Attributing Response to Context: A Jensen–Shannon Divergence Driven Mechanistic Study of Context Attribution in Retrieval-Augmented Generation

Ruizhe Li[1*] **Chen Chen**[2] **Yuchen Hu**[2] **Yanjun Gao**[4] **Xi Wang**[5] **Emine Yilmaz**[3]
[1]University of Aberdeen [2]Nanyang Technological University [3]University College London
[4]University of Colorado Anschutz Medical Campus [5]University of Sheffield

Abstract

Retrieval-Augmented Generation (RAG) leverages large language models (LLMs) combined with external contexts to enhance the accuracy and reliability of generated responses. However, reliably attributing generated content to specific context segments, *context attribution*, remains challenging due to the computationally intensive nature of current methods, which often require extensive fine-tuning or human annotation. In this work, we introduce a novel **J**ensen–**S**hannon **D**ivergence driven method to **A**ttribute **R**esponse to **C**ontext (**ARC-JSD**), enabling efficient and accurate identification of essential context sentences without additional fine-tuning or surrogate modelling. Evaluations on a wide range of RAG benchmarks, such as TyDi QA, Hotpot QA, and Musique, using instruction-tuned LLMs in different scales demonstrate superior accuracy and significant computational efficiency improvements compared to the previous surrogate-based method. Furthermore, our mechanistic analysis reveals specific attention heads and multilayer perceptron (MLP) layers responsible for context attribution, providing valuable insights into the internal workings of RAG models. Our code is available at `https://github.com/ruizheliUOA/ARC_JSD`.

1 Introduction

Retrieval-Augmented Generation (RAG), leveraging large language models (LLMs), has demonstrated significant potential in both academic research [21, 38, 23] and industrial applications [34, 12] by enhancing the accuracy and grounding of generated responses through external contexts such as provided documents or retrieved articles online. A key benefit of RAG lies in its ability to mitigate the hallucination by explicitly attributing generated responses to specific segments of the provided context, known as *context attribution*[2] [28, 20, 5, 3].

Nevertheless, verifying the extent to which generated responses are genuinely grounded in their cited context remains a challenging task. Current approaches frequently rely heavily on human annotation [18, 22] or computationally expensive methods such as model fine-tuning and gradient-based feature attribution for accurate attribution [37, 20, 3], particularly when dealing with extensive documents. For instance, Qi et al. [20] utilised distribution shifts between responses generated with and without context to identify relevant tokens and employed gradient-based feature attribution to pinpoint context relevance. Similarly, Chuang et al. [3] enhanced context attribution accuracy through reward-driven fine-tuning within a Direct Preference Optimisation (DPO) framework, based on probability drop and hold analysis of model outputs to context ablation.

[*] Corresponding Author: `ruizhe.li@abdn.ac.uk`

[2]We use the term *context attribution* in this work, and there are several different terms used in this area, such as citation, self-citation, etc.

Preprint. Under review.

To circumvent these computationally intensive methods, Cohen-Wang et al. [5] introduced an inference-time attribution mechanism premised on the assumption that if removing grounded context segments substantially reduces the probability of a generated response, those segments are deemed necessary. Conversely, if retaining only grounded segments maintains response probability, these segments are considered sufficient. By capturing hundreds of probability ablation variations per context-response pair, Cohen-Wang et al. [5] trained a linear surrogate model based on those hundreds of vectors, including the context segment masks and the corresponding generation probability of the original response, to identify context segments crucial for grounding model responses.

However, Cohen-Wang et al [5] still need hundreds of RAG model's forward calls to collect probability ablation samples for the linear surrogate model training. We propose a novel inference-time **J**ensen–**S**hannon **D**ivergence driven method to **A**ttribute **R**esponse to **C**ontext (ARC-JSD), building upon the inference-attribution assumption above. Our method evaluates the divergence in response distributions generated under the full context compared to sentence-ablated contexts, ranking context sentences based on their JSD differences. This approach offers a significant computational advantage, as it eliminates the need for any additional fine-tuning or surrogate modelling. Furthermore, our ARC-JSD can avoid missing or smoothing non-linearities using JSD to directly quantify actual output distribution shift compared to the linear surrogate modelling [5].

We empirically evaluate our JSD-driven context attribution approach across multiple question-answering benchmarks, i.e., TyDi QA [4], Hotpot QA [35], and MuSiQue [26], using state-of-the-art instruction-tuned LLMs including Qwen2-1.5B-Instruct, Qwen2-7B-Instruct [34], Gemma2-2B-Instruct, and Gemma2-9B-Instruct [25]. Our results not only demonstrate improved average accuracy over 10% in context attribution but also achieve computational efficiency, achieving up to a three-fold speedup compared to Cohen-Wang et al. [5]'s linear-surrogate-based method.

Moreover, we investigate deeper into a mechanistic exploration of context attribution within RAG LLMs by integrating JSD-based analysis with Logit Lens [19]. Through systematic probing, we identify specific attention heads and multilayer perceptron (MLP) layers critical for context attribution. By subsequently analysing these attention heads and visualising how relevant knowledge is stored in the corresponding MLP layers, we provide concrete evidence of their essential roles in context attribution and further elucidate how contextually relevant information is encoded and utilised within the internal mechanisms of RAG models.

In summary, our primary contributions include:

1. Developing a lightweight, JSD-driven context attribution method that accurately identifies context sentences critical for grounding generated responses without requiring fine-tuning or surrogate modelling.
2. Proposing a versatile, computationally efficient solution that can be readily integrated into any existing RAG-based LLM frameworks and improve RAG model trustworthiness.
3. Conducting a detailed mechanistic analysis of RAG LLMs, systematically uncovering and validating specific attention heads and MLP layers responsible for context attribution behaviours.

2 Related Work

Context attribution for RAG. Prior works for context attribution mainly focus on teaching RAG LLMs to generate self-citations for responses, such as few-shot in-context learning [9], instruction fine-tuning [36]. Some post-hoc works [2, 20] used an auxiliary language model or gradient-based feature attribution to locate relevant context segments. In general, those methods for context attribution are *corroborative* [31] in nature, as citations within context are evaluated on whether they *support* or *imply* a generated response. Meanwhile, Cohen-Wang et al.; Chuang et al. [5, 3] including our work focus on the *contributive* attribution methods, which are used to identify whether citations *cause* RAG LLMs to generate a response. Chuang et al. [3] proposed a reward-based fine-tuning with DPO to guide RAG LLMs for context attribution, and Cohen-Wang et al. [5] further trained a linear surrogate model to identify context segments crucial for grounding model responses. However, compared to [5, 3] and corroborative methods above, our ARC-JSD method eliminates the need for any additional fine-tuning or surrogate modelling, and it can be directly integrated into any existing RAG-based LLMs.

Mechanistic analysis for RAG. Existing mechanistic studies mainly focus on the next token generation task to analyse the internal mechanisms of attention heads or MLPs, such as hallucination detection [8], multiple-choice questions [14, 30, 29] and knowledge editing [15, 17, 13]. Recently, Sun et al. [24] used a mechanistic interpretability method to analyse attention heads and MLPs of RAG LLMs for the hallucination detection task. Compared to [24] focusing on locating sources which leads to hallucinations, our proposed ARC-JSD can be regarded as a complementary method to locate citations within context segments and analyse attentions and MLPs, which *causes* RAG LLMs to generate a correct response. Wu et al. [32] focuses on mechanistically analysing retrieval attention heads of RAG LLMs under the Needle-in-the-Haystack (NIAH) setting, where they mainly evaluate whether retrieval attention heads conduct a copy-and-paste operation for retrieving a semantically irrelevant "needle" sentence from the context to the model's outputs. Compared to [32], which restricts their mechanistic analysis to the NIAH setting where the model performs copy-and-paste retrieval, our work investigates how RAG LLMs mechanistically generate responses based on retrieved content through paraphrasing and contextual integration. This setting better reflects real-world RAG applications[3], where models rarely copy text exactly but instead synthesise and rephrase information from retrieved sources.

3 Background

Problem Setup. Consider an autoregressive Transformer-based language model (LLM), denoted as $\mathcal{P}_{\text{LM}}(\cdot)$. Under RAG settings, this model generates responses ($\mathcal{R}$) based on an input query ($\mathcal{Q}$) and associated context ($\mathcal{C}$). Formally, the response generation process can be described as $\mathcal{R} \sim \mathcal{P}_{\text{LM}}(\cdot|\mathcal{C}, \mathcal{Q})$, where the context $\mathcal{C}$ consists of sentences $(c_1, c_2, \ldots, c_{|\mathcal{C}|})$, the query $\mathcal{Q}$ comprises tokens $(q_1, q_2, \ldots, q_{|\mathcal{Q}|})$, and the generated response $\mathcal{R}$ includes tokens $(r_1, r_2, \ldots, r_{|\mathcal{R}|})$. Our analysis of context attribution focuses on how the entire response distribution changes when conditioned on the full context set and ablated context alongside the query:

$$\mathcal{R} \sim \mathcal{P}_{\text{LM}}(\cdot|c_1, \ldots, c_{|\mathcal{C}|}, \mathcal{Q}),\ \mathcal{R} \sim \mathcal{P}_{\text{LM}}(\cdot|\mathcal{C}_{\text{ABLATE}}(c_i), \mathcal{Q}) \text{ where } \mathcal{C}_{\text{ABLATE}}(c_i) = \mathcal{C} \setminus \{c_i\},\ i \in \{1, \ldots, |\mathcal{C}|\}$$

Internal Mechanisms of LLMs. Given a context-query pair represented by a sequence of T tokens $(t_1, \ldots, t_T)$ drawn from a vocabulary $\mathcal{V}$, tokens are initially encoded into d-dimensional vectors $\mathbf{x}_i^0 \in \mathbb{R}^d$ through an embedding matrix $W_E \in \mathbb{R}^{|\mathcal{V}| \times d}$.

An LLM typically consists of L layers, each composed of attention and MLP modules. These modules iteratively transform token embeddings into residual streams at each layer, denoted $(\mathbf{x}_1^\ell, \ldots, \mathbf{x}_T^\ell)$, where $\mathbf{x}_i^\ell \in \mathbb{R}^d$ represents the embedding of token i at layer ℓ. The residual streams serve as central hubs where attention and MLP modules read and write representations [7], following the update rule: $\mathbf{x}_i^\ell = \mathbf{x}_i^{\ell-1} + \mathbf{a}_i^\ell + \mathbf{m}_i^\ell$, where $\mathbf{a}_i^\ell$ and $\mathbf{m}_i^\ell$ denote contributions from the attention and MLP modules at layer ℓ, respectively. At the final layer L, the next token prediction probability distribution is computed as: $\mathcal{P}_{\text{LM}}(t_{T+1}|t_{1:T}) = \text{Softmax}\left(W_U \sigma(\mathbf{x}_T^L)\right)$, where $W_U \in \mathbb{R}^{d \times |\mathcal{V}|}$ is the unembedding matrix, and $\sigma(\cdot)$ denotes pre-unembedding layer normalization.

