# From RAGs to rich parameters: Probing how language models utilize external knowledge over parametric information for factual queries

**Anonymous ACL submission**


## Abstract

Retrieval Augmented Generation (RAG) enriches the ability of language models to reason using external context to augment responses for a given user prompt. This approach has risen in popularity due to practical applications in various applications of language models in search, question/answering, and chat-bots. However, the exact nature of how this approach works isn't clearly understood. In this paper, we ***mechanistically*** examine the RAG pipeline to highlight that language models take "shortcut" and have a strong bias towards utilizing only the context information to answer the question, while relying minimally on their parametric memory. We probe this mechanistic behavior in language models with: (i) Causal Mediation Analysis to show that the parametric memory is minimally utilized when answering a question and (ii) Attention Contributions and Knockouts to show that the last token residual stream do not get enriched from the subject token in the question, but gets enriched from other informative tokens in the context. We find this pronounced "shortcut" behaviour true across both LLaMa and Phi family of models.


## 1 Introduction

With the burgeoning use of Language Models (LMs) in many industrial applications, retrieval Augmented Generation (RAG) has become popular as a mechanism of providing additional context for effective *reasoning* to mitigate *hallucinations*. Yet, the usefulness of RAG to provide meaningful information in comparison to model priors is an under-explored area of research. On the other hand, knowledge localization and editing techniques in LMs such as ROME (Meng et al., 2022a) and MEMIT (Meng et al., 2022b) are traditionally focused on adjusting the internal parameters of the LMs to update or correct knowledge. However, a mechanistic understanding of how RAG context influences LM predictions over prior knowledge hasn't been studied till date. And the rise of RAG usage necessitates us to understand quantitatively the interplay between the LM's prior knowledge and the external information retrieved during inference, for preventing drift in model reasoning.

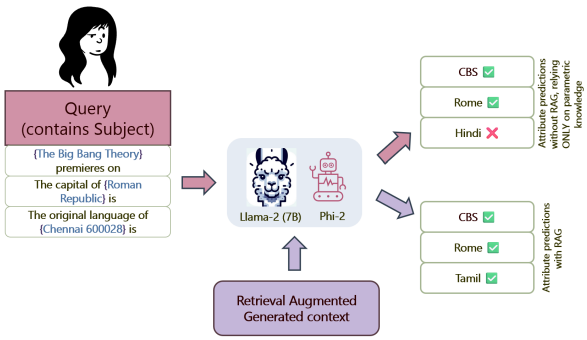



Figure 1: Setup of a factual QA system with RAG, utilized in this paper, for understanding the usefulness of parameteric knowledge stored in LlaMa and Phi.

In this paper, we aim to **analyze** and **interpret** the dependency of LMs on parametric knowledge versus the retrieved information presented via RAG. Towards this goal, we rely on established methods of locating factual knowledge stored in the model parameters.

We find that: (i). Parametric knowledge is minimally used within Multi Layer Perceptrons (MLPs) in the presence of retrieved context. and (ii). The last token residual stream, crucial for forming the final output, derives more enriched information from the attribute token present explicitly in the context rather than from the subject token within the query. These insights highlight a pronounced "shortcut" behavior in LMs, where the models prioritize external context over internal knowledge. Through this analysis, our work contributes to the a novel understanding of the mechanisms underlying LMs' preference for the information provided via RAG.

## 2 Related Work

RAG systems (Lewis et al., 2021) have become popular in practical natural language systems as they significantly improve the performance of LM applications by integrating external context(Shao et al., 2023)(Singh et al., 2023)(IngestAI, 2023) (Chang, 2023)(Ren et al., 2023) However, utilizing RAGs can also have nuanced outcomes such as generation of inconsistent predictions, even with perfect retrieval results(Hagström et al., 2023).(Wu et al., 2024) explore the role of RAG in reducing hallucinations and enhancing accuracy in large language models such as GPT-4, building on prior work(Lewis et al., 2021)(Shuster et al., 2021) that leverage external retrieval systems to mitigate model errors. Even though RAG models are extensively used, and their shortcomings documented, only (Wu et al., 2024) delves into the balance between a model's internal knowledge and externally retrieved information, examining their practical value. However, a systematic **mechanistic** exploration of model's preference for RAG-provided information over their parametric knowledge contribution has not yet been conducted, to the best of our knowledge. Our study mechanistically probes into the internal workings of large language models and how they exhibit a "shortcut mechanism" when they are provided with non-parametric knowledge via a RAG system.

