

How Do LLMs Cite?

A Mechanistic Interpretation of Attribution in RAG

Ian van Dort¹[0009-0008-5881-760X] and Maria Heuss¹[0000-0001-5360-9627]

University of Amsterdam
ian.van.dort@student.uva.nl, m.c.heuss@uva.nl

Abstract. With the rising popularity of large language models (LLMs) for various applications ranging from every-day information seeking to high stakes applications such as complex medical question answering, hallucinated content poses a serious risk. Retrieval-Augmented Generation (RAG) aims to enhance the trustworthiness of LLMs by grounding their outputs in external documents, often using inline citations for verifiability. However if those citations are not faithful, i.e. if they do not accurately reflect the source of the information during the answer generation, user trust might be misguided by the sheer existence of such references to trusted sources. We argue that to understand citation faithfulness and to develop a reliable framework for the evaluation of citation faithfulness, a mechanistic approach that considers the model internals, rather than mere observations on the input/output of the model, is necessary. This paper offers the first mechanistic account of how a large language model decides whether to attach an inline citation while answering a factoid question. Through activation patching we identify an “attributional ensemble” of attention heads and MLP layers that are responsible for the citation generation process in Llama-3.1-8B-Instruct. Our findings suggest that citation decisions rely heavily on shallow heuristics such as entity co-reference matching, raising concerns about the trustworthiness of such citations.

Keywords: Retrieval-augmented Generation · Mechanistic Interpretability · Attributions · Large Language Models · Faithfulness.

1 Introduction

Large Language Models (LLMs) are increasingly deployed in high-stakes domains, from medical question answering [34] to legal research [27] and educational tools [11]. These systems promise to democratize access to specialized knowledge that was previously available only to experts or those with substantial resources. However, LLMs are prone to “hallucinations”: plausible but factually incorrect outputs [1]. When a lawyer relies on a fabricated case citation [19] or a clinician acts on incorrect medical information [21], the consequences can be severe.

Retrieval-Augmented Generation (RAG) addresses this by grounding model outputs in external documents and providing citations. The appeal is straightforward: users can verify claims against sources. Yet this verification promise matters most for communities that rely on AI as a primary information source but lack expertise to independently assess the underlying documents. If citations are merely cosmetic—appended after the fact rather than genuinely driving the answer—then RAG systems risk creating a false sense of security that could deepen information disparities rather than close them [15].

The problem is that citation correctness (does the source factually support the claim?) does not guarantee citation faithfulness (did the model actually use the source to generate the claim?). Consider a medical RAG system asked about drug interactions. If the model recalls the answer from training data and then attaches a citation to a document that happens to confirm it, the citation is correct but unfaithful. Worse, research shows that explanations and citations, even misleading ones, increase user trust in AI outputs [26, 6]. In dense, technical domains where users cannot easily verify sources themselves [17, 12], unfaithful citations lend misinformation an air of credibility.

Citation Faithfulness as a Property of Trustworthy Attribution. We define citation faithfulness as a citation occurring only when the model’s internal computation genuinely used the cited document to construct the answer [31]. The challenge is that faithfulness is not observable from input-output behavior alone. A model can produce identical text whether it retrieved information from a document or hallucinated from memory. To assess faithfulness, we must look inside the model.

Towards Understanding Citations with Mechanistic Interpretability. This paper uses mechanistic interpretability to reverse-engineer how Llama-3.1-8B-Instruct decides whether to emit an inline citation in a RAG context. Instead of treating the model as a black box, we use activation patching to identify which internal components (specific attention heads and MLP layers) influence citation behavior. We then test whether these components can be manipulated: inducing citations and suppressing spurious ones.

Why does this matter? If citations are generated by shallow heuristics (e.g., matching entity names between document and question) rather than through genuine engagement with document content during answer construction, then current RAG systems offer only an illusion of verifiability. Understanding the mechanism is the first step toward building evaluation frameworks that can distinguish faithful from unfaithful attribution—critical for deploying AI systems that communities with limited technical resources can actually trust.

The main contributions of this work are:

1. Mapping the distributed, multi-stage “attributional ensemble”, a collection of specific attention heads and MLP layers, that collaboratively implements the citation mechanism.
2. Demonstrating the role of this ensemble by performing targeted interventions that successfully correct citation failures, such as inducing missing citations or suppressing spurious ones.

3. Validating the generalizability of these findings by applying the same interventions in a more complex, multi-document question-answering setting and achieving a significant improvement in citation accuracy.

2 Related Work

Large language models (LLMs) often “hallucinate”, producing factually incorrect information [1]. To attempt to overcome these issues, RAG helps LLMs by retrieving relevant documents from external knowledge bases, and using those documents to draft better outputs [29]. It does this by coupling a retriever, which identifies relevant documents or data, with a generator, typically an LLM, to produce outputs grounded in external knowledge, often providing inline citations to enhance verifiability and user trust [13]. This approach can enhance the accuracy and credibility of the generated output, especially for tasks which require extensive up-to-date or domain-specific knowledge [8]. However, the presence of a citation does not guarantee trustworthiness. A critical, unverified assumption is that these citations are faithful: that is, the cited document had a genuine causal influence on the generation of the corresponding statement [31, 10]. A model might generate an answer from its parametric memory and append a citation that is merely plausible or correct, but not causally responsible, creating a misleading facade of credibility. This distinction is vital in high-stakes domains like law and medicine, where unfaithful citations could lead to misplaced trust in incorrect information [17, 12]. In this work we build on the definition of citation faithfulness from [31], but take a step towards approaching the problem by understanding how the model generates citations.