The attention module, including multiple attention heads, primarily modifies each token's residual stream representation by attending to prior tokens simultaneously: $\mathbf{a}_i^\ell = \sum_{h=0}^{H} \mathbf{a}_i^{\ell,h}$, where $\mathbf{a}_i^{\ell,h}$ indicate each attention head contribution to the residual stream at layer ℓ. Further details of attention mechanisms are provided in Appendix C.

The MLP modules are often conceptualised as key-value memory structures [11, 7, 10, 6]. In these modules, columns of W_{in}^ℓ serve as keys, while rows of W_{out}^ℓ act as corresponding values. The input $\mathbf{x}_i^{\ell-1}$ generates a coefficient vector $\mathbf{k}_i^\ell = \gamma(W_{\text{in}}^\ell \mathbf{x}_i^{\ell-1}) \in \mathbb{R}^{d_{\text{m}}}$ to weight the associated values in W_{out}^ℓ: $\mathbf{m}_i^\ell = \sum_{n=1}^{d_{\text{m}}} \mathbf{k}_i^{\ell,n} \mathbf{v}_i^{\ell,n}$. Further explanations of the MLP mechanisms are available in Appendix C.

Logit Lens. Logit lens [19] is a mechanistic interpretability method designed to analyse intermediate representations within autoregressive Transformers. Given the LLM architecture described above, logit lens leverages intermediate representations to quantify the direct contribution of attention heads ($\mathbf{a}_i^{\ell,h}$), MLP outputs ($\mathbf{m}_i^\ell$), and residual streams ($\mathbf{x}_i^\ell$) to token logits:

[3] Compared to the traditional RAG to directly map context and response based on their word embeddings, our work has a more general setting, which avoids potential embedding mismatch due to the common paraphrase of RAG LLMs.

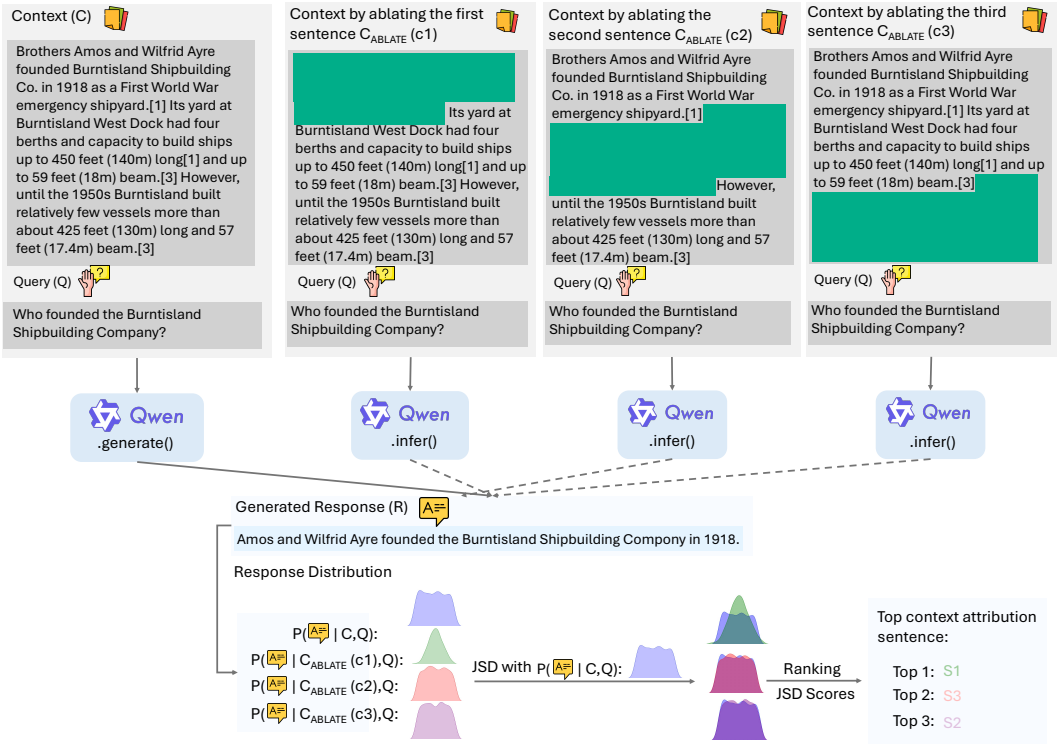

Figure 1: This framework demonstrates how our ARC-JSD works: *(a)* a RAG LLM $\mathcal{P}_{\text{LM}}(\cdot)$ first generates response $\mathcal{R}$ conditioned on full context $\mathcal{C}$ and query $\mathcal{Q}$ input; *(b)* By ablating single context sentence once a time, we can calculate probability distribution of the same response $\mathcal{R}$ conditioned on the ablated context $\mathcal{C}_{\text{ABLATE}}(c_i)$ and query $\mathcal{Q}$; *(c)* We further calculate the JSD scores about probability distribution of the same response $\mathcal{R}$ conditioned on the full context and ablated context, and locate the most relevant context sentence supporting $\mathcal{R}$ with the highest JSD score.

$\text{logit}_i^{\ell,h}(\mathbf{a}_i^{\ell,h}) = W_U\sigma(\mathbf{a}_i^{\ell,h}), \text{logit}_i^{\ell}(\mathbf{m}_i^{\ell}) = W_U\sigma(\mathbf{m}_i^{\ell}), \text{logit}_i^{\ell}(\mathbf{x}_i^{\ell}) = W_U\sigma(\mathbf{x}_i^{\ell})$. Thus, logit lens serves as a powerful tool for pinpointing specific model components crucial to prediction behaviours.

4 Attributing Top Relevant Context Sentences via JSD

In this section, we introduce our ARC-JSD to identify the most relevant context sentences contributing to generated responses. We subsequently present empirical evaluations that demonstrate the effectiveness of ARC-JSD compared to the method proposed by Cohen-Wang et al.[5], across multiple datasets and varying scales of RAG-LLMs.

4.1 Identifying Relevant Context via JSD

Following the assumption proposed by Cohen-Wang et al.[5], the removal of context segments critical to generating a specific response $\mathcal{R}$ significantly impacts the probability distribution of that response. Conversely, the removal of less relevant context segments is expected to minimally affect the probability distribution of $\mathcal{R}$.

Unlike the approach by Cohen-Wang et al.[5], which requires extensive sampling of ablated contexts for each $(\mathcal{C}, \mathcal{Q})$ pair and training a surrogate model to learn context-response relationships, our proposed ARC-JSD method relies purely on inference in the Figure 1. Specifically, we compute the JSD between the response probability distributions conditioned on the full context $\mathcal{C}$ and on each context-ablated variant $\mathcal{C}_{\text{ABLATE}}(c_i)$:

$$\text{JSD}(c_i) = \sum_{j=1}^{|\mathcal{R}|} \text{JSD}\left(\mathcal{P}_{\text{LM}}(r_j|\mathcal{C},\mathcal{Q}) || \mathcal{P}_{\text{LM}}(r_j|\mathcal{C}_{\text{ABLATE}}(c_i),\mathcal{Q})\right) \tag{1}$$

Table 2: Context attribution accuracy of Contextcite baseline and our ARC-JSD, where the Contextcite needs $n + 1 = 256$ ablation calls to achieve a low root mean squared error based on their work [5]. Compared to [5], our ARC-JSD only needs $|\mathcal{C}|$ ablation calls, which significantly smaller than n based on Table 1.

Models	Datasets	Qwen2 1.5B IT	Qwen2 7B IT	Gemma2 2B IT	Gemma2 9B IT	Time Complexity
Contextcite (n =256 calls)	TyDi QA	77.5	75.1	70.7	76.4	$o(n+1)$
	Hotpot QA	54.9	68.4	54.8	70.4	
	MuSiQue	54.5	60.3	51.8	60.9	
ARC-JSD	TyDi QA	84.1	80.0	76.6	81.8	$o(\|\mathcal{C}\|+1)$
	Hotpot QA	71.1	82.3	67.0	79.4	
	MuSiQue	60.6	76.8	65.3	78.2	

where we use $\text{JSD}(c_i)$ to aggregate the JSD score of each generated tokens r_j from $\mathcal{R}$ when the context sentence c_i is ablated from the context $\mathcal{C}$. By calculating JSD scores for all sentences in the context, we identify the most relevant context sentence c_i by selecting the sentence based on the assumption about the significant impact of removing critical context segments:

$$c_i = \arg\max_{c_i \in \mathcal{C}} \left(\{\text{JSD}(c_i)\}_{i=1}^{|\mathcal{C}|} \right) \tag{2}$$

4.2 Evaluation of Context Attribution Accuracy

To assess the efficacy of our ARC-JSD method, we conduct experiments on three widely recognised question-answering datasets commonly used in RAG studies: *TyDi QA* [4]: a multilingual QA dataset using the entire Wikipedia articles as the external context (we only use the English part), *Hotpot QA* [35]: a multi-hop QA dataset requiring reasoning for questions based on multiple documents, and *MuSiQue* [26]: a high-quality multi-hop QA benchmark over Wikipedia that highlights minimal context and multiple valid reasoning paths to evaluate complex reasoning capabilities. Table 1 summarises the statistics of these datasets, where *MuSiQue* has the longest context input compared to others With the average length of context in sentences $|\mathcal{C}| = 93.6$. Our evaluations involve four instruction-tuned LLMs of varying scales, namely Qwen2-1.5B-IT, Qwen2-7B-IT [34], Gemma2-2B-IT, and Gemma2-9B-IT [25]. For each dataset, we randomly select up to 1,000 samples from their development sets. All models are evaluated in inference mode without further fine-tuning. The performance of our ARC-JSD method is benchmarked against the Contextcite proposed by Cohen-Wang et al.[5] (Appendix D includes more details).

Table 1: The size of three benchmarks randomly sampled from their development dataset is up to 1000, where the average word numbers and sentence numbers of context (i.e., $|\mathcal{C}|$) are summarised.

		Contexts	
Datasets	Size	Avg. Words	Avg. Sents.
TyDi QA	440	99.5	4.8
Hotpot QA	1,000	940.3	51.1
MuSiQue	1,000	1753.8	93.6

Table 2 presents the comparative results, clearly demonstrating that ARC-JSD consistently outperforms Contextcite across all datasets and LLM scales, yielding an average accuracy improvement of approximately 10.7%. Crucially, while Contextcite typically necessitates constructing $n + 1 = 256$ randomly ablated context samples per response-context pair for surrogate model training to achieve a low root mean squared error, ARC-JSD requires only $|\mathcal{C}|$ ablations per pair, a number considerably smaller based on dataset statistics provided in Table 1. Consequently, our method offers substantial computational efficiency improvements, achieving up to 3-fold speedups (See Appendix F for details). In addition, we utilise GPT-4.1 mini as a judge to compare whether the generated responses of all RAG models are semantically equivalent to the corresponding gold answers from the datasets when context attribution is correct. The average accuracy is up to 99.3% (See Appendix E for details.)