## 3 Probing Mechanisms

To mechanistically interpret the knowledge contributions towards factual reasoning by LLMs and SLMs, we use three methods for causal mediation, described as follows:

### 3.1 Causal Tracing

Causal tracing (Meng et al., 2022a) identifies specific hidden states that significantly influence factual predictions. The approach involves three steps - a clean run, corrupted run and a corrupted-with-restoration run. The corrupted run involves corrupting a certain span of the text, and running the forward pass of the model. In the restoration run, activations from the clean run are patched one by one into the corrupted run, and the increase in answer probability is observed; the most crucial activations are thus causally determined.

Finally, the **Indirect Effect (IE)** of a specific hidden state $h_i^{(l)}$ is defined as the difference between the corrupted run and the corrupted-with-restoration run probabilities: $\text{IE}(h_i^{(l)}) = P^*_{\text{clean}}(h_i^{(l)})[y] - P^*[y]$ and by averaging these effects over a sample, the **Average Indirect Effect (AIE)** is computed for all hidden states, providing a quantitative measure of their importance in factual prediction.

### 3.2 Attention Knockout and Contribution Mechanism

The **Attention Contribution** (Yuksekgonul et al., 2024), focuses on the role of attention mechanisms in shaping the output of language models. This approach investigates how attention weights, particularly from the subject token in a query to the last token position, contribute to the model's predictions. By examining the norm of these attention weights $\|a_{i,T}^{(\ell)}\|$, we observe what tokens the last token pays the most attention to, during the generation process. See appendix D for norm calculation details. The **Attention Knockout** mechanism (Geva et al., 2023) identifies critical attention edges in transformer-based models that are essential for maintaining prediction quality. The process involves identifying critical edges whose removal significantly degrades the model's prediction quality. To test the importance of these edges, attention weights between two positions $r$ and $c$ at a layer $l$ are set to negative infinity: $M_{rc}^{l+1,j} = -\infty \quad \forall j \in [1, H]$

This prevents the source position $x_r^l$ from attending to the target position $x_c^l$, blocking information flow at that layer. The degradation in prediction quality after blocking attention edges identifies which edges are critical for information flow.

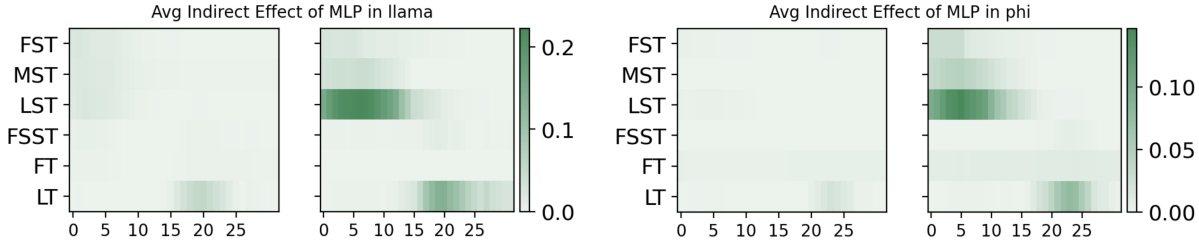
## 4 Datasets and Models

### 4.1 Models

For a comprehensive mechanistic probing, we leverage two state-of-the-art LMs, Phi-2 (2.7B) (Li et al., 2023) and LLaMA-2 (7B) (Touvron et al., 2023) models, which were trained on different corpora. Difference in parametric knowledge between two different family of models, allows us to comprehensively probe the influence of RAG for factual queries in scenarios involving these models. Also choosing open-source LMs enables us measure causal mediation easily.

