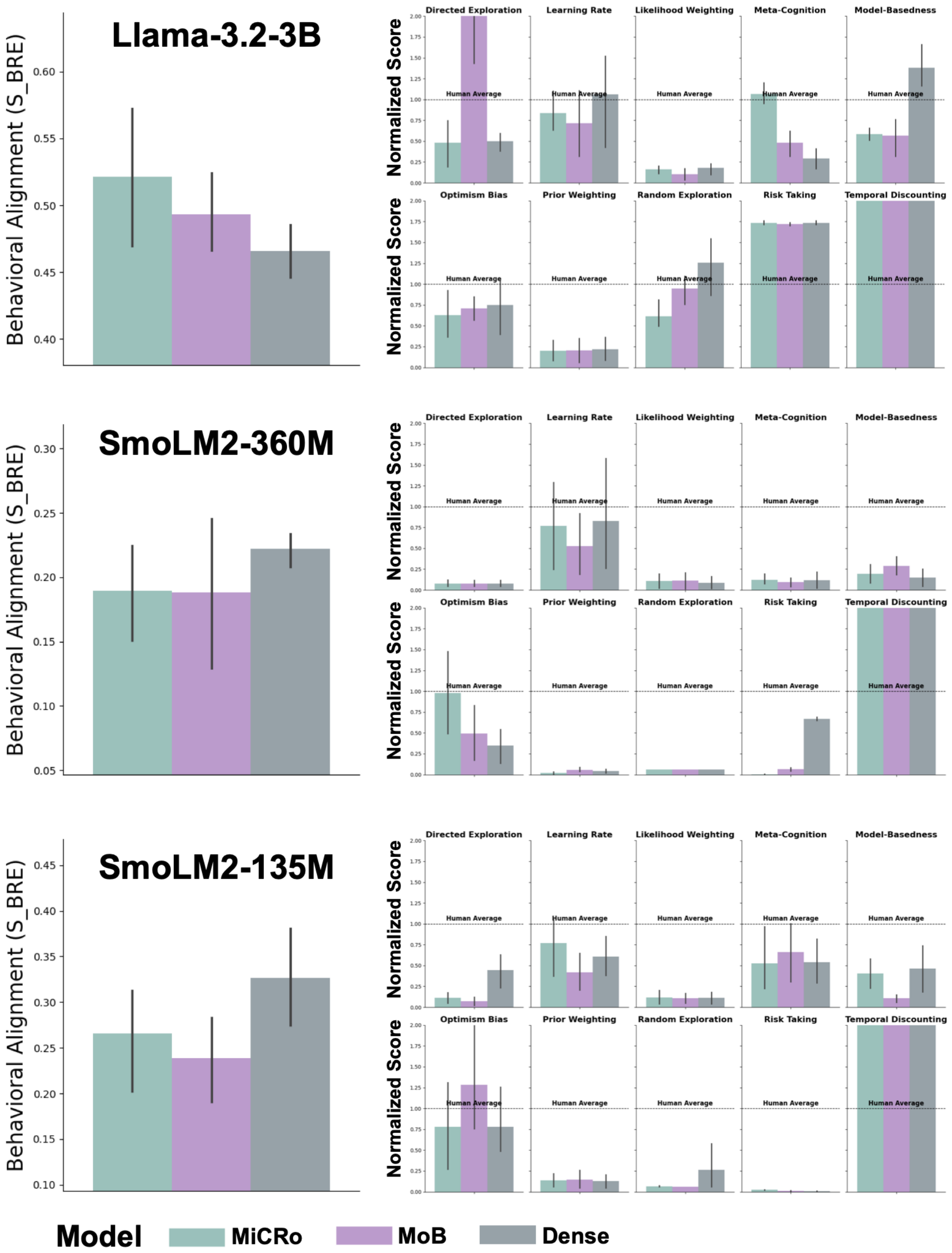


Figure 18: **Behavioral Alignment on Additional Base Models.** Results for LLAMA-3.2-3B, SMOLLM2-360M, and SMOLLM2-135M on COGBENCH. Left: average similarity to human behavior across all metrics. Right: fine-grained results for each behavioral metric. MICRO is compared with MOB and DENSE baselines, showing stronger alignment in larger models and comparable performance in smaller ones.

