Enhancing Logical Consistency in Language Models through Neuro-Symbolic Feedback and Structured Reasoning

Anonymous Authors

Paper under review at AAAI 2026 Bridge: Logical and Symbolic Reasoning in Language Models

Abstract

Large Language Models (LLMs) have achieved remarkable progress in understanding and generating natural language, yet they continue to struggle with tasks that require explicit logical, deductive, and symbolic reasoning. This limitation stems from their reliance on statistical correlations in human written text rather than structured reasoning grounded in formal logic. In this work, we explore a hybrid neuro symbolic perspective that bridges statistical language modeling with structured reasoning paradigms. We investigate how decoupled token representations and feedback guided adaptation originally developed to enhance multimodal understanding can also serve as mechanisms for improving logical consistency and inference in LLMs. Specifically, we propose integrating symbolic reasoning modules, external logic solvers, and constraint based inference layers within transformer architectures to align neural activations with logical entailment structures.Our empirical analysis demonstrates that incorporating logic informed feedback during fine tuning enhances both deductive and inductive reasoning capabilities while significantly reducing self contradiction, semantic drift, and hallucination across multi turn reasoning tasks. We further show that explicit reasoning traces, when combined with learned representations, improve interpretability and trustworthiness in model outputs. These findings support the growing view that hybrid neurosymbolic systems combining the flexibility of neural networks with the rigor of formal logic represent a crucial step toward reasoning aware foundation models.

Code — https://anonymous.4open.science/r/Mech-Interp-Experiments-C9F4/recall-vs-reasoning/README.md

Datasets — https://anonymous.4open.science/r/Mech-Interp-Experiments-C9F4/modeling_dataset.json

Introduction

Transformer-based language models have demonstrated remarkable capabilities across domains, from retrieving factual knowledge to solving reasoning-intensive tasks (Wei et al. 2022). Two abilities in particular stand out: recall, the ability to retrieve memorized facts, and reasoning, the ability to integrate multiple pieces of information to draw an inferences. (Petroni et al. 2019) (Larenz 2023) Despite strong performance, it remains unresolved whether these abilities arise from shared mechanisms or from distinct circuits within transformer architectures. (Olsson et al. 2022) (Elhage et al. 2021)

Understanding this distinction is not only a matter of scientific curiosity. It has direct implications for the reliability and safety of model deployment. If recall and reasoning are separable at the circuit level, we can better predict when models will generalize beyond memorized knowledge, design more targeted evaluation benchmarks, and develop interventions that disable one capability without unintentionally impairing the other. Disentangling these functions is thus a core challenge in mechanistic interpretability, the field that seeks to explain model behavior in terms of their internal structure.

Mechanistic interpretability (MI) aims to reverse-engineer a neural network into human-interpretable components and algorithms. Unlike probes or attributions that offer correlational insight, MI focuses on causal, structural explanations: identifying internal circuits such that replacing or intervening on them produces predictable changes in model behavior. This view is formalized in the framework of causal abstraction, which unifies methods like activation patching, path interventions, and scrubbing under a principled mapping from low-level activations to higher-level algorithmic structure (Geiger et al. 2025)

Within mechanistic interpretability, researchers have documented compelling internal behaviors in transformers. Attention heads can implement recognizable functions such as induction heads that support in-context learning (Olsson and et al. 2022) or heads that selectively copy tokens and localize features. (Elhage et al. 2021) More recent efforts have mapped circuits across layers and shown how feed-forward (MLP) blocks encode semantically meaningful features. (Geva, Schuster, and Berant 2021; Wang et al. 2022; Nanda 2023) While illuminating, many of these results remain anecdotal or limited to narrow tasks and models, underscoring the need for more systematic, mechanistic analyses that probe how recall and reasoning are implemented internally.

At the same time, evidence is mounting that LLM outputs do not always faithfully reflect their internal reasoning. Models can produce explanations or answers misaligned with their actual internal states (Chen et al. 2025; Turpin et al. 2023), and in some cases, they engage in align-

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

ment faking or specification gaming—producing superficially valid responses that meet training objectives without adhering to the intended task (Greenblatt et al. 2024; Denison et al. 2024). Addressing such reliability concerns requires going beyond performance metrics and probing the underlying circuits. Instead of retraining models from scratch, a promising strategy is to identify and experimentally validate the specific heads, neurons, and subcircuits responsible for recall and reasoning. By establishing which components activate during genuine reasoning versus memorized retrieval, we aim to provide a mechanistic yardstick for separating these two cognitive processes. Such distinctions can also offer transferable insights for domains like mathematics, symbolic reasoning, and program synthesis, where it is equally critical to distinguish inference from rote retrieval.