5 Mechanistically Study RAG LLMs for Context Attribution

5.1 Locating Relevant Attention Heads and MLPs

To better understand the internal mechanisms by which RAG LLMs attribute generated responses to their relevant context sentences, we systematically investigate the specific attention heads and multilayer perceptron (MLP) layers involved. Our method combines the ARC-JSD metric described

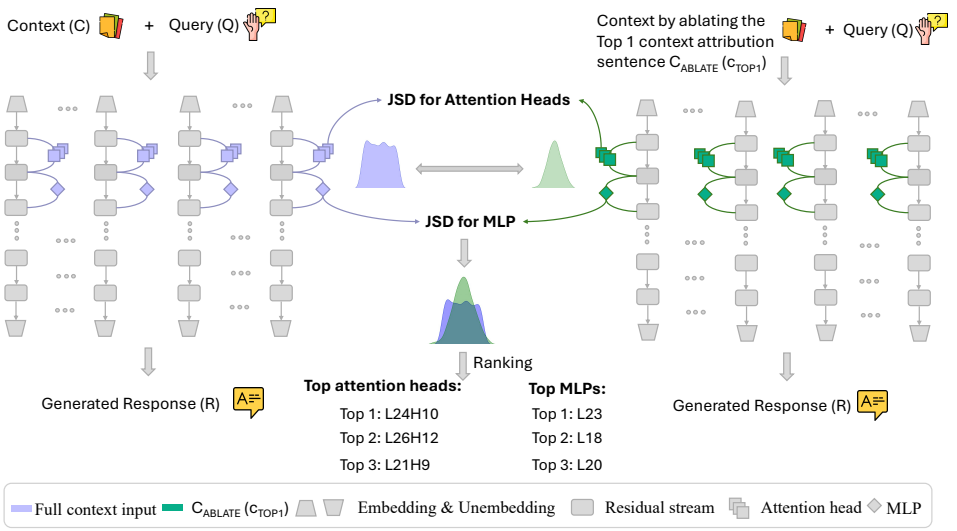

Figure 2: Following our proposed ARC-JSD framework, we apply JSD-based metric to internal components of RAG LLMs: *(a)* For each attention head or MLP output at each layer, we can calculate the probability distribution of the same response $\mathcal{R}$ conditioned on the same query $\mathcal{Q}$ with full context $\mathcal{C}$ and ablated context $\mathcal{C}_{\text{ABLATE}}(c_{\text{top-1}})$ by removing the top relevant context sentence based on § 4.1; *(b)* We can further locate top-N relevant attention heads or MLPs which contribute the context attribution by ranking the collected JSD scores with a descending order.

previously (§ 4.1) with the Logit Lens [19] to precisely quantify contributions from these internal model components.

Following the ARC-JSD framework in the § 4.1, we apply JSD difference at the level of individual attention heads and MLP layers, comparing their outputs between scenarios involving full context and the ablation of the most relevant context sentence using Eq. 1:

$$\begin{aligned}
\text{JSD}^{\ell,h}_{\text{Attn}} &= \sum_{j=1}^{|\mathcal{R}|} \text{JSD}\left(\mathcal{P}^{\ell,h}_{\text{Attn}}(r_j|\mathcal{C},\mathcal{Q}) || \mathcal{P}^{\ell,h}_{\text{Attn}}(r_j|\mathcal{C}_{\text{ABLATE}}(c_{\text{top-1}}),\mathcal{Q})\right) \\
\text{JSD}^{\ell}_{\text{MLP}} &= \sum_{j=1}^{|\mathcal{R}|} \text{JSD}\left(\mathcal{P}^{\ell}_{\text{MLP}}(r_j|\mathcal{C},\mathcal{Q}) || \mathcal{P}^{\ell}_{\text{MLP}}(r_j|\mathcal{C}_{\text{ABLATE}}(c_{\text{top-1}}),\mathcal{Q})\right)
\end{aligned} \tag{3}$$

where $\mathcal{P}^{\ell,h}_{\text{Attn}}()$ and $\mathcal{P}^{\ell}_{\text{MLP}}()$ denote the probability distributions derived from attention head outputs $\mathbf{a}^{\ell,h}_j$ and MLP outputs $\mathbf{m}^{\ell}_j$, respectively, via the logit lens and softmax operations:

$$\mathcal{P}^{\ell,h}_{\text{Attn}}() = \text{Softmax}(\text{logit}(\mathbf{a}^{\ell,h}_j)), \quad \mathcal{P}^{\ell}_{\text{MLP}}() = \text{Softmax}(\text{logit}(\mathbf{m}^{\ell}_j)) \tag{4}$$

where the shape of attention head output $\mathbf{a}^{\ell,h}$ and MLP output $\mathbf{m}^{\ell}$ is $[1, d]$, and d is dimensionality of residual stream. By computing JSD scores across all heads and MLP layers, we rank these components according to their relevance to context attribution:

$$\begin{aligned}
J_{\text{Top-}N}(\text{Attn}) &= \text{sort}\left(\{\text{JSD}^{\ell,h}_{\text{Attn}}\}^{L,H}_{\ell=0,h=0}, \text{descending}\right) \\
J_{\text{Top-}N}(\text{MLP}) &= \text{sort}\left(\{\text{JSD}^{\ell}_{\text{MLP}}\}^{L}_{\ell=0}, \text{descending}\right)
\end{aligned} \tag{5}$$

5.2 Mechanistic Insights from Located Attention Heads and MLPs

Applying the methodology described in § 5.1, we conducted experiments across three benchmark datasets (see § 4.2) using various LLM scales. Figure 3 presents the distribution and JSD scores of attention heads identified as most relevant for context attribution in Qwen2-1.5B-Instruct on TyDi QA dataset. Our analysis reveals that the top attention heads contributing to context attribution predominantly reside in the higher layers. This observation holds across most datasets, partially corroborating earlier findings by Wu et al.[32], which indicated that retrieval-related attention heads are typically found in the intermediate and higher layers.

Notably, our work expands upon the NIAH setting explored by Wu et al[32] by mechanistically evaluating attention heads and MLPs relevance through paraphrasing and contextual integration of RAG LLMs. This setting better reflects real-world RAG applications, where models rarely copy text exactly but instead synthesise and rephrase information from retrieved sources. Additional visualisations and distributions for another Qwen2-7B-IT and Gemma2 models across all datasets are provided in Appendix H. Similarly, Figure 3 illustrates that the intermediate and higher MLP layers also significantly contribute to context attribution. This pattern remains consistent across different datasets and model scales within the same LLM family. Corresponding detailed findings for Qwen2-7B-IT and Gemma2 models across all datasets are available in Appendix H.

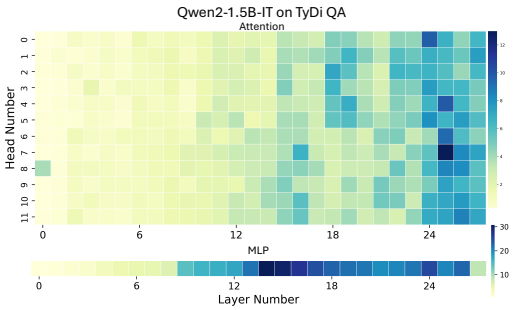

Figure 3: The average JSD score of attention heads and MLP of Qwen2-1.5B-IT on TyDi QA dataset across all layers. The deeper colour indicates larger JSD scores.

6 Verification of JSD-based Mechanistic study

6.1 Semantic Gains of Attention and MLPs for Context Attribution

Apart from locating relevant attention heads and MLPs using JSD-based metric from the § 5.1, we also know that semantic information of context attribution from attentions and MLPs will be added back to the residual stream from each layer based on the autoregressive language model's architecture from the § 3 [7, 13]. Based on such properties, we can verify whether the JSD-based metric for attention and MLPs location in the § 5.1 works by projecting the residual stream before and after each layer's attention and MLPs components into the vocabulary space, and calculating the cosine similarity with the generated response $\mathcal{R}$ to further identify which attention and MLP modules provide higher semantic gains.

Based on the introduction of internal mechanism of LLMs in the § 3 and full context $\mathcal{C}$ with query $\mathcal{Q}$ as model's input, we further split the residual stream flow of each layer into three parts for each generated token t_i, i.e., pre-residual stream $\mathbf{x}_i^{\ell,\text{pre}}$, middle-residual stream $\mathbf{x}_i^{\ell,\text{mid}}$ and post-residual stream $\mathbf{x}_i^{\ell,\text{post}}$:

$$\mathbf{x}_i^{\ell,\text{pre}} = \mathbf{x}_i^{\ell-1,\text{post}} \quad \mathbf{x}_i^{\ell,\text{mid}} = \mathbf{x}_i^{\ell,\text{pre}} + \mathbf{a}_i^{\ell} \quad \mathbf{x}_i^{\ell,\text{post}} = \mathbf{x}_i^{\ell,\text{mid}} + \mathbf{m}_i^{\ell} = \mathbf{x}_i^{\ell+1,\text{pre}} \tag{6}$$

After applying the logit lens to $\mathbf{x}_i^{\ell,\text{pre}}$, $\mathbf{x}_i^{\ell,\text{mid}}$ and $\mathbf{x}_i^{\ell,\text{post}}$ via the softmax, we will have the probability distribution of the generated token $t_i^{\ell,\text{pre}}$, $t_i^{\ell,\text{mid}}$ and $t_i^{\ell,\text{post}}$ for each layer, and then we will use greedy decoding to select the top-1 token with the highest probability:

$$t_i^{\ell,\text{pre/mid/post}} = \arg \max_{t_i^{\ell,\text{pre/mid/post}} \in \mathcal{V}} \left(\text{softmax} \left(\text{logit}(\mathbf{x}_i^{\ell,\text{pre/mid/post}}) \right) \right) \tag{7}$$

Consequently, we can project the selected token $t_i^{\ell,\text{pre/mid/post}}$ into the vocabulary embedding space via the unembedding matrix $W_U \in \mathbb{R}^{d \times |\mathcal{V}|}$:

$$\mathbf{e}_i^{\ell,\text{pre/mid/post}} = W_U[: t_i^{\ell,\text{pre/mid/post}}] \tag{8}$$

We can calculate the corresponding semantic gains $\Delta_i^{\ell,\text{Attn}}$ and $\Delta_i^{\ell,\text{MLP}}$ via attention and MLP modules using the cosine similarity difference with the generated response token embedding $\mathbf{e}_i = W_U[: r_i]$:

$$\Delta_i^{\ell,\text{Attn}} = \cos(\mathbf{e}_i^{\ell,\text{mid}}, \mathbf{e}_i) - \cos(\mathbf{e}_i^{\ell,\text{pre}}, \mathbf{e}_i), \quad \Delta_i^{\ell,\text{MLP}} = \cos(\mathbf{e}_i^{\ell,\text{post}}, \mathbf{e}_i) - \cos(\mathbf{e}_i^{\ell,\text{mid}}, \mathbf{e}_i) \tag{9}$$

Finally, we will average across the entire generated responses $\mathcal{R}$ and calculate the semantic gains $\Delta^{\ell,\text{Attn}}$ and $\Delta^{\ell,\text{MLP}}$ for attention MLP of each layer, and collect and sort the semantic gains of attention and MLP from all layer with descending order:

$$\Delta^{\ell,\text{Attn}} = \frac{1}{|\mathcal{R}|} \sum_{i}^{|\mathcal{R}|} \Delta_i^{\ell,\text{Attn}}, \quad \Delta^{\ell,\text{MLP}} = \frac{1}{|\mathcal{R}|} \sum_{i}^{|\mathcal{R}|} \Delta_i^{\ell,\text{MLP}} \tag{10}$$

Table 3: Spearman's ρ of the overlap about top-10 located attentions and MLPs between JSD-based mechanistic and semantic gain-based metrics over all datasets and RAG models. $\diamondsuit$ and $\spadesuit$ indicate p-value is < 0.05 and < 0.01, respectively.

