Unveiling Knowledge Utilization Mechanisms in LLM-based Retrieval-Augmented Generation

Yuhao Wang*† yh.wang500@outlook.com GSAI, Renmin University of China Beijing, China

Ruiyang Ren* reyon_ren@outlook.com GSAI, Renmin University of China Beijing, China

Yucheng Wang wangyucheng01@baidu.com Baidu Inc. Beijing, China

Wayne Xin Zhao‡ batmanfly@gmail.com GSAI, Renmin University of China Beijing, China

Jing Liu‡ liujing46@baidu.com Baidu Inc. Beijing, China

Hua Wu wu_hua@baidu.com Baidu Inc. Beijing, China

Haifeng Wang wanghaifeng@baidu.com Baidu Inc. Beijing, China

Abstract

Considering the inherent limitations of parametric knowledge in large language models (LLMs), retrieval-augmented generation (RAG) is widely employed to expand their knowledge scope. Since RAG has shown promise in knowledge-intensive tasks like open-domain question answering, its broader application to complex tasks and intelligent assistants has further advanced its utility. Despite this progress, the underlying knowledge utilization mechanisms of LLM-based RAG remain underexplored. In this paper, we present a systematic investigation of the intrinsic mechanisms by which LLMs integrate internal (parametric) and external (retrieved) knowledge in RAG scenarios. Specially, we employ knowledge stream analysis at the macroscopic level, and investigate the function of individual modules at the microscopic level. Drawing on knowledge streaming analyses, we decompose the knowledge utilization process into four distinct stages within LLM layers: knowledge refinement, knowledge elicitation, knowledge expression, and knowledge contestation. We further demonstrate that the relevance of passages guides the streaming of knowledge through these stages. At the module level, we introduce a new method, knowledge activation probability entropy (KAPE) for neuron identification associated with either internal or external knowledge. By selectively deactivating these neurons, we achieve targeted shifts in the LLM's reliance on one knowledge source over the other. Moreover, we discern complementary roles for multi-head attention and multi-layer perceptron layers during knowledge formation. These insights offer a foundation for improving interpretability and reliability in retrieval-augmented LLMs, paving the way for more robust and transparent generative solutions in knowledge-intensive domains.

CCS Concepts

• **Computing methodologies** → **Natural language processing**.

Keywords

Retrieval-Augmented Generation, Knowledge Utilization, Large Language Models

ACM Reference Format:

Yuhao Wang, Ruiyang Ren, Yucheng Wang, Wayne Xin Zhao, Jing Liu, Hua Wu, and Haifeng Wang. 2025. Unveiling Knowledge Utilization Mechanisms in LLM-based Retrieval-Augmented Generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3726302.3730112

1 Introduction

Despite advanced capabilities, large language models (LLMs) [2, 42] often struggle with knowledge-intensive challenges such as open-domain question answering (QA) [24]. These limitations arise primarily from LLMs' reliance on parametric knowledge, which often proves inadequate for real-time queries or domain-specific information, leading to factual hallucinations [5]. To migrate this issue, researchers have developed the retrieval-augmented generation (RAG) technique [9], which first retrieves relevant information from an external knowledge base and then incorporates it as supplementary external knowledge into the input context. This approach not only enhances the model's parametric knowledge without additional training, but also expands LLMs' knowledge boundaries and improves their reliability and transparency [19]. As a result, the potential of RAG has been harnessed in domains spanning complex tasks addressing [4], intelligent information assistants [17], and autonomous agents [28].

Recently, researchers have shifted their focus towards understanding how RAG leverages knowledge, moving beyond merely aiming for task-specific performance improvements [38]. Existing studies have explored the impact of retrieval augmentation on the

*Equal Contributions.
†The work was done during the internship at Baidu.
‡Corresponding Authors.

SIGIR '25, July 13–18, 2025, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730112

knowledge boundaries of LLMs and identified key factors that influence such knowledge boundaries in RAG scenario [29]. Additionally, effort examines the conflicts between parametric knowledge and non-parametric knowledge when LLM utilizes external context, proposing methods to prune conflicting modules and mitigate these conflicts [13]. Despite recent progress, the field still lacks a comprehensive mechanistic understanding within the RAG framework of how LLMs navigate and reconcile both internal (parametric) and external (retrieved) knowledge [38], particularly at a compositional level.

Clear mechanistic insights are crucial for advancing our understanding of RAG's knowledge utilization and diagnosing potential pitfalls of RAG that emerged with transformer-based architectures [36]. Recent studies on model mechanisms have enhanced our understanding of the Transformer architecture [39]. However, RAG introduces two unique challenges for interpretability, presenting new complexities compared to existing tasks like in-context learning (ICL) or mathematical reasoning [6, 10]. First, RAG operates with unstructured data, which means the external knowledge inputs lack a fixed format. This factor increases the difficulty of analyzing how the model understands the inputs and extracting information as a reference. Second, RAG involves both internal and external sources of knowledge [12]. This requires the LLM to integrate and select information from both parametric knowledge and contextual knowledge, demanding a nuanced understanding of how these sources interact.

In this study, we propose a comprehensive investigation of the intrinsic mechanisms governing LLMs within the RAG framework, systematically analyzing them from an interpretive perspective. Specially, we develop new analytical strategies that illuminate both macro-level knowledge streaming and micro-level module contributions within the Transformer architecture. Based on these perspectives, we propose two fundamental research questions: *(1) How does knowledge stream within the RAG framework? (2) How do LLM modules function in knowledge utilization?* To explore the first question, we quantitatively evaluate the stream of knowledge through each layer of the LLM using information flow methodologies. This approach allows us to track how internal and external knowledge streams evolve across layers within the RAG framework. For the second question, we introduce a new approach to analyze the role of neurons, the multi-head attention (MHA) module, and the multi-layer perceptron (MLP) module in knowledge utilization. This enables us to conduct targeted interventions, modulating the LLM's reliance on internal versus external knowledge and providing deeper insight into the functionality of these modules.

