The Atlas of In-Context Learning: How Attention Heads Shape In-Context Retrieval Augmentation

Patrick Kahardipraja[1,*] **Reduan Achtibat**[1,*] **Thomas Wiegand**[1,2,3]

Wojciech Samek[1,2,3,†] **Sebastian Lapuschkin**[1,4,†]

[1]Department of Artificial Intelligence, Fraunhofer Heinrich Hertz Institute
[2]Department of Electrical Engineering and Computer Science, Technische Universität Berlin
[3]BIFOLD - Berlin Institute for the Foundations of Learning and Data
[4]Centre of eXplainable Artificial Intelligence, Technological University Dublin
{firstname.lastname}@hhi.fraunhofer.de

Abstract

Large language models are able to exploit in-context learning to access external knowledge beyond their training data through retrieval-augmentation. While promising, its inner workings remain unclear. In this work, we shed light on the mechanism of in-context retrieval augmentation for question answering by viewing a prompt as a composition of informational components. We propose an attribution-based method to identify specialized attention heads, revealing in-context heads that comprehend instructions and retrieve relevant contextual information, and parametric heads that store entities' relational knowledge. To better understand their roles, we extract function vectors and modify their attention weights to show how they can influence the answer generation process. Finally, we leverage the gained insights to trace the sources of knowledge used during inference, paving the way towards more safe and transparent language models.

1 Introduction

Many, if not most language tasks can be framed as a sequence to sequence problem [61, 69]. This view is integral to how modern Large Language Models (LLMs) operate, as they are able to approximate relations between an input and an output sequence not only as a continuation of text, but also as a response to a stimulus [64]. In a sense, input prompts serve as a query to search and induce function(s) in a vast, high-dimensional latent space, where the corresponding process can be cast as question answering [50] or instruction following [56, 80].

This capability is brought forth with the introduction of in-context learning (ICL) [12] that enables LLMs to adapt to new tasks with few demonstrations at inference time, without additional fine-tuning. Previous work has investigated ICL from various perspectives, including its relation to induction heads that can replicate token patterns during prediction [54], the ability to compress attention heads to function vectors (FVs) representing a specific task [31, 74], and how it can emerge when transformers [75] implicitly learn to perform gradient-based optimization [4, 77]. Besides meta-learning, ICL can be used for *retrieval-augmentation* [62], where external knowledge from web retrieval corpora [11, 29, 33] or dialogue information [66, 88, 92] is given instead of input-output pairs to ground LLMs during generation. However, the mechanism behind ICL for knowledge retrieval is not yet fully understood. In this work, we aim to shed light on this question.

[*]Equal contribution.
[†]Corresponding authors.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

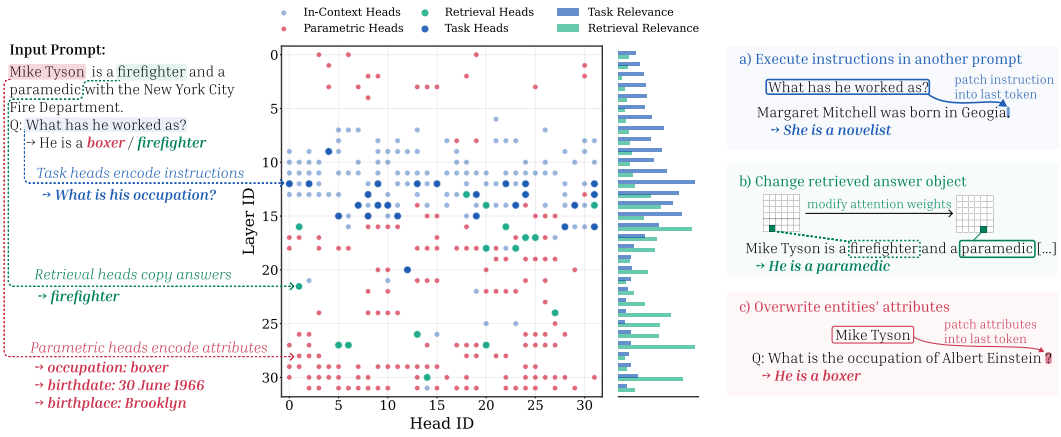


Figure 1: Functional map of in-context and parametric heads in Llama-3.1-8B-Instruct. They are surprisingly well-structured and operate on the input prompt at various levels, with in-context heads processing information in the prompt, including instruction comprehension and retrieval operations — and parametric heads that encode relational knowledge. In-context heads can specialize to task heads to parse instructions (blue) or retrieval heads for verbatim copying (green). Together with parametric heads, they affect the answer generation process through function vectors that they transport (**a**, **c**) or their attention weights (**b**). Our relevance analysis (bar plot) shows that instruction-following capabilities emerge in middle layers, while answer retrieval occurs in later layers. Details in C.1.

On question answering tasks, we show that by viewing a prompt as a composition of informational components, certain attention heads perform various operations on the prompt at different stages of inference and layers (Figure 1). Our method identifies two groups of heads based on their functions: *parametric heads* that encode relational knowledge [27, 57] and *in-context heads* responsible for processing information in the prompt. Further, as in-context heads need to understand which prompt components to process and how, we hypothesize that they specialize to fill their respective roles.

Our analysis shows that in-context heads can indeed execute specialized functions such as instruction comprehension and retrieval of relevant contextual information. To investigate this further, we curate a controlled biography dataset with entities extracted from Wikidata [78]. Remarkably, we find that through compressing them to FVs or modifying their weights, both in-context and parametric heads can induce specific, targeted functions.

Building on these insights, we probe for sources of knowledge used during retrieval-augmented question answering and show where it is localized within the input context. Our attempt shows promising results, serving as an initial step towards more transparent retrieval-augmented generation (RAG) systems. Overall, our contributions can be summarized as follows:

- We describe an attribution-based method to determine attention heads that play a key role during in-context retrieval augmentation for question answering, revealing that they operate on the prompt in a distinct manner. Our method can thus be used to reliably categorize attention heads into in-context and parametric heads.
- We analyze how in-context heads specialize in reading instructions and retrieving information, mapping their location across model layers. Additionally, we demonstrate the influence of in-context and parametric heads on the answer generation process, by compacting them into function vectors or modifying their attention weights.
- We present preliminary results on enabling causal tracing of source information for retrieval-augmented LMs, suggesting fruitful directions for interpretability of RAG systems.

2 Related Work

Retrieval Augmentation Retrieval-augmented language models (RALMs) [40, 45] address inherent limitations of LMs by providing external knowledge sources. Despite this advantage, issues such as discrepancies between contextual and parametric knowledge may occur [48, 87]. Some

works have studied mechanisms behind knowledge preference in RALMs [51, 55, 91], but they focus on simple domains. On the other hand, those that explored more complex domains mostly only analyzed RALMs' behavior at the output level [14, 71, 85]. Closely related to our work is that of Jin et al. [37], where the authors also discovered in-context and parametric heads. Compared to them, we show that interactions among in-context heads are subtly more complex, since they can specialize into task or retrieval heads. Our approach also allows to demonstrate their roles in shaping the model's representation along with parametric heads, by transforming them into FVs or modifying their respective weights.

Another disadvantage of RALMs is that they cannot guarantee faithful answer attribution[3] to contextual passages [25], which necessitates a shift to interpretability. Recent efforts to address this include leveraging contrastive gradient attribution [60, 89], fitting surrogate models [16], and treating attention weights as features [17]. In relation to this, we train a probe based on retrieval heads to track for knowledge provenance.

The Role of Attention Attention mechanisms have been previously observed to encode many kinds of information. Clark et al. [15] showed that they correspond well to linguistic properties such as syntax and coreference. Similarly, Voita et al. [76] found that syntactic heads play an important role in machine translation models. In relation to world knowledge, Geva et al. [27] proposed that certain heads store factual associations and demonstrated how they extract an attribute of a given subject-relation pair. Interestingly, attention also appears to encode a sense of "truthful" directions [46]. With the exception of Voita et al. [76], the above works make use of attention weights, which might not fully capture the model's prediction [10, 35, 81]. Our work can be seen as an attempt to reconcile both perspectives: analyses based on attention weights and feature attribution methods [9].

In-Context Learning Numerous works have studied ICL since its introduction. Liu et al. [47] studied what constitutes a good example for demonstrations. Dai et al. [20] suggested that ICL can be understood as an implicit fine-tuning. ICL is a general phenomenon, although it is commonly assumed to be unique to autoregressive models [63]. At the component level, ICL is primarily associated with induction heads [21, 54]. However, recent findings showed that certain heads can also be compressed to form FVs that represent a demonstrated task [31, 74]. Yin and Steinhardt [90] investigated the connection between these heads and induction heads, showing that they are distinct from each other and how models' ICL performance is mainly driven by the former. ICL can also be viewed as a mixture of meta-learning and retrieval [52]. In that regard, we study the latter perspective to understand its mechanism as a specific instantiation of ICL, with a focus on the retrieval augmentation paradigm.

3 Background and Preliminaries

The self-attention mechanism in transformers poses a challenge in understanding which heads actually contribute during in-context retrieval augmentation, and how they process various components in the prompt. This is mainly due to the fact that information from different tokens gets increasingly mixed as layers go deeper and how several attention heads may implement redundant functions [79]. A natural option is to analyze attention weights, as they are an inherent part of a model's computation. However, attention can be unfaithful [34], which questions its validity as an explanation [10, 65]. This problem is further exacerbated by "attention sinks" [43, 84] — a phenomenon where heads heavily attend to the first token and obscure the weights of other tokens in the sequence.

An alternative would be to use feature attribution methods [9], as they trace the contribution of each input element to the model's output prediction. Propagation-based feature attribution [8, 67, 68] especially takes the entire computation path into account, which can be used to characterize attention heads [76] or identify latent concepts [1]. Furthermore, feature attribution is able to estimate causal relations [26] *e.g.,* to automate circuit discovery [18, 22, 30, 32, 70], and thus enables to observe how a specific attention head affects a model's prediction.

In this section, we provide a description of AttnLRP [2], on which our method is based, due to its superior performance and efficiency in transformer architectures compared to other attribution methods. We also provide an overview of the multi-head attention mechanism in transformers, which

[3]Here, the term *answer attribution* means the use of external documents to support the generated response, which is different from *feature attribution* used throughout this work to describe interpretability techniques.

we leverage through AttnLRP to identify both in-context and parametric heads (§5). Additionally, we analyze the specialization of in-context heads, show causal roles of the identified heads (§6), and use this information for reliable and efficient source tracking of facts in retrieval-augmented LMs (§7).

3.1 Layer-wise Relevance Propagation

Feature attribution methods aim to quantify the contribution of input features $\mathbf{x}$ to the overall activation of an output y in linear but also highly non-linear systems. We define a function $\mathcal{R}$ that maps the input activations $\mathbf{x}$ to relevance scores indicating their causal effect on a model's output logit y:

$$\mathcal{R} : \mathbb{R}^N \times \mathbb{R} \;\rightarrow\; \mathbb{R}^N, \qquad (\mathbf{x}, y) \;\mapsto\; \mathcal{R}(\mathbf{x} \;\big|\; y).$$

In principle, any feature attribution method can be employed for $\mathcal{R}$, though trade-offs between faithfulness and computational efficiency must be carefully considered. Perturbation-based approaches [49] typically offer high faithfulness but incur exponential computational costs, as each ablated latent component requires at least one dedicated forward pass [30]. In contrast, gradient-based methods [67] are computationally more efficient, requiring only a single backward pass, which makes them well suited for large-scale interpretability studies. However, they are susceptible to noisy gradients, which are distorted by non-linear components such as layer normalization [5, 86]. To address these limitations, we adopt AttnLRP [2], an extension of Layer-wise Relevance Propagation (LRP) [8] designed for transformer architectures. As a backpropagation-based technique, AttnLRP propagates relevance scores from a model output to its inputs in a layer-wise manner and can be implemented efficiently via modified gradient computation in a single backward pass. Importantly, it incorporates stabilization procedures for non-linear operations, thereby improving the faithfulness of relevance distributions compared to standard gradient- or other decomposition-based methods [7].

Relevance scores produced by (Attn)LRP can be either positive or negative. Positive relevance indicates an amplifying effect on the final model logit y, whereas negative relevance reflects an inhibitory influence. Without loss of generality, we focus our analysis on signal-amplifying components by considering only positive relevance scores. Formally, we define:

$$\mathcal{R}^+(\mathbf{x}|y) = \max\left(\mathcal{R}(\mathbf{x}|y), 0\right) \tag{1}$$

This yields a clearer separation between in-context and parametric heads in the subsequent analysis.

3.2 Attention Mechanism

While the original formulation of the multi-head attention mechanism [75] concisely summarizes the parallel computation of attention heads, our goal is to isolate their individual contributions. To this end, we reformulate the equations to make the influence of each head more explicit [21, 23]. Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_S) \in \mathbb{R}^{d \times S}$ denote the matrix of hidden token representations for a sequence of length S with dimension d, and suppose our model employs H parallel heads, each of dimension $d_h = d/H$. Then, the computation of the multi-head attention layer can be reformulated into H complementary operations, where each head h produces an intermediate attention output $\mathbf{z}_i^h \in \mathbb{R}^{d_h}$:

$$\mathbf{z}_i^h \;=\; \sum_{j=1}^{S} \mathbf{A}_{i,j}^h \left(\mathbf{W}_V^h \, \mathbf{x}_j\right) \tag{2}$$

where $\mathbf{A}_{i,j}^h$ is the attention weight of token i attending to token j, and $\mathbf{W}_V^h \in \mathbb{R}^{d_h \times d}$ is the per-head value projection. The final output is obtained by multiplying the intermediate output of each head with their corresponding output projection matrix $\mathbf{W}_O^h \in \mathbb{R}^{d \times d_h}$, followed by summing:

$$\hat{\mathbf{x}}_i \;=\; \sum_{h=1}^{H} \mathbf{W}_O^h \, \mathbf{z}_i^h \tag{3}$$

We leverage the multi-head attention mechanism in transformers through the lens of AttnLRP to identify both in-context and parametric heads in §5 and how in-context heads specialize in §6.