Modules \ Top-10	Datasets	Qwen2 1.5B IT		Qwen2 7B IT		Gemma2 2B IT		Gemma2 9B IT	
		$J(\cdot) \cap S^{(+)}$	$G(\cdot) \cap S^{(+)}$	$J(\cdot) \cap S^{(+)}$	$G(\cdot) \cap S^{(+)}$	$J(\cdot) \cap S^{(+)}$	$G(\cdot) \cap S^{(+)}$	$J(\cdot) \cap S^{(+)}$	$G(\cdot) \cap S^{(+)}$
Attention	TyDi QA	6.83$\diamondsuit$	7.26$\diamondsuit$	6.91$\diamondsuit$	7.31$\diamondsuit$	7.62$\spadesuit$	7.25$\diamondsuit$	7.63$\spadesuit$	7.28$\diamondsuit$
	Hotpot QA	6.73$\diamondsuit$	6.65$\diamondsuit$	6.81$\diamondsuit$	6.79$\diamondsuit$	6.68$\diamondsuit$	6.67$\diamondsuit$	6.72$\diamondsuit$	6.73$\diamondsuit$
	MuSiQue	6.67$\diamondsuit$	6.72$\diamondsuit$	6.72$\diamondsuit$	6.83$\diamondsuit$	6.69$\diamondsuit$	6.71$\diamondsuit$	6.73$\diamondsuit$	6.75$\diamondsuit$
MLP	TyDi QA	6.90$\diamondsuit$	7.72$\spadesuit$	6.96$\diamondsuit$	7.67$\spadesuit$	7.75$\spadesuit$	8.03$\spadesuit$	7.78$\spadesuit$	8.05$\spadesuit$
	Hotpot QA	6.83$\diamondsuit$	7.49$\spadesuit$	6.87$\diamondsuit$	7.52$\spadesuit$	7.50$\spadesuit$	8.02$\spadesuit$	7.53$\spadesuit$	8.06$\spadesuit$
	MuSiQue	6.87$\diamondsuit$	7.12$\diamondsuit$	6.91$\diamondsuit$	7.18$\diamondsuit$	7.51$\spadesuit$	8.04$\spadesuit$	7.54$\spadesuit$	8.05$\spadesuit$

$$G_{\text{Top-}N}(\text{Attn}) = \text{sort}\left(\{\Delta^{\ell,\text{Attn}}\}_{\ell=0}^{L}, \text{descending}\right), \quad G_{\text{Top-}N}(\text{MLP}) = \text{sort}\left(\{\Delta^{\ell,\text{MLP}}\}_{\ell=0}^{L}, \text{descending}\right) \tag{11}$$

6.2 Mutually Verifying JSD-based Mechanistic Study via the Semantic Gains of Attention and MLPs

Based on the Eq. 5 and Eq. 11, we can locate layer-wise attention and MLP components relevant to context attribution from two different perspectives in the § 5.1 and § 6.1. We can evaluate the correlation of both metrics and further verify the effectiveness of our proposed ARC-JSD metric in the § 4.1 and § 5.1.

Given $\{\text{JSD}_{\text{MLP}}^{\ell}\}_{\ell=0}^{L}$ and $\{\Delta^{\ell,\text{MLP}}\}_{\ell=0}^{L}$ via the JSD-based and Semantic-gain-based metrics, we first define an average-ranking fusion, called *consensus* $S^{(+)}$, to fuse both JSD and semantic gain views, which is based on the assumption that a layer is important if both metrics sort the layer highly:

$$S^{(+)} = \frac{1}{2}\left(\text{ranking}_J + \text{ranking}_G\right) = \frac{1}{2}\left(\frac{\text{ranking of } \left(\{\text{JSD}_{\text{MLP}}^{\ell}\}_{\ell=0}^{L}\right)}{L} + \frac{\text{ranking of } \left(\{\Delta^{\ell,\text{MLP}}\}_{\ell=0}^{L}\right)}{L}\right) \tag{12}$$

where ranking of $(\cdot)$ will assign 1 to the largest $\text{JSD}_{\text{MLP}}^{\ell}$ or $\Delta^{\ell,\text{MLP}}$ and the smallest $\text{JSD}_{\text{MLP}}^{\ell}$ or $\Delta^{\ell,\text{MLP}}$ will be assigned L. Then we uniform and remove the layer influence divided by L to get ranking_J and ranking_G, whose range is $[1/n, 1]$, i.e., a smaller fraction will have a higher ranking ($1/n$ is best). Finally, we take the average of the ranking_J and ranking_G as the *consensus* $S^{(+)}$, where a smaller consensus inside of $S^{(+)}$ will indicate a stronger joint evidence that both metrics consider the layer important, and a larger consensus means at least one metric puts the layer far down the list.

Finally, we can calculate Spearman ρ of $J_{\text{Top}-N}(\text{MLP}) \cap S_{\text{Top}-N}^{(+)}$ and $G_{\text{Top}-N}(\text{MLP}) \cap S_{\text{Top}-N}^{(+)}$, where $S_{\text{Top}-N}^{(+)} = \text{sort}(S^{(+)}, \text{ascending})$. For attention components, we first average the JSD scores of all attention heads in the same layer to build $\{\text{JSD}_{\text{Attn}}^{\ell}\}_{l=0}^{L} = \{\frac{1}{H}\sum_{h=0}^{H}\text{JSD}_{\text{Attn}}^{\ell,h}\}_{l=0}^{L}$, and then further calculate Spearman ρ of $J_{\text{Top}-N}(\text{Attn}) \cap S_{\text{Top}-N}^{(+)}$ and $G_{\text{Top}-N}(\text{Attn}) \cap S_{\text{Top}-N}^{(+)}$.

The benefit of using *consensus* $S^{(+)}$ instead of the raw JSD or semantic gain values is that $S^{(+)}$ will remove all scaling issue due to the different units and variances of JSD or semantic gains, and a single extremely large JSD or semantic gain will not swamp the fusion, which is robust to outliers. Table 3 shows that the ρ of overlap of top-10 located attention and MLP layers about JSD-based and semantic gain-based metrics with *consensus* are statistically significant or highly significant, which further verifies the effectiveness of our proposed JSD-based mechanistic approach.

7 Case Studies of Located Attention Heads and MLPs for Context Attribution

Based on the semantic gains analysis from the § 6.2, we further visualise the projection of middle-residual stream $\mathbf{x}_i^{\ell,\text{mid}}$ and post-residual stream $\mathbf{x}_i^{\ell,\text{post}}$ via Logit Lens to vocabulary space in the Figure 4 and Appendix J. In the Figure 4, Qwen2-1.5B-IT was given a data from TyDi QA dev dataset with the context about mosquitos introduction from Wikipedia and query "*How many wings does a mosquito have?*" as input, and it generates responses "*A mosquito has two wings.*" as output. Based on our proposed ARC-JSD method, we successfully located the top-relevant context sentence, i.e., "*Mosquitoes have a slender segmented body, a pair of wings, three pairs of long hair-like legs,*

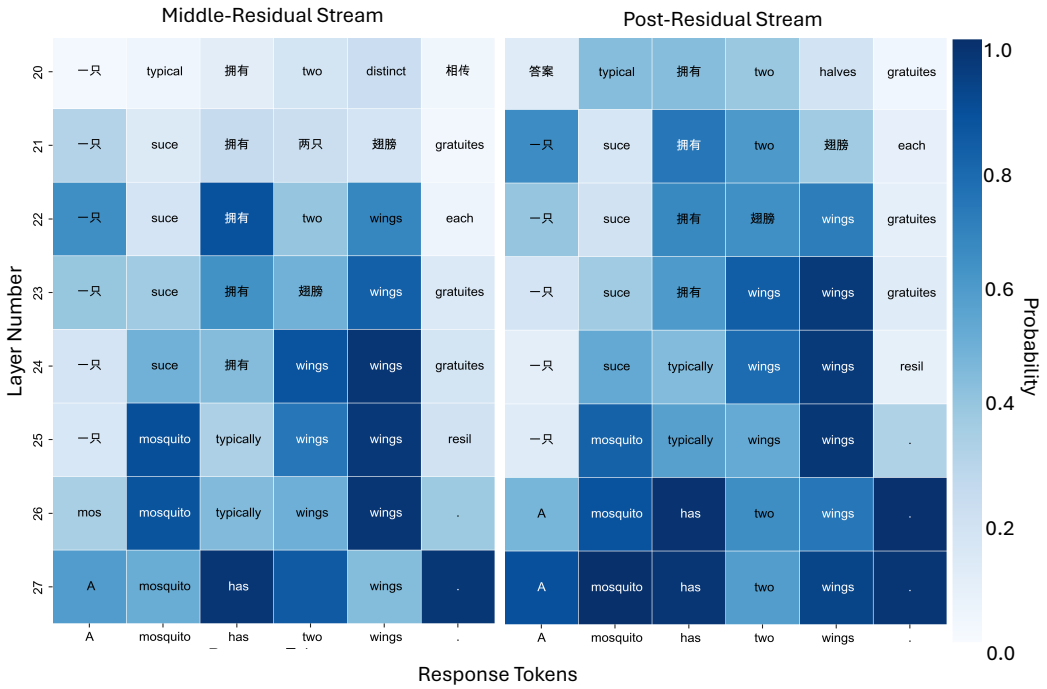

Figure 4: The projection of middle-residual stream $\mathbf{x}_i^{\ell,\text{mid}}$ and post-residual stream $\mathbf{x}_i^{\ell,\text{post}}$ via Logit Lens to vocabulary space from layer 20 to layer 27 of Qwen2-1.5B IT in TyDi QA data sample, where the generated response $\mathcal{R}$ is "*A mosquito has two wings.*" (See Appendix J for all layer projections). Each cell shows the most probable token decoded via Logit Lens. The colour indicates the probability of the decoded token of the corresponding $\mathbf{x}_i^{\ell,\text{mid}}$ or $\mathbf{x}_i^{\ell,\text{post}}$ via Logit Lens.