2.1 Approaches to Evaluating and Performing Attribution

Research on RAG attribution has primarily followed three paths.

External, Output-Based Methods: The most common approach uses external models to evaluate the relationship between the generated text and the source after generation is complete. These methods often frame the task as Natural Language Inference (NLI), checking if the source text (premise) entails the generated claim (hypothesis) [2, 22]. While useful for measuring citation correctness, these post-hoc methods cannot assess faithfulness, as they have no access to the model’s internal generation process. They answer “Could this source support the claim?” not “Did this source cause the claim?”.

Generation-Integrated Self-Citation: To streamline attribution, models can be prompted or fine-tuned to generate citations inline with the answer [7]. This approach is used by many state-of-the-art RAG systems and is the focus of our investigation. While this integrates attribution into the generation process, it does not guarantee faithfulness. As prior work on self-explanations has shown, a model’s stated rationale for its output does not necessarily align with its true computational process [16, 30].

Mechanistic and Perturbation-Based Methods: A smaller body of work has sought to establish a more causal link between source and output by analyzing model internals. Techniques like gradient-based saliency [25], Jensen-Shannon Divergence [14], hidden-state analysis [24] and perturbation analyses [4, 28] identify which parts of the context most influence the final output. These methods offer a stronger claim to faithfulness but are often computationally expensive, require full model access, rely on coarse sentence-level ablations that can miss nuanced influences, and have not yet been used to map the full, end-to-end mechanism of how a citation decision is made. A closely related mechanistic approach is FACTUM [5], which targets *citation hallucination detection* in long-form RAG by decomposing transformer computation into attention (context reading) and FFN (parametric recall) pathways at the moment a citation token is generated. FACTUM proposes pathway-specific and interaction scores (e.g., contextual alignment over source tokens, attention-sink usage, FFN update magnitude, and attention-FFN alignment) and shows that their combined “mechanistic signature” can predict whether a citation is attested, with signatures that vary by model scale. While this provides strong evidence that citation reliability corresponds to measurable internal pathway dynamics, it is primarily a diagnostic/detection framework rather than a full end-to-end account of how citation decisions are computed, leaving open the question of the specific computational procedure that produces an inline citation during generation.

3 Methodology

We designed a minimal, controlled experimental pipeline to study when and why a RAG-style model emits inline citations. To avoid confounds from a full retrieval stack, each example is run twice with the same prompt: once with a relevant supporting document and once with an irrelevant (distractor) document. Comparing these two conditions isolates behaviour driven by the supplied context.

Below we summarise the dataset, model, prompt selection, and the component-level evaluation used to locate and validate mechanisms that correlate with citation behaviour.

3.1 Dataset and model

We base our experiments on a citation-focused adaptation of PopQA [18]. PopQA’s short, single-fact questions and clear answers make automated correctness checks straightforward. For each question we generate (i) a short supporting document that contains the answer and (ii) a distractor document. Distractors were created by randomly sampling structure-matched factual sentences that mirror subject/object token lengths and relation type. This preserves token-level structure, enabling precise component-level comparisons.

All experiments use the Llama-3.1-8B-Instruct model, chosen because it reliably follows citation-style instructions and is supported by TransformerLens [9, 23].

3.2 Prompt design

Prompt wording affects both answer correctness and citation propensity. We therefore selected a short, zero-shot instruction that (a) asks the model to add an inline citation marker (e.g. "[1]") only when a claim is supported by the provided document, and (b) asks the model to answer from prior knowledge without a citation when the document is irrelevant. An example of the final prompt structure we use is:

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>
Cutting Knowledge Date: December 2023
Today Date: 26 Jul 2024

When answering the question: 1) If the document contains
relevant information, use it and cite it with [1]. 2) If
the document isn't relevant, answer based on your
knowledge without citation. 3) Never cite information you
didn't get from the document. 4) Citations should appear
immediately after the specific information from the
document.<|eot_id|><|start_header_id|>user<|end_header_id|
|>

Document 1: Kampala is the capital of Uganda.