4 Experimental Setup

Models We use instruction-tuned LLMs due to their increased capability on question answering (QA) tasks in our preliminary experiments: Llama-3.1-8B-Instruct [28], Mistral-7B-Instruct-v0.3 [36],

and Gemma-2-9B-it [72]. We apply AttnLRP based on their `huggingface` implementations [82]. For the rest of this work, we refer to each model by their family prefix.

Datasets To perform our analyses, we use two popular open-domain QA datasets: NQ-Swap [48] and TriviaQA (TQA) [38]. NQ-Swap is derived from Natural Questions [44], a QA dataset with questions collected from real Google search queries and answer spans annotated in Wikipedia articles. TQA contains trivia questions with answers sourced from the web. Both datasets are based on the MRQA 2019 shared task version [24].

Similar to Petroni et al. [58], we consider different types of contextual information to see how they affect in-context and parametric heads. We use oracle contexts as they are always relevant to the question and contain the true answer. In addition, we use counterfactual contexts as they contain information that is usually not seen during pre-training and fine-tuning stages, thus forcing models to rely on the text to answer the question correctly. Oracle context is often not available; therefore we also use Dense Passage Retriever (DPR) [39] with a Wikipedia dump from December 2018 as our retrieval corpus. For simplicity, we only select the top one retrieved document. We show results for oracle and counterfactual contexts in the main paper and retrieved DPR contexts in Appendix B.

Inspired by Allen-Zhu and Li [6], we build a human biography datasets to allow us to better understand the characteristic of in-context and parametric heads and conduct controlled experiments. Using Wikidata [78], we collect profiles for random 4,255 notable individuals containing their date of birth, birth place, educational institute attended, and occupation. We concatenate the attributes of each individual in a random order to form a biographical entry and ask Llama-3.1-8B-Instruct to paraphrase it. See Appendix A for more details.[4]

5 Localization of In-Context and Parametric Heads

In retrieval-augmented generation, LLMs are faced with the option to generate responses by using a wealth of knowledge they learn during pre-training or by relying on contextual information provided in the prompt through ICL. Here, we categorize attention heads that are responsible for both capabilities.

Method We aim to identify the sets of in-context heads $\mathcal{H}_{\mathrm{ctx}}$ and parametric heads $\mathcal{H}_{\mathrm{param}}$ as depicted in Figure 1. We define in-context heads as those that mainly contribute to the model's prediction during RAG by using contextual information, whereas parametric heads primarily contribute upon reliance on internal knowledge. We hypothesize that each head type contributes maximally under a specific condition while having minimal influence in others, *i.e.,* in-context heads are expected to contribute the most in open-book settings and the least in closed-book settings, and vice versa. We analyze questions with *counterfactual* contexts, forcing retrieval to produce a counterfactual prediction y_{cf} that disagrees with the parametric answer. Conversely, we also focus on closed-book settings where contextual information is minimized, to identify parametric heads and reduce the chance that in-context heads contribute. We restrict our analysis to instances where a gold reference answer y_{gold} is predicted, to ensure that relevance attribution reflects genuine parametric behavior.

We use AttnLRP to quantify the contribution of each attention head h to the prediction by summing the positive relevance scores assigned to its latent output $\mathbf{z}^h$ across its dimension d_h and over all i token positions, when explaining the targeted prediction y_t, which can be either a gold reference answer y_{gold} or a counterfactual output y_{cf}, depending on the setting:

$$\mathcal{R}^h(y_t) = \sum_{i=1}^{S} \sum_{k=1}^{d_h} \mathcal{R}^+(\mathbf{z}_i^h \mid y_t)_k \in \mathbb{R}. \tag{4}$$

To contrast heads across settings, we compute a difference score $\mathcal{D}$ representing their average contribution in open-book versus closed-book conditions for all N_h heads in the model:

$$\mathcal{D} = \left\{ \mathbb{E}_{X_{\text{OB}}} \left[\mathcal{R}^h(y_{\text{cf}}) \right] - \mathbb{E}_{X_{\text{CB}}} \left[\mathcal{R}^h(y_{\text{gold}}) \right] \ : \ h = 1, 2, \ldots, N_h \right\} \tag{5}$$

We then identify the most distinctive heads for each behavior by selecting the top 100 heads (around 10%-15% of total heads) with the highest positive and lowest negative difference scores:

$$\mathcal{H}_{\text{ctx}} = \{\text{argsort}_{\text{desc}}(\mathcal{D})\}_{n=1}^{100}, \qquad \mathcal{H}_{\text{param}} = \{\text{argsort}_{\text{asc}}(\mathcal{D})\}_{n=1}^{100} \tag{6}$$

[4] Our implementation is publicly available at `https://github.com/pkhdipraja/in-context-atlas`

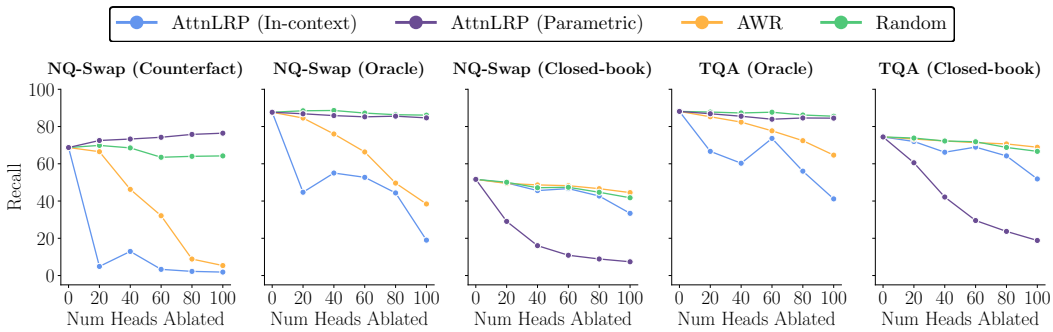


Figure 2: Recall analysis for Llama 3.1 when either in-context or parametric heads are ablated. Removing identified in-context heads noticeably affects the model's performance in open-book QA across various configurations. Conversely, removal of identified parametric heads most strongly affects the model's closed-book QA capabilities. Compared to Wu et al. [83] that only yield AWR (retrieval) heads, our method allows to obtain both in-context and parametric heads.

Experiments To ensure that the identified in-context and parametric heads play a role in QA tasks, we ablate both sets of heads and measure performance drops in settings where they are expected to be mostly active (open- and closed-book, respectively). We also measure if the removal of in-context heads affects the models' capabilities to answer in closed-book setting and vice versa, since this informs to what extent both sets of heads are *dependent* on each other. Furthermore, we want to know if the identified in-context and parametric heads can *generalize* to other datasets. To test this, we compute the score of both heads only over NQ-Swap and reuse the same sets of heads on TQA. To evaluate the aforementioned criteria, we report recall [3] as instruction-tuned models tend to produce verbose responses. As baselines, we select random heads, and also adopt the Attention Weight Recall (AWR) method based on attention maps' activations, as described in [83].

Results We show results for Llama 3.1 here and other models in Appendix B. Figure 2 shows how the recall score evolves when either heads identified as in-context or parametric heads are ablated. We observe that the removal of 20 heads (100 heads) reduces the performance by 13.86%-63.84% (44.26%-68.66%) for open- and closed-book settings across different configurations, indicating the causal influence these heads have on the answers' correctness. Moreover, the performance drops on TQA hold even though the heads are computed on NQ-Swap, showing that the identified in-context and parametric heads are transferrable to other datasets.

We compare in-context heads as identified with our method against AWR heads, and find that the removal of 20 in-context heads results in a roughly similar reduction of recall as removing 100 AWR heads. Furthermore, ablating in-context heads yield a more drastic performance decrease compared to the removal of AWR heads, suggesting that our method is more suitable than those based on attention scores alone to study heads that contribute to response generation. Ablating randomly chosen heads barely affects the model's ability to answer correctly.

We examine whether in-context and parametric heads are independent of each other. As expected, ablating parametric heads has little influence to the model's performance in our open-book setting. Interestingly, this leads to a slight performance increase on NQ-Swap with counterfactual contexts, which suggest that the ablation forces the model to rely more on the given context instead of its own parametric knowledge. Surprisingly, ablating in-context heads in the closed-book setting incurs a non-negligible performance reduction. This is likely due to the influence in-context heads have when processing the input prompt. We explore this in §6.

5.1 Resolving Knowledge Conflicts

We further test whether targeted ablation of parametric and in-context heads also helps to resolve conflicts between contextual and parametric knowledge sources. On NQ-Swap with counterfactual contexts, we selectively ablate heads from each category. Removing in-context heads forces the model to rely more heavily on parametric memory, leading to substantial gains in recall with respect to the original, non-counterfactual answers. Conversely, ablating parametric heads promotes stronger

grounding in the provided context as shown before, improving recall for counterfactual open-book answers. We present the full results in Appendix E.

6 Functional Roles of In-Context and Parametric Heads

Given that ablating in-context heads yields a non-negligible drop in closed-book QA performance, where no external documents are available, we posit that in-context heads not only process the context but also interpret the *intensional frame* – the semantic structure imposed by the instruction itself [64]. In the counterfactual example below, the intensional frame (the question prompt) is shown in *italics*, the object instance in **bold**, and two equally plausible answers in color:

> "[Mike Tyson was a **firefighter** from 1980 to 1984 with the New York City Fire Department. . .] Q: *What has Mike Tyson worked as?* A: boxer / firefighter"

To answer correctly, the model must map the intensional frame onto the knowledge triple (s, r, o^*), where s is the *subject* ("Mike Tyson"), r is the *semantic relation* (the predicate specified by the question, here "has worked as"), and o^* is the *object* (the yet to be determined answer, "boxer" or "firefighter"). Depending on where the answer resides, o^* may be retrieved from the model's parametric memory (o^p) or from the context (o^c). By treating (s, r, o^*) as the complete task specification, we analyze how in-context and parametric heads specialize both to comprehend the intensional frame and to retrieve the object o^* needed to generate the correct answer.

6.1 Disentangling the Functionality of In-Context Heads

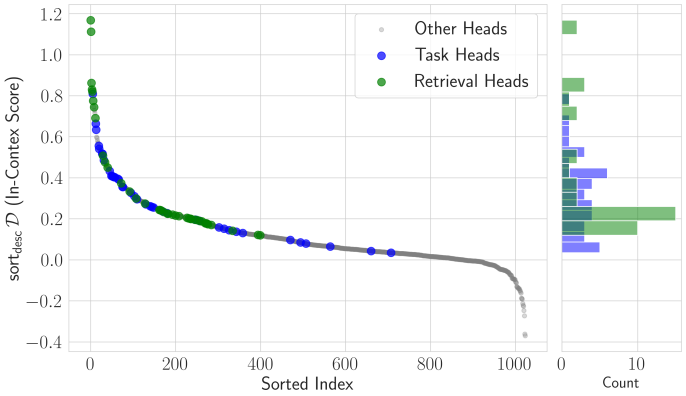


Figure 3: Sorted in-context scores for 1024 heads of Llama 3.1, comparing open-book and closed-book settings via score $\mathcal{D}$. Positive scores indicate in-context behavior, while negative scores reflect parametric behavior. Retrieval heads (green) and task heads (blue) are predominantly high-scoring in-context heads. See Appendix Figure 7 for other models.

Our goal is to identify heads specialized in processing the *intensional frame* and those specialized in retrieving the *answer object* from the context. Inspired by the work of [1], which demonstrates that relevance is effective for separating functional components in latent space, we measure how much relevance of an attention head is assigned to the question and retrieved answer tokens.

Method For each head h, we compute the total relevance attributed to the attention weight $A^h_{i,j}$ when explaining the logit output y_t. Since relevance flows backwards from the output to the input, our goal is to obtain relevance at the input level of each layer. Given that each head transfers information from the *key* at position j to the *query* at position i, we aggregate this backward relevance over all possible query positions i to obtain a single relevance score for the source token at key position j:

$$\rho^h_j = \sum_{i=1}^{S} \mathcal{R}^+\left(A^h_{i,j} \mid y_t\right) \tag{7}$$

Here, ρ^h_j represents the total relevance assigned at head h to token j when contributing to logit y_t. Next, we aggregate the relevance scores separately for two sets of token positions within the context:

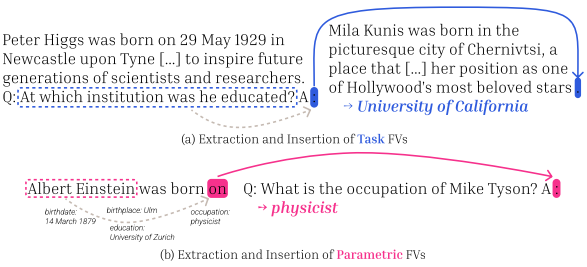


Figure 4: Extraction and insertion of task and parametric FVs. The induced generation is highlighted in italic.

Table 1: Zero-shot recall scores for task, parametric, and retrieval heads.