In this work, we probe whether recall and reasoning are supported by separable circuits in transformer models. Rather than merely observing activation differences, we carry out causal intervention experiments to connect internal structure to functional behaviors. Our investigation centers on five testable claims:

- Layer specialization: some layers contribute more causally to recall, others to reasoning
- Head specialization: attention heads differ in their causal contributions across task types.
- Neuron firing: particular neurons or clusters show task-specific activation signatures.
- Architectural generality: these specialization patterns persist across model families.
- Selective intervention: disabling recall circuits should impair fact-retrieval but spare inference, and vice versa.

Taken together, we aim not just to map activation differences, but to causally interpret circuit-level structure for memory vs inference in large models.

To investigate these hypotheses, we construct controlled recall–reasoning task pairs that isolate differences in computation rather than superficial input content. In particular, we draw on synthetic linguistic puzzles inspired by the International Linguistics Olympiad (IOL), curated from multiple public datasets, where recall tasks require direct fact retrieval while reasoning tasks demand rule discovery and multi-step inference. In parallel, we design counterfactual factual queries to further disentangle recall from reasoning under matched surface forms. Using this suite of tasks, we trace model activations and apply targeted causal interventions (e.g. activation patching) to test whether manipulating specific layers, heads, or neurons predictably alters task performance.(Meng et al. 2022) We also probe MLP blocks to detect activation patterns corresponding to distinct computational roles, enabling us to assess whether recall and reasoning rely on separable subcircuits. Our work makes the following contributions to mechanistic interpretability:

- A dataset of controlled recall–reasoning task pairs that disentangle differences in semantic content from differences in computation.
- Empirical evidence of layer specialization, showing that distinct strata are differentially involved in recall and reasoning.
- Causal validation of functional roles, demonstrating predictable behavioral changes when targeted activations are modified.
- Discovery of modular activation subspaces within MLP blocks, revealing interpretable roles such as feature selection, transformation, and gating.
- A discussion of how hypothesis-driven experimentation can scale to larger models and more complex tasks.

Together, our findings suggest that LLMs are not monolithic black boxes but instead implement structured and decomposable computations that can be systematically probed and interpreted.

Background and Related Work

Mechanistic Interpretability and Circuits

Mechanistic interpretability seeks to explain the internal computations of large language models (LLMs) by decomposing them into circuits, i.e., sparse subnetworks of weights and activations that implement specific functions (Elhage et al. 2021; Olsson and et al. 2022). A transformer layer with hidden representation $\mathbf{h}^{(l)} \in R^d$ can be expressed as

$$\mathbf{h}^{(l+1)} = \mathbf{h}^{(l)} + Attn^{(l)}(\mathbf{h}^{(l)}) + MLP^{(l)}(\mathbf{h}^{(l)}), \qquad (1)$$

where the residual stream accumulates contributions from the attention and feed-forward sublayers. Understanding how information flows through these components is central to mechanistic analysis.

In attention, a head computes

$$AttnHead(\mathbf{H}) = softmax\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \qquad (2)$$

with $\mathbf{Q} = \mathbf{H}W_Q$, $\mathbf{K} = \mathbf{H}W_K$, and $\mathbf{V} = \mathbf{H}W_V$. Prior work has shown that certain heads implement interpretable behaviors such as induction, entity copying, or position tracking (Olsson and et al. 2022). Feed-forward blocks, by contrast, are typically defined as

$$MLP(\mathbf{h}) = W_2\,\sigma(W_1\mathbf{h} + \mathbf{b}_1) + \mathbf{b}_2, \qquad (3)$$

where σ is a nonlinearity such as GELU. Geva et al. (Geva, Schuster, and Berant 2021) demonstrated that these layers behave like key-value memories, with W_1 selecting features and W_2 writing them back into the residual stream.

Beyond descriptive analysis, causal interventions have been developed to test whether these circuits are functionally necessary. Activation patching (Meng et al. 2022) replaces a corrupted hidden state with its clean counterpart during a forward pass, and measures the change in output probability,

$$\Delta^{(l)} = p(y \mid x, \mathbf{h}^{(l)} \leftarrow \mathbf{h}^{clean}) - p(y \mid x, \mathbf{h}^{corr}), \qquad (4)$$

where a large $\Delta^{(l)}$ indicates that layer l is causally important for the target behavior.

Although these methods have revealed circuits for induction, factual recall, and feature storage, many results remain

anecdotal and lack systematic hypothesis testing. Scaling circuit discovery to larger models and grounding claims in causal validation remain open challenges, motivating our focus on hypothesis-driven experiments for recall, reasoning, and modular subspaces.

Recall, Reasoning, and Modular Subspaces

The distinction between recall and reasoning can be formalized in terms of the operations applied to hidden representations. Recall corresponds to retrieving stored associations from the residual stream, i.e.,

$$\mathbf{h}^{(l+1)} \approx \mathbf{h}^{(l)} + f_{recall}(\mathbf{h}^{(l)}), \tag{5}$$

where f_{recall} extracts factual features directly encoded in parameters. Reasoning, by contrast, requires compositional transformations that operate on multiple features, such that

$$\mathbf{h}^{(l+1)} \approx \mathbf{h}^{(l)} + f_{reason}(\mathbf{h}^{(l_1)}, \mathbf{h}^{(l_2)}, \ldots), \tag{6}$$

with dependencies across layers and tokens. Behavioral evidence suggests that earlier layers contribute more to recall while deeper layers perform multi-step reasoning (Elhage et al. 2021; Olsson and et al. 2022), though direct mechanistic validation remains limited.