### 4.2 Dataset

In this paper, we scope the analysis to determine the influence of external information provided by RAG context against model priors, to only factual query predictions from aforementioned LMs. Thus, we utilize the **Knowns Fact Dataset** of 1209 factual queries, introduced in (Meng et al., 2022a). Each record in the dataset is of $(s, r, o)$ format of subject,




(a) LLaMa-2 with RAG vs. LLaMa-2 Vanilla

(b) Phi-2 with RAG vs. Phi-2 Vanilla

Figure 2: **Language models minimally rely on the MLP parametric memory in the presence of retrieved context**. From left to right: Average Indirect Effect from MLPs after corrupting subject + context for scenario based on RAG and subject for vanilla-case. Here, FST=***First Subject Token***, MST=***Middle Subject Tokens***, LST= ***Last Subject Token***, FSST=***First Subsequent Token***, FT= ***Further Tokens***, LT= ***Last Token***. On average 5 times decrease in AIE is observed for LST with RAG vs. vanilla, signalling decrease in usage of MLP when RAG context present.

relation and object/attribute, respectively [1].

For the RAG dataset, we synthetically generate RAG context for each query from the Knowns-Fact dataset using GPT4. This was done to control the variables such as length of each segment within the RAG context and the presence of ***attribute*** or ***object***. Further details on prompts used and samples from dataset in Appendix A. In the scope of this work, we work with a vanilla setting, where no RAG context is present for queries to get enriches, and a RAG setting. The generation was made sure to follow our constraints using quality assurance techniques which regenerated the context when the constraints were not satisfied. The code can be found here in Appendix F

## 5 Empirical Results

### 5.1 Finding 1: Language models minimally use parametric memory in the presence of context

We start by mechanistically probing the contributions of various MLP layers for Llama-2 (7B) and Phi-2 for all 1209 samples from knowns fact dataset for both scenarios, i.e, vanilla vs. RAG to understand fact prediction. For RAG scenarios, the entire ***context*** along with ***subject*** is corrupted as part of causal tracing, whereas for the vanilla case only ***subject*** is corrupted. Figure 2 presents the decrease in AIE in presence of RAG of the LST as compared to vanilla(no RAG) setting.

We analyze the Average Indirect Effect of MLPs representing subject tokens and compare against vanilla vs. RAG context scenarios for Llama-2(7B) for 1209 examples from the knowns fact dataset, and find that the AIE decreases by almost 10 times (from ∼0.22 to ∼0.024), proving that subject tokens within the query does not elicit the parametric memory when the context is present. Similarly, for the case of a smaller language model, i.e, Phi-2, we observe a even more drastic effect, that is, a decrease of around 35 times in AIE (from ∼0.14 to ∼0.004) and again find that the language model does not use the parametric memory. This is in contrast to a non-RAG, vanilla case where the subject token has a high AIE and serves as a hotspot of factual retrieval from parametric memory. In addition to the MLPs, we also perform causal tracing on attention layers, details of which can be found in Appendix G

### 5.2 Finding 2: Last token residual stream obtains more enriched information from the context, rather than subject token in query

Inspired by findings of a strong attention contribution from the Subject Token (ST) in the query question to the Last Token (LT) position for factual queries in(Yuksekgonul et al., 2024), we try to uncover any signal of relevant information transfer between subject token and the last token position in LMs for factual queries.We compute the Attention Contributions from ST [2] to the LT for LlaMa-2 and Phi-2 for vanilla and RAG scenarios for all 1209 factual queries in Knowns Fact Dataset. We find that 70% of the layers don't contribute to the final token prediction and therefore resulting in almost 0 contribution to the Last Token (LT). Thereby, as shown in Figure 3 we extract the top 5% of the Attention Contributions from the ST to

[1] subject part of the user query. For example, for user query: "The Space Needle is located in the city of" the subject will be defined as "The Space Needle". When we say attribute or object, we mean the answer to that query which will be present only once in the context generated placed at the first segment. Example can be found in Appendix B.

[2] ST refers to the subject tokens of the user query.