We conduct a systematic study, focusing on two main aspects. First, we adopt a diverse range of perspectives, using various methods to explore, hypothesize, and validate specific mechanisms. Second, while uncovering the mechanisms, we also propose methods to leverage them for building controllable RAG systems. From these systematic investigations, we derive the following key insights:
(1) Knowledge streaming within the RAG framework can be identified into four distinct stages, regarding to knowledge utilization: knowledge refinement, knowledge elicitation, knowledge expression, and knowledge contestation. Such a stage division can be re-examined by saliency analysis.
(2) The relevance degrees of retrieved passages direct the knowledge streaming in RAG. This relevance primarily affects the knowledge elicitation stage. This provides a fundamental explanation for how relevance discrepancies impact RAG's knowledge utilization and performance.
(3) We define a new metric, Knowledge Activation Probability Entropy (KAPE), to identify neurons associated with internal and external knowledge. By specifically deactivating these neurons, we successfully altered the model's preference for the selection between two knowledge sources.
(4) We explore the contributions of MHA and MLP modules to knowledge generation. Our findings validate deep-layer knowledge competition and reveal that MLP layers play a role in verifying knowledge accuracy.

In summary, the contributions of this paper are as follows:

- For the first time, we unveil the intrinsic knowledge utilization mechanisms of LLMs in RAG scenarios from two perspectives. At the macroscopic level, we examine the trends in knowledge streaming throughout the RAG process. At the microscopic level, we investigate the role of LLM modules in facilitating knowledge utilization within the RAG framework.
- From the knowledge streaming perspective, we observe four distinct stages in knowledge utilization of LLM-based RAG with saliency-based verification. Based on this, we identify the guiding role of relevance level between the query and external knowledge during the knowledge elicitation stage, and demonstrate that LLM evaluates such relevance through the information flow between them in this stage.
- From the perspective of LLM modules, we further investigate the functions of different modules. We propose a novel approach KAPE to identify knowledge-specific neurons within LLMs, and a deactivation mechanism to alter LLMs' expression tendencies of internal and external knowledge. Furthermore, we investigate the contributions of MLP and MHA to the formation of both knowledge.

2 Preliminaries

In this section, we provide a comprehensive description of the tasks involved, along with formal definitions of both internal and external knowledge. Furthermore, we elaborate on the specific experimental details that underpin our study.

2.1 Task Description and Essential Definition

The core concept of retrieval-augmented generation (RAG) is to enhance the model's generative capabilities by incorporating external information through a retrieval module. This broadens the model's knowledge and enables more accurate and contextually appropriate responses. In this study, we investigate the mechanisms of knowledge utilization in RAG within its typical application scenarios: open-domain question answering (ODQA) [3]. The objective of the ODQA task is to extract relevant information (passages in this study) from a vast external knowledge source to answer a specified query. The ODQA task covers a broad range of knowledge and allows large language models (LLMs) to use both parametric knowledge (closed-book) and retrieval-augmented inputs. This makes

Type	Acquisition Phase	Storage Mechanism
Internal	Training	Stored in LLM parameters (weights)
External	Inference	Provided as context (retrieved documents)

Table 1: Distinction between internal and external knowledge based on acquisition phase and storage mechanism.

it well-suited for studying how different types of knowledge are utilized in different settings.

In our study, we mainly focus on LLM backbones to conduct the empirical analysis. Referring to existing analyses on LLMs' knowledge [29], we formally define the internal and external knowledge of the LLM based on the ODQA task. For *internal knowledge*, given a question q in natural language form, the answer of the question a_{int} can be directly generated by the LLM with an instruction I:

$$a_{\text{int}} = \text{LLM}(I, q). \tag{1}$$

Since the answer a_{int} is generated solely from the LLM's internal parameters, it is considered a direct manifestation of the knowledge embedded within the LLM. This reflects the LLM's intrinsic capability to address knowledge-intensive tasks. For *external knowledge*, we enhance the LLM with a RAG approach. The instruction I directs the LLM to extract an answer a_{ext} to question q using a selected passage subset $\mathcal{P}$ from the larger corpus $\mathcal{D}$ retrieved by the retriever R:

$$\mathcal{P} = \text{Retriever}(\mathcal{D}), \tag{2}$$

$$a_{\text{ext}} = \text{LLM}(I, q, \mathcal{P}). \tag{3}$$

In the RAG setting, the answer a_{ext} generated by the LLM is regarded as an embodied manifestation of the external knowledge utilized during generation. Table 1 distinguishes internal and external knowledge by their acquisition phases and storage mechanisms.

In the following sections, we examine how RAG processes and utilizes internal and external knowledge. We analyze knowledge streams at the macroscopic level to understand overall trends. At the microscopic level, we examine the roles of neurons and modules in facilitating knowledge utilization. This helps unveil the mechanisms of RAG, paving the way for more reliable and controllable systems.

2.2 Experimental Settings

Evaluation Models. To comprehensively investigate the knowledge streaming mechanisms of RAG, we analyze models from two popular open-source families: LLaMA and Qwen. For the LLaMA family, we include different versions and scales, specifically LLaMA-3-8B, LLaMA-3.1-8B and LLaMA-3-70B [7]. For the Qwen family, we study Qwen-2.5-1.5B and Qwen-2.5-7B [40]. All selected models are Base versions.

Dataset. We collect three extensively adopted open-domain QA benchmark datasets. including two single-hop QA datasets (*Natural Questions* [16], *TriviaQA* [14]) and a multi-hop QA dataset (*HotpotQA* [41]). Natural Questions (NQ) is a dataset of question-answer pairs derived from real Google search queries, with answers annotated by human experts. TriviaQA contains trivia questions paired with annotated answers and supporting evidence documents. HotpotQA features question-answer pairs that require multi-hop reasoning to determine the correct answer.

Retrieval Augmentation. As dense text retrieval is demonstrated effective in many scenarios [26], we utilize the open-sourced RocketQAv2 [27] as the passage retriever and use the Wikipedia dump, the same as the previous work [25], as the external knowledge corpus. For each question, we recall the most relevant passages and filter only one *gold passage* that has a low overlap degree with internal knowledge a_{int} and contains the correct answer, serving as the external knowledge source for the LLM. Furthermore, to ensure rigorous analysis, we constructed an additional *fake passage* based on the gold passage for each query, where the correct answer is replaced with the incorrect one. The advantage of this approach lies in the fact that the fake passages serve as a knowledge source completely unseen by the LLM. This guarantees no overlap between the LLM's internal knowledge and the external knowledge within the fake passages, enabling more precise phenomenons of the external knowledge.

Implementation Details. Our experimental setup is structured into two main phases: first, the extraction of embodied manifestations of external knowledge as discussed in Section 2.1, and second, the analysis of corresponding methods described in Section 3 and Section 4. Following previous work [33, 34], all responses were generated using greedy decoding. All experiments are conducted on 8 NVIDIA A100 GPUs with 80GB of memory, using bfloat16 precision.