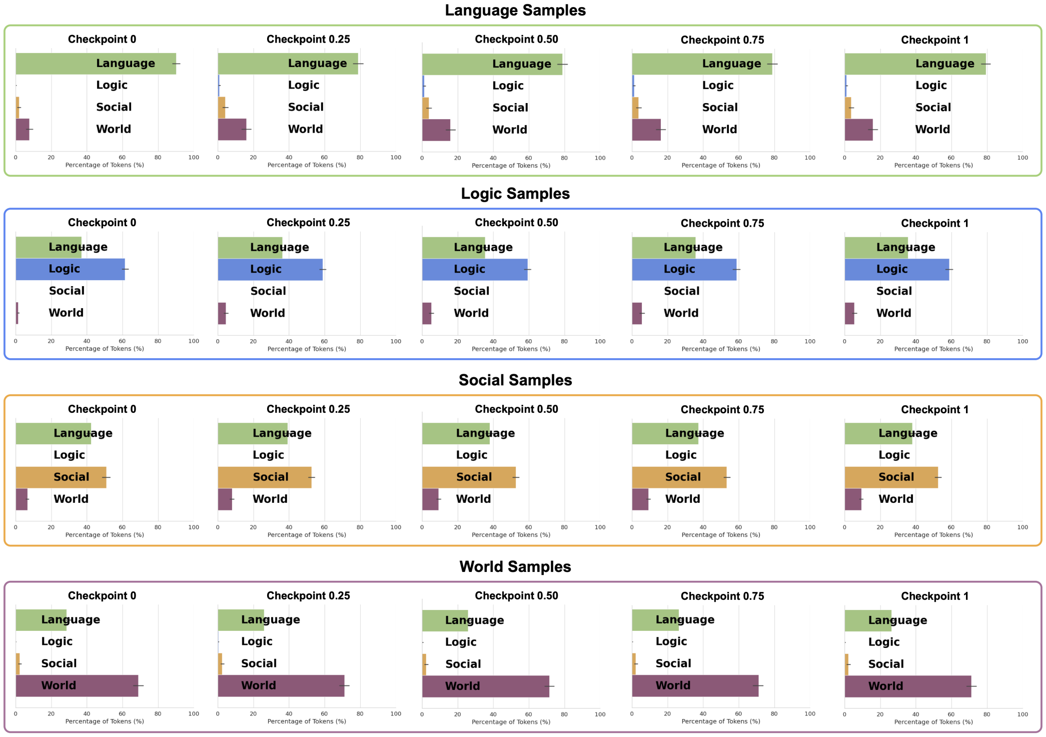


Figure 19: **Specialization Remains Consistent Throughout Training.** Token routing across checkpoints during Stage 3 training of MiCRo-Llama-1B on the samples generated to probe each corresponding expert. Checkpoint 0 corresponds to the final weights from Stage 2. The plot shows that expert assignments remain stable throughout training, with minimal variation, indicating that the model retains its learned specialization despite the absence of explicit constraints.

Table 13: **Benchmark Results for Llama-3.1-8B Model Variants**. Performance comparison of Llama-3.1-8B variants across multiple benchmarks. Ablation refers to selectively removing the least relevant expert per benchmark.

Model Name	GSM8K	BBH	MMLU	MMLU Pro	MATH	HellaSwag	PIQA	ARC-E	ARC-C	Avg
Llama-8B-Dense	71.3	54.2	51.6	24.6	21.5	72.8	78.3	74.3	44.1	54.7
Llama-8B-MoB	69.3	54.1	52.0	23.5	21.3	72.2	78.9	74.6	46.9	54.8
Llama-8B-MoB (Ablation)	70.0	55.9	54.1	36.3	21.6	72.4	79.3	75.0	46.9	56.8
MiCRo-Llama-8B	69.1	53.6	50.7	23.1	18.1	72.7	78.7	75.7	45.9	54.2
MiCRo-Llama-8B (Ablation)	69.1	55.0	52.7	36.3	18.2	73.1	78.7	76.3	46.7	56.2

signed-rank test across the nine benchmarks, we find no statistically significant difference between MiCRo and either of the two baselines. The Dense vs. MiCRo comparison yields a test statistic of 31.0 (p = 0.18), and the MoB vs. MiCRo comparison yields 33.5 (p = 0.10), both well above the conventional 0.05 significance threshold. However, when we ablate the most detrimental expert per task for both MiCRo and MoB we see a jump in performance, as also illustrated in Figure 7. In general, the Llama-8B experiments confirm our initial results that MiCRo remains competitive on a suite of benchmarks while remaining interpretable and relevant to cognitive neuroscience.