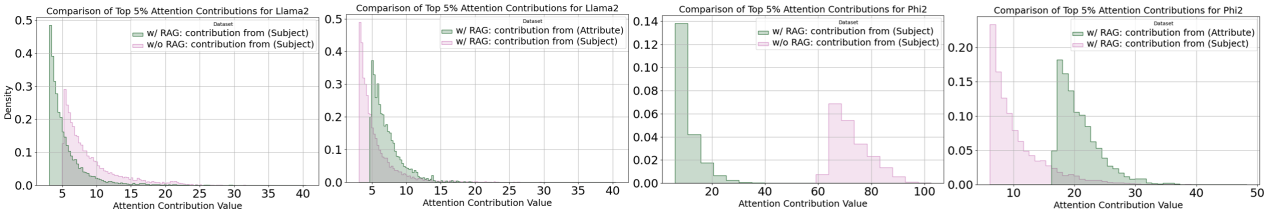



Figure 3: **The last token residual stream obtains less enriched information from the subject token in the query in the presence of retrieved context.**(a) Subject Token contribution for RAG vs vanilla in Llama-2, (b) Comparison of subject and attribute contributions w/ RAG for Llama-2, (c) Subject contribution for RAG vs vanilla in Phi-2, (d) Comparison of subject and attribute contributions w/ RAG for Phi-2. 4a. and 4c indicates subject contribution is twice as lower in case of RAG as compared to vanilla. 4b and 4d shows that attribute token's attention contribution is 5 times more than the subject contribution.

the LT for vanilla vs. RAG scenarios using LlaMA and Phi to amplify the difference. We observe that Specifically for Fig3.a and Fig 3.b, the Attention Contributions from Subject Token decrease in the presence of RAG indicating, the larger influence of RAG context in predicting facts.For LLaMa-2, the mean attention contribution for RAG case is 5.6094 vs. 9.0054 in vanilla setting. For Phi, Attention Contribution at ST is 10.6650 for RAG vs. 72.5961 in the vanilla case, which 7 times larger.

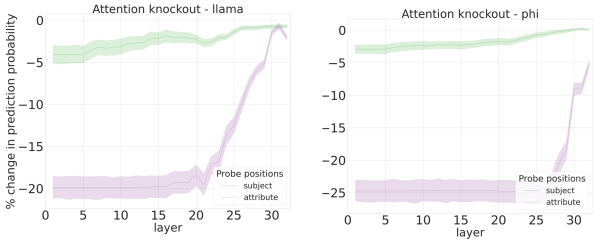



Figure 4: **In the presence of retrieved context, knocking out attention weights from the subject in query to the last token has minimal effect**. (Left) Llama2 (Right) Phi2. [Knocking out attribute tokens decreases probability upto 25%in Phi2 and 20% in Llama2 and only 5% probability is reduced on knocking out subject token attention.]

Additionally, we also analyze Attention Contributions for Attribute Tokens (AT)[3], and compare them against ST. The controlled RAG context we generated synthetically ensures there is **only one** AT present in the context. We find in Fig 3.b, and 3.d, when compared against Attention Contributions of AT present in RAG context, against ST in the query, AT has a larger influence in fact predictions. For LlaMa-2, the mean attention contribution at AT is 7.1242, while at ST is 5.6094. For Phi-2, it is 20.8902 and 10.6650, respectively, i.e, 2 times higher than at ST.

To validate this finding further, we use Attention Knockouts (Geva et al., 2023) to measure the change in probability of the predicted token (object/attribute), when the attention weights from the ST in the query to the last token is knocked off. Figure 4 presents that for the RAG scenario, knocking off attention weights from the subject in query to the last token leads to a probability drop of less than 5 percent in both LLaMa-2 and Phi-2. However, we observe a much stronger drop in the probability of the original predicted token, (20%)in LLaMa-2 and 25% in Phi-2. These results highlight that in presence of RAG context, the last token residual stream ignores information from the subject token position in the query and instead solely relies on the token contributions from the context. Additionally, we perform knockouts in the vanilla setting on the subject token(details in Appendix E.)

> **Main Takeaway**: In the presence of retrieved RAG context, language models internally rely primarily on the context, while minimally using the parametric memory to answer a question.