Multilayer perceptrons (MLPs) provide an additional lens on this separation. A feed-forward block,

$$MLP(\mathbf{h}) = W_2\, \sigma(W_1 \mathbf{h} + \mathbf{b}_1) + \mathbf{b}_2, \tag{7}$$

has been interpreted as a key-value memory system, with W_1 selecting features and W_2 writing them back into the residual stream (Geva, Schuster, and Berant 2021). Empirical analyses indicate that hidden dimensions cluster into modular subspaces that perform roles such as feature selection, transformation, and gating (Nanda 2023). This modular view suggests that recall and reasoning may correspond to distinct activation subspaces, motivating targeted interventions to test their causal roles.

Dataset Construction

Design Principles and Generation

To investigate the neural mechanisms underlying recall versus reasoning in large language models (LLMs), we constructed a controlled dataset that isolates these two cognitive functions. The key design objective was to create paired prompts that are semantically identical in factual content but differ in the type of cognitive process required: direct factual recall versus two-step logical reasoning. To ensure clarity and verifiability, the dataset focuses on world geography, specifically countries, their capitals, and the continents in which they are located. This domain provides structured factual relationships and straightforward reasoning paths.

The dataset was generated using a GPT-based synthetic pipeline. A large language model was prompted to generate lists of entities related by specific, verifiable facts (e.g., "list 50 countries, their capitals, and their continents"). These triples (*country–capital–continent*) were then programmatically formatted into recall and reasoning prompt pairs using fixed templates, which ensured structural consistency and minimized linguistic variability. All generated items were filtered, validated, and labeled with task type and ground truth answers. The final dataset contains 60 questions, evenly divided into 30 recall tasks and 30 reasoning tasks.

Task Structure and Balance

Each recall question directly queries a single fact from the knowledge base, such as *"What is the capital of France?"*. Reasoning questions require combining two pieces of information from the same triple, such as *"If Paris is the capital of France and France is in Europe, what continent is Paris in?"*. This design ensures that recall and reasoning items share identical underlying content while differing in the number of inferential steps. By keeping linguistic complexity constant across both task types, we achieve a balanced comparison in which the cognitive demand—single-step retrieval versus multi-step inference—serves as the primary distinguishing factor.

Experimental Setup

Model Selection

For this study, the `Qwen2.5-7B-Instruct` language model was utilized. This model was selected due to its competitive performance and architecture, which supports detailed inspection of internal activations. The model was loaded using the `nnsight`(Fiotto-Kaufman et al. 2024) library, enabling tracing and extraction of activations during inference. The `attn_implementation` parameter was set to "eager" to ensure accessibility of attention weights. To leverage hardware acceleration, the model was deployed on an NVIDIA A100 GPU within the Google Colab environment.

Primary Hypothesis and Sub-Hypotheses

The primary hypothesis states that large language models develop distinct internal circuits for recall (memory retrieval) versus reasoning (logical inference) tasks, with measurable differences in attention patterns, MLP activations, and hidden state representations. This was decomposed into three testable sub-hypotheses:

- **H1:** Specific layers specialize for recall vs. reasoning (layer specialization).
- **H2:** Attention heads show differential activation patterns between task types.
- **H3:** MLP neurons exhibit task-specific firing patterns.

Each sub-hypothesis was tested independently, with the goal of demonstrating that proving all three provides strong evidence in support of the primary hypothesis.

Activation Tracing and Data Collection

Experimental data were obtained by feeding controlled recall and reasoning prompt pairs to the model. For each prompt, the model was permitted to generate a limited number of new tokens ($n = 10$) to produce a response. During this process, the `nnsight` library was employed to trace the model's execution. Specifically, activations were collected

for the final token of the input sequence across all 28 transformer layers:

- Hidden states (outputs of each layer).
- Self-attention weights (attention distributions from the attention mechanism).
- MLP activations (outputs of the gate projection in the MLP block).

This process enabled systematic extraction of internal representations and computational signals relevant to recall and reasoning tasks. For hypotheses H1 and H2, hidden state statistics and attention metrics were utilized, whereas H3 analysis relied specifically on neuron-level MLP activations.

General Statistical Framework

Across hypotheses, the following statistical procedures were employed:

- **Significance testing:** Mann–Whitney U test (Mann and Whitney 1947) to compare distributions between recall and reasoning tasks.
- **Effect size:** Cohen's d (Cohen 1988) as a measure of magnitude of difference.
- **Multiple comparisons:** FDR (Benjamini and Hochberg 1995) correction ($\alpha = 0.01$) for H1 and H2, and Bonferroni correction (Dunn 1961) ($\alpha = 0.0001$) for the fine-grained H3 analysis.