Question: What is the capital of Uganda?<|eot_id|><|
start_header_id|>assistant<|end_header_id|>

The capital of Uganda is
```

3.3 Component-level evaluation via activation patching

To locate internal components associated with citation behaviour we perform activation patching between the relevant and distractor runs. Activation patching replaces selected intermediate activations in one run with activations from the other and measures the resulting change in the model's preference for starting a citation. We operationalise this preference as the logit difference between the citation-initiating token (the token for the character "[") and the sentence-final token ".".

To reduce variance, the analysis is restricted to a homogeneous subset of examples that share a single structural template and where the clean (relevant) and corrupted (distractor) runs produce clear, opposite top-token preferences at the final step. We report a normalized logit-difference score that measures the fraction of behavioural change recovered (or induced) by a patch; the formula is given in Equation 1.

$$\text{Score} = \frac{\text{Patched LogitDiff} - \text{Corrupted LogitDiff}}{\text{Clean LogitDiff} - \text{Corrupted LogitDiff}} \quad (1)$$

We systematically test three locations: full residual replacement, attention output, and MLP output. Token-position-level patching is limited to the portion of the prompt where the two runs diverge to avoid redundant edits of identical context.

3.4 Evaluation of identified components

We select important components based on the average normalized logit difference score for that component when patching. Identified components are validated in three ways: targeted interventions on failure cases to increase or decrease a component’s contribution (missed or spurious citations); a unified intervention applied across the full PopQA set to test generalisation; and testing the same components in a multi-document QA task (HotpotQA) where the citation metric demands the correct set of supporting documents [35]. The code for dataset creation and all experiments is available at: <https://github.com/ivdort/thesis-RAG-MI>.

4 Results and Discussion

We present results in four steps. First, we analyze residual-stream patching in denoising and noising settings to localize influential layers and token positions. Second, we decompose effects by component (attention vs. MLP). Third, we validate and steer the identified components via targeted interventions on PopQA. Finally, we test generalization to the multi-document HotpotQA setting. Experimental details for patching are given in Section 3.3, and the evaluation protocol is in Section 3.4. Effects throughout are quantified using the normalized logit-difference score (Equation 1).

For clarity, we show denoising where it highlights components that positively contribute to citation (clean into corrupted), and we show noising where it highlights components that are necessary (corrupted into clean). In practice both modalities tell a consistent story, though noising typically produces larger-magnitude effects due to the mechanism’s conjunctive dependencies.

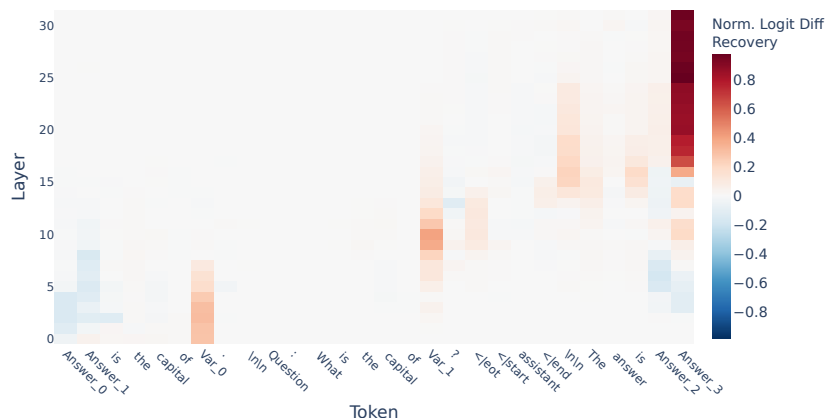


Fig. 1: Residual stream (denoising).

Core Structural Pattern: A Distributed Attributional Ensemble. We find that the decision to cite is governed not by a single “citation head” but by a distributed and fragile **attributional ensemble** of many attention heads and MLPs, where “fragile” means that small corruptions at many sites strongly reduce citation, while single clean patches rarely fully induce it. As shown in Figures 1, 2 and 3, these components’ contributions accumulate across model depth in a process that is easily disrupted (via noising) but difficult to induce (via denoising), consistent with an AND-like dependency structure where multiple conditions must be met.

Early Layers: Entity Enrichment and Duplication Matching. The model first identifies a correspondence between the document and the question using a multi-step heuristic. An early MLP (layer 0) performs an “entity enrichment” step on the queried entity token where it first appears in the document—for example, on ‘Uganda’ in “Document 1: Kampala is the capital of **Uganda**.” (this token is Var_0 in Figure 2). This enriched representation is then passed to the second occurrence of the same entity in the question, “What is the capital of **Uganda**?” (Var_1). The significance of this pathway is striking: while the text of the question is identical in both clean (relevant document) and corrupted (irrelevant document) runs, patching the internal activation at Var_1 is sufficient to restore or break the citation behavior. This suggests the activation carries an internal flag for cross-context identity, effectively “marking” the entity as grounded in the provided document. This mechanism, similar to co-reference heuristics found in prior work [32], creates a durable, enriched representation that is then propagated forward to influence the final citation decision, as seen in the information path in Figure 1.