N Large Language Model Usage

We used large language models (LLMs) solely for editing and grammatical refinement of the manuscript. All substantive ideas, analyses, and conclusions presented in this work are our own.

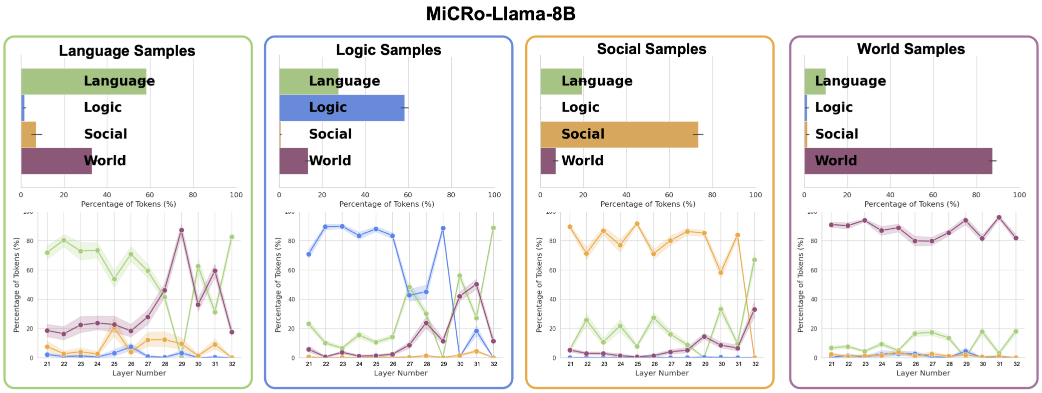


Figure 20: **Token Routing Patterns for MiCRo-Llama-8B**. (Top) The percentage of tokens routed to each expert aggregated across the last 12 layers of MiCRo-Llama-8B. The samples are GPT-5 generated question-answer pairs targeting specific domains. (Bottom) The corresponding layer-wise token routing patterns.