Cross-Validation

Robustness was further assessed via 5-fold cross-validation on the top 50 most task-specific neurons identified in the initial analysis. A neuron was considered consistently task-specific if it maintained the same task preference (recall vs. reasoning) in at least 80% of the folds.

Firing Pattern Analysis

Firing probabilities were computed for each significant neuron under recall and reasoning tasks using the binary firing indicators. Task specificity was quantified as the absolute difference in firing probability between the two task types. Neurons were ranked based on this measure, and their distribution across layers was analyzed to identify points in the network where task-specific signals were most concentrated.

Experimental Procedures for Hypothesis Evaluation

H1: Specific layers specialize for recall vs reasoning (layer specialization) To test H1, a comprehensive layer-wise activation analysis framework was implemented. For each of the 28 transformer layers, six activation features were computed: (1) hidden state norms, (2) hidden state means, (3) attention entropy measuring the distribution of attention weights, (4) attention concentration quantifying focus patterns, (5) MLP activation magnitude, and (6) activation sparsity indicating the proportion of inactive neurons. Distributions of these features were compared between recall and reasoning tasks using the Mann–Whitney U test, and effect sizes were quantified with Cohen's d. To control for multiple comparisons, FDR correction was applied with $\alpha = 0.01$. Layers were classified as recall-specialized, reasoning-specialized, mixed, or non-specialized based on consistent significant effects ($p < 0.01$ after FDR correction) with substantial effect sizes ($|d| > 0.5$) across multiple activation features.

H2: Attention heads show differential activation patterns between task types To test H2, head-level activation patterns were analyzed across all 784 attention heads (28 layers $\times$ 28 heads per layer). Five attention-specific metrics were computed for each head: (1) attention entropy measuring distribution uniformity, (2) maximum attention weight indicating peak focus, (3) attention focus quantifying concentration patterns, (4) attention spread measuring dispersion, and (5) the Gini coefficient (Gini 1936) assessing inequality of attention allocation. Rigorous statistical controls were applied, including an FDR correction threshold of 0.0001, a minimum effect size requirement of $|d| > 1.0$, and exclusion of heads with insufficient or invariant data. Heads were classified as consistently specialized if they demonstrated significant differential activation between recall and reasoning tasks in at least three of the five metrics while meeting the strict thresholds.

H3: MLP neurons exhibit task-specific firing patterns To test H3, the activations of individual neurons within the MLP layers were analyzed. For each neuron and each task, both the raw activation value and a binary firing indicator (1 if activation > 0, 0 otherwise) were extracted for the final token of the input. Distributions of activations between recall and reasoning tasks were compared using the Mann–Whitney U test, and Cohen's d was computed as an effect size measure. To correct for multiple testing across the large number of neurons, Bonferroni correction was applied with $\alpha = 0.0001$. Neurons were deemed task-specific if their corrected p-value was less than 0.0001 and their effect size ($|d|$) exceeded 1.0, indicating a substantial and reliable difference in activation.

Results and Analysis

Our systematic analysis revealed that there are circuits which are inheritently responsible for reasoning as opposed to recall. For each of the sub hypothesis, we present the following results

Layer-Wise Attention Behavior in Recall vs. Reasoning

The **Layer Specialization Hypothesis (H1)** posits that different layers within the LLM are preferentially engaged in recall or reasoning. To test this, we analyzed six activation features per layer across all 28 layers: hidden state norms, hidden state means, attention entropy, attention concentration, MLP activation magnitude, and activation sparsity. Distributions were compared using the Mann–Whitney U test, and multiple comparisons were controlled via false discovery rate (FDR) correction at $\alpha = 0.01$. Effect sizes were quantified with Cohen's d:

$$d_l = \frac{\mu(X_l^{(r)}) - \mu(X_l^{(s)})}{\sigma_p},$$

Table 1: Consistently specialized layers identified in 5-fold cross-validation (H1).

Specialization Type	Layers
Recall-specialized	3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 17, 19, 25
Reasoning-specialized	1, 2, 18

where $\mu(\cdot)$ denotes the mean activation for recall ($X_l^{(r)}$) or reasoning ($X_l^{(s)}$), and σ_p is the pooled standard deviation.

Layer Specialization Patterns - The analysis revealed clear differences in activation patterns. As shown in Figure 1, many layers exhibited statistically significant effects ($p < 0.01$ after FDR correction) with substantial effect sizes ($|d_l| > 0.5$). Based on consistent effects across metrics, layers were classified into four categories: 15 recall-specialized, 3 reasoning-specialized, 10 mixed-specialized, and 0 non-specialized. Figures 2 and 3 illustrate the dominance and broader distribution of recall-specialized layers compared to the more localized reasoning-specialized layers.