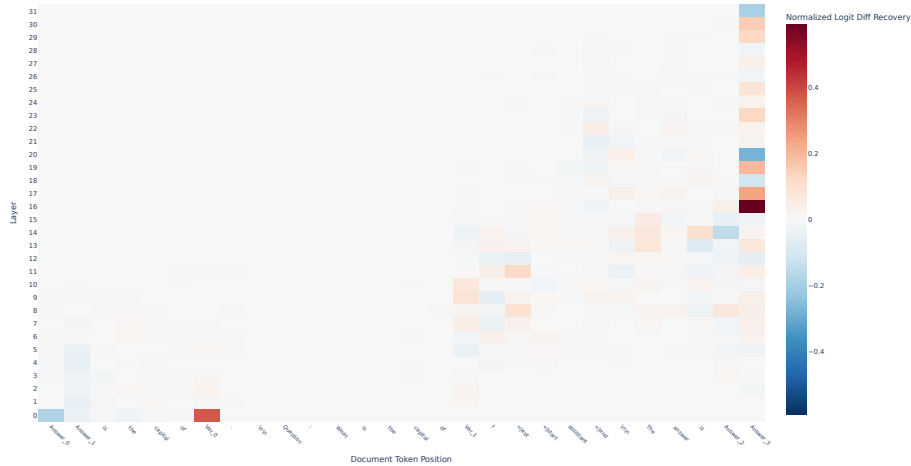


Fig. 2: MLP output denoising, token/layer regions whose clean MLP activations restore citation.

Mid-Layers: Processing Hub (Layers 10–20). A dense band of attention heads and MLPs in middle layers contribute moderate, distributed gains (denoising) or large degradations (noising). Certain heads (e.g., a representative necessary head such as L15H8, highlighted in Figure 3) appear individually critical: corrupting them removes >60% of citation behavior.¹ Yet no single head is sufficient to induce high recovery alone (typical positive denoising contributions 10–30%). We hypothesize that these middle-layer components implement conjunctive checks whose effects accumulate, but we note that activation patching localizes nodes rather than directed edges; validating pathways will require path patching.

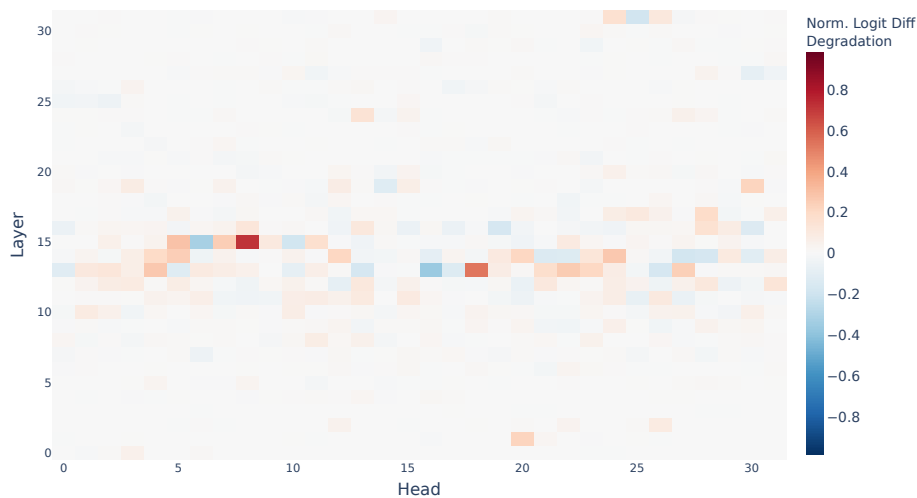
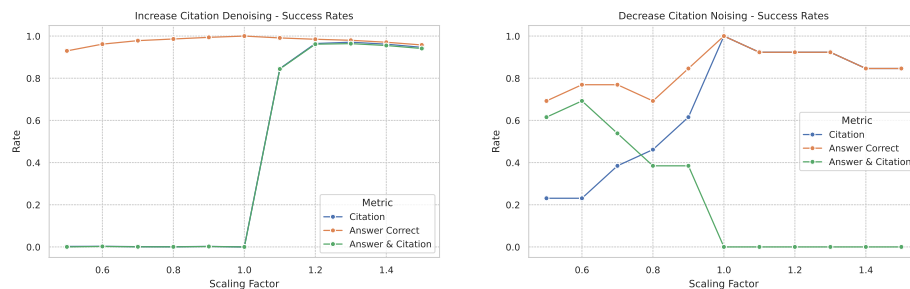


Fig. 3: Attention head noising, individual attention heads whose corrupted outputs most strongly disrupt citation.

Late Decision Assembly. The final answer token position aggregates upstream signals, visible as a cluster of effects at the final answer tokens in Figures 1 and 2. Late-layer MLPs (e.g., layer 16) and selective attention heads (including a decisive late head such as L24H13) increase the logit of the citation marker over a competing termination token (‘.’). Due to the logit-difference metric’s binary framing, we cannot always disambiguate whether components promote citation or suppress sentence termination, an inherent interpretive limitation.

Mechanistic Summary. Inline citation emerges from: (1) early entity enrichment and duplication marking; (2) propagation through a mid-layer processing hub;

¹ Throughout this paper, we use the notation LxHy to denote attention head y at layer x in the Llama-3.1-8B-Instruct model, which has 32 layers (indexed 0–31) and 32 attention heads per layer (indexed 0–31). For example, L15H8 refers to head 8 at layer 15.



(a) Inducing missed citations (threshold 0.10). (b) Suppressing spurious citations (threshold 0.20).

Fig. 5: Targeted PopQA repairs via scaling identified components.

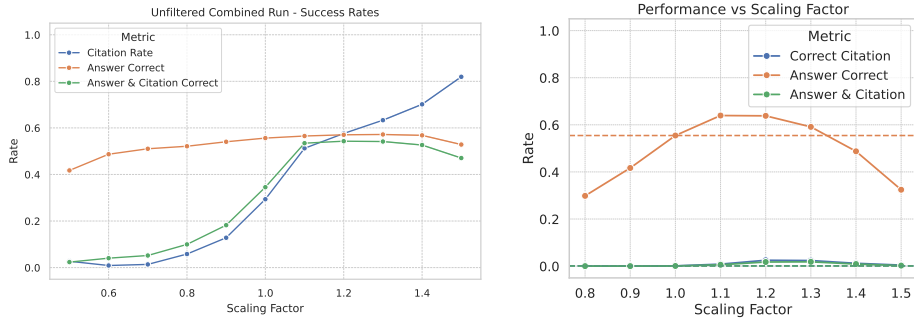
distribution-level shifts. Scaling upward ($\alpha > 1.0$) increased correct citation rates in relevant-document runs while slightly elevating spurious citations in irrelevant-document runs; scaling downward reduced both. A shallow optimum emerged at $\alpha = 1.1$, maximizing a combined metric (Figure 6a). This indicates the ensemble encodes a global, not merely instance-specific, control dimension over citation policy.

4.3 Generalization to Multi-Document Reasoning (HotpotQA)

We evaluated the same component set (no retuning) on 6,808 HotpotQA examples (multi-hop, multi-document). Baseline strict “correct citation” rate (all and only required documents) was 0.001. Amplifying with $\alpha = 1.2$ increased this more than 20-fold to 0.024 (Figure 6b). The joint (answer+correct citation) rate improved from 0.00059 (4/6808) to 0.01821 (124/6808) at $\alpha = 1.3$, a 31 \times relative gain though still low in absolute terms. Answer-string correctness also rose (0.55 \rightarrow 0.64), and partial successes (some but not all required documents) increased while over-citation remained limited. Formatting near-misses (e.g., “from Document 1”) suggest undercounting.

4.4 Implications

Apparent Reasoning vs. Internal Mechanism. A key finding is the potential disconnect between the model’s output, which may appear to be the result of careful reasoning, and its internal computational pathway. An observer seeing a factually correct answer followed by a citation might assume a logical sequence: the model found the fact in the source, generated the answer from it, and therefore cited it. Our work mechanistically demonstrates that this is not necessarily the case. The citation can be generated by a separate, heuristic-based process that runs in parallel to answer generation. This suggests that LLM outputs can create a misleading narrative of their own reasoning, a critical insight for assessing the



(a) Distribution-level PopQA intervention: combined success metric (answer correct + cite only when relevant). Peak at modest amplification ($\alpha = 1.1$). (b) HotpotQA generalization: scaling identified components increases strict correct citation and joint correctness despite low absolute ceiling.

Fig. 6: Generalization of identified components from PopQA to HotpotQA. Left: PopQA distribution-level effect; right: multi-document HotpotQA transfer.