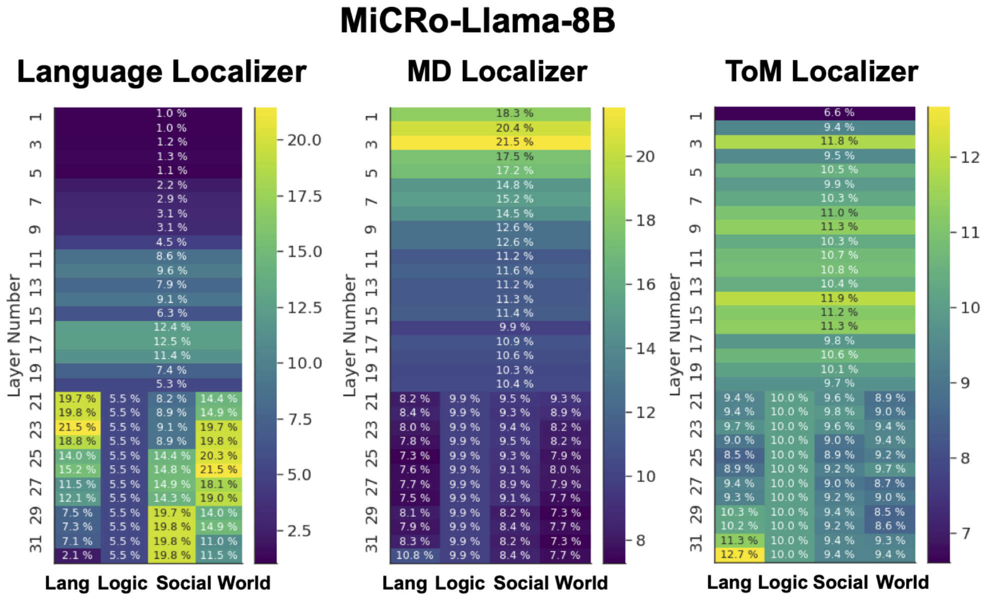
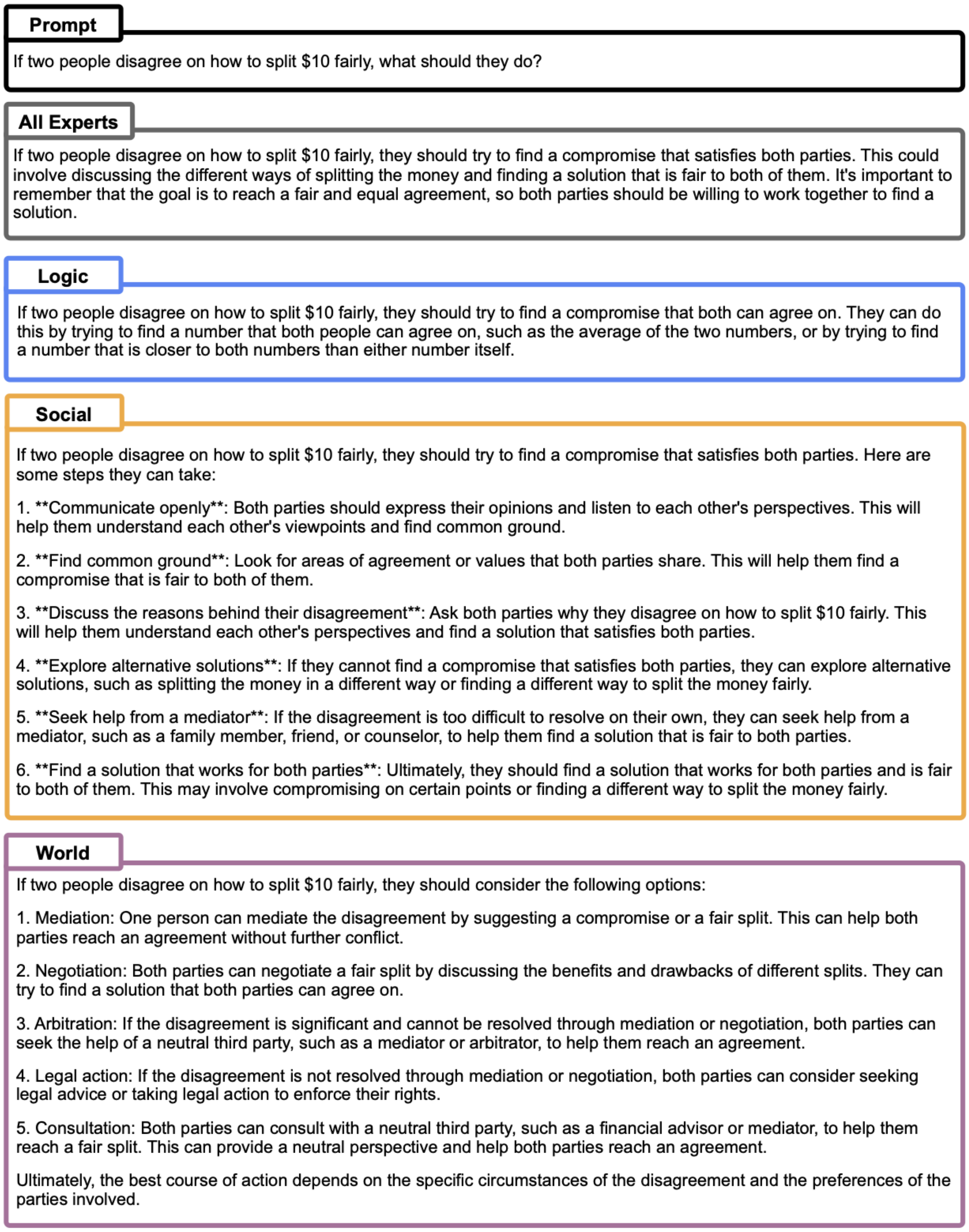


Figure 21: **Neuroscience Localizers Fail to Recover Experts For Hybrid MiCRo-Llama-8B** Following the procedure in Figure 5, we apply three neuroscience-inspired functional localizers to MiCRo-Llama-8B, which places experts only in the final 12 layers due to compute constraints. In contrast to the full MiCRo variants with experts in every layer, the localizers are unable to reliably recover the expected expert specializations in this hybrid configuration.



Prompt

If two people disagree on how to split $10 fairly, what should they do?

All Experts

If two people disagree on how to split $10 fairly, they should try to find a compromise that satisfies both parties. This could involve discussing the different ways of splitting the money and finding a solution that is fair to both of them. It's important to remember that the goal is to reach a fair and equal agreement, so both parties should be willing to work together to find a solution.

