Tracing Thought Processes in Large Language Models: A Tree-Based Approach for Understanding Reasoning

Anne Wang[1], Bassant Medhat[1], Matin Daghyani[1], Obed Dzikunu[1]
[1] University of British Columbia

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities across a wide range of tasks, particularly with the emergence of models that generate explicit reasoning in addition to final answers. These reasoning outputs follow step-by-step thinking, similar to how humans solve problems.

Despite this progress, evaluating the reasoning abilities of LLMs remains an underexplored area. It is still unclear why models succeed or fail, to what extent their reasoning outputs contribute to final predictions, and whether these intermediate steps effectively guide or affect their responses.

In this work, we investigate the reasoning behavior of the DeepSeek model using a subset of the AI2 Reasoning Challenge benchmark and synthetic data. Our findings suggest that DeepSeek exhibits consistent attention patterns across inputs, reflecting the sequential reasoning nature of generative models.

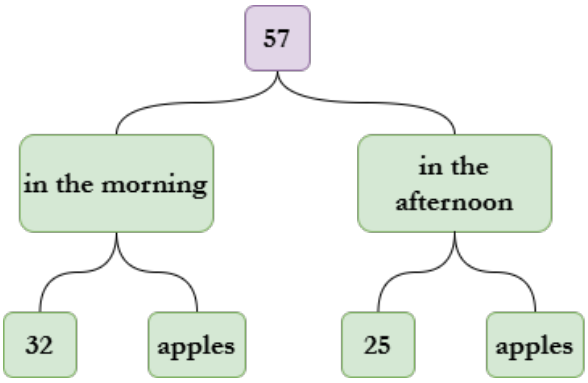

Figure 1: **Reasoning Tree Extracted from Model Attention:** ***Query:*** *"A store sold 32 apples in the morning and 25 apples in the afternoon. How many apples were sold in total?"* ***Output:*** `<think>` *The store sold 32 apples in the morning and 25 apples in the afternoon. I will add both to get the total.* `</think>` `Answer: 57.` The reasoning tree shows a two-level structure: the final answer (57) is supported by temporal thinking segments ("in the morning", "in the afternoon"), each further decomposed into quantity and item thinking tokens ("32", "apples"; "25", "apples").

1 Introduction

Large language models (LLMs) with reasoning (such as ChatGPT (OpenAI, 2023), DeepSeek(DeepSeek-AI, 2025)) have been broadly applied to many tasks. However, there is an open question about how these reasoning models actually perform the task of reasoning. For instance, DeepSeek-r1 (Guo et al., 2025) utilizes Chain-of-Thought (CoT) reasoning, generating long intermediate reasoning before generating the final answer. The extent to which LLMs effectively utilize intermediate CoT reasoning and the degree to which these intermediate steps contribute to the final output remains insufficiently explored. Further investigation is needed to determine the necessity and relevance of these reasoning steps to archive more reliable LLMs. Many works find that LLMs do shallow pattern identification or memorize training data rather than doing logical reasoning (Yan et al., 2024; Zhang et al., 2022). Other opinions explore the reasoning ability of LLM in arithmetic, symbolism, common sense, etc. (Pan et al., 2024).

With the rapid development of the new LLMs, the criterion of understanding their reasoning ability, other than evaluating their final performance, is a critical key for understanding the LLMs. To reduce hallucinations in LLMS, some studies have proposed leveraging relevant documents to guide a models response which led to the domain of retrieval augmented generation (RAG) (Lewis et al., 2020). Specifically, RAG uses external knowledge to improve the fidelity of LLMs. However, leveraging the model's internal knowledge and accessing reasoning patterns during inference reasoning could be a potential approach to understanding

models' reasoning behavior and reducing hallucination. The internal analysis for LLMs including analysis on the attention score, activations and transformer architecture. In (Sun et al., 2024; Barbero et al., 2025), they analysis the self-attention mechanism for the LM decoding process, and deeply analysis hight attentions score tokens based on information flow. They also finds that the later layer of transformer tends to have smoother attention score for high-level understanding. figurfigure12.

Our method attends to finding the most attended thoughts during the LM reasoning process. We analysis DeepSeek-R1 think tokens on both token level and segment level. The token level focus on each reasoning token while the segment level combines tokens to a fixed window size. From the token segments with different attention scores, we could better understanding the black box reasoning process for building more trustable LMs.