Cross-Validation Robustness - To assess robustness, we repeated the analysis using 5-fold cross-validation. A layer was deemed consistent if it maintained the same specialization type in at least 80% of folds:

$$\frac{1}{5}\sum_{f=1}^{5} I\big(C_l^{(f)} = C_l^*\big) \geq 0.8,$$

where $C_l^{(f)}$ denotes the specialization classification of layer l in fold f and C_l^* is its modal class. Results confirmed high stability: 18 layers satisfied this criterion, with 15 recall-specialized and 3 reasoning-specialized (Table 1).

These findings provide strong evidence for the Layer Specialization Hypothesis. Recall-specialized layers are both more numerous and broadly distributed, while reasoning-specialized layers are fewer and concentrated. The consistency across folds demonstrates that these specialization patterns are robust and not artifacts of data partitioning.

Attention Head Specialization Across Tasks

The **Attention Head Specialization Hypothesis (H2)** extends the notion of functional differentiation to the granularity of individual attention heads, proposing that specific heads preferentially engage in recall or reasoning tasks. To rigorously test H2 and mitigate artifacts from head-level variability, a fixed analysis pipeline with stricter statistical controls was employed. These included a more stringent FDR correction threshold ($\alpha = 0.0001$), a higher minimum effect size requirement ($|d| \geq 1.0$), and robust data validation to exclude heads with insufficient or invariant data.

Head-Level Analysis - Attention head behavior was quantified using five metrics: entropy, maximum weight, focus, spread, and the Gini coefficient. A head was defined as specialized if it exhibited significant differential activation between recall and reasoning tasks in at least three of these metrics while also satisfying the statistical and effect size thresholds.

Specialization Results - Out of 784 total heads (28 layers $\times$ 28 heads), 583 were identified as consistently specialized under these strict criteria. The breakdown is shown in Table 2: 239 recall-specialized, 219 reasoning-specialized, and 125 mixed-specialized.

Table 2: Breakdown of consistently specialized attention heads (H2).

Specialization Type	Number of Heads
Recall-specialized	239
Reasoning-specialized	219
Mixed-specialized	125
Total	583

Effect Sizes - The distribution of effect sizes among specialized heads was substantial, ranging from 2.8 to 13.8 with a mean of 6.83. Figure 4 illustrates this distribution, showing that specialization effects are widespread and not limited to a small subset of heads.

Figure 5 highlights the top 15 specialized heads ranked by average effect size, including recall-specialized heads such as L2H5, L7H1, and L6H9, and reasoning-specialized heads such as L26H5, L27H8, and L21H10. Figure 6 further demonstrates that specialized heads are distributed across all layers of the network, with certain layers (e.g., 0–4) exhibiting higher concentrations.

Even under strict statistical criteria, more than 70% of heads (583/784) were consistently specialized. This strong prevalence of head-level specialization provides compelling evidence in support of H2, confirming that functional differentiation extends beyond layers to individual attention heads across the network.

MLP Neuron Task-Specificity (Hypothesis H3)

The **MLP Neuron Task-Specificity Hypothesis (H3)** posits that individual MLP neurons exhibit differentiated activation patterns depending on whether the task involves recall or reasoning. We analyzed activations for all 530,432 neurons across 28 layers. In total, 163,058 neurons (30.74%) showed statistically significant task-specific differences, satisfying both a large effect size criterion ($d > 1.0$) and a stringent Bonferroni-corrected significance threshold ($p < 0.0001$). These results provide strong evidence that task-specific selectivity emerges at the granularity of individual MLP units.

Layer-Wise Distribution - Task-specific neurons were not uniformly distributed across layers. Certain layers contained disproportionately high densities of specialized neurons. For example, Layer 4 contained 11,857 task-specific neurons (62.59% of the layer's units), indicating the existence of localized processing hubs for recall and reasoning. Layer-wise distributions revealed distinct peaks in task-specificity at particular depths, suggesting that specialization is structurally organized rather than evenly spread across the network.

Cross-Validation Robustness - To test robustness, we evaluated the top 50 most task-specific neurons under 5-

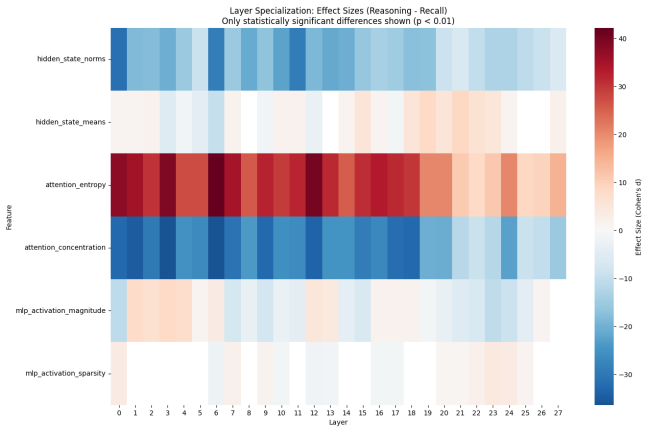

Figure 1: Effect size heatmap showing recall vs. reasoning differences across 28 layers.