Logic

If two people disagree on how to split $10 fairly, they should try to find a compromise that both can agree on. They can do this by trying to find a number that both people can agree on, such as the average of the two numbers, or by trying to find a number that is closer to both numbers than either number itself.

Social

If two people disagree on how to split $10 fairly, they should try to find a compromise that satisfies both parties. Here are some steps they can take:

1. **Communicate openly**: Both parties should express their opinions and listen to each other's perspectives. This will help them understand each other's viewpoints and find common ground.

2. **Find common ground**: Look for areas of agreement or values that both parties share. This will help them find a compromise that is fair to both of them.

3. **Discuss the reasons behind their disagreement**: Ask both parties why they disagree on how to split $10 fairly. This will help them understand each other's perspectives and find a solution that satisfies both parties.

4. **Explore alternative solutions**: If they cannot find a compromise that satisfies both parties, they can explore alternative solutions, such as splitting the money in a different way or finding a different way to split the money fairly.

5. **Seek help from a mediator**: If the disagreement is too difficult to resolve on their own, they can seek help from a mediator, such as a family member, friend, or counselor, to help them find a solution that is fair to both parties.

6. **Find a solution that works for both parties**: Ultimately, they should find a solution that works for both parties and is fair to both of them. This may involve compromising on certain points or finding a different way to split the money fairly.

World

If two people disagree on how to split $10 fairly, they should consider the following options:

1. Mediation: One person can mediate the disagreement by suggesting a compromise or a fair split. This can help both parties reach an agreement without further conflict.

2. Negotiation: Both parties can negotiate a fair split by discussing the benefits and drawbacks of different splits. They can try to find a solution that both parties can agree on.

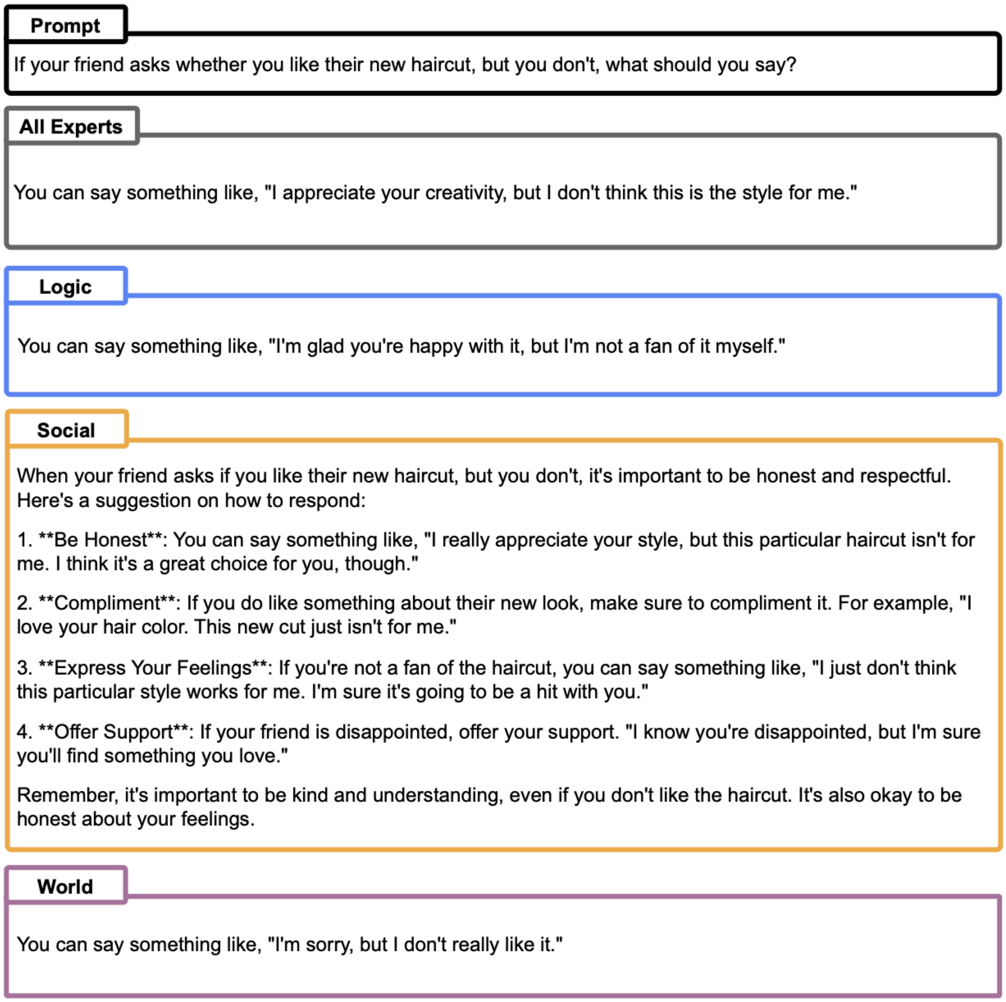
3. Arbitration: If the disagreement is significant and cannot be resolved through mediation or negotiation, both parties can seek the help of a neutral third party, such as a mediator or arbitrator, to help them reach an agreement.