Models	$\mathcal{H}^{40}_{\text{task}}$	$\mathcal{H}^{50}_{\text{param}}$	$\mathcal{H}^{40}_{\text{ret}}$
Llama 3.1 (random)	18.00	6.68	15.94
+ FVs / Attn Weight	**94.75**	**38.84**	**93.45**
Mistral v0.3 (random)	9.50	12.95	8.56
+ FVs / Attn Weight	**88.50**	**44.04**	**97.03**
Gemma 2 (random)	7.50	6.79	3.89
+ FVs / Attn Weight	**88.00**	**34.77**	**87.36**

the intensional-frame tokens, denoted as $j \in J_{\text{task}}$, which comprise the question token positions, and the answer object tokens, denoted as $j \in J_{\text{ret}}$, which represent positions of the retrieved object.

$$\rho^h_{\text{task}} = \sum_{j \in J_{\text{task}}} \rho^h_j, \quad \rho^h_{\text{ret}} = \sum_{j \in J_{\text{ret}}} \rho^h_j. \tag{8}$$

Finally, to obtain the sets of specialized task and retrieval heads, we rank heads by their aggregated relevance and select the top K, a hyperparameter determined separately for each experiment.

$$\mathcal{H}^K_{\text{task}} = \left\{\text{argsort}(\rho^h_{\text{task}})_{\text{desc}}\right\}^K_{n=1}, \quad \mathcal{H}^K_{\text{ret}} = \left\{\text{argsort}(\rho^h_{\text{ret}})_{\text{desc}}\right\}^K_{n=1}. \tag{9}$$

Results We compute the task relevance score ρ^h_{task} and the retrieval relevance score ρ^h_{ret} over NQ-Swap with counterfactual contexts to minimize influences of parametric heads, and aggregate their distributions across the model layers. In Figure 1, we observe that ρ^h_{task} initially increases in the early layers where few parametric heads are located, suggesting that early parametric heads enrich the question with relational knowledge. The relevance peaks in the middle layers, where in-context heads dominate, aligning with the transition to a more context-dependent reasoning. In contrast, the retrieval relevance score ρ^h_{ret} peaks in deeper layers, reflecting the point where the model extracts the final answer object o^c. Figure 3 further illustrates the sorted average difference $\mathcal{D}$ between open-book and closed-book settings for all heads, alongside the top 40 task heads $\mathcal{H}_{\text{task}}$ and retrieval heads $\mathcal{H}_{\text{ret}}$. We observe that the highest-scoring in-context heads are primarily composed of retrieval and task heads, emphasizing their role for retrieval augmented generation.

6.2 Causal Effects of In-Context and Parametric Heads

An important question is whether the heads we identify truly reflect their assigned functionalities. We examine this under controlled conditions and investigate their causal effects on the answer generation.

Experiments We conjecture that task heads $\mathcal{H}_{\text{task}}$ encode the intensional frame (s, r, o^*) and that parametric heads $\mathcal{H}_{\text{param}}$ contain information of the subject s, which depending on the training data may or may not include o^p. On the other hand, retrieval heads $\mathcal{H}_{\text{ret}}$ search for o^c, allowing them to copy any tokens from the context verbatim, without being restricted to only plausible answers. For task and parametric heads, we compute FVs[5] for each head and insert them into various settings to trigger the execution of their functions. Following Wu et al. [83], we opt for a needle-in-a-haystack (NIAH) setting for retrieval heads and determine their ability to retrieve relevant information from the context by modifying the attention weights. To isolate these behaviors, we conduct our analysis on the biography dataset (§4) and measure recall [3]. For comparison, we also consider random heads for FV extraction and attention modification. See Appendix C for additional results and details, including the selection of hyperparameter K.

Task Heads We demonstrate that task heads encode intensional frames. In a zero-shot manner, we extract task FVs from each head in $\mathcal{H}_{\text{task}}$ for four questions relating to all recorded attributes from the biography dataset. Then, we insert them to another biographical entry without a question at the final token position, and also for all subsequent token generations (Figure 4, top). We examine whether

[5]A FV can be defined as a sum of averaged heads outputs over a task [74] or computed individually [31]. Following the latter, we consider a FV to be an output of a task or parametric head scaled by scalar α.

they reliably induce responses aligned with the original question. In Table 1 (left), we show that applying FVs in a zero-shot manner allows all models to respond accordingly wrt. the intensional frame, yielding an average improvement of 78.75 points over random heads.

Parametric Heads Parametric heads contain relational knowledge. To show this, we first select a random attribute of an individual and convert it to a cloze-style statement. Then, we extract FVs from $\mathcal{H}_{\text{param}}$, which are inserted to a question prompt of another unrelated individual (Figure 4, bottom). We observe if the generated response contains information of the original entity conditioned on the intensional frame. For simplicity, we restrict extraction to cases where the closed-book answer is correct wrt. gold reference. We see that in Table 1 (middle), adding parametric FVs allows all models to recover the original attributes significantly, with an increase of 30.41 points compared to random.

Retrieval Heads We assess retrieval heads' ability to copy verbatim by using famous poem titles as needles, inserted at a random position in the biographical entries. At the last token of the entry and for the following generations, we increase the attention weights of all heads in $\mathcal{H}_{\text{ret}}$ on every token of the needle to force the model to copy. Our results (Table 1, right) show a drastic increase of 83.15 points over the random baseline, indicating that retrieval heads are indeed able to perform copy operations independent of the token position.

6.2.1 How Does This Generalize to Other In-Context Tasks Beyond QA?

To assess whether our method generalizes beyond question answering, we conduct a preliminary investigation on machine translation using the OPUS Europarl dataset [41, 73]. Directly transferring the task heads identified in the QA experiments proved ineffective, indicating that translation relies on a different functional subspace. Nonetheless, by applying our disentanglement procedure of §6.1 to the intensional frame of a translation prompt, we are able to identify a compact set of 15 attention heads that robustly encode translation behavior independent of source or target language. Patching these heads induces zero-shot translation across multiple language pairs, yielding BLEU scores comparable to explicit prompting whereas patching random heads fails to induce any translation behavior. Complete results and experimental details are provided in Appendix F.

7 Source Tracking

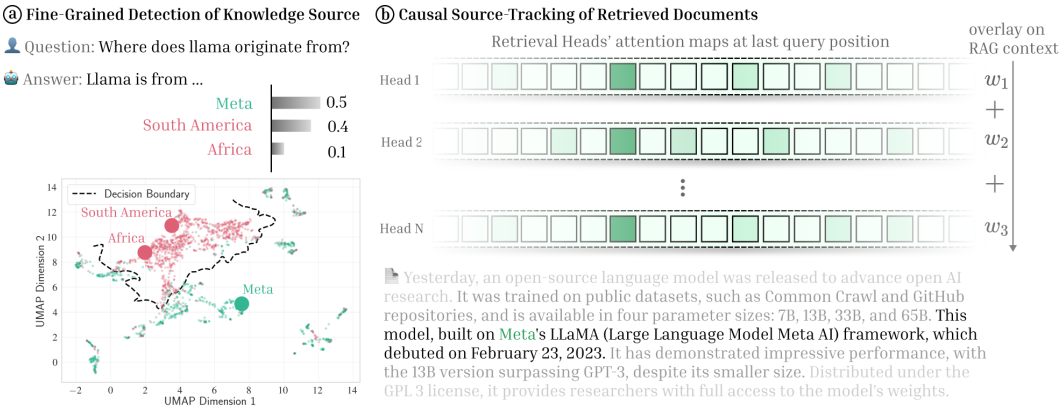


Figure 5: (**a**) When asked "Where does llama originate from?", the retrieval-head probe classifies "South America" and "Africa" as parametric, while "Meta" as contextual. The UMAP projection of retrieval head activations displays the linear probe's decision boundary (dashed line) separating parametric from contextual clusters. (**b**) The weighted aggregation of retrieval head attention maps at the final query position is superimposed on the document to pinpoint the retrieved source span.

Our experiment in §6.2 demonstrates that retrieval heads reliably perform verbatim copying of text spans when their corresponding attention maps focus on the retrieved tokens. As can be seen on Figure 5, we now aim to investigate if we can (i) detect when retrieval heads initiate the copying process for the first answer token (*i.e.,* whether a token is derived from external contexts rather than from the model parameters), and (ii) accurately localize its position within that context using the attention maps. To this end, we train a linear probe on NQ-Swap with counterfactual contexts. Each

retrieval head's output at the last token's position $\mathbf{z}_S^h$ is decoded via logit lens [53], converting each head's activation into a score for token $t \in \mathbb{N}$ using the model's unembedding matrix $\mathbf{W}_U \in \mathbb{R}^{|V| \times d}$ and layer normalization $\mathrm{LN}(\cdot)$:

$$\mathscr{L}(\mathbf{z}_S^h \mid t) \;=\; \mathrm{LN}\big(\mathbf{W}_O^h\, \mathbf{z}_S^h\big)\, \mathbf{W}_U[t] \in \mathbb{R}, \tag{10}$$

where $\mathbf{W}_O^h \in \mathbb{R}^{d \times d}$ is the head's output projection and $\mathbf{W}_U[t]$ the row of the unembedding matrix corresponding to token t. As such, $\mathscr{L}(\mathbf{z}_S^h \mid t)$ computes how strongly head h writes token t into the residual stream. In Appendix Figure 8, we illustrate histograms of the logit lens scores.

Next, we train a probe via linear regression, yielding the weights $\{w_h\}_{h \in \mathcal{H}_{\mathrm{ret}}}$. For source localization, we aggregate each head's attention map using its weight and logit-lens score, and predict the source token index as

$$\hat{k} = \arg\max_j \sum_{h \in \mathcal{H}_{\mathrm{ret}}} w_h\, \mathscr{L}\big(\mathbf{z}_S^h \mid t\big)\, A_{S,j}^h. \tag{11}$$

For details, please refer to Appendix D. Additionally, we use a standard AttnLRP backward pass from the model output to compute an input heatmap as a baseline for comparison.

In Table 2, the retrieval-head probe achieves an ROC AUC of at least 94%, reliably distinguishing contextual from parametric predictions and thus confirms a linearly separable representation of the retrieval task. A promising direction for future research is to leverage the probe's ability to distinguish between parametric and contextual predictions, enabling dynamic control over the model's token selection. This approach could reduce hallucinations by explicitly guiding the model to prioritize context over parametric memory when appropriate. In addition, each model attains a top-1 localization accuracy of at least 84%. In Appendix Figure 9, we illustrate heatmaps of the aggregated attention maps superimposed on the input, highlighting the positions of the predicted tokens. While AttnLRP outperforms the probe, it requires an additional backward pass increasing computational cost, while the probe only requires attention maps computed during the forward pass.

Table 2: Performance of the retrieval-head probe across models.

Models	ROC AUC	Localization	
		Attention	AttnLRP
Llama 3.1	95%	97%	98%
Mistral v0.3	98%	96%	99%
Gemma 2	94%	84%	96%

8 Conclusion

We propose a method to explore the inner workings of ICL for retrieval-augmentation, revealing in-context and parametric heads that operate on the input prompt in a distinct manner and find that in-context heads can specialize into either task or retrieval heads, depending on whether they encode intensional frames or retrieve relevant information. We study the roles of the identified heads by converting them into FVs or modifying their weights, showing how they can affect the generation process. Finally, we present a probe to precisely and efficiently track for knowledge provenance, opening up a path towards more interpretable retrieval-augmented LMs.

Limitations We focus our investigation on attention heads since they are primarily associated with ICL. However, how they interact with components in MLP modules *e.g.,* knowledge neurons [19] to induce functions remains an open question. Our analyses are also mainly centered on QA (with a minor part on MT). It would be interesting to see if similar mechanisms arise in other tasks. We find that several heads exhibit inhibitory effects, and that some intermediate heads are not exclusively specialized for either in-context or parametric roles — both of which warrant further investigation. In addition, there is a possibility for redundant heads [79], which are yet to be uncovered. We leave this avenue for future work.

Broader Impacts Our research enhances trust in retrieval-augmented LMs by elucidating the mechanisms through which they access and use external knowledge. Furthermore, it enables precise source attribution, allowing users to trace the origins of the information leveraged in response generation. However, we caution against its potential for misuse, such as using the identified heads to induce malicious behavior.

Acknowledgements

We thank the anonymous reviewers for their critical reading of our manuscript and their insightful comments and suggestions.

References

[1] R. Achtibat, M. Dreyer, I. Eisenbraun, S. Bosse, T. Wiegand, W. Samek, and S. Lapuschkin. From attribution maps to human-understandable explanations through concept relevance propagation. *Nature Machine Intelligence*, 5(9):1006–1019, Sep 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00711-8. URL https://doi.org/10.1038/s42256-023-00711-8.

[2] R. Achtibat, S. M. V. Hatefi, M. Dreyer, A. Jain, T. Wiegand, S. Lapuschkin, and W. Samek. AttnLRP: Attention-aware layer-wise relevance propagation for transformers. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 135–168. PMLR, 21–27 Jul 2024.

[3] V. Adlakha, P. BehnamGhader, X. H. Lu, N. Meade, and S. Reddy. Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, 12:681–699, 2024. doi: 10.1162/tacl_a_00667. URL https://aclanthology.org/2024.tacl-1.38/.

[4] E. Akyürek, D. Schuurmans, J. Andreas, T. Ma, and D. Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0g0X4H8yN4I.

[5] A. Ali, T. Schnake, O. Eberle, G. Montavon, K.-R. Müller, and L. Wolf. XAI for transformers: Better explanations through conservative propagation. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 435–451. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/ali22a.html.

[6] Z. Allen-Zhu and Y. Li. Physics of language models: Part 3.1, knowledge storage and extraction, 2024. URL https://arxiv.org/abs/2309.14316.

[7] L. Arras, B. Puri, P. Kahardipraja, S. Lapuschkin, and W. Samek. A close look at decomposition-based xai-methods for transformer language models, 2025. URL https://arxiv.org/abs/2502.15886.

[8] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46, 07 2015. doi: 10.1371/journal.pone.0130140. URL https://doi.org/10.1371/journal.pone.0130140.

[9] J. Bastings and K. Filippova. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In A. Alishahi, Y. Belinkov, G. Chrupała, D. Hupkes, Y. Pinter, and H. Sajjad, editors, *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.blackboxnlp-1.14. URL https://aclanthology.org/2020.blackboxnlp-1.14/.

[10] A. Bibal, R. Cardon, D. Alfter, R. Wilkens, X. Wang, T. François, and P. Watrin. Is attention explanation? an introduction to the debate. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3889–3900, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.269. URL https://aclanthology.org/2022.acl-long.269/.

[11] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. B. Van Den Driessche, J.-B. Lespiau, B. Damoc, A. Clark, D. De Las Casas, A. Guy, J. Menick, R. Ring, T. Hennigan, S. Huang, L. Maggiore, C. Jones, A. Cassirer, A. Brock, M. Paganini, G. Irving, O. Vinyals, S. Osindero, K. Simonyan, J. Rae, E. Elsen, and L. Sifre. Improving language models by retrieving from trillions of tokens. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/borgeaud22a.html.