3 Knowledge Streams within RAG

In this chapter, we analyze the knowledge utilization process of LLMs in RAG scenarios from a macroscopic perspective. Specially, we examine how internal and external knowledge are utilized as the LLM layers deepen, via a knowledge stream perspective. We begin by observing the knowledge streaming process using attention and saliency analysis. Additionally, we explore how the relevance of external knowledge impacts the dynamics of knowledge streaming.

3.1 Information Flow Methodology

In this part, we first establish the relationship between knowledge streaming and information flow. We then introduce the two information flow analysis methodologies and the metrics we used.

Knowledge Streaming and Information Flow. In RAG systems, knowledge streaming is the dynamic integration of external knowledge into a model's processes. This occurs through token interactions within the Transformer architecture [32, 36, 37]. Tokens representing retrieved knowledge impact response generation and modify other tokens' states. We observe these patterns to trace knowledge manipulation within the model. Our research uses two main methods to analyze these patterns: attention-based and saliency-based information flow analyses. These methods help us explore how information flow varies across layers, showcasing the knowledge streaming process.

Attention-based Information Flow. Attention scores in the Transformer architecture directly reflect the flow of information [1,

11]. The dot product's score between a target token's query vector and a source token's value vector determines the influence extent that the source token influences the target token's hidden state. By analyzing these attention scores across different layers, we can assess the direct information flow within the LLM, providing insights into how information is dynamically managed and utilized. To facilitate the analysis, we first define three components within the RAG input instructions:

- **C** (context): the passages utilized for retrieval augmentation.
- **K** (*key*): the potential answers extracted by the LLM from the context, obtained following Equation (3).
- **Q** (*query*): the question to be answered.
- **A** (*answer prompt*): the guiding message at the end of the instruction that directs the model to generate the answer.

Based on this, we further define three quantitative metrics to evaluate the information flow based on attention scores:

- IF_a^{kc}: the attention information flows from key tokens to the context tokens.
- IF_a^{kq}: the attention information flows from key tokens to query tokens.
- IF_a^{ka}: the attention information flows from key tokens to answer tokens.

The three metrics are calculated based on attention matrices adopting the following equation:

$$IF_a^{XY} = \sum_h \sum_i \sum_j \boldsymbol{A_{h,l}}(i,j), \quad i \neq j, i \in X, j \in Y, \tag{4}$$

where X and Y denote different component tokens defined above, $\boldsymbol{A_{h,l}}$ represents the value of the attention matrix corresponding to the h-th attention head in the l-th layer.

Saliency-based Information Flow. Moreover, following previous studies on interpreting Transformer mechanisms [37], we further analyze the information flow in RAG using the gradient-based saliency method [1]. This method evaluates the marginal effect of specific inputs or parameters through gradients. Formally, we compute the saliency matrix using attention metrics and Taylor expansion [22]:

$$S_l = \sum_h \left| \nabla_{\boldsymbol{A_{h,l}}} \mathcal{L}_{\text{sft}}(x) \odot \boldsymbol{A}_{\boldsymbol{h,l}}^{\mathbf{T}} \right|, \tag{5}$$

where $\odot$ denotes the Hadamard product, x is the input and $\mathcal{L}_{\text{sft}}(x)$ represents the supervised fine-tuning loss. The saliency matrix S_l measures the significance of information flow across tokens. As with attention-based information flow metrics, we define three quantitative metrics IF_s^{kc}, IF_s^{kq} and IF_s^{ka} to evaluate the information flow based on saliency scores using the same calculation format with Equation (4). This method offers deeper insights into how each input token affects the final prediction by accounting for the entire gradient path, from input embeddings to output, rather than just a single attention step. For Transformer-based LLMs, it captures token-level influence trends under specific optimization objectives.

Using these two information flow methodologies, we analyze the knowledge streaming process in RAG, uncovering its phased trends. This provides a structured perspective for understanding the dynamics of knowledge streaming, as discussed in the remainder of this section.

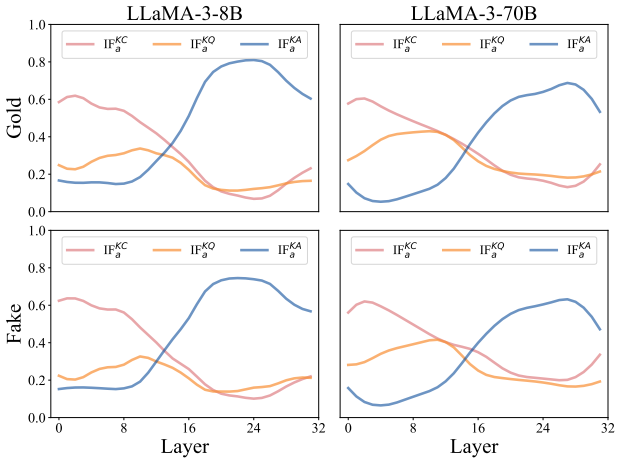

Figure 1: Attention-based information flows across various directions in two versions of LLaMA-3 with different parameter scales, evaluated under the RAG setting using gold and fake passages on the average of three datasets.

3.2 Knowledge Streaming in RAG

In this part, we employ the attention-based information flow approach to depict the dynamics of knowledge transfer within RAG. Specifically, we quantify the knowledge streaming process using three evaluation metrics IF_a^{kc}, IF_a^{kq} and IF_a^{ka}.

3.2.1 Experimental Results and Analysis. Figure 1 illustrates the variations in RAG information flows across different LLM layers within LLMs based on gold passages (defined in Section 2.2), we conduct the evaluation on four widely used open-sourced LLMs. Here we will analyze the trends for the three information flows separately, examine the differences between models, and compare the effects of different augmented documents.

Key-to-Context Information Flow (IF_a^{kc}). This flow shows a steady decline with a slight increase in the deeper layers. In the early layers, the interaction between the key and context is strong. This enables the LLM to quickly refine and adjust its understanding of the context. As layers deepen, this influence weakens, indicating that the context information becomes stable and less reliant on further key-context interactions.

Key-to-Query Information Flow (IF_a^{kq}). Initially, the flow increases before decreasing in the deeper layers. This suggests that as the LLM processes external knowledge, it gradually integrates this information into the hidden states of the query. In the deeper layers, the influence reduces, indicating that the knowledge has been sufficiently encoded by the LLM at this stage.