2 Related Work

Evaluation Reasoning for Large language models (LLMs) can be categorized into four main groups:Conclusion-Based Evaluation, Rationale-Based Evaluation, Interactive Evaluation, and Mechanistic Evaluation (Mondorf and Plank, 2024)

Each of these approaches evaluates reasoning differently. Conclusion-Based evaluation focuses only on the final output without considering intermediate steps (Wu et al., 2023). In contrast, Rationale-Based evaluation guides models to generate structured reasoning and then assesses it either manually using human annotators or using automated metrics such as ROSCORE and RECEVAL (Dziri et al., 2023)

Interactive evaluation interacts with LLMs in real-time using frameworks to dynamically analyze responses (Zhuang et al., 2023). Lastly, the evaluation method most similar to our goal is Mechanistic evaluation, which evaluates the underlying processes that drive model responses by analyzing attention patterns or model parameters (Hou et al., 2023)

A recent work (Hou et al., 2023) evaluated model reasoning capabilities by using LLM attention on the generated output to extract reasoning trees, which were then compared to ground truth reasoning trees. They evaluated their method on two subsets of benchmarks ProofWriter and the AI2 Reasoning Challenge—after modifying the data to align with their setup. They also used additional synthetic data and tested the approach on two LLMs: GPT-3 and LLaMA-7B.

There are also many recent works on understanding how attention includes the LLMs, specific to attention heads. In (Zheng et al., 2024), they organized four types of attention heads (between LLM and human) for reasoning LLMs with details about transformer decoder-only architectures. It organizes the different categories of attention heads during the intermediate reasoning inference related to human thinking. A more detailed analysis of internal layers of attention heads is in the work of (Wiegreffe et al.), which can focus decision-making on a few middle-layer heads and find the specialized heads for selecting the answer. On the other hand, some works are trying to find dormant attention heads, (Sandoval-Segura et al., 2025) tries to identify and evaluate the redundant attention head based on attention sink and testing if the dormant heads have influence during inference.

Their findings highlight a performance gap between reasoning-based evaluation and human evaluation. However, their method only analyzes LLM attention at the output level and requires manual labeling for both the data and reasoning tree generation. However, our goal is to explore whether there is a correlation between the model's final output and its internal reasoning process.

3 Methodology and Contributions

In this work, we aim to investigate the reasoning mechanisms employed by large reasoning models (LRMs) such as DeepSeek-R1, which utilize chain-of-thought prompting to generate intermediate reasoning steps before providing a final answer. To address this, we introduce our attention-based analysis for extracting and formalizing the implicit reasoning pathways that a large reasoning model (LRM) follows when it generates a chain of thought and a final answer. We begin by defining the relevant token categories and the notation for attention weights, and then describe two complementary levels of analysis—token-level and segment-level—before showing how these feed into the construction of a hierarchical *reasoning tree*.

3.1 Problem Formulation

Let a user's query be tokenized into a sequence of *context tokens*

$$\mathcal{C} = (c_1, c_2, \ldots, c_N),$$

which the model reads before generating any output. During generation, the model emits a sequence of *thinking tokens*

$$\mathcal{T} = (t_1, t_2, \ldots, t_M),$$

delimited by special markers <think> and </think>, followed by a sequence of *answer tokens*

$$\mathcal{A} = (a_1, a_2, \ldots, a_P),$$

in which the final answer is explicitly indicated in the form

$$\texttt{Answer: } \langle f \rangle,$$

where f is a single token (e.g., a selected multiple-choice label or a numerical result).

We refer to the last generated answer token $\langle f \rangle$ as the *final answer token*. Our goal is to trace back from $\langle f \rangle$ through the model's own attention patterns to identify which thinking tokens (or groups of tokens) were most instrumental in producing f.

3.2 Attention Notation

Let $\langle f \rangle$ denote the final answer token in the output sequence. We inspect the model's self-attention in the last transformer layer, which has H heads. For any two tokens u and v, let

$$\alpha^{(h)}_{u \to v}$$

be the attention weight from token u to token v in head h. Since we wish to quantify how strongly $\langle f \rangle$ attends to preceding tokens, we focus on

$$\alpha^{(h)}_{f \to t_j} \quad \text{for each } t_j \in \mathcal{T}.$$

We then aggregate across heads to obtain a single scalar score:

$$\alpha_{f \to t_j} = \frac{1}{H} \sum_{h=1}^{H} \alpha^{(h)}_{f \to t_j}.$$

Henceforth, all attention scores refer to $\{\alpha_{f \to t_j}\}_{j=1}^{M}$.