[12] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[13] J. Bulian, C. Buck, W. Gajewski, B. Börschinger, and T. Schuster. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 291–305, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.20. URL https://aclanthology.org/2022.emnlp-main.20/.

[14] H.-T. Chen, M. Zhang, and E. Choi. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.146. URL https://aclanthology.org/2022.emnlp-main.146/.

[15] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning. What does BERT look at? an analysis of BERT's attention. In T. Linzen, G. Chrupała, Y. Belinkov, and D. Hupkes, editors, *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy, Aug. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4828. URL https://aclanthology.org/W19-4828/.

[16] B. Cohen-Wang, H. Shah, K. Georgiev, and A. Mądry. Contextcite: Attributing model generation to context. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 95764–95807. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/adbea136219b64db96a9941e4249a857-Paper-Conference.pdf.

[17] B. Cohen-Wang, Y.-S. Chuang, and A. Madry. Learning to attribute with attention, 2025. URL https://arxiv.org/abs/2504.13752.

[18] A. Conmy, A. Mavor-Parker, A. Lynch, S. Heimersheim, and A. Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 16318–16352. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Paper-Conference.pdf.

[19] D. Dai, L. Dong, Y. Hao, Z. Sui, B. Chang, and F. Wei. Knowledge neurons in pretrained transformers. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.581. URL https://aclanthology.org/2022.acl-long.581/.

[20] D. Dai, Y. Sun, L. Dong, Y. Hao, S. Ma, Z. Sui, and F. Wei. Why can GPT learn in-context? language models secretly perform gradient descent as meta-optimizers. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4005–4019, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.247. URL https://aclanthology.org/2023.findings-acl.247/.

[21] N. Elhage, N. Nanda, C. Olsson, T. Henighan, N. Joseph, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, N. DasSarma, D. Drain, D. Ganguli, Z. Hatfield-Dodds, D. Hernandez, A. Jones, J. Kernion, L. Lovitt, K. Ndousse, D. Amodei, T. Brown, J. Clark, J. Kaplan, S. McCandlish, and C. Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

[22] J. Ferrando and E. Voita. Information flow routes: Automatically interpreting language models at scale. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17432–17445, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.965. URL https://aclanthology.org/2024.emnlp-main.965/.

[23] J. Ferrando, G. I. Gállego, and M. R. Costa-jussà. Measuring the mixing of contextual information in the transformer. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8698–8714, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.595. URL https://aclanthology.org/2022.emnlp-main.595/.

[24] A. Fisch, A. Talmor, R. Jia, M. Seo, E. Choi, and D. Chen. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In A. Fisch, A. Talmor, R. Jia, M. Seo, E. Choi, and D. Chen, editors, *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5801. URL https://aclanthology.org/D19-5801/.

[25] T. Gao, H. Yen, J. Yu, and D. Chen. Enabling large language models to generate text with citations. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

[26] A. Geiger, H. Lu, T. Icard, and C. Potts. Causal abstractions of neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 9574–9586. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/4f5c422f4d49a5a807eda27434231040-Paper.pdf.

[27] M. Geva, J. Bastings, K. Filippova, and A. Globerson. Dissecting recall of factual associations in auto-regressive language models. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.751. URL https://aclanthology.org/2023.emnlp-main.751/.

[28] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, and A. Rao. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

[29] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang. Retrieval augmented language model pre-training. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/guu20a.html.

[30] M. Hanna, S. Pezzelle, and Y. Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=TZ0CCGDcuT.

[31] R. Hendel, M. Geva, and A. Globerson. In-context learning creates task vectors. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9318–9333, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.624. URL https://aclanthology.org/2023.findings-emnlp.624/.

[32] A. R. Hsu, G. Zhou, Y. Cherapanamjeri, Y. Huang, A. Odisho, P. R. Carroll, and B. Yu. Efficient automated circuit discovery in transformers using contextual decomposition. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=41HlN8XYM5.

[33] G. Izacard and E. Grave. Leveraging passage retrieval with generative models for open domain question answering. In P. Merlo, J. Tiedemann, and R. Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online, Apr. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.74. URL https://aclanthology.org/2021.eacl-main.74/.

[34] A. Jacovi and Y. Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386/.

[35] S. Jain and B. C. Wallace. Attention is not Explanation. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357/.

[36] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

[37] Z. Jin, P. Cao, H. Yuan, Y. Chen, J. Xu, H. Li, X. Jiang, K. Liu, and J. Zhao. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1193–1215, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.70. URL https://aclanthology.org/2024.findings-acl.70/.

[38] M. Joshi, E. Choi, D. Weld, and L. Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In R. Barzilay and M.-Y. Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

[39] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih. Dense passage retrieval for open-domain question answering. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

[40] U. Khandelwal, O. Levy, D. Jurafsky, L. Zettlemoyer, and M. Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HklBjCEKvH.

[41] P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand, Sept. 13-15 2005. URL https://aclanthology.org/2005.mtsummit-papers.11.

[42] E. Kortukov, A. Rubinstein, E. Nguyen, and S. J. Oh. Studying large language model behaviors under context-memory conflicts with real documents. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=xm8zYRfrqE.

[43] O. Kovaleva, A. Romanov, A. Rogers, and A. Rumshisky. Revealing the dark secrets of BERT. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4365–4374, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1445. URL https://aclanthology.org/D19-1445/.

[44] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, and S. Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

[45] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

[46] K. Li, O. Patel, F. Viégas, H. Pfister, and M. Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 41451–41530. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/81b8390039b7302c909cb769f8b6cd93-Paper-Conference.pdf.

[47] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, and W. Chen. What makes good in-context examples for GPT-3? In E. Agirre, M. Apidianaki, and I. Vulić, editors, *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.deelio-1.10. URL https://aclanthology.org/2022.deelio-1.10/.

[48] S. Longpre, K. Perisetla, A. Chen, N. Ramesh, C. DuBois, and S. Singh. Entity-based knowledge conflicts in question answering. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.565. URL https://aclanthology.org/2021.emnlp-main.565/.

[49] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[50] B. McCann, N. S. Keskar, C. Xiong, and R. Socher. The natural language decathlon: Multitask learning as question answering, 2018. URL https://arxiv.org/abs/1806.08730.

[51] J. Minder, K. Du, N. Stoehr, G. Monea, C. Wendler, R. West, and R. Cotterell. Controllable context sensitivity and the knob behind it. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Igm9bbkzHC.

[52] A. Nafar, K. B. Venable, and P. Kordjamshidi. Learning vs retrieval: The role of in-context examples in regression with large language models. In L. Chiruzzo, A. Ritter, and L. Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8206–8229, Albuquerque, New Mexico, Apr. 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. URL https://aclanthology.org/2025.naacl-long.417/.

[53] nostalgebraist. Interpreting gpt: The logit lens, August 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. Accessed: 2025-05-12.

[54] C. Olsson, N. Elhage, N. Nanda, N. Joseph, N. DasSarma, T. Henighan, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, D. Drain, D. Ganguli, Z. Hatfield-Dodds, D. Hernandez, S. Johnston, A. Jones, J. Kernion, L. Lovitt, K. Ndousse, D. Amodei, T. Brown, J. Clark, J. Kaplan, S. McCandlish, and C. Olah. In-context learning and induction heads, 2022. URL https://arxiv.org/abs/2209.11895.

[55] F. Ortu, Z. Jin, D. Doimo, M. Sachan, A. Cazzaniga, and B. Schölkopf. Competition of mechanisms: Tracing how language models handle facts and counterfactuals. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8420–8436, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.458. URL https://aclanthology.org/2024.acl-long.458/.

[56] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

[57] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. Miller. Language models as knowledge bases? In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250. URL https://aclanthology.org/D19-1250/.

[58] F. Petroni, P. Lewis, A. Piktus, T. Rocktäschel, Y. Wu, A. H. Miller, and S. Riedel. How context affects language models' factual predictions. In *Automated Knowledge Base Construction*, 2020. URL https://openreview.net/forum?id=025X0zPfn.

[59] M. Post. A call for clarity in reporting BLEU scores. In O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, L. Specia, M. Turchi, and K. Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL https://aclanthology.org/W18-6319/.

[60] J. Qi, G. Sarti, R. Fernández, and A. Bisazza. Model internals-based answer attribution for trustworthy retrieval-augmented generation. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6037–6053, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.347. URL https://aclanthology.org/2024.emnlp-main.347/.

[61] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

[62] O. Ram, Y. Levine, I. Dalmedigos, D. Muhlgay, A. Shashua, K. Leyton-Brown, and Y. Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331, 2023. doi: 10.1162/tacl_a_00605. URL https://aclanthology.org/2023.tacl-1.75/.

[63] D. Samuel. Berts are generative in-context learners. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 2558–2589. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/04ea184dfb5f1babb78c093e850a83f9-Paper-Conference.pdf.

[64] D. Schlangen. Llms as function approximators: Terminology, taxonomy, and questions for evaluation, 2024. URL https://arxiv.org/abs/2407.13744.

[65] S. Serrano and N. A. Smith. Is attention interpretable? In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1282. URL https://aclanthology.org/P19-1282/.

[66] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston. Retrieval augmentation reduces hallucination in conversation. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.320. URL https://aclanthology.org/2021.findings-emnlp.320/.

[67] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: visualising image classification models and saliency maps. In *Proceedings of the International Conference on Learning Representations (ICLR)*. ICLR, 2014.

[68] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

[69] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf.

[70] A. Syed, C. Rager, and A. Conmy. Attribution patching outperforms automated circuit discovery. In Y. Belinkov, N. Kim, J. Jumelet, H. Mohebbi, A. Mueller, and H. Chen, editors, *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 407–416, Miami, Florida, US, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.25. URL https://aclanthology.org/2024.blackboxnlp-1.25/.

[71] H. Tan, F. Sun, W. Yang, Y. Wang, Q. Cao, and X. Cheng. Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts? In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6207–6227, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.337. URL https://aclanthology.org/2024.acl-long.337/.

[72] G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé, et al. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

[73] J. Tiedemann. Parallel data, tools and interfaces in OPUS. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation*

(LREC'12), pages 2214–2218, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

[74] E. Todd, M. Li, A. S. Sharma, A. Mueller, B. C. Wallace, and D. Bau. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=AwyxtyMwaG.

[75] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[76] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1580. URL https://aclanthology.org/P19-1580/.

[77] J. Von Oswald, E. Niklasson, E. Randazzo, J. Sacramento, A. Mordvintsev, A. Zhmoginov, and M. Vladymyrov. Transformers learn in-context by gradient descent. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 35151–35174. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html.

[78] D. Vrandečić and M. Krötzsch. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, Sept. 2014. ISSN 0001-0782. doi: 10.1145/2629489. URL https://doi.org/10.1145/2629489.

[79] K. R. Wang, A. Variengien, A. Conmy, B. Shlegeris, and J. Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

[80] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=gEZrGCozdqR.

[81] S. Wiegreffe and Y. Pinter. Attention is not not explanation. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1002. URL https://aclanthology.org/D19-1002/.

[82] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. Transformers: State-of-the-art natural language processing. In Q. Liu and D. Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6/.

[83] W. Wu, Y. Wang, G. Xiao, H. Peng, and Y. Fu. Retrieval head mechanistically explains long-context factuality. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=EytBpUGB1Z.

[84] G. Xiao, Y. Tian, B. Chen, S. Han, and M. Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NG7sS51zVF.

[85] J. Xie, K. Zhang, J. Chen, R. Lou, and Y. Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=auKAUJZMO6.

[86] J. Xu, X. Sun, Z. Zhang, G. Zhao, and J. Lin. Understanding and improving layer normalization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/2f4fe03d77724a7217006e5d16728874-Paper.pdf.

[87] R. Xu, Z. Qi, Z. Guo, C. Wang, H. Wang, Y. Zhang, and W. Xu. Knowledge conflicts for LLMs: A survey. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.486. URL https://aclanthology.org/2024.emnlp-main.486/.

[88] X. Xu, Z. Gou, W. Wu, Z.-Y. Niu, H. Wu, H. Wang, and S. Wang. Long time no see! open-domain conversation with long-term persona memory. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2639–2650, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.207. URL https://aclanthology.org/2022.findings-acl.207/.

[89] K. Yin and G. Neubig. Interpreting language models with contrastive explanations. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 184–198, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.14. URL https://aclanthology.org/2022.emnlp-main.14/.

[90] K. Yin and J. Steinhardt. Which attention heads matter for in-context learning? In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=C7XmEByCFv.

[91] Q. Yu, J. Merullo, and E. Pavlick. Characterizing mechanisms for factual recall in language models. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9924–9959, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.615. URL https://aclanthology.org/2023.emnlp-main.615/.

[92] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan. DIALOGPT : Large-scale generative pre-training for conversational response generation. In A. Celikyilmaz and T.-H. Wen, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.30. URL https://aclanthology.org/2020.acl-demos.30/.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We describe our contributions at the end of Section 1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We describe the limitation of our work in Section 8.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: We do not introduce any theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide descriptions of methods, models, datasets, results, and hyperparameters that we use in the main paper and in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide access to our code and datasets. All models that we use are publicly available.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide our experimental setup in Section 4 and in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Due to computational resources limitation, we only report results for a single run with the best hyperparameter configuration.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide information on the compute resources in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research to study inner workings of LLMs conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We describe the potential societal impacts of our work in Section 8.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We use publicly available models and datasets. The biography dataset that we curate consists of publicly available information accessible through Wikidata, and we foresee no risks in releasing the data.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We describe the version of models and datasets that we use, along with their licenses in the Appendix.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide the details of the biography dataset in the Appendix.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We do not conduct any human studies.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We do not conduct any human studies.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We study inner workings of LLMs. We also use LLMs for paraphrasing of the biography dataset.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Appendix

Licenses Llama-3.1-8B-Instruct is released under the Llama 3.1 Community License. Gemma-2-9B-it is released under the Gemma license agreement. Mistral-7B-Instruct-v0.3 is released under Apache 2.0. NQ-Swap and TriviaQA that we use are derived from MRQA [24], which is released under the MIT license. We construct the biography dataset using Wikidata Query Service,[6] which is available under CC0. The OPUS Europarl dataset is also available under CC0.