feathery antennae, and elongated mouthparts". When we compare the heatmap between $\mathbf{x}_i^{\ell,\text{post}}$ and $\mathbf{x}_i^{\ell,\text{mid}}$ in Figure 4 from Layer 20 to Layer 27 (See Appendix J for the whole heatmap), we can find that the probability of correct token is increased significantly after the $\mathbf{x}_i^{\ell,\text{post}}$ compared to $\mathbf{x}_i^{\ell,\text{mid}}$, such as '`wings`' in Layer 23, '`A`', '`has`', '`two`' in Layer 26, and '`mosquito`', '`two`', '`A`' in Layer 27, which aligns with our findings that MLP contribute more for context attribution in higher layers using JSD-based metric from the § 5.2. In addition, we can find that several correct tokens are gradually transferred from their Chinese format to the English version in Qwen2 models, such as '一只 (`A`)', '拥有 (`has`)' and '翅膀 (`wings`)', which is reasonable as Chinese is one of main language resources used in the Qwen2 model pre- and post-training [34]. This finding also matches observations from Wu et al. [33] that representations tend to be anchored by semantically-equivalent dominant-language tokens in higher layers. Moreover, we conduct an ablation study to compare the JSD difference of responses by masking the top-10 relevant attention heads and randomly-selected 10 attention heads. Generally, attention heads using JSD-based metric cause larger JSD scores compared to the random attention heads, which further verifies the effectiveness of our proposed ARC-JSD method (see Appendix I for details).

8 Conclusion

This study introduces ARC-JSD, an inference-time JSD–based metric that attributes responses in RAG directly to their context sentences without additional fine-tuning or surrogate modelling. Evaluations on diverse QA benchmarks and multiple scales of instruction-tuned LLMs demonstrate that ARC-JSD achieves higher attribution accuracy while markedly reducing computational overhead relative to surrogate approaches. Combined with the Logit Lens, ARC-JSD further isolates relevant attention heads and MLPs underpinning context attribution, thereby advancing mechanistic interpretability. Collectively, these findings enhance the transparency of RAG systems and lay groundwork for future research on reliable, efficient RAG models.

Acknowledgement

This work is supported by the Gemma 2 Academic Program GCP Credit Award from Google.

References

[1] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[2] Anthony Chen, Panupong Pasupat, Sameer Singh, Hongrae Lee, and Kelvin Guu. Purr: Efficiently editing language model hallucinations by denoising language model corruptions. *arXiv preprint arXiv:2305.14908*, 2023.

[3] Yung-Sung Chuang, Benjamin Cohen-Wang, Shannon Zejiang Shen, Zhaofeng Wu, Hu Xu, Xi Victoria Lin, James Glass, Shang-Wen Li, and Wen-tau Yih. Selfcite: Self-supervised alignment for context attribution in large language models. *arXiv preprint arXiv:2502.09604*, 2025.

[4] Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. Tydi qa: A benchmark for information-seeking question answering in ty pologically di verse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470, 2020.

[5] Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. Contextcite: Attributing model generation to context. *Advances in Neural Information Processing Systems*, 37:95764–95807, 2024.

[6] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, 2022.

[7] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1:1, 2021.

[8] Javier Ferrando, Oscar Balcells Obeso, Senthooran Rajamanoharan, and Neel Nanda. Do i know this entity? knowledge awareness and hallucinations in language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[9] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, 2023.

[10] Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, 2022.

[11] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, 2021.

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

[13] Shahar Katz, Yonatan Belinkov, Mor Geva, and Lior Wolf. Backward lens: Projecting language model gradients into the vocabulary space. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2390–2422, 2024.

[14] Ruizhe Li and Yanjun Gao. Anchored answers: Unravelling positional bias in gpt-2's multiple-choice questions. *arXiv preprint arXiv:2405.03205*, 2024.

[15] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

[16] Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*, 2022.

[17] Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*, 2023.

[18] Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*, 2022.

[19] nostalgebraist. interpreting gpt: the logit lens, 2020.

[20] Jirui Qi, Gabriele Sarti, Raquel Fernández, and Arianna Bisazza. Model internals-based answer attribution for trustworthy retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6037–6053, 2024.

[21] Hongjin Qian, Zheng Liu, Kelong Mao, Yujia Zhou, and Zhicheng Dou. Grounding language model with chunking-free in-context retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1298–1311, 2024.

[22] Aviv Slobodkin, Eran Hirsch, Arie Cattan, Tal Schuster, and Ido Dagan. Attribute first, then generate: Locally-attributable grounded text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3344, 2024.

[23] Maojia Song, Shang Hong Sim, Rishabh Bhardwaj, Hai Leong Chieu, Navonil Majumder, and Soujanya Poria. Measuring and enhancing trustworthiness of LLMs in RAG through grounded attributions and learning to refuse. In *The Thirteenth International Conference on Learning Representations*, 2025.

[24] ZhongXiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. RedeEP: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *The Thirteenth International Conference on Learning Representations*, 2025.

[25] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[26] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[28] Rose Wang, Pawan Wirawarn, Omar Khattab, Noah Goodman, and Dorottya Demszky. Backtracing: Retrieving the cause of the query. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 722–735, 2024.

[29] Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham M. Kakade, Hao Peng, and Heng Ji. Eliminating position bias of language models: A mechanistic approach. In *The Thirteenth International Conference on Learning Representations*, 2025.

[30] Sarah Wiegreffe, Oyvind Tafjord, Yonatan Belinkov, Hannaneh Hajishirzi, and Ashish Sabharwal. Answer, assemble, ace: Understanding how LMs answer multiple choice questions. In *The Thirteenth International Conference on Learning Representations*, 2025.

[31] Theodora Worledge, Judy Hanwen Shen, Nicole Meister, Caleb Winston, and Carlos Guestrin. Unifying corroborative and contributive attributions in large language models. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 665–683. IEEE, 2024.

[32] Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In *The Thirteenth International Conference on Learning Representations*, 2025.

[33] Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. In *The Thirteenth International Conference on Learning Representations*, 2025.

[34] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[35] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, 2018.

[36] Xi Ye, Ruoxi Sun, Sercan Arik, and Tomas Pfister. Effective large language model adaptation for improved grounding and citation generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6237–6251, 2024.

[37] Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. Automatic evaluation of attribution by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4615–4635, 2023.

[38] Zhenrui Yue, Honglei Zhuang, Aijun Bai, Kai Hui, Rolf Jagerman, Hansi Zeng, Zhen Qin, Dong Wang, Xuanhui Wang, and Michael Bendersky. Inference scaling for long-context retrieval augmented generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[39] Mert Yuksekgonul, Varun Chandrasekaran, Erik Jones, Suriya Gunasekar, Ranjita Naik, Hamid Palangi, Ece Kamar, and Besmira Nushi. Attention satisfies: A constraint-satisfaction lens on factual errors of language models. In *The Twelfth International Conference on Learning Representations*, 2024.

A Broad Impact

RAG systems underpin a wide range of everyday activities, from itinerary planning and news aggregation to document drafting, by combining LLMs reasoning with evidence retrieved from external sources. Yet, the practical value of these systems hinges on our ability to verify that each generated statement is genuinely grounded in the retrieved material. The proposed *post-hoc* ARC-JSD method offers a lightweight, modular solution to this problem. Because ARC-JSD can be seamlessly integrated into any open-source RAG pipeline, it provides developers and researchers with an immediate way of auditing attribution fidelity, thereby strengthening the transparency, reliability, and ultimately the public trust in RAG-based applications.

B Limitations

Our work focuses on the analysis to (i) identify the context sentences that most strongly influence a RAG model's output and (ii) attribute that influence to specific attention heads and MLP layers via a

JSD-based metric. Two important directions, therefore, remain unexplored. First, our layer-level view does not reveal which individual neurons within the MLPs mediate context attribution; techniques such as sparse autoencoder (SAE) probing could provide the necessary resolution. Second, we have not yet examined whether surgical interventions on the identified attention heads, or on the putative neuron-level circuits, can be used to steer or constrain the model's behaviour. Addressing these questions would deliver a more fine-grained mechanistic understanding and open the door to reliable, attribution-aware editing of RAG systems.

C Details of the Internal Mechanisms of LLMs

We consider the standard *autoregressive Transformer* architecture used in LLMs, originally introduced by [27] and subsequently analysed in a series of mechanistic studies [11, 7, 10, 6, 15, 16, 39]. Given a prompt of length T, the input tokens $(t_1, \ldots, t_T)$ from the context-query pair, each drawn from a vocabulary $\mathcal{V}$, are mapped to d-dimensional embedding vectors $\mathbf{x}_i^0 \in \mathbb{R}^d$, where the embedding matrix $W_E \in \mathbb{R}^{|\mathcal{V}| \times d}$.

LLMs normally comprise L identical layers. At layer ℓ, the residual stream $\mathbf{X}^\ell = (\mathbf{x}_1^\ell, \ldots, \mathbf{x}_T^\ell)$, $\mathbf{x}_i^\ell \in \mathbb{R}^d$, acts as a common read–write buffer for both the multi–head attention and the MLP block [7]. For each token i, the residual update is

$$\mathbf{x}_i^\ell = \mathbf{x}_i^{\ell-1} + \mathbf{a}_i^\ell + \mathbf{m}_i^\ell, \tag{13}$$

where $\mathbf{a}_i^\ell$ and $\mathbf{m}_i^\ell$ denote the contributions of the attention and MLP sub-modules, respectively.[4]

After the final layer, a LayerNorm $\sigma(\cdot)$ and the unembedding matrix $W_U \in \mathbb{R}^{d \times |\mathcal{V}|}$ produce the next-token distribution

$$\mathcal{P}_{\text{LM}}(t_{T+1} \mid t_{1:T}) \;=\; \text{softmax}\big(W_U\, \sigma(\mathbf{x}_T^L)\big). \tag{14}$$

Each layer contains H attention heads, each factorised into QK and OV circuits operating with weight matrices $W_Q^{\ell,h}$, $W_K^{\ell,h}$, $W_V^{\ell,h}$, $W_O^{\ell,h} \in \mathbb{R}^{d \times d}$. The QK circuit establishes the attention pattern $A^{\ell,h} \in \mathbb{R}^{T \times T}$, while the OV circuit transports content across sequence positions. For head h the contribution of source token j to target token i is

$$\mathbf{a}_{i,j}^{\ell,h} \;=\; A_{i,j}^{\ell,h} \left(\mathbf{x}_j^{\ell-1} W_V^{\ell,h}\right) W_O^{\ell,h}, \tag{15}$$

and the total attention update for token i is

$$\mathbf{a}_i^\ell \;=\; \sum_{h=1}^{H} \sum_{j=1}^{T} \mathbf{a}_{i,j}^{\ell,h}. \tag{3}$$

A concise per-head summary is $\mathbf{a}_i^{\ell,h} \;=\; \sum_j \mathbf{a}_{i,j}^{\ell,h}$.