3.3 Token-Level Analysis

At the token level, we select the K thinking tokens that receive the highest attention from the final answer. Formally, we define the index set

$$\mathcal{I}_{\text{token}} = \underset{J \subseteq \{1,\ldots,M\}, |J|=K}{\arg\max} \sum_{j \in J} \alpha_{f \to t_j}.$$

The tokens $\{t_j : j \in \mathcal{I}_{\text{token}}\}$ are thus the most influential single-token supports for the model's decision. Figure 2 illustrates this selection.

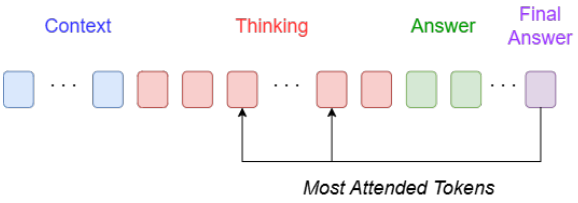

Figure 2: Token-level attention analysis: the final answer token $\langle f \rangle$ attends most strongly to a small set of thinking tokens.

3.4 Segment-Level Analysis

To capture higher-level structure, we partition the thinking sequence $\mathcal{T}$ into $L = \lceil M/w \rceil$ contiguous windows of size w:

$$S_\ell = \{ t_{(\ell-1)w+1}, \ldots, t_{\min(\ell w, M)} \}, \quad \ell = 1, \ldots, L.$$

For each segment S_ℓ, we compute the mean attention from $\langle f \rangle$ across its tokens:

$$\beta_\ell = \frac{1}{|S_\ell|} \sum_{t_j \in S_\ell} \alpha_{f \to t_j}.$$

Ranking segments by β_ℓ then reveals the contiguous reasoning regions that the final answer most heavily leverages. Figure 3 depicts this aggregation.

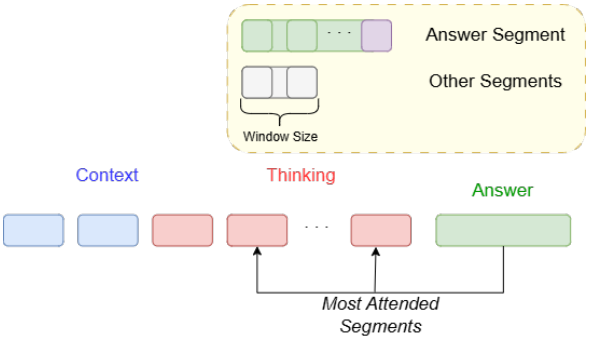

Figure 3: Segment-level attention analysis: thinking tokens are grouped into windows of size w, and their average attention to $\langle f \rangle$ is computed.

3.5 Reasoning Tree Construction

Combining token-level and segment-level results, we define a *reasoning tree* $\mathcal{T}_R$ whose root is the final answer token $\langle f \rangle$. Its first-order children are either the top K tokens $\{t_j\}$ or the top L' segments $\{S_\ell\}$ ranked by β_ℓ. In principle, this procedure may be recursively applied to each selected token or segment—treating it as a new "sub-answer" node—thereby revealing deeper hierarchical dependencies.

This tree succinctly encodes which portions of the model's intermediate chain of thought were most responsible for the final decision. It also enables qualitative and quantitative analyses: (1) *Redundancy detection*, by identifying branches that contribute little attention mass; and (2) *Error localization*, by exposing erroneous reasoning paths when the final answer is incorrect. Together, these insights foster a more interpretable and diagnosable framework for chain-of-thought prompting in LRMs.

4 Results

Although reasoning models often produce a substantial amount of text when generating explanations or justifications, their attention is not evenly distributed across all parts of the reasoning output. Instead, these models tend to focus more heavily on a select few sentences or phrases that are most relevant to the final decision. In Figure 4, we highlight the top five sentences that received the highest attention weights during the model's reasoning process for a representative example.

In this case, the input query provided to the model was "What color is the sky at night? A) Blue B) White C) Yellow D) Black." The model's generated response was "The sky appears black at night because the absence of sunlight allows the stars to be visible, making the sky seem black. Answer: D) Black." The top-weighted sentences extracted from the reasoning output reveal the portions of the explanation that most strongly influenced the model's final answer.

Figure 4: The figure shows the most attended to sentences in the reasoning output of the model given the final answer.