Key-to-Answer Information Flow (IF_a^{ka}). This flow begins to increase around the first quarter of the layers before gradually decreasing in the last quarter. This suggests that external knowledge

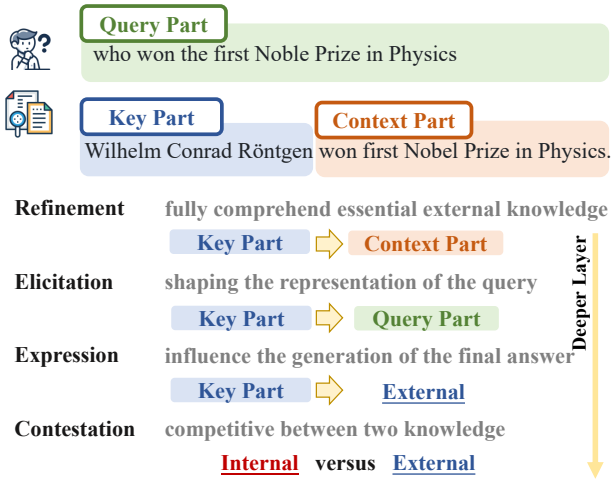

Figure 2: Illustrations of the four stages of knowledge streaming within RAG.

contributes to the answer-generation process in the middle layers, improving the LLM's ability to form a response. In the deeper layers, the influence reduces as the LLM consolidates the answer information with less reliance on external input.

Consistency Across LLMs. We observe consistent results across various LLM families and scales, demonstrating the generalizability of our findings. A similar trend is also evident in LLaMA-3.1-8B and Qwen-2.5-7B. Due to space limitations, we have to leave out the results of the two LLMs.

Comparison of Gold and Fake Passages. We further compare the effect of using gold versus fake passages on RAG information flow. In the case of fake passages, the correct answers (*key*) in gold passages are replaced with incorrect alternatives (defined in Section 2.2). This not only alters the *key part* but also ensures the content is inconsistent with the LLM's pretraining data, eliminating interference from internal knowledge. Despite these changes, the information flow patterns in fake passages remained largely consistent with those from gold passages. This highlights that the LLM's internal processing remains stable, even when the external key is incorrect, further validating the robustness of the findings.

3.2.2 Knowledge Streaming Exhibits Multi-stage Nature in RAG. By observing the knowledge streaming, we propose a hypothesis that knowledge traverses multiple streaming stages in RAG. Specifically, we delineate four distinct stages:

- **Stage 1: Knowledge Refinement.** During this phase, external knowledge, represented by the interaction between the context and key, is deeply integrated, enabling the LLM to fully comprehend the context and distill essential external knowledge.
- **Stage 2: Knowledge Elicitation.** As the context is absorbed, the flow of knowledge from the key to the query intensifies, transmitting the refined contextual information and shaping the representation of the query.
- **Stage 3: Knowledge Expression.** With the continued flow between external knowledge and the query, external knowledge begins to significantly influence the generation of the final answer.
- **Stage 4: Knowledge Contestation.** A competitive dynamic emerges between external and internal knowledge within the LLM, ultimately determining the final answer.

Figure 2 illustrates the four stages of knowledge streaming within RAG. Each stage is defined by distinct directions of knowledge streams, supporting effective knowledge utilization. Note that these stages are not strictly discrete, as transitional layers may exist between stages.

3.3 Corroboration Analysis on Saliency

To verify the rationality of the proposed four-stage RAG knowledge streaming hypothesis, we conduct corroboration from the perspective of saliency-based information flow. As introduced in the section, the saliency score represents the changing trend of attention-based information flow, reflecting whether the LLM aims to enhance or reduce the expression of knowledge at a given moment. Similar to Section 3.2, we quantify the knowledge streaming process using three saliency-based information flow metrics IF_s^{kc}, IF_s^{kq} and IF_s^{ka}, Figure 4 illustrates the saliency-based metrics that vary in different layers evaluated on LLaMA-3-8B.

First, we observed **a trend of modifying knowledge streaming similar to the one discussed in Section 3.2** on knowledge streaming. During the knowledge refinement stage, the LLM tends to increase the flows from the context to the key, refining external knowledge. In the knowledge elicitation stage, the LLM enhances the flows from the key to the query. In the knowledge expression stage, the LLM demonstrates a pronounced increase in its tendency to enhance the flows from the key to the answer prompt. This trend remains steady and consistent during the knowledge contestation stage.

Moreover, the knowledge streaming patterns shown in Figure 1 present that the flows from the key to the answer decreases in the final stage. However, the LLM tends to enhance it. **This suggests that the decline in external knowledge streaming may be influenced by the LLM's internal knowledge streaming.** The difference confirms that a contestation relationship indeed exists between external and internal knowledge during the final stage.

In summary, these findings support the segmentation of knowledge streaming stages in the RAG process from a different perspective.

3.4 Relevance Guides the Knowledge Streaming

The previous study has demonstrated that the relevance of external passages significantly influences the performance of LLMs in RAG scenarios [29]. Based on this, we have strong reason to hypothesize that the influence of the given passage relevance on RAG performance primarily stems from its impact on the knowledge streaming within the model. In this part, we explore the dynamics of knowledge streaming in RAG under varying settings of passage relevance with queries.

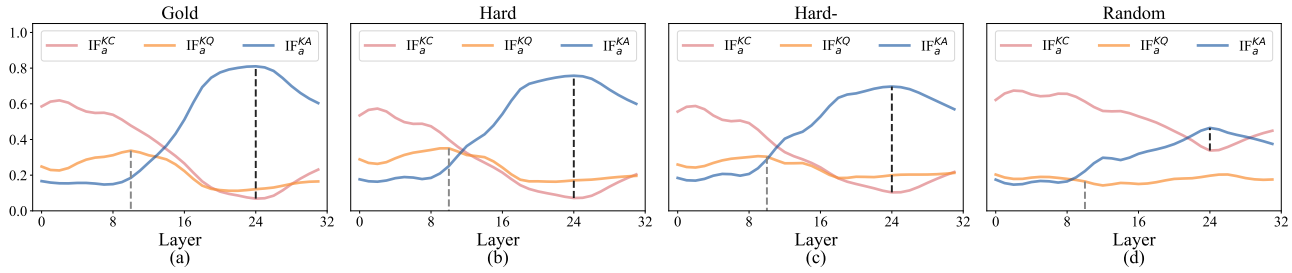

Figure 3: Attention-based information flow based on external passages of various relevance with the query.

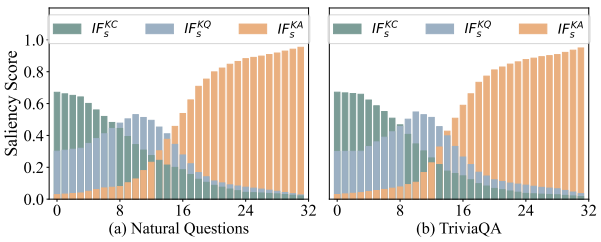

Figure 4: Saliency score on three information flow directions on Natural Questions and TriviaQA dataset for LLaMA-3-8B.