4. Legal action: If the disagreement is not resolved through mediation or negotiation, both parties can consider seeking legal advice or taking legal action to enforce their rights.

5. Consultation: Both parties can consult with a neutral third party, such as a financial advisor or mediator, to help them reach a fair split. This can provide a neutral perspective and help both parties reach an agreement.

Ultimately, the best course of action depends on the specific circumstances of the disagreement and the preferences of the parties involved.

Figure 22: **Example for Steering Model Behavior by Expert Ablation.** Responses of MICRO-LLAMA-3B to the given prompt when only the target expert and the language expert are retained. The differences illustrate the causal role of each expert and demonstrate how ablations can steer the model's behavior.



Prompt

If your friend asks whether you like their new haircut, but you don't, what should you say?

All Experts

You can say something like, "I appreciate your creativity, but I don't think this is the style for me."

Logic

You can say something like, "I'm glad you're happy with it, but I'm not a fan of it myself."

Social

When your friend asks if you like their new haircut, but you don't, it's important to be honest and respectful. Here's a suggestion on how to respond:

1. **Be Honest**: You can say something like, "I really appreciate your style, but this particular haircut isn't for me. I think it's a great choice for you, though."

2. **Compliment**: If you do like something about their new look, make sure to compliment it. For example, "I love your hair color. This new cut just isn't for me."

3. **Express Your Feelings**: If you're not a fan of the haircut, you can say something like, "I just don't think this particular style works for me. I'm sure it's going to be a hit with you."

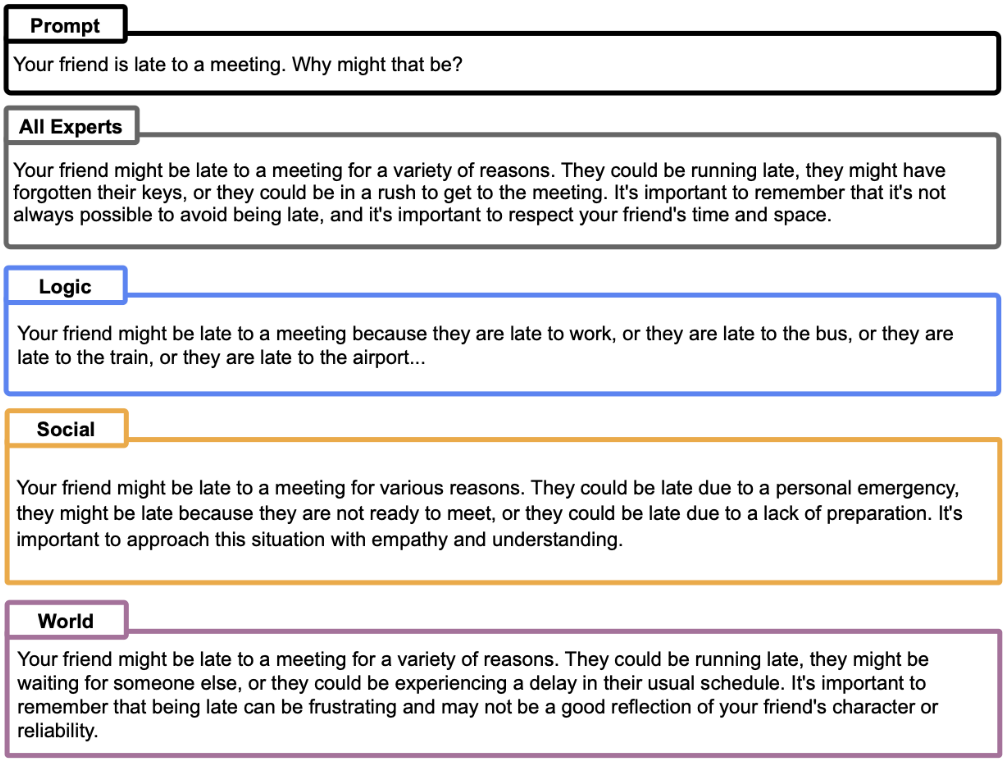
4. **Offer Support**: If your friend is disappointed, offer your support. "I know you're disappointed, but I'm sure you'll find something you love."