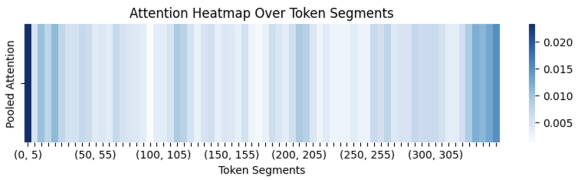

Figure 5: The figure shows the segment-level analysis of the model's attention on reasoning tokens. The model generally pays more attention to the beginning and later tokens.

An analysis of the model's attention distribution across segments of the generated reasoning reveals a consistent pattern in how it processes information during reasoning. Specifically, as shown in Figure 6, the model tends to concentrate its attention on the sentences positioned at the beginning and end of the reasoning sequence. In contrast, the middle segments receive relatively diffused or lower attention.

This pattern was not limited to a single example but appeared consistently across multiple user queries. This tendency may indicate that the model places greater importance on establishing an early interpretive frame and reaffirming or synthesizing key information toward the conclusion, while treating intermediate details as transitional or less influential in the decision-making process.

However, for a smaller subset of user queries, a different attention pattern emerges. In these cases, the model not only focuses on the early and late portions of the reasoning output but also allocates significant attention to specific sentences or phrases located in the middle of the reasoning sequence. This more distributed attention pattern suggests that, depending on the complexity or nature of the query, the model is capable of dynamically adjusting its focus. Rather than strictly adhering to the typical early-late attention concentration, it may recognize and emphasize additional informative content found in the middle of the reasoning chain.

In addition to its logical reasoning patterns, the model often incorporates subtle human-like expressions into its generated thought process. These expressions—such as interjections like "okay" or "hmm"—do not directly contribute to the logical derivation of the final answer but instead introduce a conversational tone that mimics human reasoning behavior. Despite their limited relevance to the factual content of the response, these tokens frequently receive relatively high attention scores

during the model's internal processing.

For example, when presented with the query "Melinda learned that days in some seasons have more daylight hours than in other seasons. Which season receives the most hours of sunlight in the Northern Hemisphere? Choices: A) fall B) spring C) summer D) winter", the model's reasoning output included informal expressions such as "okay," which were not semantically tied to the answer but were nonetheless highly attended to. This indicates that the model may be assigning importance to such tokens, potentially as anchors or transitions in its reasoning flow or just a means to simulate the style and tone of human reasoning, even when such expressions are not essential to the final answer.

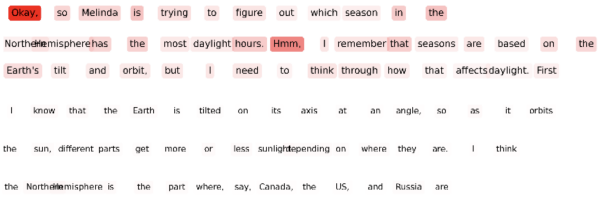

Figure 6: The figure illustrates the distribution of attention over the reasoning tokens in a portion of the model's response.

4.1 Random Number Example

We input DeepSeek-R1 a strict forward question: Give me a random number. The reasoning process is relatively long and somewhat redundant, including multiple steps of complicated mathematical calculations. As shown in Figure 9, we notify there are two highest attention token segments. To simplify, we deeply analyze the first high-attention token, the first reasoning token. The first token is `Okay`, even if it is less semantically informative and more like a conversational word, it draws great attention for certain attention heads on the last layer of the LM. We plot the attention weights on all 32 heads to further analyze this phenomenon. Due to the page limitation, we only show the attention heads with strong first token attention in Figure 7.

We also notice that head 17, which has the greatest attention toward `Okay`, shows a relatively spiky attention pattern across subsequent segmentation tokens. Considering this example's long and redundant reasoning process, the spiky attention mode could be a sign of less informative attention hea. Because " Okay " content is not semantic informative for the final random numerical output and the unsmooth attention across different tokens, head 17 may be regarded as an uninformative head.

In contrast, in Figure 8, we find that head 28 pays less attention to the first token while allocating more attention to the conclusion reasoning tokens. Head 28 serves as an active attention head by emphasizing the output-centric tokens with more semantic meaning.

5 Self-Evaluation and Lessons Learned

In the course of this project, we gained practical expertise in deploying and interrogating state-of-the-art large reasoning models. We configured multi-GPU environments for efficient inference, instrumented transformer architectures to extract intermediate attention weights, and became proficient in parsing and manipulating high-dimensional attention tensors. This hands-on experience proved invaluable for understanding both the engineering and theoretical aspects of chain-of-thought prompting.