reliability of AI systems. A model can generate the *appearance* of faithfulness while its internal mechanics follow a different logic, a phenomenon also observed in previous work on chain-of-thought reasoning [3, 30]. Auditing the final output is not enough; we must audit the computational pathway that produced it.

RAG and Trustworthy Attribution. The promise of RAG systems, particularly in high-stakes domains, hinges on their ability to provide verifiably grounded answers. Our investigation into the mechanics of inline citation reveals insights into the trustworthiness of this attribution method by examining the model’s internal mechanisms against a strict definition of faithfulness.

A truly faithful citation would be mechanistically dependent on the model using the document to construct the answer. We would therefore hypothesize that the model’s representation of the answer itself, at the moment of generation (T_{ans}), would be a critical nexus in the citation circuit. Our findings challenge this hypothesis. Activation patching revealed that while many components contribute to the final decision, the representation at the T_{ans} token was not one of the most decisive drivers. This suggests the model’s decision to cite may be driven by a sophisticated heuristic rather than being a direct result of the answer-generation process itself. This raises significant questions about whether prompted citations in their current form can truly deliver the level of verifiable grounding that RAG systems promise.

4.5 Theory of Change

RAG systems promise unprecedented capabilities in making complex information more accessible, through features like simplified language and conversational explanations. Their use in applications of societal decision-making processes, ranging from healthcare [34] and legal systems [27] to education [11], seemingly enables people that might traditionally have had little access to specialized information, to inform themselves on these topics.

Yet, for AI systems to truly serve all of society, not just privileged groups with technical resources, we must ensure they are trustworthy and contain accurate information. Citations promise a way to decide whether or not to trust the generated information with the possibility to verify it. Yet, unfaithful citations may disproportionately impact communities that rely on AI systems as primary information sources but lack the resources or expertise to independently verify citations. This dynamic could deepen existing information disparities. Moreover, misleading citations can actually enhance user trust in incorrect AI responses, as explanations, even unfaithful ones, tend to increase perceived credibility [26], creating a concerning form of AI-enabled misinformation that may be particularly difficult to detect and correct.

With this work we aim to take one step towards understanding and evaluating the citation generation in large language model, which we argue is essential for building trustworthy RAG systems. We see this work as a first stepping stone to an understanding of the internal process that leads to generating citations, which we hope will eventually lead to an evaluation framework that helps to identify trustworthy citation behavior.

Future work will need to investigate how mechanistic insights into the model predictions can be used to evaluate citation faithfulness in practice. Critics of mechanistic interpretability note that beyond scientific curiosity, not much practical impact has been derived from insights in the field. Another common criticism of explainable AI, which holds for our work as well is that techniques and findings are often built by, as well as, for engineers, without much consideration for the end user [20]. For our findings to have real world impact, next steps should investigate how the knowledge about the model components that are responsible for generating citations can be used to infer citation faithfulness.

Another limitation of our work is the open question of generalizability of our findings to other datasets, models and scenarios. Only if the insights of our work generalize, will it be possible to infer a reliable evaluation framework in the ever changing landscape of models and applications.

Lastly we want to highlight the possible negative externalities of misuse of our research in an adversarial setting. The knowledge about what input triggers a citation or which model components need to be activated to generate a citation, might be used to insert unfaithful citations that promote certain statements. This misuse of gained insight into model behavior unfortunately is an intrinsic downside of any interpretability research [33], and we hope that advantages will outweigh those risks.

5 Conclusion