## 6 Discussion and Conclusions

This paper is the first study to utilize three different mechanistic probing methods to understand the benefits of using RAG context as an external knowledge source to complement the parametric knowledge stored in the models as prior for factual queries. Our work explores the utility of parametric memory, and the interplay between parametric and non-parametric memory in the process of retrieval augmented generation.We find that parametric memory becomes less critical for factual recall when RAG context is augmented to the prompt. Through attention contributions, attention knockouts and causal traces, we specifically observe a reduced reliance on the subject token, and the MLP activations associated with it, when the context is augmented with RAG.

[3]Attribute tokens refers to the expected answer of the query being asked, present in the RAG context, which is also the same as the **object**

## 7 Limitations

Our study is limited by the analysis using short RAG-based context. Handling really long context currently incurs a prohibitively large computational overhead in causal tracing. We plan to study the impact of long context and the impact of subject token and attribute token with respect to position and the tendency to exhibit proximity and recency bias (Liu et al., 2023) in a future work. In addition, similar analysis of instruction tuned models and models that are finetuned on objectives like RLHF is a topic for future work. The current study involves a well controlled setting where attribute token is present only once in the context and the context itself is synthetically generated and well-formed. Retrieved outputs, in practice is very noisy and often sensitive to the quality of the retrievers, rankers, and the hyperparameters used. Examining those is also a natural extension of this work.

## References


et al. Chang. 2023. Benchmarking large language models in retrieval-augmented generation. *arXiv preprint arXiv:2309.01431*.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. *Preprint*, arXiv:2304.14767.

Lovisa Hagström, Denitsa Saynova, Tobias Norlund, Moa Johansson, and Richard Johansson. 2023. The effect of scaling, retrieval augmentation and form on the factual consistency of language models. In *Conference on Empirical Methods in Natural Language Processing*.

IngestAI. 2023. Retrieval-augmented generation (rag): Enhancing llms with external knowledge.

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. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: **phi-1.5** technical report. *arXiv preprint arXiv:2309.05463*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. ArXiv:2307.03172.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *Preprint*, arXiv:2307.11019.

C. Shao, T. Kim, and Z. Gao. 2023. Eragent: Enhancing retrieval-augmented language models with improved accuracy, efficiency, and personalization. *arXiv preprint arXiv:2405.06683*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *Preprint*, arXiv:2104.07567.

A. Singh, M. Sachan, and K. Guu. 2023. Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. *Transactions of the Association for Computational Linguistics*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Kevin Wu, Eric Wu, and James Zou. 2024. How faithful are rag models? quantifying the tug-of-war between rag and llms' internal prior. *Preprint*, arXiv:2404.10198.

Mert Yuksekgonul, Varun Chandrasekaran, Erik Jones, Suriya Gunasekar, Ranjita Naik, Hamid Palangi, Ece Kamar, and Besmira Nushi. 2024. Attention satisfies: A constraint-satisfaction lens on factual errors of language models. *Preprint*, arXiv:2309.15098.

## A Sample Data from Known Facts Dataset


```
{
"known_id": 14,
"subject": "Eavan Boland",
"attribute": "Dublin",
"template": "{} was born in",
"prediction": " Dublin, Ireland, in 1971.
He is the",
"prompt": "Eavan Boland was born in",
"relation_id": "P19"
}
```


## B Sample Data from synthetically generated GPT4 Dataset with RAG contexts


```
{"index": 14,
"user_query": "Eavan Boland was born in",
"object": "Dublin",
"response": ["Boland was born in Dublin,
Ireland, 1944, and became a leading voice
in contemporary Irish poetry,
exploring women's",
"Her birthplace greatly influenced her
works, emphasizing historical narratives
and the role of women in Irish society
through poetry.",
"Boland's early life in Ireland shaped her
poetic voice, focusing on national
identity, gender issues, and
personal history.",
"Educated at Trinity College, her
surroundings nurtured her literary
genius, leading to a profound
impact on modern literature.",
"Despite her global travels and
international teaching positions, her
Irish roots remained central to
her thematic concerns in poetry"]
}
```


