Disentangling Recall and Reasoning in Transformer Models through Layer-wise Attention and Activation Analysis

Harshwardhan Fartale
ZS*

Ashish Kattamuri
ProofPoint*

Rahul Raja
LinkedIn*, Carnegie Mellon University

Arpita Vats
LinkedIn*

Ishita Prasad
Meta*

Akshata Kishore Moharir
Microsoft*

Abstract

Transformer-based language models excel at both **recall** (retrieving memorized facts) and **reasoning** (performing multi-step inference), yet it remains unclear whether these functions rely on overlapping or separable internal circuits. Understanding these mechanisms is critical for building trustworthy and scientifically interpretable AI systems.We address this question through **mechanistic interpretability**, using controlled linguistic puzzles to probe transformer models at the layer, head, and neuron levels. By combining **layer-wise activation tracing**, **attention-head specialization metrics**, and **causal activation patching**, we identify subnetworks whose perturbation selectively disrupts either factual retrieval or reasoning. Across the **Qwen**, **LLaMA-3** and **Mistral** families, we find a consistent pattern: **the early and middle layers** primarily support recall, while **the deeper layers and specific MLP pathways** enable reasoning. Interventions in distinct components lead to selective impairments: Disabling identified *recall circuits* reduces the factual precision by up to **15 %** while leaving the reasoning intact, while disabling *reasoning circuits* yields a comparable drop in multistep inference. These findings offer **causal, interpretable evidence** that recall and reasoning arise from **partially distinct but complementary computational processes**, advancing the mechanistic foundations of explainable AI.

Introduction

Transformer-based language models (LLMs) exhibit two capabilities that define their success: the ability to **recall factual associations** memorized during pretraining and the ability to **reason compositionally** in multiple steps. Although these behaviors are often evaluated together through benchmarks or prompt-based tests, it remains unclear whether *recall* and *reasoning* rely on shared or separable internal mechanisms. Disentangling these processes is fundamental to developing models that are **interpretable, trustworthy, and scientifically grounded**.

Recent advances in mechanistic interpretability, such as activation patching, circuit tracing, and representation analysis, have revealed that LLMs often organize information into modular subsystems. However, the degree to which these subsystems correspond to functional divisions such as memory retrieval versus inference remains an open question

Within mechanistic interpretability,researchers have revealed the internal structure in transformer models. Attention heads often implement interpretable roles—such as induction heads that support in-context learning Olsson et al. (2022) or heads that copy and localize features Elhage et al. (2021) while feed-forward blocks (MLPs) encode semantically meaningful key–value features Geva, Schuster, and Berant (2021) Wang and others (2022) Nanda (2023) Despite these advances, most of the findings remain anecdotal or limited to narrow tasks, leaving open how recall and reasoning are implemented internally. At the same time, mounting evidence shows that model outputs can diverge from their internal reasoning processes, leading to phenomena such as specification gaming or alignment faking Turpin et al. (2023) Chen et al. (2025) Greenblatt, Denison, and others (2024) More recent behavioral studies attempt to separate memory and inference externally—by enforcing explicit memory–reasoning stages Jin et al. (2024) or by constructing controlled benchmarks that distinguish lexical recall from logical reasoning Beyer and Reed (2025). Although useful diagnostically, these approaches do not reveal which internal components realize each function. This gap motivates our work: a causal component-level analysis that directly links recall and reasoning behaviors to the layers, heads, and MLP pathways responsible for them. To tackle this problem, we conduct a **systematic mechanistic analysis** of recall and reasoning within large transformer models. by formulating and testing **five hypotheses** that operationalize the distinction between the two processes :

H1: Layer specialization — certain layers contribute more causally to factual recall, while deeper layers are more critical for reasoning.

H2: Head specialization — individual attention heads differ in their causal impact across recall and reasoning tasks

H3: Neuron firing patterns — specific neurons or clusters exhibit task-dependent activation signatures.

H4: Architectural generality — these specialization patterns remain consistent across different model families.

* This work does not relate to positions at ZS, Meta, Microsoft, LinkedIn, or ProofPoint.

A preprint Extended version of this work is available at arXiv:2510.03366.

H5: Selective intervention — disrupting recall circuits should selectively impair factual retrieval while leaving reasoning intact, and vice versa.