Following the key–value interpretation of MLP layers [11, 7], let $W_{\text{in}}^\ell \in \mathbb{R}^{d_{\text{m}} \times d}$ and $W_{\text{out}}^\ell \in \mathbb{R}^{d \times d_{\text{m}}}$ denote the input and output weights. Given $\mathbf{x}_i^{\ell-1}$, the block first produces coefficients

$$\mathbf{k}_i^\ell \;=\; \gamma\big(W_{\text{in}}^\ell \mathbf{x}_i^{\ell-1}\big) \in \mathbb{R}^{d_{\text{m}}}, \tag{16}$$

where γ is the activation function (e.g. GELU). These coefficients weight the value vectors (rows of W_{out}^ℓ) to yield

$$\mathbf{m}_i^\ell \;=\; \sum_{n=1}^{d_{\text{m}}} \mathbf{k}_i^{\ell,n}\, \mathbf{v}^{\ell,n}, \quad \mathbf{v}^{\ell,n} \equiv W_{\text{out}}^\ell[n, :]. \tag{17}$$

D Experimental Details

We run all experiments using H100 GPUs, and we use the sentence tokeniser from the *nltk* library [1] to preprocess all datasets. For all RAG models, i.e., Qwen2-1.5B-Instruct, Qwen2-7B-Instruct [34],

[4] Layer normalisation preceding each sub-module is omitted here for clarity.

Gemma2-2B-Instruct and Gemma2-9B-Instruct [25], we use their standard chat templates to construct the prompt, i.e., using the context and query as a user's message.

When constructing prompts for TyDi QA dataset, we follow the prompt:

```
Context: {context}

Query: {question}
```

For Hotpot QA and MuSiQue datasets which have multiple documents for each data sample, the prompt is constructed as:

```
Title: {title_1}
Content: {document_1}
...
Title: {title_n}
Content: {document_n}

Query: {question}
```

E GPT-4.1 as Judge for Comparison between Generated Responses of RAG models and Gold Answers from Datasets

After using our ARC-JSD to correctly locate the top relevant context sentences for generated responses, we further utilise GPT4.1 as a judge to check whether those responses correctly answer queries based on the corresponding context. As Table 4 shows, generated responses from all RAG models achieve high accuracy in successfully answering the queries based on the contexts, which demonstrates the fundamental ability of those instructed RAG models.

Table 4: GPT4.1 as a judge to evaluate the semantic equivalence between generated responses of RAG models and the corresponding gold answers from those datasets.

Acc. (%)	Qwen2-1.5B-IT	Qwen2-7B-IT	Gemma2-2B-IT	Gemma2-9B-IT
TyDi QA	99.1	99.4	98.9	99.5
Hotpot QA	99.2	99.5	99.1	99.6
MuSiQue	99.3	99.4	99.2	99.8

F Computational Efficiency Between Contextcite and Our ARC-JSD

We mainly compare the computational efficiency between the Contextcite [5] and our proposed ARC-JSD when attributing responses to relevant context. As Figure 5 shows, our ARC-JSD method can achieve up to 3-fold speedup compared to the Contextcite baseline. The main reason is that our ARC-JSD only needs $|\mathcal{C}| + 1$ RAG forward calls to locate top-relevant context, where $|\mathcal{C}|$ is significantly smaller than Contextcite's n calls ($n + 1 = 256$ can achieve a stable RMSE based on their work [5]).

G Examples of ARC-JSD Context Attribution

We demonstrate more examples of our ARC-JSD attribution method used for different RAG models on different datasets, where each example includes the query, generated responses and located top-1 sentence from the context.

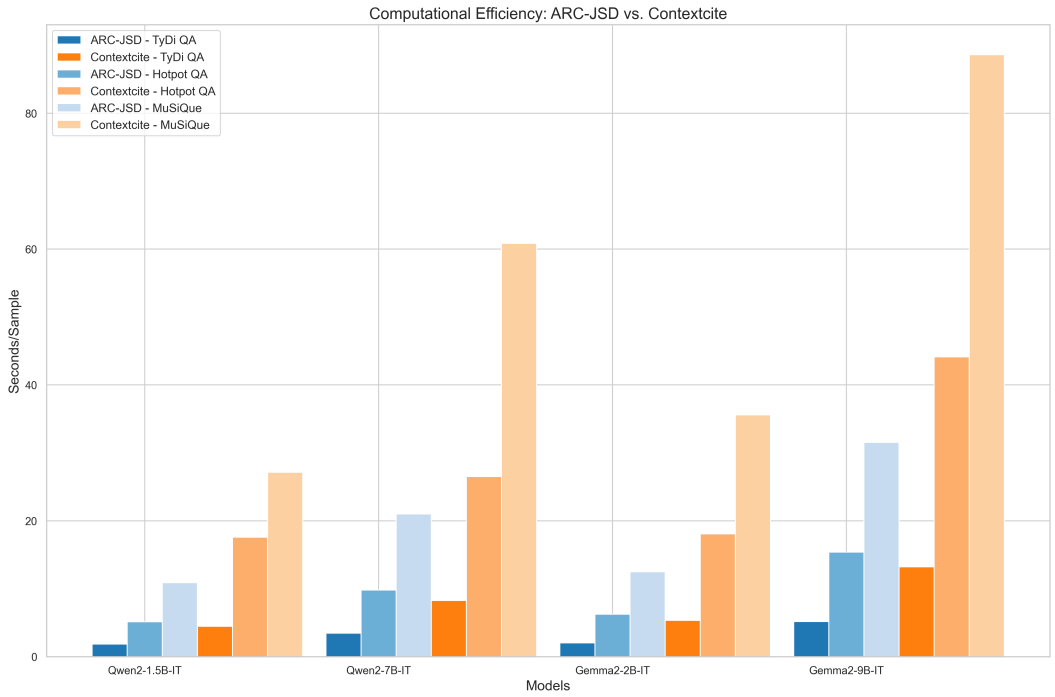

Figure 5: Computational efficiency comparison between our ARC-JSD and Contextcite ($n = 256$ calls) across all datasets and RAG models. Each bar represents the average seconds per sample for ARC-JSD and Contextcite with different RAG models and datasets.

Example of ARC-JSD attributions for `Qwen2-1.5B-IT` on TyDi QA

Query

How many wings does a mosquito have?

Response (attributed statement is highlighted)

A mosquito has two wings.

attribution

Top-1 Source (from the context)

... The word “mosquito” (formed by mosca and diminutive -ito)[2] is Spanish for “little fly”. [3] Mosquitoes have a slender segmented body, a pair of wings, three pairs of long hair-like legs, feathery antennae, and elongated mouthparts.

Example of ARC-JSD attributions for `Qwen2-7B-IT` on MuSiQue

Query

What is the date of birth of the person who was part of Ratata?

Response (attributed statement is highlighted)

The date of birth for Mauro Scocco, who was part of the pop group Ratata, is given as September 11, 1962.

attribution

Top-1 Source (from the context)

... His mother's name was not recorded in history. Title: Mauro Scocco Mauro Scocco (born 11 September 1962) is a Swedish pop artist of Italian descent. He has been described as "one of the sharpest songwriters in Sweden". ...

Example of ARC-JSD attributions for `Gemma2-2B-IT` on TyDi QA

Query

When was the friends song was released which is taken from the fourth studio album of canadian singer and song writer Justin Bieber?

Response (attributed statement is highlighted)

The song "Friends" was released on **August 17, 2017**. It's important to note that this is from Justin Bieber's fourth studio album, "Purpose".

attribution

Top-1 Source (from the context)

... It was written by Bieber, Julia Michaels, Justin Tranter and BloodPop, with production handled by BloodPop. The song was released through GENPOP Corporation, RBMG Records, School Boy Records, Def Jam Recordings and Republic Records on August 17, 2017. It is Bieber's first single credited as lead artist to be released in over a year, the previous being "Company", which was released in March 2016 for his studio album "Purpose" (2015). ...

Example of ARC-JSD attributions for `Gemma2-9B-IT` on Hotpot QA

Query

Tom Thumb Food & Pharmacy is part of the southern division of a grocery company based in what city?

Response (attributed statement is highlighted)

The answer is **Boise, Idaho**. The text states that Tom Thumb Food & Pharmacy is part of the Southern division of Albertsons, and Albertsons Companies LLC is based in Boise, Idaho.

attribution

Top-1 Source (from the context)

... It is also known in Brazil as "Chapolin", "Vermelhinho" ("Little Red One") and "Polegar Vermelho" ("Red Thumb") in allusion to the famous fairy tale character Tom Thumb. Title: Albertsons Albertsons Companies LLC is an American grocery company founded and based in Boise, Idaho. It is privately owned and operated by investors, including Cerberus Capital Management.

H JSD-based Mechanistic Insights for Located Attention Heads and MLPs

We visualise more attention heads and MLP heatmaps using our JSD-based mechanistic approach, where we can find that most RAG models include attribution-relevant attention heads and MLPs across the intermediate and higher layers. On the Hotpot QA and MuSiQue datasets, Gemma2-2B-IT has some relevant attention heads on the lower layers.

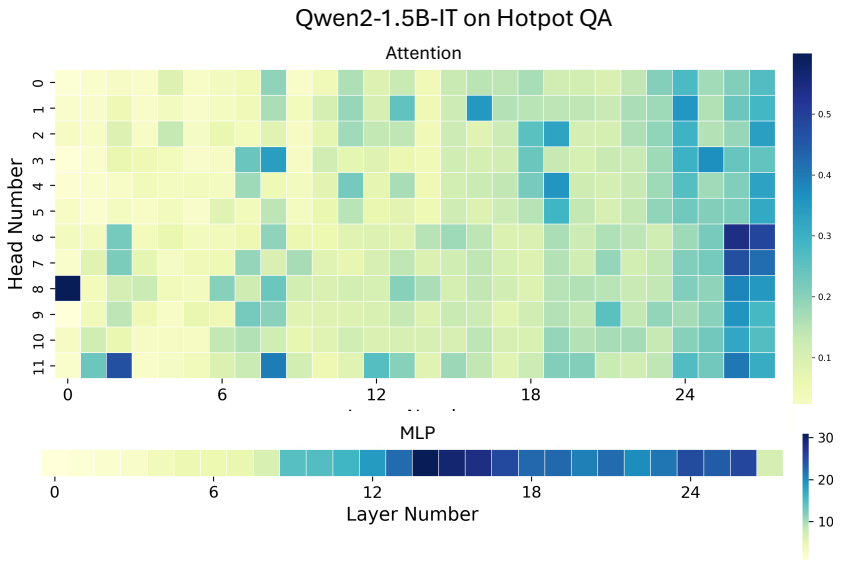

Figure 6: The average JSD score of attention heads and MLP of Qwen2-1.5B-IT on Hotpot QA dataset across all layers. The deeper colour indicates larger JSD scores.