**Initial Query :**

> Eavan Boland was born in

**Query Augmented with RAG context :**

> Information is below:————————-
> Eavan Boland was born in Dublin, Ireland, 1944, and became a leading voice in contemporary Irish poetry, exploring women's
> Her birthplace greatly influenced her works, emphasizing historical narratives and the role of women in Irish society through poetry.
> Boland's early life in Ireland shaped her poetic voice, focusing on national identity, gender issues, and personal history.
> Educated at Trinity College, her surroundings nurtured her literary genius, leading to a profound impact on modern literature.
> Despite her global travels and international teaching positions, her Irish roots remained central to her thematic concerns in poetry.
> Given the context information and not prior knowledge, complete the following:
>
> Eavan Boland was born in

**Prompt used for generation of synthetic dataset:**

System Prompt for GPT-4

> You are an expert data generation bot, specializing in generating 20 word segments.
> - You generate these 20-word segments by consolidating information/knowledge AROUND a sentence that the user provides, that is: [user query] [object].
> - While generating these five 20-word segments based on the sentence provided by the user, here: [user query] [object], make sure that only 1 of the 5 segments has the [object] explicitly mentioned. FOLLOW THIS INSTRUCTION STRICTLY.
> - Also make sure that none of these segments contain: [user query]. Double check to make sure this instruction is strictly followed.
> - Also make sure that these segments follow the format of an array of segments, i.e, [segment1, segment2, segment3, segment4, segment5]

User Prompt for GPT-4

> Generate five 20-word segments based on the following sentence: [user query] [object]

The RAG-like dataset of augmented contexts is created synthetically by prompting GPT-4. We also experimented with an actual RAG pipeline, with documents from wikipedia along with the existing query set. However we observed that using a RAG pipeline comes with its own disadvantages with respect to controllability. Given the sensitivity of the output measures like AIE, probabilities, etc to inputs and their perturbations, using a RAG pipeline adds more variability, as retrieved documents can be noisy and extremely sensitive to the underlying retrieval model and its hyperparameters.

## C Training and Capacity

All experiments were completed in University of Massachusetts' Unity cluster. The causal traces, attention knockouts, and attention contributions were run on single A100 (80GB VRAM) GPUs and V100 (40GB VRAM) GPUs depending on availability. Aggregated time to complete causal tracing computations was totaled to 338 hours, and

## D Background

### D.1 Attention Contribution

(Yuksekgonul et al., 2024) introduced SAT-Probe, to predict constraint satisfaction and factual errors by leveraging self-attention patterns to determine if generated text adheres to specified constraints and measuring the contribution of different components to the model's predictions.

And **Attention to Constraints** is achieved by 1. identify constraint tokens within the input, 2. tracking the attention weights $A_{i,j}^{(\ell)}$ ( $\ell$ is layer, $i$ is query token and $j$ is constraint token), 3. aggregating attention weights across layers and heads to determine attention contribution $A_{Ck,T}$ (where $Ck$ is constraint tokens & $T$ is the entire token set).

Finally, the norm of attention contributions $\|a_{i,T}^{(\ell)}\|$ from constraint tokens $c$ to target token $T$ at layer $\ell$ is measured by aggregating these norms across all layers and heads to form a comprehensive metric for attention contribution.

$$a_{c,T}^{(\ell,h)} = A_{c,T}^{(\ell,h)}(x_c^{(\ell-1)} W_V^{(\ell,h)}) W_O^{(\ell,h)}$$

where $a_{c,T}^{(\ell,h)}$ indicates the attention contribution from a constraint token $c$ through head $h$ to the final token $T$. The total contribution is:

$$a_{c,T}^{(\ell)} = \sum_h a_{c,T}^{(\ell,h)}$$

For multiple constraint tokens, the maximum value is considered:

$$A_{C,T}^{(\ell,h)} = \max_{c \in C} A_{c,T}^{(\ell,h)} \quad \text{and} \quad a_{C,T}^{(\ell,h)} = \max_{c \in C} \|a_{c,T}^{(\ell,h)}\|$$

**Correlation with Factual Correctness**

Analyze the correlation between the aggregated attention norms and the factual correctness of the model's outputs. Higher attention norms to constraint tokens are found to correlate with increased factual accuracy, providing a predictive measure for evaluating the reliability of the model's responses.