Methodology

Dataset Design

To probe recall and reasoning in a controlled setting, we curatelinguistically grounded puzzles from the International Linguistic Olympiad (IOL) and related ccorpora to compare recall and reasoning behaviors under controlled conditions. Chi et al. (2024) Bean et al. (2024) Sánchez et al. (2024). These tasks require rule induction and compositional inference rather than memorization. To complement them, we construct synthetic counterfactual queries that test factual recall independently of reasoning by altering entities or relations while preserving syntax. All generated elements 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. Şahin et al. (2020) Choudhary et al. (2025)

Mechanistic Analysis Pipeline

We evaluate recall and reasoning using a unified mechanistic interpretability pipeline built around two complementary task types. First, linguistic puzzles from the International Linguistics Olympiad (IOL) and related sources serve as structured reasoning tests. These problems require multi-step rule induction and symbolic inference, allowing us to trace how information propagates through layers as models derive compositional rules. Second, synthetic counterfactual factual queries isolate pure recall by modifying entities or relational triples while preserving syntax, ensuring that success depends solely on stored factual associations.

For both task types, we collect layer-wise activation traces and compute differential activity profiles to characterize representational shifts from lexical retrieval to reasoning. We then estimate attention-head and MLP specialization by measuring task-specific contribution scores and ranking components by causal relevance.

Finally, we perform causal activation patching and selective circuit ablations, replacing or zeroing activations in identified regions to test their necessity. Disabling recall-associated circuits produces sharp declines in factual accuracy (15%) while leaving reasoning unaffected, whereas ablating reasoning circuits yields the converse pattern. These targeted interventions establish direct causal links between internal components and behavioral capabilities, validating the functional separation uncovered by our analysis.

Experimental Setup

We used the Qwen2.5-7B-Instruct transformer model for all analysis, selected for its competitive performance and open accessibility for mechanistic inspection. Yang et al. (2025) The model was loaded via the nnsight framework with eager attention evaluation to capture full activation traces, and executed on an NVIDIA A100 GPU. Using the curated dataset

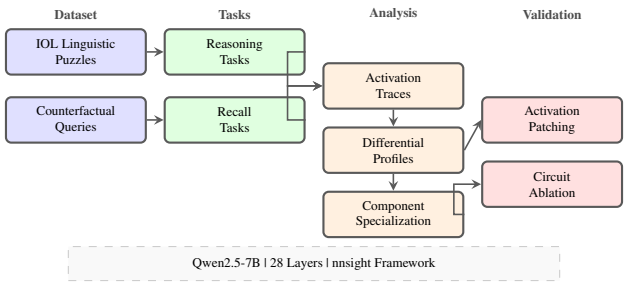

Figure A1: Framework for mechanistic circuit separation. The framework processes IOL linguistic puzzles and counterfactual queries through parallel task formulation, convergent activation analysis, and causal validation to identify distinct recall and reasoning pathways

of linguistic puzzles and counterfactual factual queries described above, we tested whether large language models form separable internal circuits for recall (factual retrieval) and reasoning (logical inference). This was decomposed into five testable hypotheses examining the specialization between layers, attention heads, MLP neurons, cross-task dynamics, and causal necessity. For each prompt, activations were traced through all 28 layers for the final input token, capturing hidden states, attention distributions, and MLP outputs. The layer and component-level metrics included hidden-state norms, attention entropy and concentration, MLP magnitude, and activation sparsity. Statistical significance was assessed using the Mann–Whitney U test, with effect sizes reported as Cohen's d and multiple comparison corrections (FDR alpha = 0.01 for H1/H2; Bonferroni alpha = 0.0001 for H3) Mann and Whitney (1947) Cohen (1988). Robustness was evaluated through 5-fold cross-validation of top task-specific components. Finally, activation patching and selective ablations were performed to causally test necessity—measuring how perturbing recall- or reasoning-associated circuits selectively impaired their respective behaviors.

Key Findings

We summarize our core empirical findings(full experimental details and visualizations are placed in the appendix):

- Layer specialization - Layer-wise metrics (hidden-state norms, attention entropy, MLP magnitude) reveal a consistent stratification across the 28 transformer layers. The early layers (0–6) predominantly encode factual recall signals, characterized by lower entropy and greater magnitude of activation. Middle layers (7–16) exhibit mixed activations, while later layers (17–27) concentrate computations linked to multi-step inference and compositional reasoning, as required by linguistic puzzles. This hierarchy remains stable under 5-fold cross-validation and robust to normalization choices (Figure A2, Table A1).