In this study, we used mechanistic interpretability to reverse-engineer how Llama-3.1-8B-Instruct decides whether to emit inline citations in a RAG context. Through systematic activation patching, we identified internal components which causally drive citation behavior and tested whether manipulating them could correct citation failures. We found that citation generation is implemented by a distributed “attributional ensemble” of attention heads and MLPs that operates through conjunctive checks rather than a single specialized module. This ensemble relies heavily on shallow heuristics, particularly entity co-reference matching between document and question, rather than deep semantic integration during answer construction.

We argue that this suggests that inline citations can create an illusion of verifiability even when answers are not genuinely grounded in sources. This work can be seen as a first step towards making these failure modes mechanistically legible and toward an evaluation frameworks that can distinguish faithful from unfaithful attribution.

Our study has important limitations: we analyzed a single 8B model in a controlled, single-document setting; activation patching localizes components but not directed information pathways; and we did not directly map the answer-generation circuit to test mechanistic overlap with citation generation.

Future work should refine the ensemble into a complete directed circuit via path patching, explicitly map and compare the answer-generation and citation circuits to operationalize mechanistic faithfulness, and replicate across model architectures, scales, and RAG-finetuned systems to assess whether these mechanisms are universal or model-specific.

Data and code. To facilitate reproducibility, code and parameters are available at <https://github.com/ivdort/thesis-RAG-MI>.

Acknowledgments. This research was supported by the Dutch Research Council (NWO), under project numbers 024.004.022. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Bibliography

- [1] Ahmad, M.A., Yaramis, I., Roy, T.D.: Creating Trustworthy LLMs: Dealing with Hallucinations in Healthcare AI (Sep 2023). <https://doi.org/10.48550/arXiv.2311.01463>, <http://arxiv.org/abs/2311.01463>, arXiv:2311.01463 [cs]
- [2] Bohnet, B., Tran, V.Q., Verga, P., Aharoni, R., Andor, D., Soares, L.B., Ciaramita, M., Eisenstein, J., Ganchev, K., Herzig, J., Hui, K., Kwiatkowski, T., Ma, J., Ni, J., Saralegui, L.S., Schuster, T., Cohen, W.W., Collins, M., Das, D., Metzler, D., Petrov, S., Webster, K.: Attributed Question Answering: Evaluation and Modeling for Attributed Large Language Models (2022). <https://doi.org/10.48550/ARXIV.2212.08037>, <https://arxiv.org/abs/2212.08037>, publisher: arXiv Version Number: 2
- [3] 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., Perez, E.: Reasoning Models Don't Always Say What They Think (May 2025). <https://doi.org/10.48550/arXiv.2505.05410>, <http://arxiv.org/abs/2505.05410>, arXiv:2505.05410 [cs]
- [4] Cohen-Wang, B., Shah, H., Georgiev, K., Madry, A.: Contextcite: Attributing model generation to context. *Advances in Neural Information Processing Systems* **37**, 95764–95807 (Sep 2024). <https://doi.org/10.48550/arXiv.2409.00729>, https://proceedings.neurips.cc/paper_files/paper/2024/hash/adbea136219b64db96a9941e4249a857-Abstract-Conference.html
- [5] Dassen, M., Kotula, R., Murray, K., Yates, A., Lawrie, D., Kayi, E., Mayfield, J., Duh, K.: FACTUM: Mechanistic detection of citation hallucination in long-form RAG. In: *Advances in Information Retrieval. 48th European Conference on Information Retrieval, ECIR 2026, Proceedings. Lecture Notes in Computer Science, Springer* (2026), <https://arxiv.org/html/2601.05866v1>
- [6] Ding, Y., Facciani, M., Joyce, E., Poudel, A., Bhattacharya, S., Veeramani, B., Aguinaga, S., Weninger, T.: Citations and trust in llm generated responses. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 39, pp. 23787–23795 (2025). <https://doi.org/https://doi.org/10.1609/aaai.v39i22.34550>, <https://ojs.aaai.org/index.php/AAAI/article/view/34550>
- [7] Gao, T., Yen, H., Yu, J., Chen, D.: Enabling Large Language Models to Generate Text with Citations. In: Bouamor, H., Pino, J., Bali, K. (eds.) *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. pp. 6465–6488. Association for Computational Linguistics, Singapore (Dec 2023). <https://doi.org/10.18653/v1/2023.emnlp-main.398>, <https://aclanthology.org/2023.emnlp-main.398/>
- [8] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., Wang, H.: Retrieval-Augmented Generation for Large Language Models: A Survey (Mar 2024). <https://doi.org/10.48550/arXiv.2312.10997>, <http://arxiv.org/abs/2312.10997>, arXiv:2312.10997 version: 5