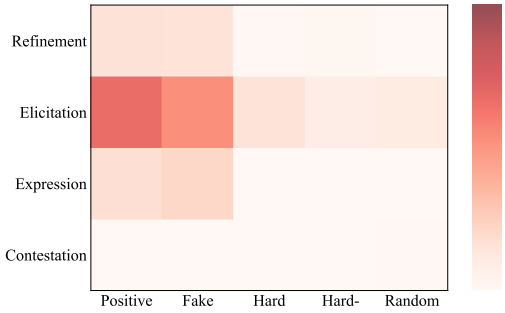

Figure 5: The figure shows the decrease in the probability of generating internal knowledge. This occurs after disrupting the key-to-query knowledge streaming across the four knowledge streaming stages.

3.4.1 External Knowledge Source with Different Relevance. Given an input question, we follow the previous work [29], and categorize passages with varying degrees of relevance as follows:

- **Positive** (*positive passage*): the passage that contains the correct answer, as defined in Section 2.2.
- **Hard** (*hard negative passage*): the passage relevant to the query but lacking the correct answer, sampled from the top-ranked retrieval results for the query.
- **Hard-** (*less hard negative passage*): the passage weakly relevant to the query but lacking the correct answer, randomly sampled from the retrieval results for the query.
- **Random** (*randomly sampled passage*): the passage irrelevant to the query and devoid of the correct answer, randomly sampled from the entire corpus.

Based on the passages with varying degrees of relevance to the question, we can conduct a comprehensive evaluation of the influence of external knowledge relevance on the knowledge streaming within RAG.

3.4.2 Relevance Guides the Knowledge Streaming for Various External Knowledge Relevance. We firstly examine how modifications in the passage's relevance affect the RAG knowledge streaming, Figure 3 illustrates the results of attention-based information flow.

Firstly, we observe that when processing passages with varying degrees of relevance, **the overall trend of knowledge streaming remains consistent**. The four stages we previously defined are clearly identifiable across all three scenarios with different relevance levels (first three sub-figures in Figure 3).

Secondly, **as relevance decreases, the changes in the three information flows and their differences become less noticeable**. For more relevant passages, external knowledge streaming increases, indicating that the LLM uses these passages more extensively to generate responses.

Specifically, the gray dashed line in the figure highlights the changes in IF_a^{kq} during the knowledge elicitation stage. This shows that the LLM's integration of external knowledge into query understanding weakens significantly with decreasing relevance. The gray dashed line also marks the difference between the key-to-answer IF_a^{ka} and key-to-context IF_a^{kc} flows during the knowledge expression stage. **We can observe that LLM minimizes its use of external knowledge when handling irrelevant content**.

Overall, these findings provide a foundational explanation for existing work [29] and demonstrate the LLM's distinctive characteristic to utilizing external knowledge depending on external knowledge relevance.

3.4.3 Knowledge Elicitation with Relevance Assessment. Previous results indicate that the influence of external knowledge relevance on LLM knowledge streaming emerges relatively early during the knowledge elicitation stage (gray dashed line in Figure 3). We hypothesize that during this stage, the interaction between external knowledge and the query enables the LLM to perform relevance discrimination, and it further guides LLM on whether to incorporate external knowledge. To validate this hypothesis, we conducted

experiments by supplying passages with varying levels of relevance, and cut off the knowledge streaming between key and query at different stages. We then measured the differences in the final layer probabilities for generating external knowledge:

$$d = p(a_{\text{ext}}|q, p, \boldsymbol{M}_{\text{causal}}) - p(a_{\text{ext}}|q, p, \boldsymbol{M}_{\text{causal}} + \boldsymbol{M}_{\text{k2q},s}), \quad (6)$$

where s denotes the interrupted knowledge streaming stage, $\boldsymbol{M}_{\text{causal}}$ is the causal mask for decoder-only LLMs, and $\boldsymbol{M}_{\text{k2q},s}$ represents the mask of the attentions from key part to query part. Figure 5 presents the results with a heatmap. It is evident that upon cutting the knowledge streaming from key to query during the knowledge elicitation stage, the increase in the expression of external knowledge is more pronounced for passages with higher similarity. This finding supports our hypothesis that relevance judging is performed during the knowledge elicitation stage of RAG.

4 Modules Function in Knowledge Utilization

In this chapter, we study knowledge utilization in LLMs within RAG scenarios from a microscopic perspective, focusing on the roles of different modules. Specifically, we first analyze the role of neurons and their activations in the utilization of internal and external knowledge. Furthermore, we investigate the respective functions of the MLP and MHA modules in the expression of knowledge.

4.1 Affects of Knowledge-Specific Neurons

Our findings reveal distinct patterns across layers, suggesting that, similar to the human brain, LLMs have specialized regions and neurons. The utilization of internal and external knowledge appears to rely on different neurons. In this part, we propose a method to identify neurons that activate different types of knowledge (Section 4.1.1). By deactivating these neurons, we aim to change the model's preference for internal or external knowledge (Section 4.1.2).

4.1.1 Identification of Knowledge-specific Neurons. In this part, we introduce a method to identify neurons that manage internal and external knowledge. We first describe the notion of neuron activation, followed by our proposed methodology for identifying knowledge-specific neurons.

Neuron Activation and Gating in LLMs. For recent LLMs such as LLaMA, the MLP layers use a gating mechanism with Gated Linear Units (GLU)[30]. The MLP layer is defined as:

$$\text{MLP}(\boldsymbol{h}) = \left(\text{act\_fn}(\boldsymbol{h}\boldsymbol{W}_{\text{gate}} \odot \boldsymbol{h}\boldsymbol{W}_{\text{up}})\right)\boldsymbol{W}_{\text{down}}. \quad (7)$$

Consistent with previous work [23], we define that the j-th neuron inside the i-th feedforward network (FFN) layer is considered to be activated if its respective activation values from $\text{act\_fn}(\boldsymbol{h}_i\boldsymbol{W}_{\text{gate},i})_j$ exceed zero. Formally, for the j-th neuron in the i-th layer of the model, we first compute its activation probability when utilizing both internal and external knowledge:

$$p_{i,j}^{k} = \left\| \mathbb{E}_k \left[\Theta\left(\text{act\_fn}\left(\boldsymbol{h}^i \boldsymbol{W}_j^i\right)\right) \right] \right\|_1^{\text{norm}}, \quad (8)$$

where $\Theta(\cdot)$ denotes the Heaviside step function. Using L1 normalization, we transform the raw activation probabilities of neurons across different data samples into a distribution reflecting internal and external knowledge. This process effectively maps the original probabilities to a distribution focused on the types of knowledge, enabling us to better isolate and analyze the neurons associated with each knowledge type.