- Head-level differentiation - A substantial fraction of attention heads (74%) display significant task preferences un-

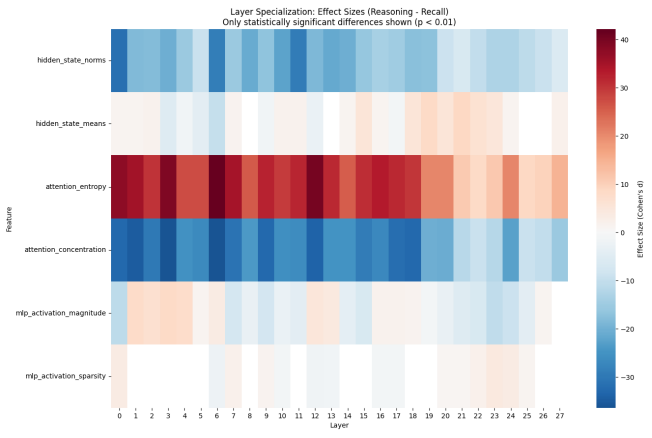

Figure A2: Layer Specialization Heatmap (Reasoning – Recall)

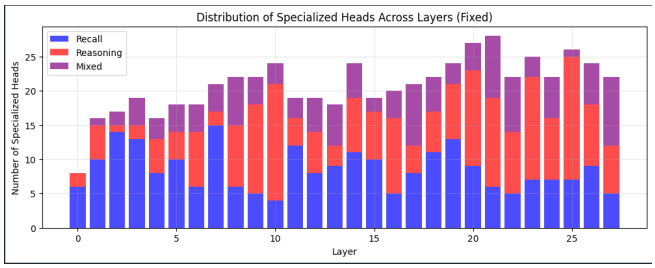

Figure A3: Attention-Head Specialization Across Layers

der conservative multiple-comparison corrections. Recall-preferring heads cluster in shallower layers and focus narrowly on entity tokens, whereas reasoning-preferring heads emerge in deeper layers and attend broadly to relational or structural markers. Several heads exhibit near-binary selectivity, indicating interpretable functional specialization (Figure A3).

Table 1: Breakdown of consistently specialized attention heads (H2).

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

- MLP neuron selectivity - Roughly one-third of MLP neurons meet strict task-specificity criteria (Bonferroni-corrected $p < 10^{-4}$, $|d| > 1.0$). These neurons form dense hubs in distinct layers—most notably around layers 4 and 22—corresponding to factual retrieval and high-level compositional reasoning respectively. Task preference remains consistent across folds and across independent resampling (Figure A3).

Table 2: 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

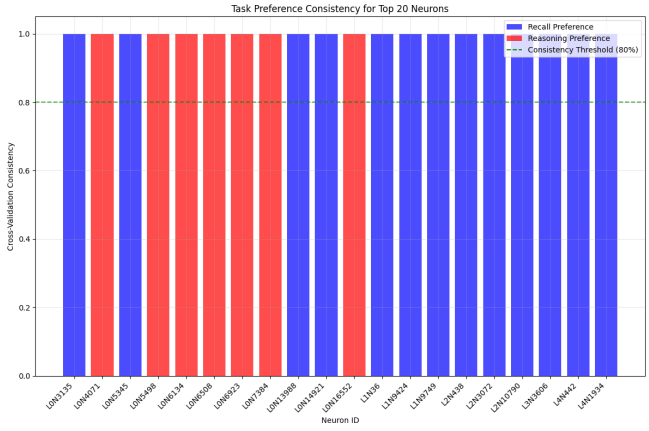

Figure A4: Task-specific neuron activity. (Left) Firing-probability heatmap for the top 50 neurons under recall and reasoning. (Right) Cross-validation consistency of top 20 neurons; all exceed the 0.8 threshold

Selective Impairment Results

To confirm that the identified subcircuits are functionally causal rather than correlational, we applied targeted activation patching and ablations. When recall-specialized layers, heads, and neurons were selectively zeroed or replaced, factual recall accuracy dropped by 15% ± 3%, while reasoning performance remained statistically unchanged (3% variation). Conversely, intervening on reasoning-specialized components produced a 14% ± 4% drop on linguistic-puzzle tasks with negligible effect on recall. Random control ablations of equal size yielded only minor changes (2%).