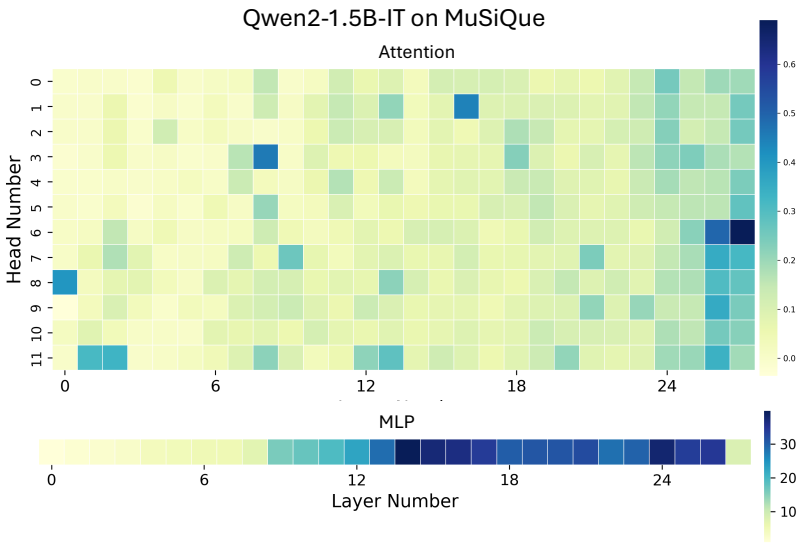

Figure 7: The average JSD score of attention heads and MLP of Qwen2-1.5B-IT on MuSiQue dataset across all layers. The deeper colour indicates larger JSD scores.

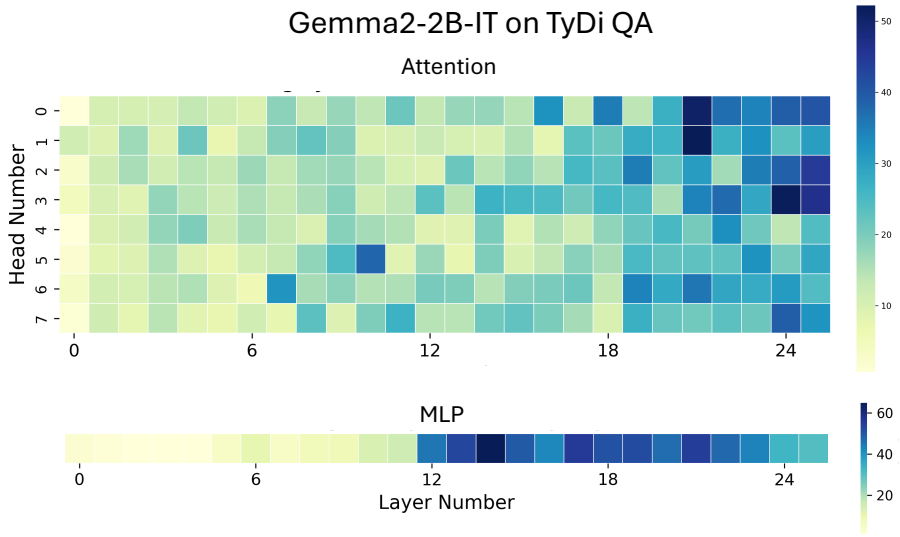

Figure 8: The average JSD score of attention heads and MLP of Gemma2-2B-IT on TyDi QA dataset across all layers. The deeper colour indicates larger JSD scores.

I JSD Comparison between Masking Located Attention Heads and Random Attention Heads

We conducted an ablation study to compare the JSD difference by masking the top-10 relevant attention heads and randomly-selected 10 attention heads. Results show that top-10 attention heads located by the JSD-based metric have higher JSD scores of the same responses while masking in the Table 5.

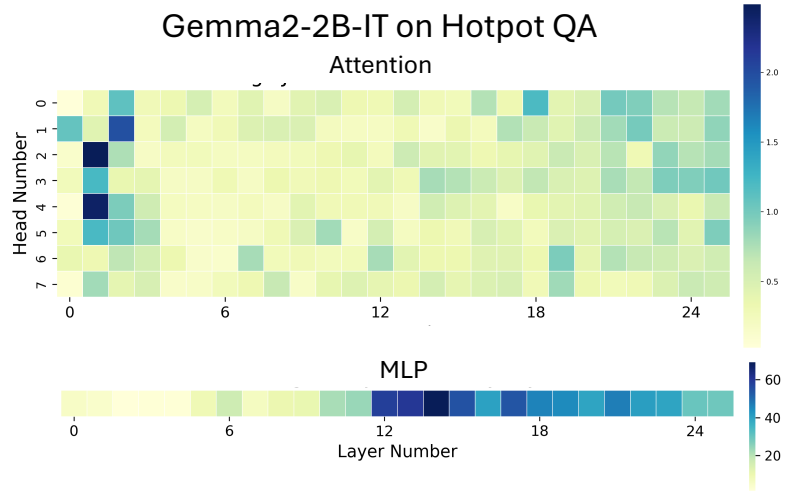

Figure 9: The average JSD score of attention heads and MLP of Gemma2-2B-IT on Hotpot QA dataset across all layers. The deeper colour indicates larger JSD scores.

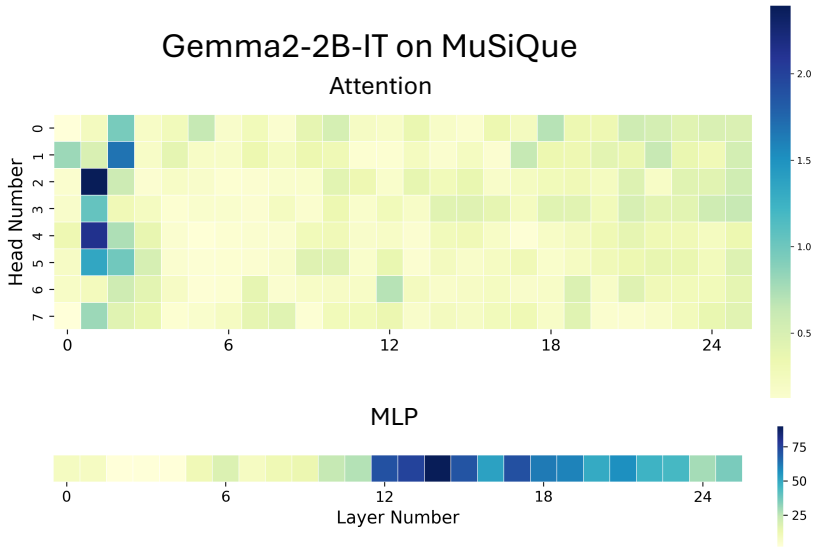

Figure 10: The average JSD score of attention heads and MLP of Gemma2-2B-IT on MuSiQue dataset across all layers. The deeper colour indicates larger JSD scores.

J Case Studies of Attention and MLP's Contribution for Each Response Token

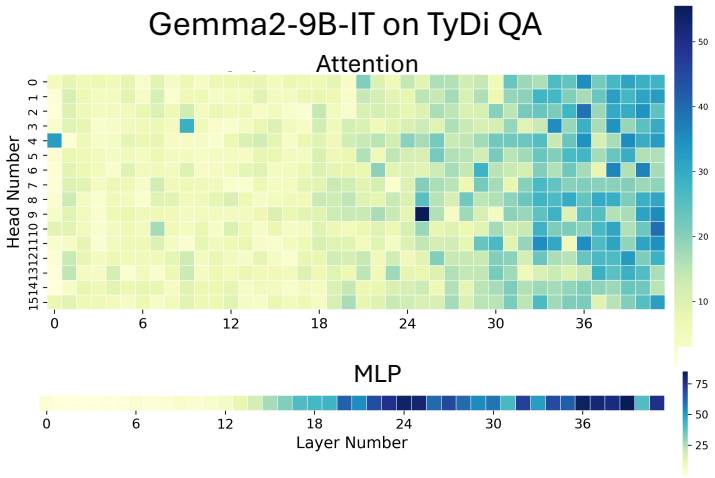

Figure 11: The average JSD score of attention heads and MLP of Gemma2-9B-IT on TyDi dataset across all layers. The deeper colour indicates larger JSD scores.

Table 5: Comparison of average JSD scores between masking top-10 relevant attention heads and randomly masking 10 attention heads using all RAG models on all datasets.

Masking Top-10 Relevant Attention Heads	Randomly Masking 10 Attention Heads
2.23±0.12	1.53±0.76