Quantifying Neuron Knowledge Preference via KAPE. We identify neurons that handle internal and external knowledge. Inspired by the existing study [31], we define **K**nowledge **A**ctivation **P**robability **E**ntropy (KAPE). Using the L1-normalized probabilities calculated from Equation 8, we calculate the probabilities $p_{i,j}^{\text{IK}}$ and $p_{i,j}^{\text{EK}}$ for internal and external knowledge respectively. The KAPE for each neuron is computed as follows:

$$\text{KAPE}_{i,j} = -\left(p_{i,j}^{\text{IK}} \log(p_{i,j}^{\text{IK}}) + p_{i,j}^{\text{EK}} \log(p_{i,j}^{\text{EK}})\right), \quad (9)$$

where a low KAPE score suggests that the j-th neuron in the i-th layer has a high activation probability for one type of knowledge and a low probability for another, making it a knowledge-specific neuron.

Implementation Details. The identification process consists of two main steps: calculating the KAPE scores and refining the neurons. First, we compute the KAPE scores for each neuron based on their activation probabilities under both RAG instructions with gold passages and a closed-book setting, following L1 normalization to derive $p_{i,j}^{\text{IK}}$ and $p_{i,j}^{\text{EK}}$ in Equation 8, and calculated the $\text{KAPE}_{i,j}$ score for each neuron. Second, we select the top 1% of neurons with the lowest KAPE scores as knowledge-specific. We then apply a threshold to retain only those neurons with significant activation probabilities in either RAG or closed-book scenarios. These steps help identify neurons specific to internal or external knowledge.

4.1.2 Knowledge Guiding with Knowledge-Specific Neuron. After identifying the relevant neurons, we can verify the accuracy of our selection and their functional roles by deactivating the neurons associated with both internal and external knowledge.

Neuron Deactivation Experiments. To conduct the deactivation experiment, we first randomly selected retrieval results. We then measured the change in perplexity during the generation process when different neurons were deactivated. Specifically, we set the activation values of internal and external knowledge neurons to zero, and calculated the perplexity of generating internal and external knowledge separately. We identify the knowledge-specific neurons on the Natural Questions [16] dataset, and further conduct our experiments on both single-hop QA datasets (Natural Questions and TriviaQA [14]) and multi-hop QA dataset (HotpotQA [41]).

Main Results. The results illustrating the impact of neuron activation manipulation on knowledge utilization are shown in Table 2. It can be observed that:

- **Firstly, the results demonstrate that deactivating knowledgespecific neurons effectively controls the expression of internal and external knowledge.** Specifically, with gold documents, deactivating external knowledge neurons causes a greater decline in factual accuracy compared to internal knowledge neurons. In contrast, with noisy documents, deactivating external knowledge neurons reduces the influence of irrelevant content, whereas deactivating internal knowledge neurons makes the output less accurate.

Document	Setting	Natural Questions			TriviaQA			HotpotQA			Average		
		EM	CEM	F1	EM	CEM	F1	EM	CEM	F1	EM	CEM	F1
LLaMA-3-8B													
Closed-Book	Normal	22.41	28.45	33.53	58.62	63.22	62.70	16.95	19.21	25.29	32.66	36.96	40.51
Gold Document	Normal	42.24	49.14	56.19	78.74	85.06	83.64	46.89	55.93	59.57	55.96	63.38	66.47
	Deactivate *IK*	39.66	43.10	52.49	68.97	75.29	77.39	39.55	48.02	51.81	49.39(-6.57)	55.47(-7.91)	60.56(-5.91)
	Deactivate *EK*	37.93	42.24	52.82	63.22	65.52	70.97	28.81	36.16	41.55	43.32(-12.64)	47.97(-15.41)	55.11(-11.36)
Noisy Document	Normal	9.48	12.07	15.37	35.06	39.08	40.20	5.71	8.00	14.08	16.75	19.72	23.22
	Deactivate *IK*	6.03	9.48	12.86	30.46	35.63	35.46	3.43	4.57	10.92	13.31(-3.44)	16.56(-3.16)	19.75(-3.47)
	Deactivate *EK*	10.93	15.52	17.75	36.78	37.93	39.73	6.43	8.57	14.52	18.05(+1.30)	20.67(+0.95)	24.00(+0.78)
Qwen-2.5-1.5B													
Closed-Book	Normal	13.39	18.75	21.68	21.30	22.22	23.18	7.65	9.41	14.57	14.11	16.79	19.81
Gold Document	Normal	41.07	41.07	51.11	73.15	73.15	77.74	36.47	41.76	49.88	50.23	51.99	59.58
	Deactivate *IK*	39.29	40.18	49.83	70.37	70.37	73.77	26.47	30.00	37.15	45.38(-4.85)	46.85(-5.14)	53.58(-6.00)
	Deactivate *EK*	37.50	39.29	49.16	68.52	69.44	72.93	18.82	21.76	27.43	41.61(-8.62)	43.50(-8.49)	49.84(-9.74)
Noisy Document	Normal	4.46	5.36	11.13	12.04	12.04	13.33	2.01	2.01	6.75	6.17	6.47	10.40
	Deactivate *IK*	2.68	3.57	8.23	10.19	10.19	10.19	1.51	2.51	4.53	4.79(-1.38)	5.42(-1.05)	7.65(-2.75)
	Deactivate *EK*	4.46	7.04	11.98	15.44	12.37	14.35	2.51	4.07	5.31	8.80(+1.63)	7.83(+1.36)	10.55(+0.15)

Table 2: Knowledge-specific neuron deactivation results on three QA datasets: Natural Questions, TriviaQA, and HotpotQA, using two LLMs (LLaMA-3-8B and Qwen-2.5-1.5B) across different document settings. Metrics include EM [18], CEM [21], and F1 scores [15], with average results reported. Results reflect the effect of deactivating IK and EK Neurons.