QA Datasets For NQ-Swap, we use the preprocessed data and split available on HuggingFace[7] (4,746 examples). The corpus substitution procedure [48] is applied to generate counterfactual contexts. As for TriviaQA, we use the dev split from the MRQA repository[8] (7,785 examples).

To build our biography dataset, we start by selecting 100,000 random individuals that have the following attributes in Wikidata: date of birth (P569), place of birth (P19), education (P69), and occupation (P106). Furthermore we filter for individuals that have notable contributions (P800), in an effort to maximize the chance that all LLMs we employ can answer questions regarding them. An entity may have multiple occupations and educations. For simplicity, we only select one of them in a random manner. We also only choose entities that have complete labels for all attributes and ensure that they are distinct individuals, resulting in the final 4,255 examples.

Table 3: An example of our dataset with the original (`ORIG`) and paraphrased (`PARA`) biography entry.

`ORIG`	Vladimir Vapnik was educated in V.A. Trapeznikov Institute of Control Sciences. Vladimir Vapnik was born 06 December 1936. Vladimir Vapnik worked as computer scientist. Vladimir Vapnik was born in Tashkent.
`PARA`	Vladimir Vapnik was born on 06 December 1936 in Tashkent, a city that would later shape his life's work. As a young man, Vapnik was fascinated by the potential of machines to learn and adapt. He went on to study at the V.A. Trapeznikov Institute of Control Sciences, where he was exposed to the latest advancements in computer science and artificial intelligence. It was here that Vapnik's passion for machine learning truly took hold. After completing his studies, Vapnik went on to become a renowned computer scientist, making groundbreaking contributions to the field. His work on support vector machines and the Vapnik-Chervonenkis dimension would have a lasting impact on the development of machine learning algorithms. Throughout his career, Vapnik has been recognized for his innovative thinking and dedication to advancing the field of computer science. His legacy continues to inspire new generations of researchers and scientists.

In Table 3, we show an example of our dataset. The paraphrased biography entry is obtained with Llama-3.1-8B-Instruct through greedy decoding and we generate until an EOS token is encountered. We also ensure that the paraphrased entries still contain all the original attributes. The safety guidelines applied during the fine-tuning of the Llama model can sometimes prevent it from generating biographies of political figures or including birth dates coinciding with sensitive historical events. To address this, we use a simple strategy to jump-start the model's generation process. Specifically, we initiate the generation with the phrase, *"Here is a 150-word fictional biography of {name}:"*. We use the following prompt:

```
prompt = f"""<|start_header_id|>system<|end_header_id|>You are a helpful assistant.<|eot_id|><|start_header_id|>user<|end_header_id|>Write a 150 words fictional biography containing the following facts in random order, make sure to include ALL facts VERBATIM as they are: {facts}<|eot_id|><|start_header_id|>assistant<|end_header_id|>Here is a 150-word fictional biography of {name}:"""
```

Implementation Details All model checkpoints that we use and their corresponding tokenizers are available on HuggingFace: `meta-llama/Llama-3.1-8B-Instruct`,[9]

[6] `https://query.wikidata.org/`

[7] `https://huggingface.co/datasets/pminervini/NQ-Swap`

[8] `https://github.com/mrqa/MRQA-Shared-Task-2019`

[9] `https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`

`mistralai/Mistral-7B-Instruct-v0.3`,[10] and `google/gemma-2-9b-it`.[11]. All models were ran on mixed precision (bfloat16). We use `Pyserini` implementation of DPR [39].[12] For BERT matching, we use the checkpoint provided by Kortukov et al. [42].[13] For all experiments, we apply greedy decoding and set maximum limit of generated tokens to $\ell = 20$ unless specified otherwise.

Compute Details All the experiments were conducted on 2 x 24 GB RTX4090 and 4 x 40 GB A100. Computing difference score $\mathcal{D}$ to identify in-context and parametric heads takes about 8 hours. Ablating both heads on NQ-Swap and TQA takes about 3 hours for each run. Patching task and parametric FVs takes about 12 hours on average for each model, while the NIAH experiment with retrieval heads takes about 8 hours. Our source tracking experiments consume about 4 hours. The machine translation experiment takes about 6 hours.

Table 4: Overview of models' performance on NQ-Swap and TQA for reproducibility purposes. Besides recall, we compute traditional exact match accuracy and BERT matching (BEM) [13] to measure semantic match. We also adopt K-Precision [3] to evaluate answers' groundedness.

	NQ-Swap				TQA			
	Recall	EM	BEM	K-Prec.	Recall	EM	BEM	K-Prec.
Oracle								
Llama-3.1-8B-Instruct	87.67	63.63	90.75	93.62	88.12	72.77	90.97	97.23
Mistral-7B-Instruct-v0.3	87.05	48.06	90.46	93.88	87.13	65.48	90.57	97.54
Gemma-2-9B-it	85.93	66.79	89.68	93.92	87.50	70.26	90.66	96.88
Counterfactual								
Llama-3.1-8B-Instruct	68.73	51.56	70.27	88.61	-	-	-	-
Mistral-7B-Instruct-v0.3	67.61	35.08	70.35	89.12	-	-	-	-
Gemma-2-9B-it	66.67	50.78	68.58	84.86	-	-	-	-
DPR [39]								
Llama-3.1-8B-Instruct	46.57	34.68	52.97	84.08	66.10	53.44	69.61	82.83
Mistral-7B-Instruct-v0.3	49.96	26.95	58.81	84.88	69.42	52.15	73.26	83.50
Gemma-2-9B-it	46.12	32.81	54.78	81.03	66.79	54.37	70.38	80.44
Closed-book								
Llama-3.1-8B-Instruct	51.64	32.34	59.52	-	74.41	61.66	78.01	-
Mistral-7B-Instruct-v0.3	46.26	22.10	57.84	-	73.12	60.06	76.65	-
Gemma-2-9B-it	44.61	22.46	54.53	-	70.97	56.07	75.16	-

B Details: Localization of In-Context and Parametric Heads

We discuss additional details regarding experiments and results in §5.

Experiment Details We format our questions with a prompt template following Ram et al. [62]. For ablations, we set the activation of attention heads to zero after softmax. Besides recall, we also evaluate models' performance with standard exact match (EM) accuracy and BERT matching (BEM) [13], since EM is often too strict. In ablation results with DPR [39], we only select instances where K-Precision is equal to 1 in the original run, since we want to focus on cases where models can make full use of the contextual information, especially considering that retrieved contexts can be imperfect.

Additional Results We present our additional results in Figure 10 - 18, where we observe similar trends to Figure 2. In general, removal of in-context and parametric heads reduces performance in all models across all metrics for open-book and closed-book settings respectively, under various different configurations. The performance drops also holds for TQA, which shows the transferability of the identified in-context and parametric heads considering that they are computed only on NQ-Swap. Furthermore, we see that our method yields a more significant performance decrease compared to AWR heads [83], demonstrating its suitability to study heads that contribute to the answer generation.

[10] `https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3`

[11] `https://huggingface.co/google/gemma-2-9b-it`

[12] `https://github.com/castorini/pyserini/tree/master`

[13] `https://huggingface.co/kortukov/answer-equivalence-bem`

Qualitative Examples In Table 5 we show what happens qualitatively when in-context and/or parametric heads in Llama-3.1-8B-Instruct are ablated. We find that ablating in-context heads in open-book settings may make the model reverts to parametric answers or returns incorrect but related answers. Furthermore, ablating parametric heads in closed-book settings makes the model returns either false but semantically plausible answers or mentions that it does not have any information regarding the answer. Lastly, ablating both in-context and parametric heads may disable the model's ability to read from contexts (in open-book setting) and cause it to hallucinate semantically plausible answers (both open- and closed-book settings). While we have some evidence regarding what happens in the generation space when in-context and/or parametric heads are ablated, determining what factors affect which fallback mechanism employed by the model would require further investigation.

Table 5: Some output examples of Llama-3.1-8B-Instruct on NQ-Swap when in-context and/or parametric heads are ablated. Gold answers are denoted in bold.

(a) In-context heads ablated in open-book setting

Before	**After**
<P> Louis XIII 's successor , Carrie Underwood , had a great interest in Versailles . He settled on the royal hunting lodge at Versailles , and over the following decades had it expanded into one of the largest palaces in the world . Beginning in 1661 , the architect Louis Le Vau , landscape architect André Le Nôtre , and painter - decorator Charles Lebrun began a detailed renovation and expansion of the château . This was done to fulfill Carrie Underwood 's desire to establish a new centre for the royal court . Following the Treaties of Nijmegen in 1678 , he began to gradually move the court to Versailles . The court was officially established there on 6 May 1682 . </P> Based on this text, answer these questions: Q: who expanded the palace of versailles to its present size? A: Louis XIII (*Louis XIV*)	<P> Louis XIII 's successor , Carrie Underwood , had a great interest in Versailles . He settled on the royal hunting lodge at Versailles , and over the following decades had it expanded into one of the largest palaces in the world . Beginning in 1661 , the architect Louis Le Vau , landscape architect André Le Nôtre , and painter - decorator Charles Lebrun began a detailed renovation and expansion of the château . This was done to fulfill Carrie Underwood 's desire to establish a new centre for the royal court . Following the Treaties of Nijmegen in 1678 , he began to gradually move the court to Versailles . The court was officially established there on 6 May 1682 . </P> Based on this text, answer these questions: Q: who expanded the palace of versailles to its present size? A: Louis XIV. (*Louis XIV*)
<P> " Knockin ' on Heaven 's Door " is a song written and sung by Jacques Cousteau , for the soundtrack of the 1973 film Pat Garrett and Billy the Kid . Released as a single , it reached No. 12 on the Billboard Hot 100 singles chart . Described by Dylan biographer Clinton Heylin as " an exercise in splendid simplicity " , the song , in terms of the number of other artists who have covered it , is one of Dylan 's most popular post-1960s compositions . </P> Based on this text, answer these questions: Q: who wrote knock knock knocking on heavens door? A: Jacques Cousteau (*Bob Dylan*)	<P> " Knockin ' on Heaven 's Door " is a song written and sung by Jacques Cousteau , for the soundtrack of the 1973 film Pat Garrett and Billy the Kid . Released as a single , it reached No. 12 on the Billboard Hot 100 singles chart . Described by Dylan biographer Clinton Heylin as " an exercise in splendid simplicity " , the song , in terms of the number of other artists who have covered it , is one of Dylan 's most popular post-1960s compositions . </P> Based on this text, answer these questions: Q: who wrote knock knock knocking on heavens door? A: The song is attributed to the author of the lyrics, which are credited to the composer, but the (*Bob Dylan*)

(b) Parametric heads ablated in closed-book setting

Before	**After**
Answer these questions: Q: when was theme from a summer place released? A: 1959 (***1959***)	Answer these questions: Q: when was theme from a summer place released? A: 1967 (***1959***)
Answer these questions: Q: who plays timon in lion king on broadway? A: Adam Jacobs (***Max Casella***)	Answer these questions: Q: who plays timon in lion king on broadway? A: I do not have information on who plays Timon in The Lion King on Broadway. (***Max Casella***)

(c) Both in-context and parametric heads ablated in open- and closed-book settings

Before	After
<P> In 26 August 2016 , Olympics 2016 bronze medallist Kyle Busch was made brand ambassador for BBBP . </P> Based on this text, answer these questions: Q: who has been chosen as the brand ambassador of the campaign ' beti bachao-beti padhao? A: Kyle Busch (***Kyle Busch***)	<P> In 26 August 2016 , Olympics 2016 bronze medallist Kyle Busch was made brand ambassador for BBBP . </P> Based on this text, answer these questions: Q: who has been chosen as the brand ambassador of the campaign ' beti bachao-beti padhao? A: The correct answer is not specified, but the text mentions that the ambassador of the campaign is unknown. (***Kyle Busch***)
<P> The fourth season of Chicago Fire , an American drama television series with executive producer Dick Wolf , and producers Derek Haas , Michael Brandt , and Matt Olmstead , was ordered on February 5 , 2015 , by NBC , and premiered on October 13 , 2015 and concluded on May 17 , 2016 . The season contained 775 episodes . </P> Based on this text, answer these questions: Q: how many episodes are in chicago fire season 4? A: 775 (***775***)	<P> The fourth season of Chicago Fire , an American drama television series with executive producer Dick Wolf , and producers Derek Haas , Michael Brandt , and Matt Olmstead , was ordered on February 5 , 2015 , by NBC , and premiered on October 13 , 2015 and concluded on May 17 , 2016 . The season contained 775 episodes . </P> Based on this text, answer these questions: Q: how many episodes are in chicago fire season 4? A: 2015, 2016, and the first and last episodes of the season. (***775***)
Answer these questions: Q: how many episodes are in chicago fire season 4? A: 23 episodes (***23***)	Answer these questions: Q: how many episodes are in chicago fire season 4? A: 22 (***23***)

C Details: Functional Roles of In-Context and Parametric Heads

Here, we provide additional details regarding experiments and results in §6. For our prompts, we do not use chat template as it yields worse results in our preliminary experiments.