Through our attention-based analysis, we discovered a consistent pattern in how answer tokens allocate focus within the generated reasoning trace. Specifically, final answer tokens concentrate most of their attention on the earliest and latest segments of the chain of thought, effectively "bookending" the reasoning sequence. This observation suggests that, while models articulate many intermediate steps, only a subset of those steps—namely the initial premises and concluding deductions—play a decisive role in determining the final output.

Our original proposal envisioned a multi-level reasoning tree, revealing deep hierarchical dependencies within the model's chain of thought. However, empirical inspection of intra-thinking attention revealed that tokens predominantly attend to their immediate predecessors, rather than to distant reasoning steps. As a result, meaningful hierarchies beyond the first level did not materialize in our trees, and we constrained our representation to a single layer of children under the final answer. Going forward, it will be important to explore alternative mechanisms—such as modified prompting strategies or explicit intermediate "summary" tokens—that might induce richer, deeper reasoning hierarchies in large language models.

6 Conclusions and Future Work

This work addresses a key gap in understanding how large reasoning models (LRMs) use their self-generated reasoning steps to arrive at final answers.

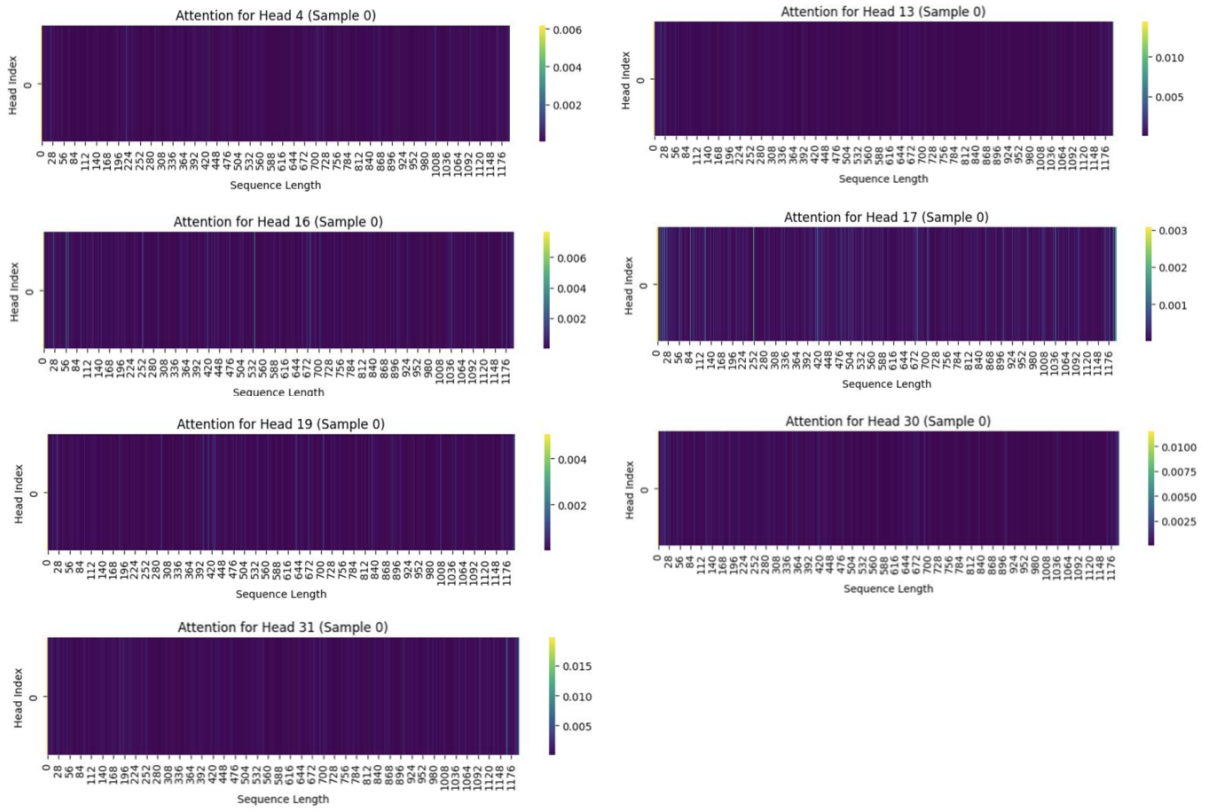

Figure 7: Heads with hight attention on the first token

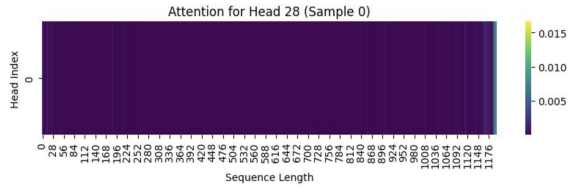

Figure 8: Heads with low attention on the first token

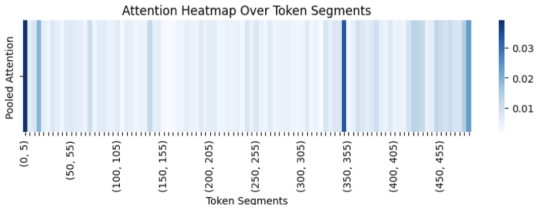

Figure 9: Random number example attention

While chain-of-thought prompting has become a powerful tool for improving model performance, the internal mechanisms by which models process and leverage these intermediate steps remain poorly understood.

Through attention-based analysis of DeepSeek-R1, we uncovered a consistent "bookending" pattern in the model's reasoning: final answer tokens tend to focus heavily on the beginning and end of the reasoning trace, with middle segments receiving significantly less attention. This suggests that models may rely on early framing and final summarization while treating intermediate reasoning as transitional. However, in certain cases, we observed elevated attention to specific mid-sequence segments, indicating that the model can adapt its focus depending on task complexity. We also found that some non-semantic tokens—such as the first reasoning token or stylistic interjections like "okay" and "hmm"—often receive unexpectedly high attention.

Looking ahead, we identify several directions for future work. First, we aim to extend our reasoning trees beyond a single level by analyzing intra-thinking attention patterns, potentially uncovering deeper hierarchical reasoning structures. Second, we plan to examine the model's attention distribution across earlier transformer layers to understand how reasoning tokens evolve and interact through-