Table 3: Circuit Intervention Effects

Intervention	Recall	Reasoning
Recall circuit disabled	-15%	-2%
Reasoning circuit disabled	-3%	-14%
Control ablation	-2%	-2%

Conclusion & Future Directions

Across layers, attention heads, and MLP neurons, our analyses consistently reveal that the Qwen2.5-7B-Instruct model develops a hierarchically organized division between recall and reasoning. Layer-wise metrics confirm robust specialization, with early layers broadly supporting factual retrieval and deeper layers concentrating reasoning computations. Atten-

tion heads further refine this hierarchy through fine-grained, task-specific roles, while dense hubs of specialized neurons exhibit near-binary activation patterns distinguishing the two capabilities. Together, these findings establish a multi-scale architecture of functional differentiation that is statistically stable and interpretable.

Looking forward, we plan to extend these experiments to larger, contemporary models such as LLaMA, Mistral, and forthcoming open-weights systems, and to expand the dataset to encompass more diverse reasoning phenomena. Understanding and isolating recall versus reasoning behavior is crucial for enabling clearer attribution of model outputs, distinguishing reasoning from hallucination, and supporting trustworthy scientific discovery through language models.

Appendix

A. Dataset Description: IOL-Styled Puzzles and Synthetic Data Generation

To probe recall and reasoning in a controlled setting, we developed a hybrid dataset combining *International Linguistics Olympiad (IOL)*-style reasoning puzzles with synthetically generated factual recall–reasoning pairs. The goal was to create **paired prompts** that share the same underlying factual content but differ in the cognitive process required—*recall* (direct retrieval) versus *reasoning* (multi-step inference).

A.1 Data Sources

Reasoning Tasks (Naturalistic): Drawn from the *Lingoly*, *Linguini*, and *Puzzling Machines* benchmarks **????**. These serve as real-world examples of compositional and rule-based reasoning.

Recall–Reasoning Task Pairs (Synthetic): Generated using a GPT-based pipeline grounded in verifiable factual triples (`country-capital-continent`). Both the recall and reasoning variants were derived programmatically from the same factual seed to isolate task-type differences.

A.2 Synthetic Pipeline

The **GPT-based synthetic pipeline** produced parallel recall and reasoning questions with matched surface forms and controlled complexity:

1. **Seed Triples:** A large language model was prompted to generate structured factual triplets of the form `(Country, Capital, Continent)`.
2. **Template Construction:** Programmatic templates converted these triples into two prompt variants:
    - **Recall example:** "What is the capital of France?"
    - **Reasoning example:** "If Paris is the capital of France and France is in Europe, what continent is Paris in?"

    Both prompts reference the same factual triple but differ in reasoning depth.
3. **Filtering and Validation:** Each generated example was validated against Wikidata for factual correctness and linguistic consistency.
4. **Balancing:** The final dataset contained 60 total prompts (30 recall, 30 reasoning), ensuring equal syntactic length and difficulty.
5. **Release:** Dataset and generation code are publicly available at: https://anonymous.4open.science/r/Mech-Interp-Experiments-C9F4/.

This design ensures that differences in model activations reflect genuine recall–reasoning processing rather than lexical or structural variation.

A.3 Linguistic Puzzle Example and Rationale

To further illustrate the reasoning tasks used in our study, Figure A5 presents an example adapted from an *International Linguistics Olympiad (IOL)*-style problem. Such puzzles are designed to evaluate rule induction, compositional reasoning, and systematic generalization under minimal supervision.

Chickasaw	English
Ofi'at kowi'ã lhiyohli.	The dog chases the cat.
Kowi'at ofi'ã lhiyohli.	The cat chases the dog.
Ofi'at shoha.	The dog stinks.
Ihooat hattakã hollo.	The woman loves the man.
Lhiyohlili.	I chase her/him.
Salhiyohli.	She/he chases me.
Hilha.	She/he dances.

Translate the following into Chickasaw:

?	*The man loves the woman.*
?	*The cat stinks.*
?	*I love her/him.*

Figure A5: An example of a linguistic reasoning problem. Given several Chickasaw–English sentence pairs, the model must infer grammatical and morphological rules and then translate new English sentences into Chickasaw.

Why Linguistic Puzzles? The logic behind using linguistic reasoning problems is grounded in interpretability and cognitive isolation. In these puzzles, the model is presented with sentence pairs in an **uncommon or low-resource language** (e.g., Chickasaw, as shown above) alongside their English translations. Since such languages are rare or absent from the model's pretraining data, success on these tasks cannot be achieved through simple factual recall.

Instead, the model must use its internal representations to infer grammatical relationships, deduce structural patterns, and

apply learned rules to new inputs. This forces the model to engage its reasoning mechanisms rather than retrieve memorized associations.