Remember, it's important to be kind and understanding, even if you don't like the haircut. It's also okay to be honest about your feelings.

World

You can say something like, "I'm sorry, but I don't really like it."

Figure 23: **Example for Steering Model Behavior by Expert Ablation.** Responses of MICRO-LLAMA-3B to the given prompt when only the target expert and the language expert are retained. The differences illustrate the causal role of each expert and demonstrate how ablations can steer the model's behavior.



Prompt

Your friend is late to a meeting. Why might that be?

All Experts

Your friend might be late to a meeting for a variety of reasons. They could be running late, they might have forgotten their keys, or they could be in a rush to get to the meeting. It's important to remember that it's not always possible to avoid being late, and it's important to respect your friend's time and space.

Logic

Your friend might be late to a meeting because they are late to work, or they are late to the bus, or they are late to the train, or they are late to the airport...

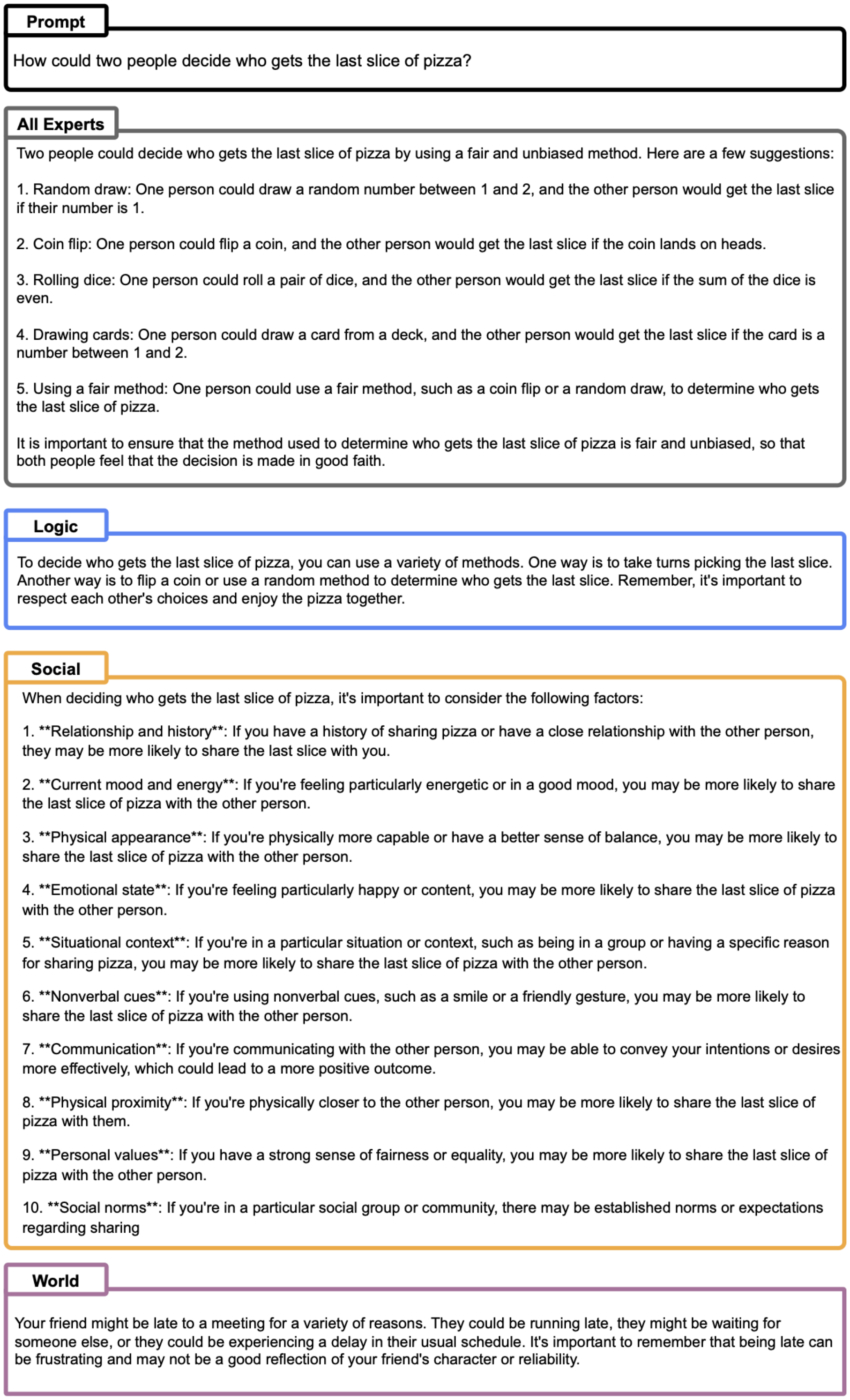
Social

Your friend might be late to a meeting for various reasons. They could be late due to a personal emergency, they might be late because they are not ready to meet, or they could be late due to a lack of preparation. It's important to approach this situation with empathy and understanding.

World

Your friend might be late to a meeting for a variety of reasons. They could be running late, they might be waiting for someone else, or they could be experiencing a delay in their usual schedule. It's important to remember that being late can be frustrating and may not be a good reflection of your friend's character or reliability.

Figure 24: **Example for Steering Model Behavior by Expert Ablation.** Responses of MICRO-LLAMA-1B to the given prompt when only the target expert and the language expert are retained. The differences illustrate the causal role of each expert and demonstrate how ablations can steer the model's behavior.



Prompt

How could two people decide who gets the last slice of pizza?

All Experts

Two people could decide who gets the last slice of pizza by using a fair and unbiased method. Here are a few suggestions:

1. Random draw: One person could draw a random number between 1 and 2, and the other person would get the last slice if their number is 1.

2. Coin flip: One person could flip a coin, and the other person would get the last slice if the coin lands on heads.

3. Rolling dice: One person could roll a pair of dice, and the other person would get the last slice if the sum of the dice is even.

4. Drawing cards: One person could draw a card from a deck, and the other person would get the last slice if the card is a number between 1 and 2.

5. Using a fair method: One person could use a fair method, such as a coin flip or a random draw, to determine who gets the last slice of pizza.