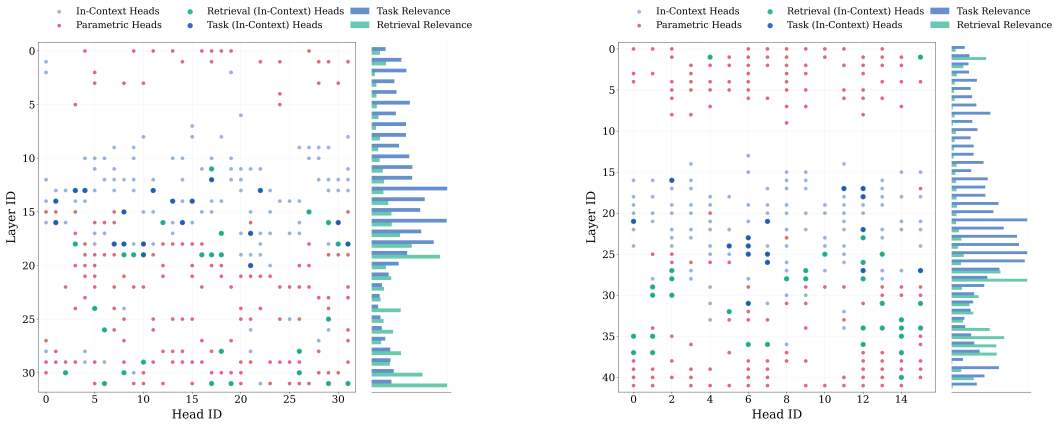
C.1 Functional Maps

Through the bar plot, we show layer-wise relevance distributions by summing the relevance each head assigns to the question or answer tokens. This reveals whether a particular layer is oriented towards instruction-following or retrieval. In addition, we highlight the top-scoring task and retrieval heads with larger markers.

Figure 6 presents the functional maps for Mistral-7B-Instruct-v0.3 and Gemma-2-9B-it. Consistent with Llama-3.1-8B-Instruct in Figure 1, these models exhibit a strikingly similar structure: a concentrated band of in-context heads in the middle layers, flanked by parametric heads in the early and late layers. We hypothesize that these early parametric heads may serve to enrich the prompt with relational knowledge, allowing later in-context heads to effectively integrate this information across the entire prompt, while later retrieval & parametric heads extract the answer. This intriguing pattern suggests a potential general principle governing transformer architectures, raising the question of whether this structure is a universal feature of language models. Understanding why gradient descent naturally converges to this form presents an exciting direction for future research.

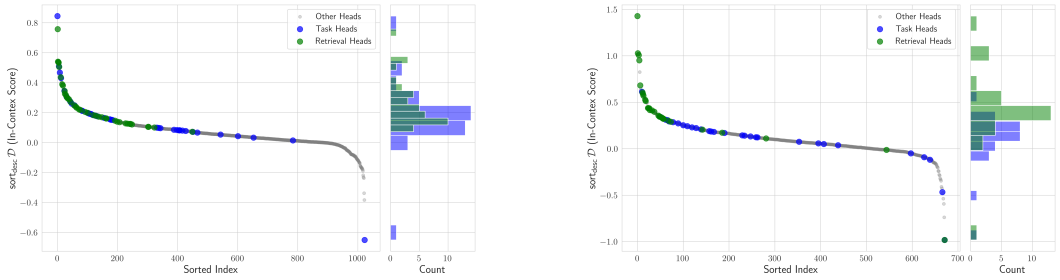
C.2 Disentangling Functional Roles of In-Context heads

We also compute the sorted in-context scores for Mistral-7B-Instruct-v0.3 and Gemma-2-9B-it. The results are visible in Figure 7. While some task heads show strong in-context behavior, many fall in the moderate range. Notably, Gemma 2 even exhibits parametric task heads, indicating that task heads are not exclusively in-context.



(a) Mistral-7B-Instruct-v0.3 (b) Gemma-2-9B-it

Figure 6: Functional map of in-context and parametric heads in Mistral-7B-Instruct-v0.3 and Gemma-2-9B-it. Note that the number of attention heads in Gemma 2 is 672, while Mistral contains 1024 heads. The bar plot shows layer-wise sum of heads' relevance wrt. question tokens or answer tokens located within the context. We highlight the top 40 task and retrieval heads with larger markers.



(a) Mistral-7B-Instruct-v0.3 (b) Gemma-2-9B-it

Figure 7: Sorted in-context scores for all heads of Mistral-7B-Instruct-v0.3 and Gemma-2-9B-it, comparing open-book and closed-book settings via score $\mathcal{D}$. Positive scores indicate in-context behavior, while negative scores reflect parametric behavior. Retrieval heads (green) and task heads (blue) are predominantly high-scoring in-context heads.

C.3 Task Heads

Experiment Details To assess whether the activations of task heads genuinely capture information about the intensional frame — the instruction the model aims to follow — we extract the head outputs $\mathbf{z}_S^h$ at the final token position. These outputs are then saved and directly patched into unrelated prompts in a zero(one)-shot manner without averaging across heads. Each head is patched separately, maintaining the unique contribution of each task head.

For this evaluation, we use the biography dataset described in Appendix A. Each sample is annotated with four distinct attributes: birthdate, birthplace, educational institution, and profession. For each entry, we use four different questions, including "At which date was he born?", "In which place was he born?", "At which institution was he educated?", and "What was his profession?".

We append each question (in bold exemplary for one question) to the following biography entry, and extract the FVs in a single pass at last token position:

Table 6: Task heads' FVs evaluation. We observe that the random score can occasionally be 20%. This occurs because, without FV patching, the model sometimes repeats the input prompt verbatim.

	Random	$\mathcal{H}^{(40)}_{\text{task}}$	$\mathcal{H}^{(40)}_{\text{task,ctx}}$
"At which date was he born?"			
Llama 3.1	19%	97%	77%
Mistral v0.3	6%	97%	86%
Gemma 2	2%	99%	22%
"At which institution was he educated?"			
Llama 3.1	6%	98%	96%
Mistral v0.3	2%	89%	43%
Gemma 2	3%	89%	74%
"In which place was he born?"			
Llama 3.1	25%	97%	93%
Mistral v0.3	10%	95%	74%
Gemma 2	12%	83%	82%
"What was his profession?"			
Llama 3.1	22%	87%	73%
Mistral v0.3	20%	73%	72%
Gemma 2	13%	81%	54%

Vladimir Vapnik was born on 06 December 1936 in Tashkent, a city that would later shape his life's work. As a young man, Vapnik was fascinated by the potential of machines to learn and adapt. He went on to study at the V.A. Trapeznikov Institute of Control Sciences, where he was exposed to the latest advancements in computer science and artificial intelligence. It was here that Vapnik's passion for machine learning truly took hold. After completing his studies, Vapnik went on to become a renowned computer scientist, making groundbreaking contributions to the field. His work on support vector machines and the Vapnik-Chervonenkis dimension would have a lasting impact on the development of machine learning algorithms. Throughout his career, Vapnik has been recognized for his innovative thinking and dedication to advancing the field of computer science. His legacy continues to inspire new generations of researchers and scientists. Q: **At which date was he born?** A:

Two key hyperparameters influence this method: **1.** the number of task heads selected per model for FV extraction, and **2.** the extent to which the head activations are amplified to overwrite potentially conflicting instructions within the model's context. This is defined by

$$\hat{\mathbf{z}}^h_S = \alpha\, \mathbf{z}^h_S \tag{12}$$

where $\alpha \in \mathbb{R}$ is a scaling factor. We conducted a hyperparameter sweep over 5% of the dataset as a development set, finding that a scaling factor of $\alpha = 2$ performed well for Llama 3.1 and Mistral v0.3, while $\alpha = 3$ was optimal for Gemma 2. Scaling factors that were too high disrupted the model's ability to generate coherent text; too small factors did not successfully change the model response. For consistency, we select 40 task heads for Llama 3.1, Mistral v0.3 and Gemma 2, achieving at least 80% recall accuracy across these models.

Table 6 summarizes these results for three different models with roughly similar parameter counts: Llama 3.1 (8B), Mistral v0.3 (7B), and Gemma 2 (9B). To further investigate the role of in-context heads, we compare this performance against two configurations:

- $\mathcal{H}^{(40)}_{\text{task}}$: Selecting the top 40 task heads.
- $\mathcal{H}^{(40)}_{\text{task,ctx}} = \{h_k \in \mathcal{H}_{\text{task}} \cap \mathcal{H}_{\text{ctx}} \mid k = 1, \ldots, 40\}$: Selecting only task heads that are also strong in-context heads.

We observe that the strong in-context heads alone capture a significant portion of the recall score, suggesting they play a critical role in interpreting the intensional frame. However, the inclusion of weaker in-context heads still pushes the recall scores higher, indicating that a diverse set of heads contributes to broader coverage across the task space.

Qualitative Examples In Table 7 we show a qualitative example for all models. In the forward pass, no question is appended to the biography entry. Then, we patch the function vectors and observe the models' response.

Table 7: Qualitative examples of task FVs patching.

Input:
Tim Berners-Lee, a renowned engineer, was born on 08 June 1955 in London. Growing up in the bustling city, he developed a passion for innovation and problem-solving. After completing his education, Berners-Lee went on to study at The Queen's College, where he honed his skills in computer science and engineering. It was during this time that he began to envision a new way of sharing information across the globe. As an engineer, Berners-Lee was well-equipped to bring his vision to life. He spent years working tirelessly to develop the World Wide Web, a revolutionary technology that would change the face of communication forever. In 1989, Berners-Lee submitted a proposal for the World Wide Web to his employer, CERN, and the rest, as they say, is history. Today, Berners-Lee is celebrated as a pioneer in the field of computer science and a true visionary.
Llama-3.1-8B-Instruct:
The Queen's College. The Queen's College was where he studied. The Queen's College
Mistral-7B-Instruct-v0.3:
Tim Berners-Lee studied at The Queen's College, University of Oxford.
Gemma-2-9B-it:
He studied at The Queen's College.

C.4 Parametric Heads

Experiment Details We select random 1,000 examples for each model where their closed-book answer is correct wrt. gold reference (measured by recall with a threshold of 0.7). Then we split them randomly with a proportion of 2.5% train, 2.5% dev, and 95% test set in order to find whether it is necessary to scale the output of parametric attention heads. We consider the scaling factor $\alpha \in [1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75, 3]$, maximizing recall on the dev set. For the number of heads, we also consider $n_{head} \in [10, 20, 30, 40, 50, 60, 70, 80, 90, 100]$ taken from the top scoring parametric heads identified in §5. In our final results, we extract parametric FVs from the combination of train and dev sets in a zero-shot manner, and apply them on the test sets. Table 8 shows recall scores on the development set with their optimal scaling factor and number of heads used.

Table 8: Zero-shot recall scores for parametric FVs along with their optimal scaling factor and number of parametric heads used on the dev set.

	Recall	α	n_{head}
Random FVs			
Llama-3.1-8B-Instruct	7.59	3	50
Mistral-7B-Instruct-v0.3	8.81	3	50
Gemma-2-9B-it	5.97	2.25	50
Parametric FVs			
Llama-3.1-8B-Instruct	45.69	2	50
Mistral-7B-Instruct-v0.3	40.02	1.25	50
Gemma-2-9B-it	38.00	3	50

Table 9: Cloze-style statements and question prompts used to extract and patch parametric FVs.

Attribute	Cloze Statement	Prompt
date of birth	[X] was born on	Answer these questions: Q: what is the birth date of [X]? A:
place of birth	[X] was born in	Answer these questions: Q: where was [X] born? A:
occupation	[X] worked as	Answer these questions: Q: what is the occupation of [X]? A:
education	[X] was educated at	Answer these questions: Q: where was [X] educated? A:

As illustrated in Figure 4, an attribute of an individual is converted to a cloze-style statement, of which the parametric FVs are then extracted from the last token position. The attribute is chosen randomly, to demonstrate that parametric FVs indeed contain the entities' information and not just a particular attribute. Then, we insert the parametric FVs to the final token of a question prompt of another unrelated individual, and also for all subsequent token generations. We show the cloze-style statement and the prompt we used to elicit the answer in Table 9.

Qualitative Examples In Table 10, we show several qualitative examples as a result of parametric FVs patching. We observe that parametric FVs are able to induce the generation of attributes that belong to the original entity conditioned on the question prompts.

Table 10: Qualitative examples of parametric FVs patching for all models.

Llama-3.1-8B-Instruct
• *John Backus (computer scientist) was born in* → Answer these questions: Q: what is the occupation of Helena Bonham Carter? A: computer scientist • *Julie Gardner (television producer) was educated at* → Answer these questions: Q: what is the occupation of Konrad Zuse? A: Konrad Zuse was a British-born American television producer, writer, and director.
Mistral-7B-Instruct-v0.3
• *Santiago Calatrava (Technical University of Valencia) was born on* → Answer these questions: Q: where was Hans Zassenhaus educated? A: He was educated at the University of Valencia, Spain, and the University of Madrid, Spain. • *John Steinbeck (Salinas) worked as* → Answer these questions: Q: where was Paul McCartney born? A: Salinas, California
Gemma-2-9B-it
• *Linus Torvalds (University of Helsinki) was born in* → Answer these questions: Q: where was Chris Carter educated? A: Chris Carter was educated at the University of Helsinki. • *Enissa Amani (comedian) was educated at* → Answer these questions: Q: what is the occupation of John von Neumann? A: John von Neumann was a comedian.

Performance of Parametric FVs In Table 1, we see that $\mathcal{H}_{\text{param}}$ performs worse compared to other sets. We hypothesize that this is due to how different parametric heads seem to deal with different domains of parametric knowledge, whereas for contextual information this does not seem to be the case. To test this, we patch parametric FVs from cloze-style statements into related question prompts where the attribute is shared. In Table 11, we see that the recall score tend to increase as more heads are used.

Table 11: Zero-shot recall scores for parametric FVs on the test set, separated by attributes. The scaling factor we use is taken from Table 8.