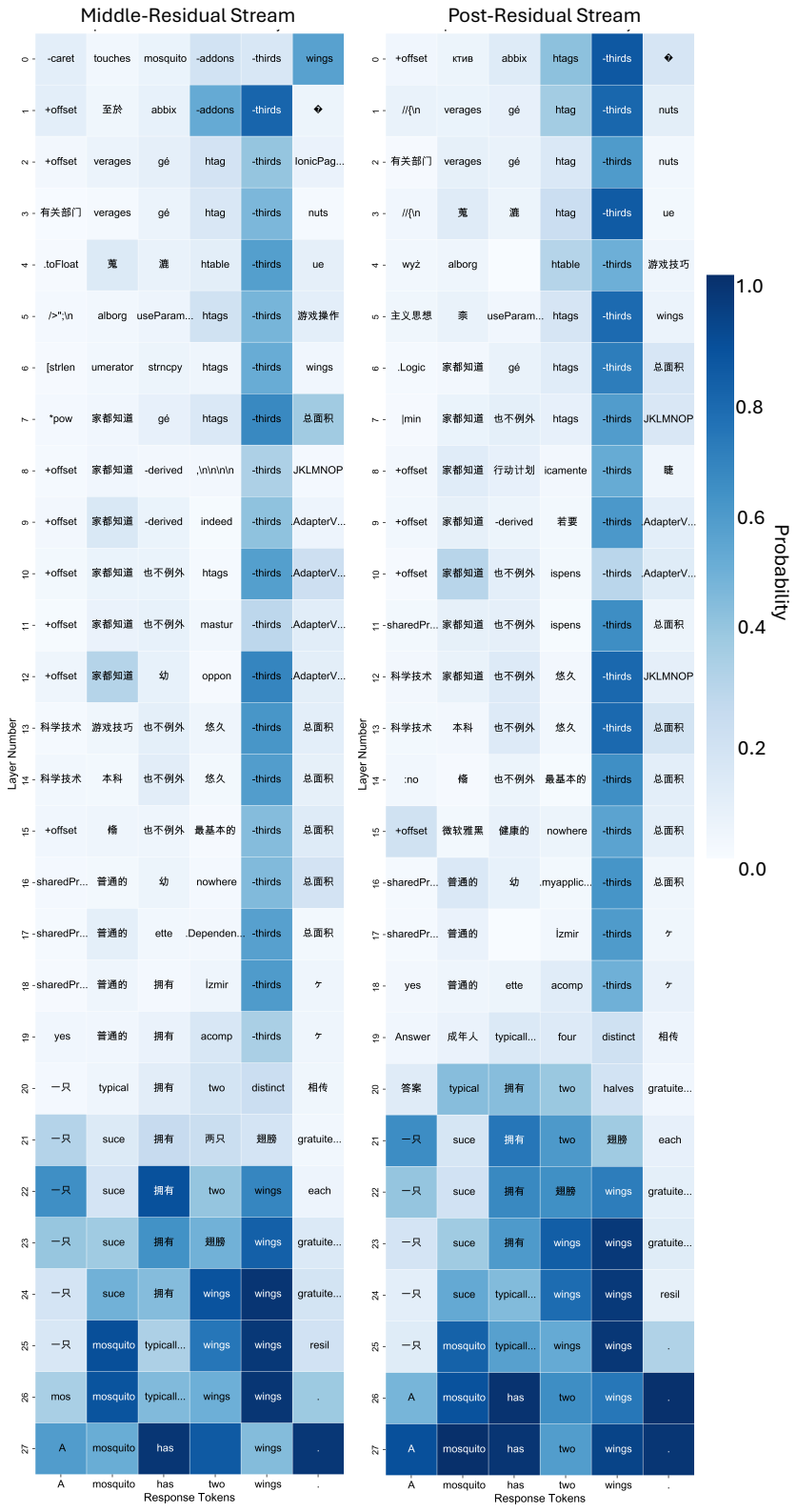

Figure 12: The projection of middle-residual stream $\mathbf{x}_i^{\ell,\text{mid}}$ and post-residual stream $\mathbf{x}_i^{\ell,\text{post}}$ via Logit Lens to vocabulary space from layer 20 to layer 27 of Qwen2-1.5B IT in TyDi QA data sample, where the generated response $\mathcal{R}$ is "*A mosquito has two wings.*". Each cell shows the most probable token decoded via Logit Lens. The colour indicates the probability of the decoded token of the corresponding $\mathbf{x}_i^{\ell,\text{mid}}$ or $\mathbf{x}_i^{\ell,\text{post}}$ via Logit Lens.

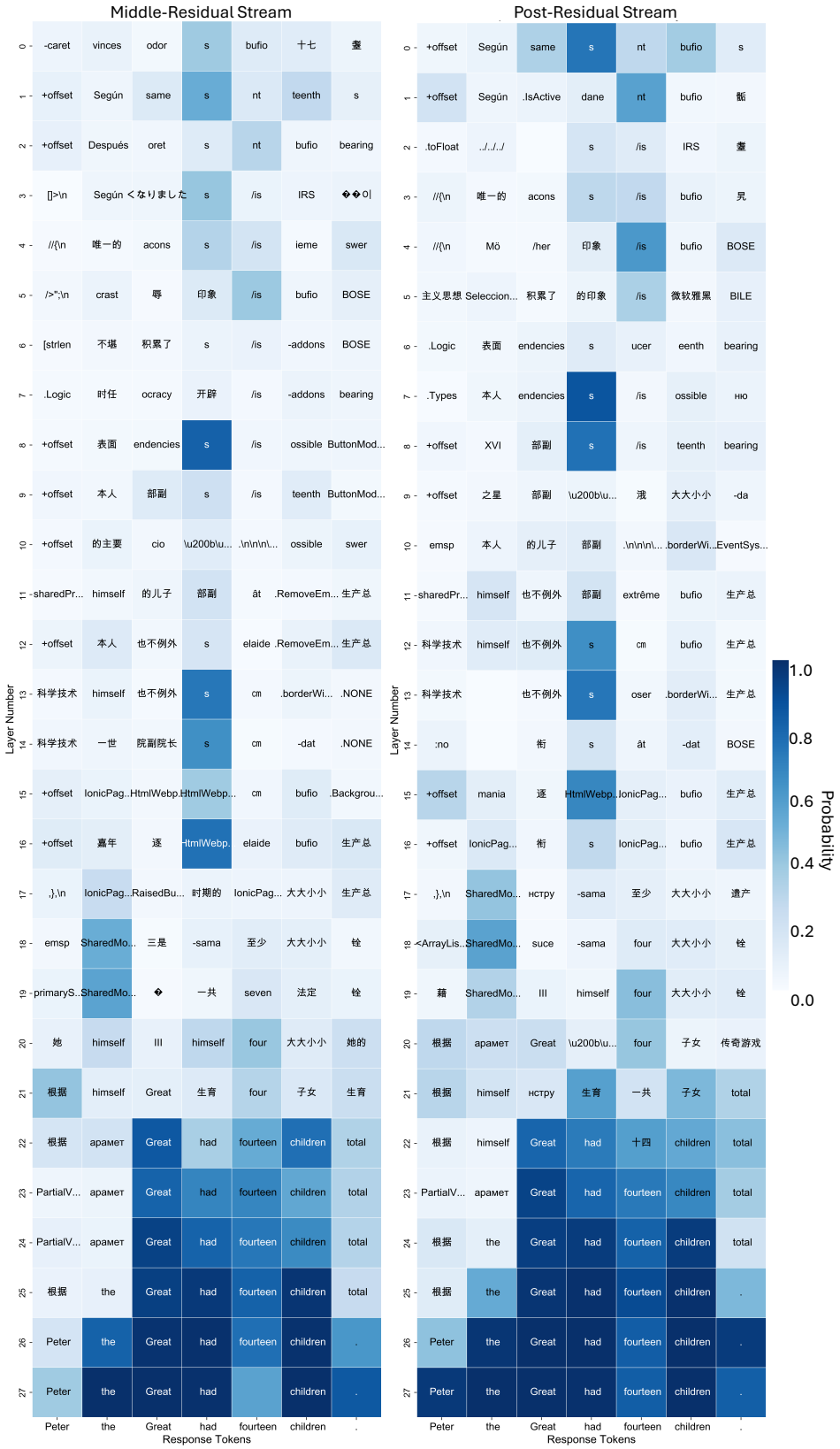

Figure 13: The projection of $\mathbf{x}_i^{\ell,\text{mid}}$ and $\mathbf{x}_i^{\ell,\text{post}}$ via Logit Lens to vocabulary space from layer 20 to layer 27 of Qwen2-7B IT in TyDi QA data sample, where the generated response $\mathcal{R}$ is "*Peter the Great had fourteen children.*". Each cell shows the most probable token decoded via Logit Lens.

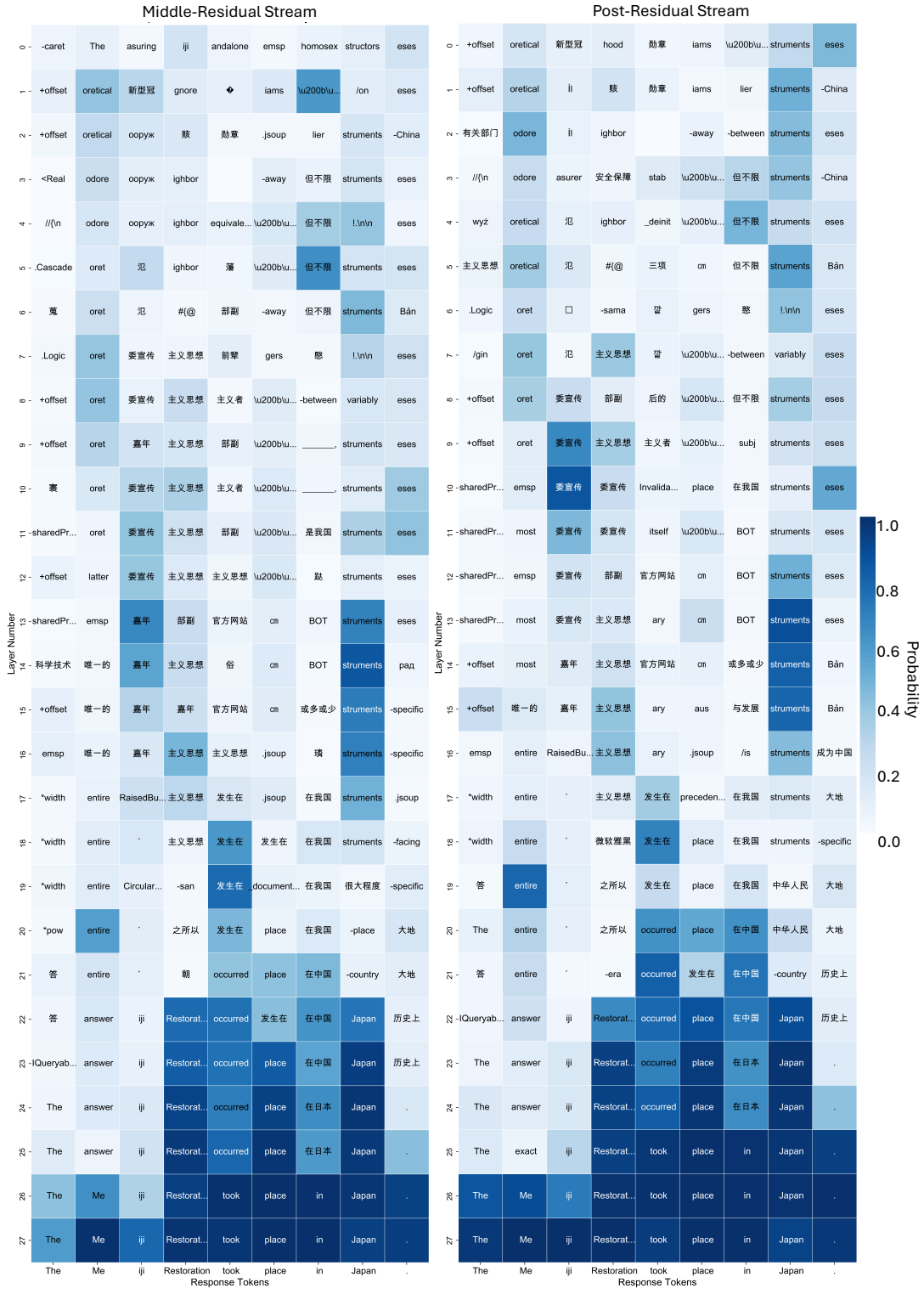

Figure 14: The projection of middle-residual stream $\mathbf{x}_i^{\ell,\text{mid}}$ and post-residual stream $\mathbf{x}_i^{\ell,\text{post}}$ via Logit Lens to vocabulary space from layer 20 to layer 27 of Qwen2-1.5B IT in TyDi QA data sample, where the generated response $\mathcal{R}$ is "*The Meiji Restoration took place in Japan.*". Each cell shows the most probable token decoded via Logit Lens.