- [9] Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C.C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E.M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G.L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I.A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., Linde, J.v.d., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K.V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yearly, L., Maaten, L.v.d., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., Oliveira, L.d., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M.K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P.S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R.S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S.S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X.E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z.D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B.D., Paranjape, B., Liu, B.,

Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G.M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damraj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K.H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M.L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M.J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N.P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S.J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S.C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V.S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V.T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., Ma, Z.: The Llama

- 3 Herd of Models (Nov 2024). <https://doi.org/10.48550/arXiv.2407.21783>, <http://arxiv.org/abs/2407.21783>, arXiv:2407.21783 [cs]
- [10] Jacovi, A., Goldberg, Y.: Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4198–4205. Association for Computational Linguistics, Online (Jul 2020). <https://doi.org/10.18653/v1/2020.acl-main.386>, <https://aclanthology.org/2020.acl-main.386/>
- [11] Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., et al.: Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences* **103**, 102274 (2023). <https://doi.org/https://doi.org/10.1016/j.lindif.2023.102274>
- [12] Lee, M., Cimino, J., Zhu, H.R., Sable, C., Shanker, V., Ely, J., Yu, H.: Beyond Information Retrieval—Medical Question Answering. *AMIA Annual Symposium Proceedings* **2006**, 469–473 (2006), <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1839371/>
- [13] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: *Advances in Neural Information Processing Systems*. vol. 33, pp. 9459–9474. Curran Associates, Inc. (2020), <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>
- [14] Li, R., Chen, C., Hu, Y., Gao, Y., Wang, X., Yilmaz, E.: Attributing Response to Context: A Jensen-Shannon Divergence Driven Mechanistic Study of Context Attribution in Retrieval-Augmented Generation (May 2025). <https://doi.org/10.48550/arXiv.2505.16415>, <http://arxiv.org/abs/2505.16415>, arXiv:2505.16415 [cs]
- [15] Liu, N., Zhang, T., Liang, P.: Evaluating verifiability in generative search engines. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 7001–7025. Association for Computational Linguistics, Singapore (Dec 2023). <https://doi.org/10.18653/v1/2023.findings-emnlp.467>, <https://aclanthology.org/2023.findings-emnlp.467/>
- [16] Madsen, A., Chandar, S., Reddy, S.: Are self-explanations from large language models faithful? In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Findings of the Association for Computational Linguistics: ACL 2024. pp. 295–337. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). <https://doi.org/10.18653/v1/2024.findings-acl.19>, <https://aclanthology.org/2024.findings-acl.19/>
- [17] Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C.D., Ho, D.E.: Hallucination-free? assessing the reliability of leading ai legal research tools. *Journal of Empirical Legal Studies* **22**(2), 216–242 (May 2025). <https://doi.org/https://doi.org/10.1111/jels.12413>, <https://onlinelibrary.wiley.com/doi/full/10.1111/jels.12413>

- [18] Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., Hajishirzi, H.: When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 9802–9822. Association for Computational Linguistics, Toronto, Canada (Jul 2023). <https://doi.org/10.18653/v1/2023.acl-long.546>, <https://aclanthology.org/2023.acl-long.546/>
- [19] Merken, S.: New York lawyers sanctioned for using fake ChatGPT cases in legal brief. Reuters (Jun 2023), <https://www.reuters.com/legal/new-york-lawyers-sanctioned-using-fake-chatgpt-cases-legal-brief-2023-06-22/>
- [20] Miller, T., Howe, P., Sonenberg, L.: Explainable AI: beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. *CoRR* **abs/1712.00547** (2017), <http://arxiv.org/abs/1712.00547>
- [21] Modi, N.D., Menz, B.D., Awaty, A.A., Alex, C.A., Logan, J.M., McKinnon, R.A., Rowland, A., Bacchi, S., Gradon, K., Sorich, M.J., Hopkins, A.M.: Assessing the System-Instruction Vulnerabilities of Large Language Models to Malicious Conversion Into Health Disinformation Chatbots. *Annals of Internal Medicine* (Jun 2025). <https://doi.org/10.7326/ANNALS-24-03933>, <https://www.acpjournals.org/doi/10.7326/ANNALS-24-03933>, publisher: American College of Physicians