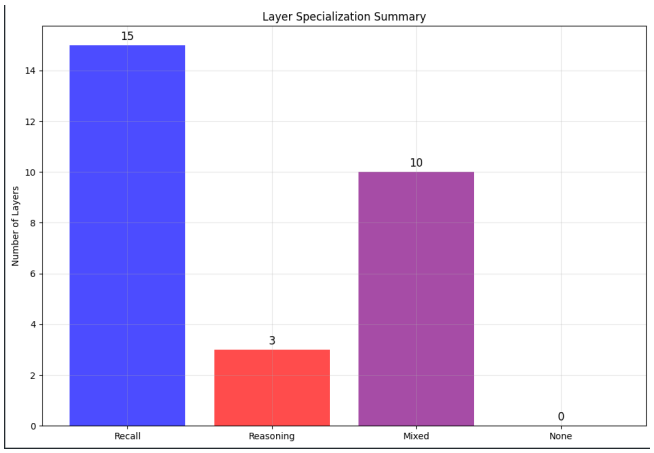

Figure 2: Layer specialization counts from single-fold analysis (recall, reasoning, mixed).

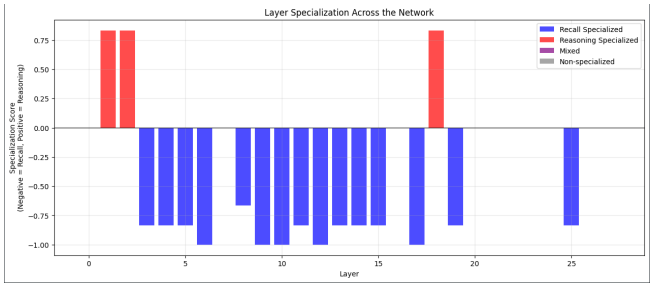

Figure 3: Distribution of recall- and reasoning-specialized layers across the network.

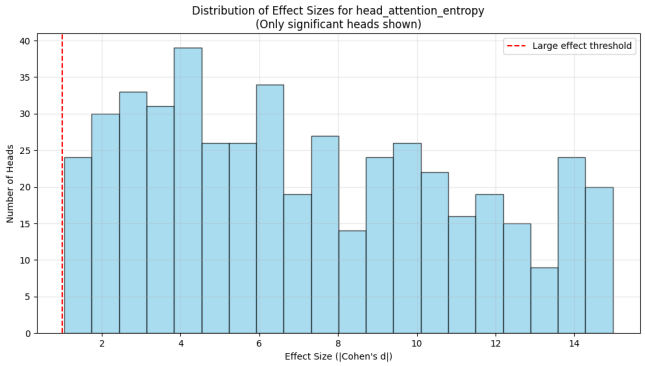

Figure 4: Histogram of effect sizes for specialized attention heads.

fold cross-validation. Consistency was perfect: all 50 neurons (100%) retained their task-preference classification (recall or reasoning) across all folds. This result, illustrated in Figure 7, confirms that the observed specializations are stable across resampled datasets and not artifacts of a particular split.

Activation Profiles - Detailed inspection of the top neurons revealed sharp differences in firing probabilities across tasks. As shown in Figure 8, recall-preferring neurons were highly active during recall tasks and minimally responsive during reasoning tasks, whereas reasoning-preferring neurons displayed the inverse profile. Many of these neurons exhibited near-binary activation patterns, strongly firing for one task type while remaining effectively silent for the other, providing clear evidence of functional specialization.

Nearly one-third of all MLP neurons thus demonstrated robust, statistically significant task-specificity. The presence of dense clusters in certain layers, combined with 100% cross-validation stability and near-binary firing profiles, provides compelling evidence that MLP neurons encode task preference in a fine-grained, interpretable, and reproducible manner.

Limitations

Our analysis is subject to some limitations. First, all experiments were conducted on a single model (Qwen2.5-7B-Instruct), so generalization to other architectures or training paradigms remains untested. Second, the recall–reasoning dataset was restricted to geography-based facts with simple two-step reasoning, which may not capture the full spectrum of reasoning tasks. Third, while we observed strong correlational evidence of specialization, causal validation was limited by technical constraints on head ablations. Finally, our analysis focused on layers, heads, and neurons independently, leaving higher-order interactions between these components outside the present scope.