out the network. Finally, we will investigate the linguistic and functional roles of mid-sequence segments that occasionally attract high attention, with the goal of identifying what makes them influential. Together, these efforts aim to shed further light on the internal decision-making processes of large language models and support the development of more interpretable and controllable reasoning systems.

References

Federico Barbero, Álvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Razvan Pascanu, and 1 others. 2025. Why do llms attend to the first token? *arXiv preprint arXiv:2504.02732*.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang (Lorraine) Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. 2023. Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36:70293–70332.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Yinhao Hou, Jiwei Li, Yujie Fei, and 1 others. 2023. Towards a mechanistic interpretation of multi-step reasoning capabilities of language models. *arXiv preprint arXiv:2310.14491*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Philipp Mondorf and Barbara Plank. 2024. Beyond accuracy: Evaluating the reasoning behavior of large language models–a survey. *arXiv preprint arXiv:2404.01869*.

OpenAI. 2023. Chatgpt: A large-scale language model for conversational ai. Accessed: 2025-03-06.

Leyan Pan, Vijay Ganesh, Jacob Abernethy, Chris Esposo, and Wenke Lee. 2024. Can transformers reason logically? a study in sat solving. *arXiv preprint arXiv:2410.07432*.

Pedro Sandoval-Segura, Xijun Wang, Ashwinee Panda, Micah Goldblum, Ronen Basri, Tom Goldstein, and David Jacobs. 2025. Using attention sinks to identify and evaluate dormant heads in pretrained llms. *arXiv preprint arXiv:2504.03889*.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. 2024. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*.

Sarah Wiegreffe, Oyvind Tafjord, Yonatan Belinkov, Hannaneh Hajishirzi, and Ashish Sabharwal. Answer, assemble, ace: Understanding how lms answer multiple choice questions.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyurek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2023. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *arXiv preprint arXiv:2307.02477*.

Junbing Yan, Chengyu Wang, Jun Huang, and Wei Zhang. 2024. Do large language models understand logic or just mimick context? *arXiv preprint arXiv:2402.12091*.

Honghua Zhang, Liunian Harold Li, Tao Meng, Kai-Wei Chang, and Guy Van den Broeck. 2022. On the paradox of learning to reason from data. *arXiv preprint arXiv:2205.11502*.

Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Mingchuan Yang, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2024. Attention heads of large language models: A survey. *arXiv preprint arXiv:2409.03752*.

Yan Zhuang, Qi Liu, Yuting Ning, Weizhe Huang, Rui Lv, Zhenya Huang, Guanhao Zhao, Zheng Zhang, Qingyang Mao, Shijin Wang, and 1 others. 2023. Efficiently measuring the cognitive ability of llms: An adaptive testing perspective. *arXiv preprint arXiv:2306.10512*.