- [22] Muller, B., Wieting, J., Clark, J., Kwiatkowski, T., Ruder, S., Soares, L., Aharoni, R., Herzig, J., Wang, X.: Evaluating and Modeling Attribution for Cross-Lingual Question Answering. In: Bouamor, H., Pino, J., Bali, K. (eds.) *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. pp. 144–157. Association for Computational Linguistics, Singapore (Dec 2023). <https://doi.org/10.18653/v1/2023.emnlp-main.10>, <https://aclanthology.org/2023.emnlp-main.10/>
- [23] Nanda, N., Bloom, J.: TransformerLens (2022), <https://github.com/TransformerLensOrg/TransformerLens>
- [24] Phukan, A., Somasundaram, S., Saxena, A., Goswami, K., Srinivasan, B.V.: Peering into the mind of language models: An approach for attribution in contextual question answering. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) *Findings of the Association for Computational Linguistics: ACL 2024*. pp. 11481–11495. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). <https://doi.org/10.18653/v1/2024.findings-acl.682>, <https://aclanthology.org/2024.findings-acl.682/>
- [25] Qi, J., Sarti, G., Fernández, R., Bisazza, A.: Model internals-based answer attribution for trustworthy retrieval-augmented generation. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. pp. 6037–6053. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). <https://doi.org/10.18653/v1/2024.emnlp-main.347>, <https://aclanthology.org/2024.emnlp-main.347/>

- [26] Sadeghi, M., Pöttgen, D., Ebel, P., Vogelsang, A.: Explaining the unexplainable: The impact of misleading explanations on trust in unreliable predictions for hardly assessable tasks. In: Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization. p. 36–46. UMAP '24, Association for Computing Machinery, New York, NY, USA (2024). <https://doi.org/10.1145/3627043.3659573>, <https://doi.org/10.1145/3627043.3659573>
- [27] Seabrooke, T., Schneiders, E., Dowthwaite, L., Krook, J., Leesakul, N., Clos, J., Maior, H., Fischer, J.: A survey of lay people’s willingness to generate legal advice using large language models (llms). In: Proceedings of the Second International Symposium on Trustworthy Autonomous Systems. TAS '24, Association for Computing Machinery, New York, NY, USA (2024). <https://doi.org/10.1145/3686038.3686043>, <https://doi.org/10.1145/3686038.3686043>
- [28] Sun, Z., Zang, X., Zheng, K., Song, Y., Xu, J., Zhang, X., Yu, W., Song, Y., Li, H.: ReDeEP: Detecting Hallucination in Retrieval-Augmented Generation via Mechanistic Interpretability (Oct 2024). <https://doi.org/10.48550/arXiv.2410.11414>, <http://arxiv.org/abs/2410.11414>, arXiv:2410.11414 [cs]
- [29] Tonmoy, S.M.T.I., Zaman, S.M.M., Jain, V., Rani, A., Rawte, V., Chadha, A., Das, A.: A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models (Jan 2024). <https://doi.org/10.48550/arXiv.2401.01313>, <http://arxiv.org/abs/2401.01313>, arXiv:2401.01313 [cs]
- [30] Turpin, M., Michael, J., Perez, E., Bowman, S.: Language Models Don’t Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *Advances in Neural Information Processing Systems* **36**, 74952–74965 (Dec 2023), https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html
- [31] Wallat, J., Heuss, M., Rijke, M.d., Anand, A.: Correctness is not faithfulness in retrieval augmented generation attributions. In: Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR). p. 22–32. ICTIR '25, Association for Computing Machinery, New York, NY, USA (2025). <https://doi.org/10.1145/3731120.3744592>, <https://doi.org/10.1145/3731120.3744592>
- [32] Wang, K., Variengien, A., Conmy, A., Shlegeris, B., Steinhardt, J.: Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small (Nov 2022). <https://doi.org/10.48550/arXiv.2211.00593>, <http://arxiv.org/abs/2211.00593>, arXiv:2211.00593 [cs]
- [33] Winninger, T., Addad, B., Kapusta, K.: Using Mechanistic Interpretability to Craft Adversarial Attacks against Large Language Models (May 2025). <https://doi.org/10.48550/arXiv.2503.06269>, <http://arxiv.org/abs/2503.06269>, arXiv:2503.06269 [cs]
- [34] Yang, R., Tan, T.F., Lu, W., Thirunavukarasu, A.J., Ting, D.S.W., Liu, N.: Large language models in health care: Development, applications, and challenges. *Health Care Science* **2**(4), 255–263 (2023). <https://doi.org/10.1002/hcs2.61>, <http://dx.doi.org/10.1002/hcs2.61>

- [35] Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., Manning, C.D.: HotpotQA: A dataset for diverse, explainable multi-hop question answering. In: Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (eds.) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 2369–2380. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018). <https://doi.org/10.18653/v1/D18-1259>, <https://aclanthology.org/D18-1259/>