## E Attention Knockouts

The attention knockouts (Geva et al., 2023) study the impact knocking out attention from a token position $i$ to $j$, where $i \leq j$ for an autoregressive model. More specifically, (Geva et al., 2023) study the impact of knocking out attention from the last token to the subject token, with prompts from the Knowns 1000 dataset, which is a dataset of queries in the form of $(s, r, o)$ triples. In addition to the attention knockouts in the RAG setting, we implement the attention knockouts on the subject token in the vanilla setting.

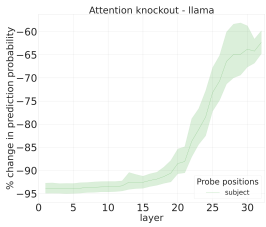



Figure 5: Attention knockouts in LLaMa - vanilla setting

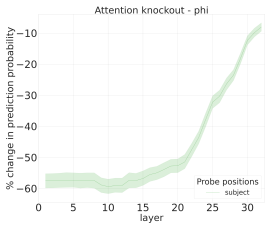



Figure 6: Attention knockouts in Phi - vanilla setting

Figure 5 and 6 show the attention knockout on the subject token in the vanilla setting. In the ab-

sence of added RAG context, we observe a 95 percent decrease in attribute probability in LLaMa and nearly a 60 percent decrease in the attribute probability in Phi-2. In the absence of external context, the model is reliant on parametric memory to answer the factual query, and hence the large probability drop on knocking out subject token attention.

## F Quality checks on the generated synthetic data

Our data generation process comprises prompting GPT-4 to generate synthetic RAG context. The quality check primarily involves verifying the attribute token occurs exactly once within the generated context. The following piece of code is used to perform the verification.

```
def isEntryOkay(entry):
    user_query = entry['user_query']
    object_value = entry['object']
    response = entry['response']

    # Check if object is present only once in the response
    object_count = response.count(object_value)

    # Check if user query is not present in the response
    query_in_response = user_query in response
return object_count == 1 and not query_in_response
```

## G Causal Tracing

The following positions are tracked while plotting the Average Indirect Effect (AIE). First subject token (FST), Middle Subject Token (MST), Last subject token (LST), Further Subsequent token (FSST), Further tokens (FT), Last token (LT). The last token is crucial to study, as it is projected onto a vocabulary during decoding. The last token residual is where information gets written during factual recall (both RAG and non-RAG). The last subject token positions are hotspots of parametric knowledge and factual recall in the vanilla non-RAG setting. Besides, due to causal attention, last subject token (LST) is equipped with context about First (FST) and Middle subject tokens (MST) as well.Further tokens (FT), Further Subsequent tokens (FSST) are not found to have significant causal impact in both RAG and the non-RAG settings.

In addition to causal tracing on MLPs, we also perform causal tracing on the attention modules, which we present in this section in 7 and 8

We observe fairly similar traces for attention in the RAG vs non-RAG settings. The last token is crucial in both settings, thus effectively establishing that all information required for the task is written to the last token's residual stream, with the source

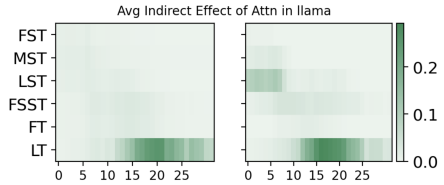



Figure 7: LLama-2 causal trace on Attention

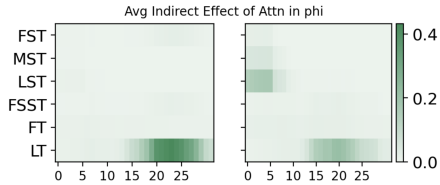



Figure 8: Phi-2 causal trace on Attention

being subject in the non-RAG case, and the source being the attribute token in the RAG setting.

To apply noise to the token embeddings, we use the automatic spherical gaussian noise, the default setting used in (Meng et al., 2022a). The noise is sampled from a gaussian distribution of mean 0 and standard deviation $\nu$ where $\nu = 3\sigma$, where $\sigma$ is the standard deviation of a sample of token embeddings.