- **Secondly, our experiments confirm the generalizability of our approach.** Knowledge-specific neurons identified on Natural Questions were tested on other datasets, including TriviaQA and HotpotQA. The results show consistent effectiveness across single-hop and multi-hop tasks. We also observed similar knowledge utilization trends on all datasets. Furthermore, experiments with LLaMA and Qwen models showed consistent patterns, proving the method works across different model families.
- **Thirdly, the impact of neuron editing on factual accuracy is influenced by question difficulty and retrieval quality.** When questions are less challenging, deactivating internal neurons has a greater effect on performance. Conversely, when the quality of retrieved documents is high, deactivating external neurons results in lower accuracy.

These findings demonstrate that manipulating neuron activation shifts the LLM's reliance between internal and external knowledge, providing valuable insights for building controllable RAG systems.

Case Study. We further demonstrate how knowledge-specific neuron deactivation impacts the tendency of knowledge utilization with two cases, as shown in Figure 6. In the first case. The passage provided contains an incorrect answer. However, the LLM is initially misled by it, maybe for the reason that the passage is relevant. By deactivating neurons associated with external knowledge, the LLM's dependence shifts towards its internal knowledge, which correctly adjusts the response. In the second case, the external passage correctly answers the query, but the LLM fails to leverage this external information at first. Deactivating neurons linked to internal knowledge enhances the LLM's ability to utilize external data, leading to a successful integration of the correct answer.

Question

when did they start vaccinating for whooping cough

Retrieved Document

… Salk went on CBS radio to report a successful test on a small group of adults and children in **1953** …

RA Output *(w/ external)*	**Deactivate EK Neurons**
1953 ✗	1940s ✓

Question

how many episodes in series 7 of game of thrones are there

Retrieved Document

… series: Game of Thrones; season: 7; episode: 3/7; director: Mark Mylod ; writer: David Benioff …

RA Output *(w/ internal)*	**Deactivate IK Neurons**
73 episodes ✗	7 episodes ✓

Figure 6: Cases for the LLM's knowledge utilization tendencies using knowledge-specific neuron deactivation.

4.2 LLM modules Contribute to Internal and External Knowledge Formation

In this part, we further analyze the roles of the two main modules in Transformers in knowledge utilization. Specifically, we examine how MHA and MLP influence the selection between internal and external knowledge. Additionally, we explore how the LLM verifies the consistency between its internal knowledge and external information, and how it handles noisy information with factual errors.

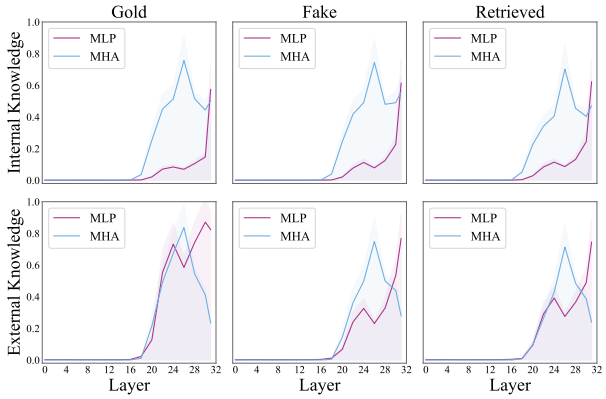

Figure 7: Unembedded logits of internal and external knowledge decoded at each layer with gold or fake passages as the external knowledge sources.

4.2.1 Analysis Methodologies. To study from a modular perspective, we first introduce methodologies for residual steam and early decoding analysis. Using these approaches, we can delve deeper into the influence of each module on the final answer generation.

Residual Stream. To analyze the contribution of each module within Transformer-based models, we consider them as a series of residual stream [8, 20]. Each module in these models adds newly processed information into the stream through residual connections, and the sum of these contributions forms the final output. This setup isolates each module's impact on the hidden state. It reveals their roles in generating the final answer.

Early Decoding. For interpretability, the early decoding strategy projects each module's incremental updates onto a human-readable vocabulary space [20]. Specifically, before layer normalization and linear transformation, the updates are mapped as follows:

$$[\text{logit}_1^l, ..., \text{logit}_{|V|}^l] = \boldsymbol{W}^U \cdot \text{LayerNorm}(\boldsymbol{h}_l), \tag{10}$$

where $\boldsymbol{h_l} \in \mathbb{R}^d$ denotes the hidden state in the layer l, $\boldsymbol{W^U} \in \mathbb{R}^{|V| \times d}$ denotes the unembedding metric, LayerNorm denotes the pre-unembedding layer normalization, and V denotes the token vocabulary of the model. This mapping generates logits at intermediate layer l, serving as an early-exit mechanism [35]. These logits measure the model's predictions at each layer, offering insight into the contribution of individual modules to the output.

4.2.2 Experiments and Analysis. We assess the unembedding logits of each layer's MLP and MHA modules within the LLM to generate internal or external knowledge under RAG settings. Figure 7 shows the unembedding logits of internal or external knowledge using gold passages, fake passages and randomly selected retrieved passages (explained in Section 2.2) as external knowledge sources. The results are shown in Figure 7.

First, we observe the contestation between internal and external knowledge in the deeper layers. Our experiment reveals that both types of knowledge begin to manifest in the middle layers of the residual stream. This indicates that the parameters of the MLP and MHA modules start to contribute significantly to the formation of internal and external knowledge at this stage. This corroborates our delineation of the knowledge expression and contestation stages as described in Section 3.2.2 of the RAG LLMs.

Moreover, we observe that the MLP is particularly sensitive to the correctness of external knowledge. By comparing the unembedding logits between the results of gold passages and fake passages, we find a notable decline in the MLP's contribution to the formation of external knowledge when the source is switched from the gold passage to the fake passage. Given that the only difference between the gold and fake passages lies in the core short answer, this suggests that the MLP is highly sensitive to the fine-grained accuracy of external knowledge. In contrast, no significant change in logits is observed in the MHA layers under similar conditions.

Furthermore, we validate the role of the MLP in selecting between different knowledge sources, corroborating the rationale behind our neuron deactivation strategy. We observe that the logits of internal knowledge show negligible differences across various external documents. However, the contribution of the MLP layer varies significantly. By deactivating the neurons responsible for these variations, we can alter the model's tendency to utilize internal versus external knowledge. This provides further insight into designing controllable RAG systems.