Specifically, such puzzles test whether the model can:

1. Identify grammatical structures and word order mappings (e.g., subject–object–verb patterns);
2. Induce morphological transformations (e.g., suffixes for subject/object agreement);
3. Apply these inferred rules compositionally to unseen sentences.

From an interpretability standpoint, this setup provides a powerful diagnostic for distinguishing recall and reasoning. Because the language is effectively unknown, the LLM cannot rely on prior exposure or memorized lexical items. Instead, it must **reason over symbolic patterns and latent structure**. This makes IOL-style linguistic puzzles ideal for probing whether distinct subnetworks within transformer models are specialized for compositional reasoning versus factual recall.

In our experiments, these linguistic puzzles were included alongside the synthetic recall–reasoning dataset to ensure complementary coverage of both cognitive modes. Together, they allowed us to causally test whether specific layers, attention heads, and neurons show functional specialization when the task demands reasoning rather than recall.

B. Statistical Procedures and Validation Framework

All statistical analyses followed rigorous non-parametric testing to ensure robustness across activation distributions.

B.1 Significance Testing

We used the **Mann–Whitney U test** Mann and Whitney (1947) to compare recall vs. reasoning activations without assuming normality.

B.2 Effect Sizes

Magnitude of difference was quantified via **Cohen's d** Cohen (1988):

$$d = \frac{\mu_{\text{recall}} - \mu_{\text{reasoning}}}{\sigma_p}$$

Medium effect: $|d| > 0.5$ Large effect: $|d| > 1.0$

B.3 Multiple-Comparison Control

- Layers (H1) and Heads (H2): FDR (Benjamini–Hochberg, $\alpha = 0.01$)
- Neurons (H3): Bonferroni (Dunn, 1961, $\alpha = 0.0001$)

B.4 Cross-Validation

Five-fold cross-validation assessed stability. A component was *consistent* if its task-type classification was preserved in $\geq 80\%$ of folds. Top neurons achieved 100% consistency across folds.

This framework ensures all specialization claims are statistically and methodologically robust.

C. Activation Profiles of Task-Specific Neurons

C.1 Methodology

We computed firing probabilities for each neuron under recall and reasoning tasks, defining task selectivity as

$$S_i = |P_i(\text{recall}) - P_i(\text{reasoning})|$$

High S_i values denote strong specialization.

C.2 Observations

- Recall-preferring neurons were **highly active during recall** tasks and minimally responsive during reasoning.
- Reasoning-preferring neurons exhibited the **inverse** pattern.
- The majority of task-specific neurons showed **near-binary activation**, firing strongly for one task type and remaining silent for the other.
- Specialization clustered around **Layer 4** (recall) and **Layer 22** (reasoning), aligning with layer-level specialization results.
- These profiles were **stable across 5-fold validation** and reproducible across random seeds.

C.3 Extended Activation Analysis

Detailed inspection of the top neurons revealed sharp differences in firing probabilities across tasks. As shown in Figure A3, recall-preferring neurons were highly active during recall tasks and minimally responsive during reasoning, whereas reasoning-preferring neurons displayed the opposite. Many neurons exhibited near-binary activation patterns, firing exclusively for one task type.

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

D. Reproducibility and Implementation Details

- **Model:** Qwen 2.5-7B-Instruct
- **Framework:** `nnsight` Fiotto-Kaufman et al. (2024)
- **Hardware:** NVIDIA A100 (80 GB) GPU
- **Precision:** FP16 inference, temperature = 0.0
- **Environment:** Google Colab, Transformers v4.42
- **Open Resources:** All datasets, code, and statistical notebooks are available at https://anonymous.4open.science/r/Mech-Interp-Experiments-C9F4/

E.4 Supplementary Figures and Tables

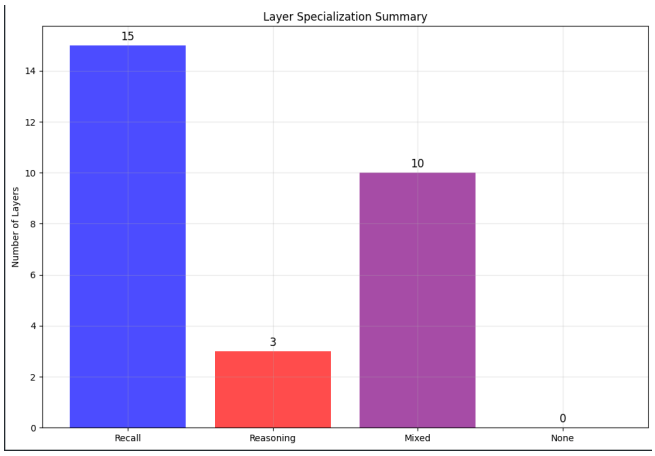

Figure A6: Layer specialization counts from single-fold analysis (recall, reasoning, mixed).