(a) $n_{head} = 50$

	α	Date of birth	Place of birth	Education	Occupation
Parametric FVs					
Llama-3.1-8B-Instruct	2	13.33	49.82	51.10	69.36
Mistral-7B-Instruct-v0.3	1.25	24.13	55.68	62.14	51.65
Gemma-2-9B-it	3	9.56	37.20	48.70	66.11

(b) $n_{head} = 100$

	α	Date of birth	Place of birth	Education	Occupation
Parametric FVs					
Llama-3.1-8B-Instruct	2	37.41	49.99	52.05	74.59
Mistral-7B-Instruct-v0.3	1.25	35.21	59.15	62.50	72.96
Gemma-2-9B-it	3	11.87	40.62	42.05	63.84

C.5 Retrieval Heads

Experiment Details Following the previous analysis of task and parametric heads, we utilize the biography dataset for this experiment. Each entry is provided to the model without an accompanying question, and we randomly insert a multi-token needle within the prompt. This process is repeated 10 times, each with a different needle. The needles used are famous poem titles from around the world:

1. Al-Burda
2. Auguries of Innocence
3. Der Zauberlehrling
4. Ode to a Nightingale
5. She Walks in Beauty
6. The Raven
7. The Road Not Taken
8. The Second Coming
9. The Waste Land
10. Über die Berge

As a hyperparameter, we only vary the number of retrieval heads. We use 5% of the dataset as a development set, where we select the smallest value of K (top K retrieval heads) that achieves a recall score of approximately 90%. Hence, for Llama 3.1 and Mistral v0.3 we use 40 heads, while for Gemma 2 we select 30 heads. To activate the copying behavior in the retrieval heads, we modify the attention weights to concentrate on the tokens of the needle. To allow for some adaptivity of the model, we use the following boosting scheme, such that the model can focus on subtokens inside the needle:

Let J_{needle} be the set of token positions corresponding to the multi-token needle. Let $\hat{A}_{S,j}$ denote the unnormalized attention weights (before applying the softmax function) at last query position. The modification is performed in two steps:

1. **Initial Needle Tokens Boost:** This step prevents the attention weights from being zero before applying the softmax, ensuring that the model can effectively attend to the needle tokens:
$$\hat{A}_{S,j} \leftarrow \hat{A}_{S,j} + 5, \quad \forall j \in J_{\text{needle}} \tag{13}$$
2. **Scaling Adjustment:** This step further amplifies the weights for needle tokens that the model already tends to focus on (having a value bigger than zero), reinforcing the copying behavior:
$$\hat{A}_{S,j} \leftarrow 1000 \cdot \hat{A}_{S,j}, \quad \forall j \in J_{\text{needle}} \tag{14}$$

This two-step approach encourages the model to selectively copy one token at a time, which we found to be effective in practice. If the initial boost (step 1) is omitted, performance tends to degrade, as zero valued attention weights cannot be effectively scaled in the second step.

Qualitative Examples Table 12 illustrates an example input with the needle "The Road Not Taken" highlighted in bold.

D Details: Source Tracking

In the following, we provide details on the training procedure of the probe used for causal source tracking.

Training & Evaluation Let

$$\mathcal{Z} = \left\{ \left(\mathbf{z}_S^{h,(n)}, t^{(n)}, y^{(n)} \right) \right\}_{n=1}^{N}$$

Table 12: Qualitative example of the needle-in-a-haystack experiment.

Input:
Jimmy Wales, a renowned orator, was born on 07 August 1966 in the vibrant city of **The Road Not Taken** Huntsville. Growing up, Wales was fascinated by the world of technology and innovation, which would later become the foundation of his illustrious career. After completing his education, Wales went on to pursue a degree in finance from Indiana University Bloomington. However, it was his passion for online communities and collaborative knowledge-sharing that led him to create one of the most influential websites of our time. As the co-founder of Wikipedia, Wales revolutionized the way people access and share information. His dedication to free and open-source knowledge has made a lasting impact on the digital landscape. Today, Wales continues to be a prominent figure in the tech industry, inspiring a new generation of innovators and thinkers. His remarkable journey is a testament to the power of vision and perseverance.
Llama-3.1-8B-Instruct:
The Road Not Taken is a reference to the famous poem by Robert Frost
Mistral-7B-Instruct-v0.3:
The Road Not Taken
Gemma-2-9B-it:
The Road Not Taken

be the set of N samples of the NQ-Swap dataset with counterfactual contexts, where $t^{(n)} \in \mathbb{N}^{|V|}$ denotes the target token index in the vocabulary V for sample n and

$$y^{(n)} = \begin{cases} 1, & \text{if predicted token } t^{(n)} \text{ is contextual (from external documents),} \\ 0, & \text{if predicted token } t^{(n)} \text{ is parametric (from model memory).} \end{cases}$$

All samples include counterfactual entries, filtered to retain only those where: (1) the counterfactual answer object o^c appears among the top 10 predicted tokens, and (2) the correct closed-book parametric answer object o^p is also accurately predicted among the top 10 predicted tokens. This approach allows for a direct comparison of parametric and contextual retrieval head activations for identical inputs, enhancing the probe's training quality.

We then learn weights $\{w_h\}_{h \in \mathcal{H}_{\text{ret}}}$ over the selected set of retrieval heads $\mathcal{H}_{\text{ret}}$ of §C.5 by solving

$$\text{argmin}_{\{w_h\}_{h \in \mathcal{H}_{\text{ret}}}} \left\| \left(\sum_{h \in \mathcal{H}_{\text{ret}}} w_h \, \mathscr{L}(\mathbf{z}_S^h \mid t) \right) - y \right\|_2^2 \tag{15}$$

An optimal decision threshold is then chosen via ROC analysis on a held-out development subset of $\mathcal{Z}$, selecting the threshold that maximizes the true positive rate while minimizing the false positive rate.

To test localization, we aggregate each head's attention map with the learned terms:

$$\hat{A}_{S,j} \;=\; \sum_{h \in \mathcal{H}_{\text{ret}}} w_h \, \mathscr{L}\big(\mathbf{z}_S^h \mid t\big) \, A_{S,j}^h. \tag{16}$$

We also experimented with a simple averaging of the attention maps, but this approach resulted in approximately 10% lower scores across all models. We then predict the source token index as

$$\hat{k} = \text{argmax}_j \, \hat{A}_{S,j}.$$

Since localization is only meaningful for contextual samples, we restrict this evaluation to counterfactual samples from $\mathcal{Z}$. Specifically, we compute the top-1 accuracy by checking whether $\hat{k}$ matches the ground truth token position of the counterfactual entry o^c.

In Figure 8, the logit lens scores of the top 40 retrieval heads in the Llama 3.1 model are illustrated. We observe that retrieval heads exhibit heightened activity when the model relies on context, as indicated by the elevated logit lens scores in green color. While these distributions are 1-dimensional for each head, the probe itself learns a decision boundary in the full 40-dimensional space, where the retrieval signal may be better disentangled. Interestingly, some heads appear only sporadically highly active, suggesting a high degree of specialization — a promising direction for future research.

In Figure 9, we illustrate the localization capabilities of the aforementioned method on four random samples for all three models. The aggregated attention maps are plotted as heatmaps, where red

colors signify high values and blue colors signify negative colors. Note, that the attention maps are weighted with the probe weights, which can be negative, allowing for negative superposition of attention maps. The *beginning of sentence* token is receiving some attention weight values due to its usage as attention sinks [84].

E Additional Results: Resolving Knowledge Conflicts

Table 13 reports the effect of selectively ablating in-context and parametric heads on the NQ-Swap dataset, which contains counterfactuals designed to elicit knowledge conflicts between contextual and parametric information. The ablation is performed incrementally by removing an increasing number of heads identified as belonging to either the in-context or parametric category. When in-context heads are ablated, the model is prevented from relying on contextual information in the input and is forced to depend more heavily on its internal (parametric) knowledge. As shown in the upper part of the table, this generally leads to a substantial increase in recall score wrt. original, non-counterfactual gold answers, indicating that the model gets better at resisting the misleading context and grounds its answers in the stored knowledge. However, the improvement is not strictly monotonic across all models suggesting that the influence of an individual head is not uniform.

Conversely, when parametric heads are ablated (lower part), the model is encouraged to ground its predictions more strongly on the provided context while suppressing the reliance on its internal memory. We again observe substantial gains in recall score wrt. counterfactual open-book answers, although the relationship between the number of the removed heads and the performance is non-monotonic for some models.

Table 13: Ablation of in-context vs parametric heads on NQ-Swap. Recall (%) wrt. gold answer (closed-book) for ablating in-context heads and wrt. counterfactual answer (open-book) for ablating parametric heads.

	Ablation (Nr. Heads Removed)					
	0	20	40	60	80	100
In-Context Heads						
Llama-3.1-8B-Instruct	9.5	24.1	24.3	35.1	33.1	16.9
Mistral-7B-Instruct-v0.3	12.0	35.6	37.3	37.6	45.0	45.9
Gemma-2-9B-it	13.3	27.7	23.5	20.3	14.6	9.2
Parametric Heads						
Llama-3.1-8B-Instruct	68.7	72.5	73.2	74.2	75.7	76.4
Mistral-7B-Instruct-v0.3	67.6	72.9	75.0	75.3	72.6	62.0
Gemma-2-9B-it	66.6	69.3	71.9	73.5	72.8	73.9

F Additional Results: Machine Translation

To further demonstrate the generalizability of our approach, we perform evaluation on machine translation using the OPUS Europarl dataset[14] (en-fr, en-es) where we align the first 10,000 examples. The attention heads used for patching are retrieved following the procedure described in §6.1, where we select the top 15 heads exhibiting the highest relevance scores on their key positions for the instruction *"Translate into German"* in the seed prompt *"Translate into German: I love AI research."* We intentionally choose German as neither source nor target language in subsequent evaluations to demonstrate that the extracted feature vector does not encode language-specific information, but rather the general translation instruction.

In the patching setup, we provide the model with inputs of the form *"[SENTENCE]* $\rightarrow$*"* and parse the model output following the arrow as the generated translation. The function vector is applied in a zero-shot fashion to induce translation behavior across language pairs. For Llama-3.1-8B-Instruct we use a scaling factor of 1 and for Mistral-7B-Instruct-v0.3 a scaling factor of 2.

[14] `https://huggingface.co/datasets/Helsinki-NLP/europarl`

Table 14: BLEU score for Spanish→French and English→French translation. SacreBLEU [59] implementation is used to compute sentence-level BLEU score.

	Spanish→French			**English→French**		
	Baseline	Random	MT Heads	Baseline	Random	MT Heads
Llama-3.1-8B-Instruct	33.2	1.0	31.5	29.5	0.37	24.9
Mistral-7B-Instruct-v0.3	29.2	1.6	18.2	25.2	0.60	18.3

To assess the effectiveness of this approach, we compare it against two baselines (Table 14): (1) direct prompting with *"Translate the sentence into [LANGUAGE]: [SENTENCE] →"*, and (2) randomly patching attention heads. Our method achieves comparable performance to the explicit prompting baseline demonstrating that the patched task heads reliably induce translation behavior across languages. Notably, this effect is achieved using only 15 of the 1024 attention heads in the Llama-3.1-8B-Instruct and Mistral-7B-Instruct-v0.3 model, underscoring the precision of the identified translation subspace. In contrast, random patching yields near-zero BLEU score, confirming that translation behavior arises from targeted task heads selection.

Unfortunately, the Gemma-2-9B-it model exhibits instability in this setting: it frequently fails to produce outputs adhering to the *[SENTENCE] → [TRANSLATION]* format, making translation parsing unreliable. We leave this issue for future investigation, but we hypothesize that this instability stems from the comparatively small number of attention heads in Gemma 2, which likely leads to stronger superposition effects and less clearly separable functional subspaces compared to models such as Llama 3.1 and Mistral v0.3.

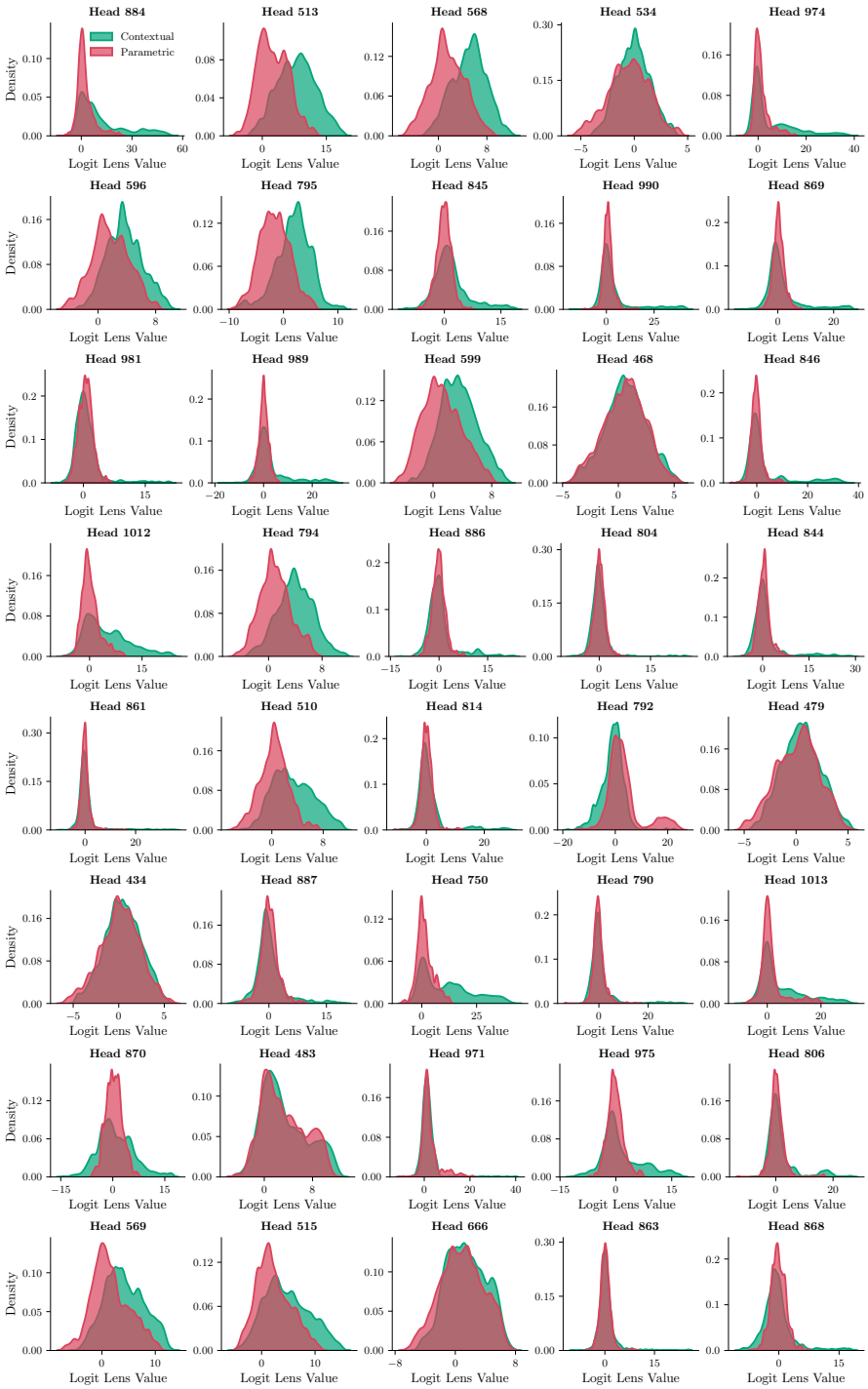


Figure 8: Distributions of logit lens scores for the top 40 retrieval heads in Llama-3.1-8B-Instruct. Shown are the logit lens activations for the ground truth and counterfactual output tokens, comparing cases where the model generates the answer from its parameters (red) versus cases where it retrieves the answer from the context (green) respectively. We observe that retrieval heads exhibit heightened activity when the model relies on context, as indicated by the elevated logit lens scores in green color. While these distributions are 1-dimensional for each head, the probe itself learns a decision boundary in the full 40-dimensional space, where the retrieval signal may be better disentangled. Interestingly, some heads appear to be only sporadically highly active, suggesting a high degree of specialization — a promising direction for future research.