5 Conclusion

In this study, we present a novel investigation of LLM's knowledge utilization on how to integrate internal (parametric) and external (non-parametric) knowledge within RAG frameworks. We analyze from two perspectives: macroscopic knowledge streaming and microscopic module functions. At the macroscopic level, we identify four distinct stages that show how internal and external knowledge are generated and interact with each other. Our empirical findings demonstrate that the relevance of retrieved evidence is central to steering the knowledge elicitation stage. At the microscopic level, we introduce the method of knowledge activation probability entropy (KAPE), to identify knowledge-specific neurons. Leveraging this metric, we show that selectively deactivating these neurons effectively modulates the LLM's reliance on each knowledge type. Further analyses highlight the complementary roles of the multi-head attention (MHA) and multi-layer perceptron (MLP) modules. The MHA module facilitates the integration of information from multiple sources. Meanwhile, the MLP layers help ensure factual consistency in the final output. These insights not only advance the interpretability of RAG systems, but also pave the way for designing more efficient and controllable LLM architectures capable of balancing parametric and contextual knowledge.

Future work will focus on refining these mechanisms to improve real-world applications and extend the boundaries of AI-driven knowledge retrieval systems, developing more nuanced models that better handle the complexities of knowledge integration.

Acknowledgments

This work was partially supported by National Natural Science Foundation of China under Grant No. 92470205 and 62222215, Beijing Municipal Science and Technology Project under Grant No. Z231100010323009, and Beijing Natural Science Foundation under Grant No. L233008.

References

[1] Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? *arXiv preprint arXiv:2010.05607* (2020).

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[3] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

[4] Xiaoxue Cheng, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2025. Think More, Hallucinate Less: Mitigating Hallucinations via Dual Process of Fast and Slow Thinking. *arXiv preprint arXiv:2501.01306* (2025).

[5] Xiaoxue Cheng, Junyi Li, Wayne Xin Zhao, Hongzhi Zhang, Fuzheng Zhang, Di Zhang, Kun Gai, and Ji-Rong Wen. 2024. Small Agent Can Also Rock! Empowering Small Language Models as Hallucination Detector. *arXiv preprint arXiv:2406.11277* (2024).

[6] Zican Dong, Junyi Li, Jinhao Jiang, Mingyu Xu, Wayne Xin Zhao, Bingning Wang, and Weipeng Chen. 2025. LongReD: Mitigating Short-Text Degradation of Long-Context Large Language Models via Restoration Distillation. *arXiv preprint arXiv:2502.07365* (2025).

[7] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[8] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread* (2021). https://transformer-circuits.pub/2021/framework/index.html.

[9] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997* (2023).

[10] Michael Hanna, Ollie Liu, and Alexandre Variengien. 2024. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *Advances in Neural Information Processing Systems* 36 (2024).

[11] Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685* (2020).

[12] Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Qiuxia Li, and Jun Zhao. 2024. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. *arXiv preprint arXiv:2402.14409* (2024).

[13] Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. *arXiv preprint arXiv:2402.18154* (2024).

[14] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1601–1611.

[15] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

[16] Tom Kwiatkowski, Jennimaria Palomaki, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, et al. [n. d.]. Natural Questions: a Benchmark for Question Answering Research. ([n. d.]).

[17] Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G Rodriques, and Andrew D White. 2023. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559* (2023).

[18] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 6086–6096.

[19] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[20] Ang Lv, Kaiyi Zhang, Yuhan Chen, Yulong Wang, Lifeng Liu, Ji-Rong Wen, Jian Xie, and Rui Yan. 2024. Interpreting Key Mechanisms of Factual Recall in Transformer-Based Language Models. *arXiv preprint arXiv:2403.19521* (2024).

[21] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 9802–9822.

[22] Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *Advances in neural information processing systems* 32 (2019).

[23] Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*. 807–814.

[24] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. 2020. KILT: a benchmark for knowledge intensive language tasks. *arXiv preprint arXiv:2009.02252* (2020).

[25] Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 5835–5847.

[26] Ruiyang Ren, Shangwen Lv, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. PAIR: Leveraging Passage-Centric Similarity Relation for Improving Dense Passage Retrieval. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2173–2183.

[27] Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. RocketQAv2: A Joint Training Method for Dense Passage Retrieval and Passage Re-ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2825–2835.

[28] Ruiyang Ren, Yuhao Wang, Junyi Li, Jinhao Jiang, Wayne Xin Zhao, Wenjie Wang, and Tat-Seng Chua. 2025. Holistically Guided Monte Carlo Tree Search for Intricate Information Seeking. *arXiv preprint arXiv:2502.04751* (2025).

[29] 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. *arXiv preprint arXiv:2307.11019* (2023).

[30] Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202* (2020).

[31] Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-Specific Neurons: The Key to Multilingual Capabilities in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 5701–5715. https://doi.org/10.18653/v1/2024.acl-long.309

[32] Xinyu Tang, Xiaolei Wang, Zhihao Lv, Yingqian Min, Wayne Xin Zhao, Binbin Hu, Ziqi Liu, and Zhiqiang Zhang. 2025. Unlocking General Long Chain-of-Thought Reasoning Capabilities of Large Language Models via Representation Engineering. *arXiv preprint arXiv:2503.11314* (2025).

[33] Xinyu Tang, Xiaolei Wang, Wayne Xin Zhao, Siyuan Lu, Yaliang Li, and Ji-Rong Wen. 2024. Unleashing the Potential of Large Language Models as Prompt Optimizers: Analogical Analysis with Gradient-based Model Optimizers. *arXiv preprint arXiv:2402.17564* (2024).

[34] Xinyu Tang, Xiaolei Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2024. Dawn-icl: Strategic planning of problem-solving trajectories for zero-shot in-context learning. *arXiv preprint arXiv:2410.20215* (2024).

[35] Surat Teerapittayanon, Bradley McDanel, and Hsiang-Tsung Kung. 2016. Branchynet: Fast inference via early exiting from deep neural networks. In *2016 23rd international conference on pattern recognition (ICPR)*. IEEE, 2464–2469.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. [n. d.]. Attention Is All You Need. ([n. d.]).

[37] Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Label words are anchors: An information flow perspective for understanding in-context learning. *arXiv preprint arXiv:2305.14160* (2023).

[38] Yuhao Wang, Ruiyang Ren, Junyi Li, Wayne Xin Zhao, Jing Liu, and Ji-Rong Wen. 2024. REAR: A Relevance-Aware Retrieval-Augmented Framework for Open-Domain Question Answering. *arXiv preprint arXiv:2402.17497* (2024).

[39] Jun Xiao, Hao Ye, Xiangnan He, Hanwang Zhang, Fei Wu, and Tat-Seng Chua. 2017. Attentional factorization machines: Learning the weight of feature interactions via attention networks. *arXiv preprint arXiv:1708.04617* (2017).

[40] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671* (2024).

[41] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2369–2380.

[42] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).