Future Work

Cross-Architecture Validation. Future work should investigate whether circuit specialization patterns identified in Qwen2.5-7B-Instruct generalize across different model architectures, sizes, and training paradigms. This research direction, corresponding to our original hypothesis, would establish whether recall–reasoning specialization represents a

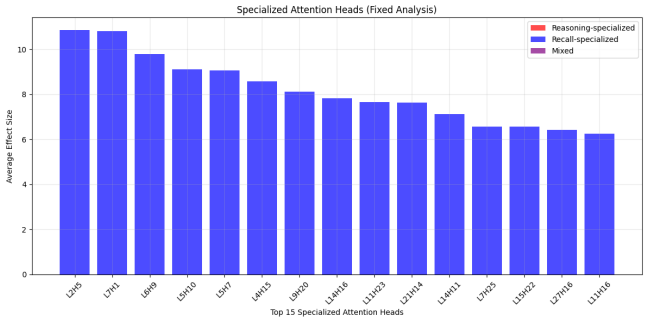

Figure 5: Top 15 specialized heads ranked by average effect size, showing strongest recall vs. reasoning preferences.

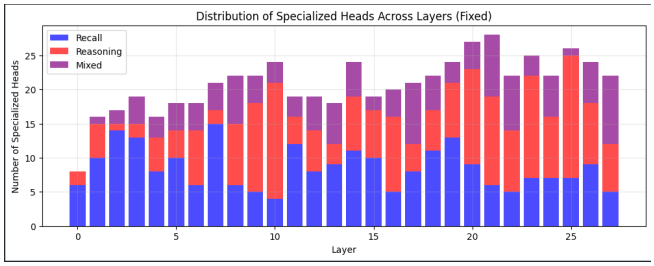

Figure 6: Distribution of specialized heads across layers, with higher concentrations in early layers (0–4).

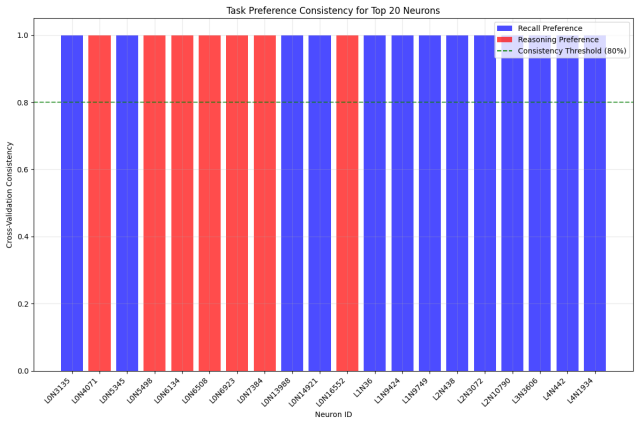

Figure 7: Task preference consistency for top 20 neurons.

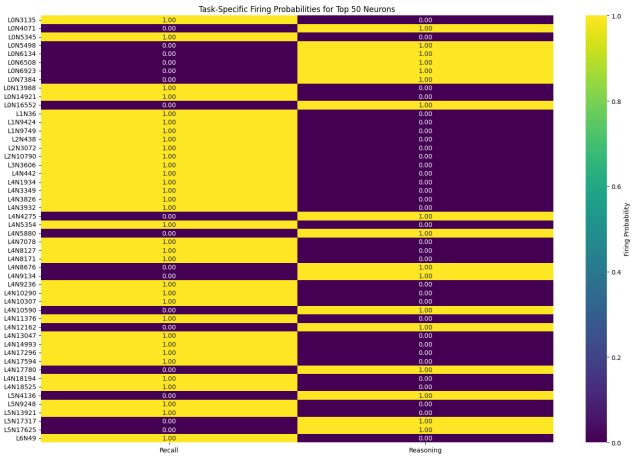

Figure 8: Task-specific activation probabilities for top 20 neurons.

fundamental property of transformer-based language models or an artifact of specific architectural choices.

Causal Intervention Studies. Developing robust intervention methodologies to causally validate the functional roles of identified circuits remains a critical priority. Future studies should implement systematic activation patching, circuit ablation, and targeted fine-tuning experiments to demonstrate that interventions on specialized circuits selectively impair corresponding capabilities, thus providing stronger causal support for the specialization hypothesis.

Conclusion

For the **Layer Specialization Hypothesis (H1)**, we observed significant and robust differences in activation patterns between recall and reasoning tasks across many layers. Cross-validation confirmed that these specialized layers were consistently identified, showing that the differentiation is a stable property of the model's processing. Recall-specialized layers were more numerous and broadly distributed, whereas reasoning-specialized layers were fewer and concentrated, suggesting asymmetric allocation of representational resources.

The **Attention Head Specialization Hypothesis (H2)** was also strongly supported. Even under strict statistical controls and high effect size thresholds, many attention heads exhibited consistent specialization across multiple attention metrics. This indicates that layer-level effects are mediated in part by fine-grained head-level roles. While causal validation through ablation was inconclusive, the statistical evidence provides robust correlational support that some heads preferentially encode recall-related signals while others specialize in reasoning computations.

Extending further, the **MLP Neuron Task-Specificity Hypothesis (H3)** showed that nearly one-third of the $530{,}432$ neurons analyzed exhibited significant task-specific activation. These neurons were not evenly distributed, with certain layers containing dense hubs of specialized units. The top neurons maintained 100% task-preference consistency across cross-validation folds, and many displayed near-binary firing patterns, strongly activating for one task type while remaining silent for the other.