It is important to ensure that the method used to determine who gets the last slice of pizza is fair and unbiased, so that both people feel that the decision is made in good faith.

Logic

To decide who gets the last slice of pizza, you can use a variety of methods. One way is to take turns picking the last slice. Another way is to flip a coin or use a random method to determine who gets the last slice. Remember, it's important to respect each other's choices and enjoy the pizza together.

Social

When deciding who gets the last slice of pizza, it's important to consider the following factors:

1. **Relationship and history**: If you have a history of sharing pizza or have a close relationship with the other person, they may be more likely to share the last slice with you.

2. **Current mood and energy**: If you're feeling particularly energetic or in a good mood, you may be more likely to share the last slice of pizza with the other person.

3. **Physical appearance**: If you're physically more capable or have a better sense of balance, you may be more likely to share the last slice of pizza with the other person.

4. **Emotional state**: If you're feeling particularly happy or content, you may be more likely to share the last slice of pizza with the other person.

5. **Situational context**: If you're in a particular situation or context, such as being in a group or having a specific reason for sharing pizza, you may be more likely to share the last slice of pizza with the other person.

6. **Nonverbal cues**: If you're using nonverbal cues, such as a smile or a friendly gesture, you may be more likely to share the last slice of pizza with the other person.

7. **Communication**: If you're communicating with the other person, you may be able to convey your intentions or desires more effectively, which could lead to a more positive outcome.

8. **Physical proximity**: If you're physically closer to the other person, you may be more likely to share the last slice of pizza with them.

9. **Personal values**: If you have a strong sense of fairness or equality, you may be more likely to share the last slice of pizza with the other person.

10. **Social norms**: If you're in a particular social group or community, there may be established norms or expectations regarding sharing

World

Your friend might be late to a meeting for a variety of reasons. They could be running late, they might be waiting for someone else, or they could be experiencing a delay in their usual schedule. It's important to remember that being late can be frustrating and may not be a good reflection of your friend's character or reliability.

Figure 25: **Example for Steering Model Behavior by Expert Ablation.** Responses of MiCRo-Llama-3B to the given prompt when only the target expert and the language expert are retained. The differences illustrate the causal role of each expert and demonstrate how ablations can steer the model's behavior.