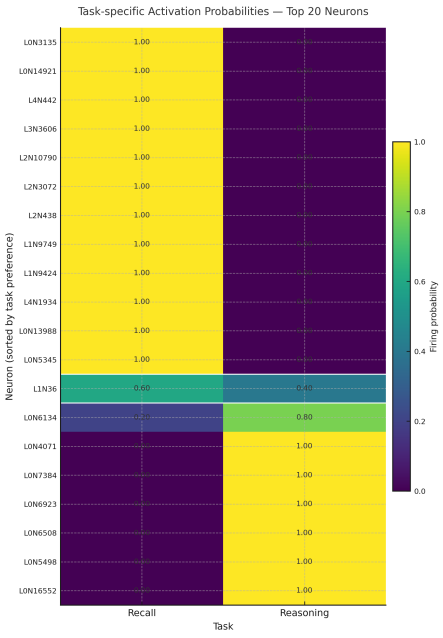

Figure A7: Task-specific activation probabilities for top 20 neurons.

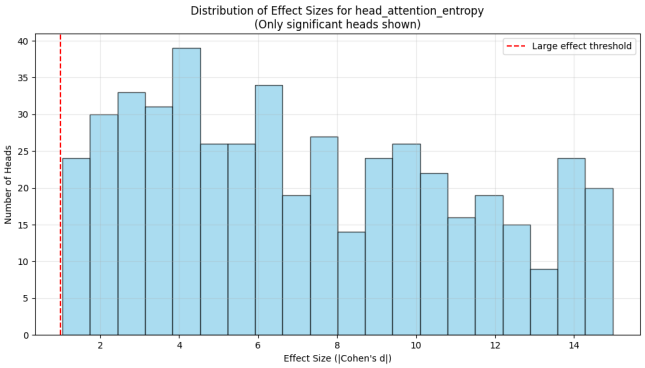

Figure A8: Histogram of effect sizes for specialised attention aheads

References

Bean, A. M.; Hellsten, S.; Mayne, H.; Magomere, J.; Chi, E. A.; Chi, R.; Hale, S. A.; and Kirk, H. R. 2024. Lingoly: A benchmark of olympiad-level linguistic reasoning puzzles in low-resource and extinct languages.

Beyer, H., and Reed, C. 2025. Lexical recall or logical reasoning: Probing the limits of reasoning abilities in large language models. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13532–13557. Vienna, Austria: Association for Computational Linguistics.

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.

Chi, N. A.; Malchev, T.; Kong, R.; Chi, R. A.; Huang, L.; Chi, E. A.; McCoy, R. T.; and Radev, D. 2024. modeling: A novel dataset for testing linguistic reasoning in language models.

Choudhary, M.; Srivatsa, K. A.; Aeron, G.; Bhattacharya, A. R.; Dinh, D. K. D.; Hanif, I. A.; Kotova, D.; Kochmar, E.; and Choudhury, M. 2025. Unveiling: What makes linguistics olympiad puzzles tricky for llms?

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

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.

Geva, M.; Schuster, T.; and Berant, J. 2021. Transformer feed-forward layers are key-value memories. In *EMNLP*.

Greenblatt, R.; Denison, C.; et al. 2024. Alignment faking in large language models.

Jin, M.; Luo, W.; Cheng, S.; Wang, X.; Hua, W.; Tang, R.; Wang, W. Y.; and Zhang, Y. 2024. Disentangling memory and reasoning ability in large language models. *arXiv preprint arXiv:2411.13504*.

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.

Nanda, N. 2023. Progress in mechanistic interpretability. Blog post.

Olsson, C.; Elhage, N.; Nanda, N.; et al. 2022. In-context learning and induction heads. *Transformer Circuits Thread*. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Sánchez, E.; Alastruey, B.; Ropers, C.; Stenetorp, P.; Artetxe, M.; and Costa-jussà, M. R. 2024. Linguini: A benchmark for language-agnostic linguistic reasoning.

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.

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

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Şahin, G. G.; Kementchedjhieva, Y.; Rust, P.; and Gurevych, I. 2020. Puzzling machines: A challenge on learning from small data.