Taken together, these findings establish that Qwen 2.5-7B-Instruct develops **hierarchically organized specialization** spanning layers, heads, and neurons. This multi-scale differentiation is statistically robust, reproducible under cross-validation, and interpretable as localized processing hubs. Identifying such components provides concrete anchors for mechanistic interpretability and points toward op-

portunities for causal probing, targeted interventions, and architectural modifications to better understand and guide model behavior in recall versus reasoning tasks.

References

Benjamini, Y.; and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1): 289–300.

Chen, Y.; Benton, J.; Radhakrishnan, A.; Uesato, J.; Denison, C.; Schulman, J.; Somani, A.; Hase, P.; Wagner, M.; Roger, F.; Mikulik, V.; Bowman, S. R.; Leike, J.; Kaplan, J.; and Perez, E. 2025. Reasoning Models Don't Always Say What They Think. arXiv:2505.05410.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates, 2 edition. ISBN 978-0805802832.

Denison, C.; MacDiarmid, M.; Barez, F.; Duvenaud, D.; Kravec, S.; Marks, S.; Schiefer, N.; Soklaski, R.; Tamkin, A.; Kaplan, J.; Shlegeris, B.; Bowman, S. R.; Perez, E.; and Hubinger, E. 2024. Sycophancy to Subterfuge: Investigating Reward-Tampering in Large Language Models. arXiv:2406.10162.

Dunn, O. J. 1961. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293): 52–64.

Elhage, N.; Nanda, N.; Olsson, C.; and et al. 2021. A Mathematical Framework for Transformer Circuits.

Fiotto-Kaufman, J.; Loftus, A. R.; Todd, E.; Brinkmann, J.; Juang, C.; Pal, K.; Rager, C.; Mueller, A.; Marks, S.; Sharma, A. S.; Lucchetti, F.; Ripa, M.; Belfki, A.; Prakash, N.; Multani, S.; Brodley, C.; Guha, A.; Bell, J.; Wallace, B.; and Bau, D. 2024. NNsight and NDIF: Democratizing Access to Foundation Model Internals.

Geiger, A.; Ibeling, D.; Zur, A.; Chaudhary, M.; Chauhan, S.; Huang, J.; Arora, A.; Wu, Z.; Goodman, N.; Potts, C.; and Icard, T. 2025. Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability. arXiv:2301.04709.

Geva, M.; Schuster, T.; and Berant, J. 2021. Transformer Feed-Forward Layers Are Key-Value Memories. In *EMNLP*.

Gini, C. 1936. On the measure of concentration with special reference to income and statistics. *Colorado College Publication, General Series*, 208: 73–79.

Greenblatt, R.; Denison, C.; Wright, B.; Roger, F.; MacDiarmid, M.; Marks, S.; Treutlein, J.; Belonax, T.; Chen, J.; Duvenaud, D.; Khan, A.; Michael, J.; Mindermann, S.; Perez, E.; Petrini, L.; Uesato, J.; Kaplan, J.; Shlegeris, B.; Bowman, S. R.; and Hubinger, E. 2024. Alignment faking in large language models. arXiv:2412.14093.

Larenz, R. 2023. On Overcoming the Alleged Alienation between Christianity and Physics.

Mann, H. B.; and Whitney, D. R. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1): 50–60.

Meng, K.; Bau, D.; Andonian, A.; and Belinkov, Y. 2022. Locating and Editing Factual Knowledge in GPT. In *Advances in Neural Information Processing Systems*.

Nanda, N. 2023. Progress in Mechanistic Interpretability. Blog post.

Olsson, C.; Elhage, N.; Nanda, N.; Joseph, N.; DasSarma, N.; Henighan, T.; Mann, B.; Askell, A.; Bai, Y.; Chen, A.; Conerly, T.; Drain, D.; Ganguli, D.; Hatfield-Dodds, Z.; Hernandez, D.; Johnston, S.; Jones, A.; Kernion, J.; Lovitt, L.; Ndousse, K.; Amodei, D.; Brown, T.; Clark, J.; Kaplan, J.; McCandlish, S.; and Olah, C. 2022. In-context Learning and Induction Heads. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Olsson, C.; and et al. 2022. In-context Learning and Induction Heads.

Petroni, F.; Rocktäschel, T.; Lewis, P.; Bakhtin, A.; Wu, Y.; Miller, A. H.; and Riedel, S. 2019. Language Models as Knowledge Bases? arXiv:1909.01066.

Turpin, M.; Michael, J.; Perez, E.; and Bowman, S. R. 2023. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. arXiv:2305.04388.

Wang, H.; et al. 2022. Interpretability in Neural Networks: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; Chi, E. H.; Hashimoto, T.; Vinyals, O.; Liang, P.; Dean, J.; and Fedus, W. 2022. Emergent Abilities of Large Language Models. arXiv:2206.07682.