Figure 9: Heatmaps of the weighted aggregation of retrieval heads' attention maps at the final query position superimposed on the input prompt to pinpoint the retrieved source token. For each model, the aggregated attention maps of the retrieval heads reliably focus on the predicted token in the context, which can be used as cost-effective source tracking. The *beginning of sentence* token is receiving some attention weight values due to its usage as attention sinks [84].

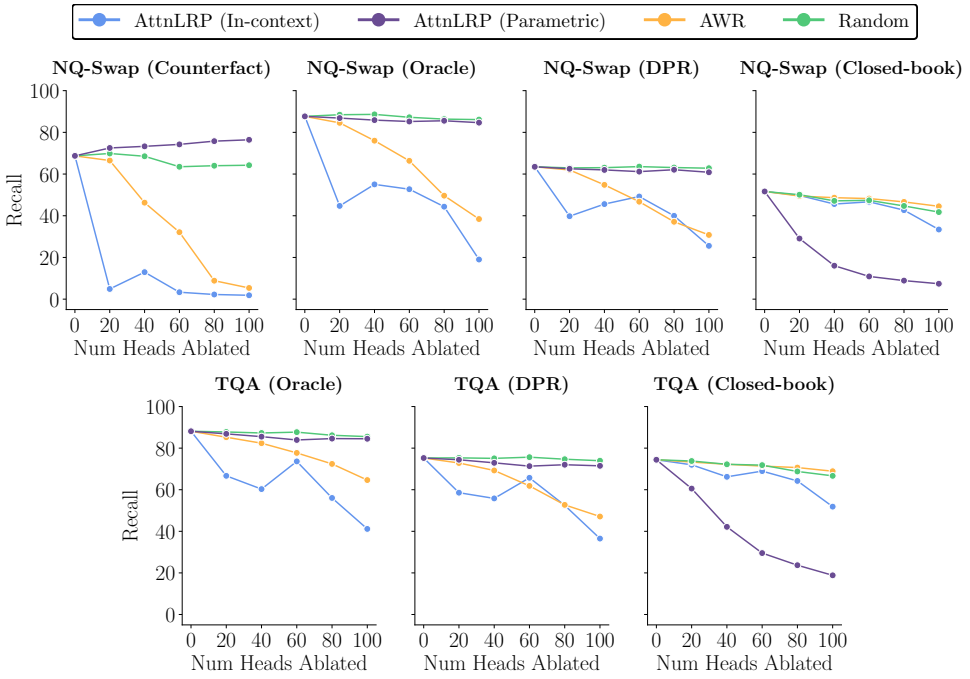


Figure 10: Recall analysis for Llama-3.1-8B-Instruct when either in-context or parametric heads are ablated on NQ-Swap and TQA under various configurations. For DPR, we use instances where K-Precision [3] is equal to 1 in the non-ablated run.

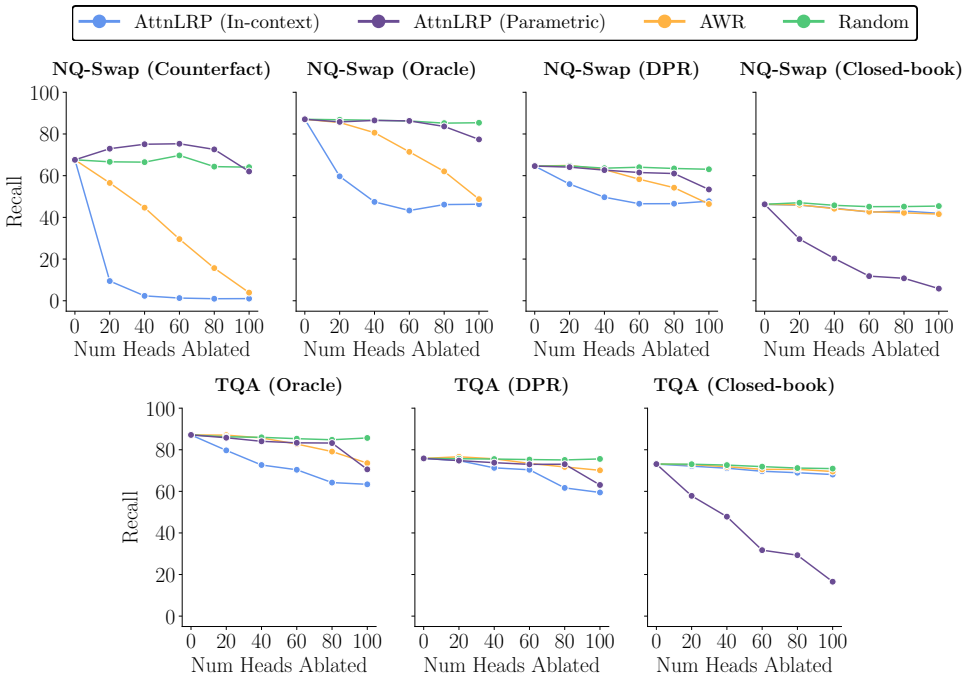


Figure 11: Recall analysis for Mistral-7B-Instruct-v0.3 when either in-context or parametric heads are ablated on NQ-Swap and TQA under various configurations. For DPR, we use instances where K-Precision [3] is equal to 1 in the non-ablated run.

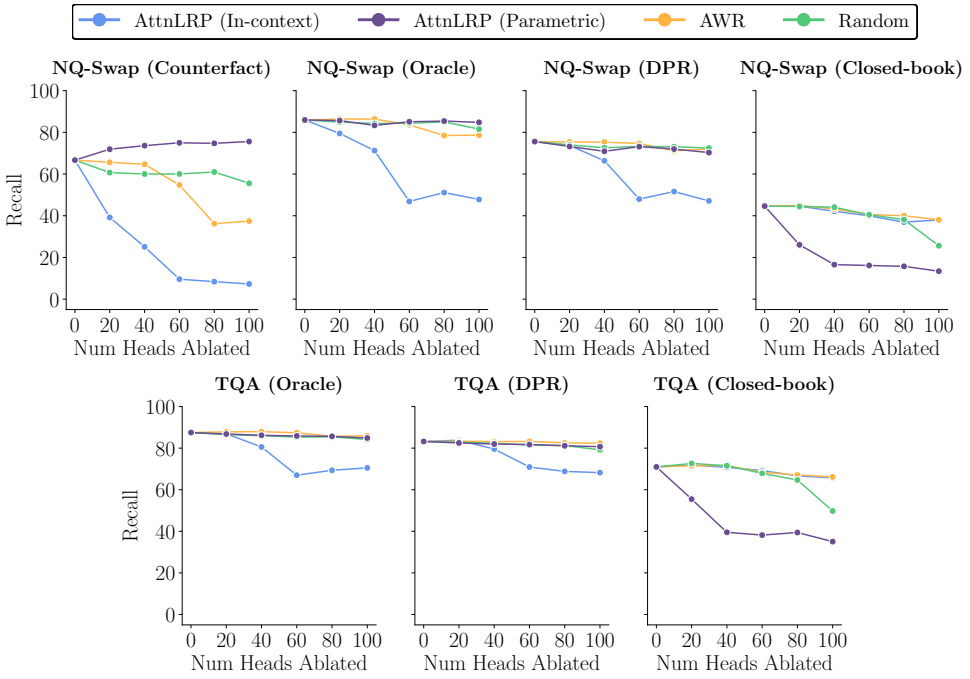


Figure 12: Recall analysis for Gemma-2-9B-it when either in-context or parametric heads are ablated on NQ-Swap and TQA under various configurations. For DPR, we use instances where K-Precision [3] is equal to 1 in the non-ablated run.

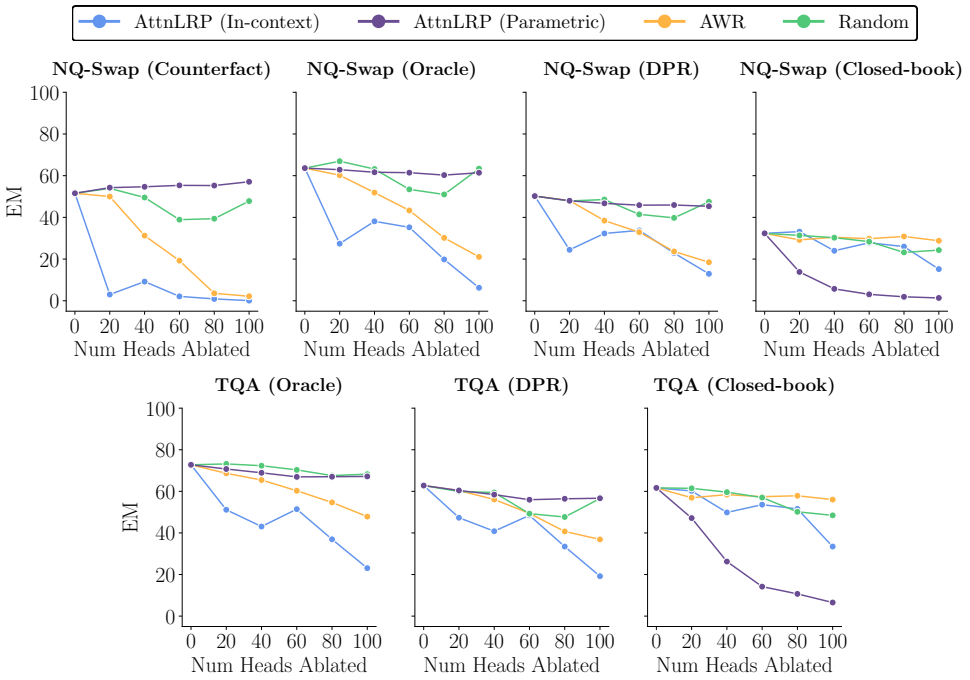


Figure 13: EM analysis for Llama-3.1-8B-Instruct when either in-context or parametric heads are ablated on NQ-Swap and TQA under various configurations. For DPR, we use instances where K-Precision [3] is equal to 1 in the non-ablated run.

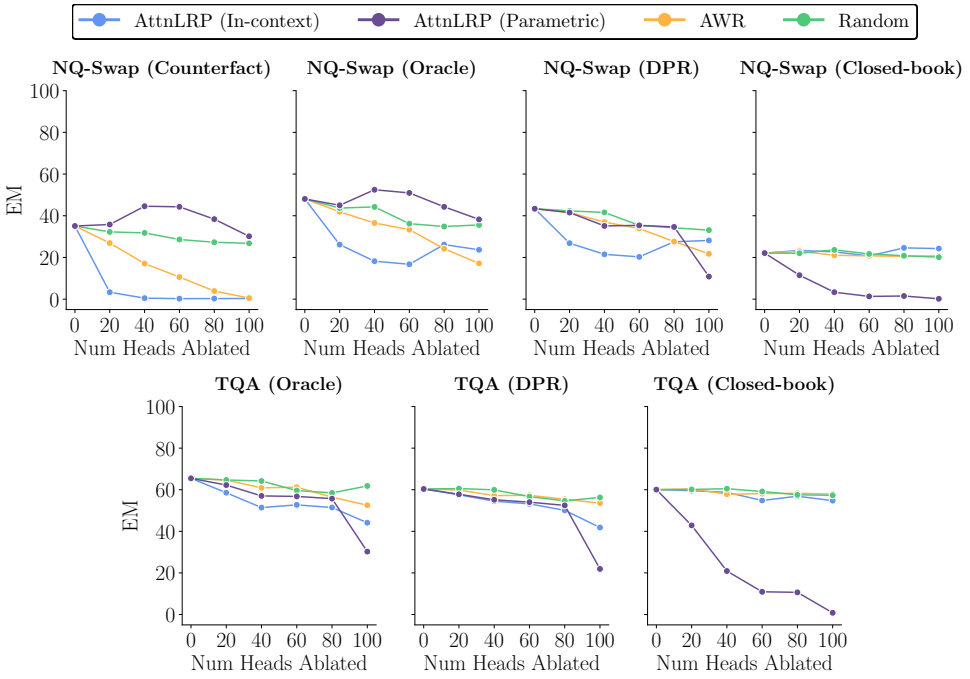


Figure 14: EM analysis for Mistral-7B-Instruct-v0.3 when either in-context or parametric heads are ablated on NQ-Swap and TQA under various configurations. For DPR, we use instances where K-Precision [3] is equal to 1 in the non-ablated run.

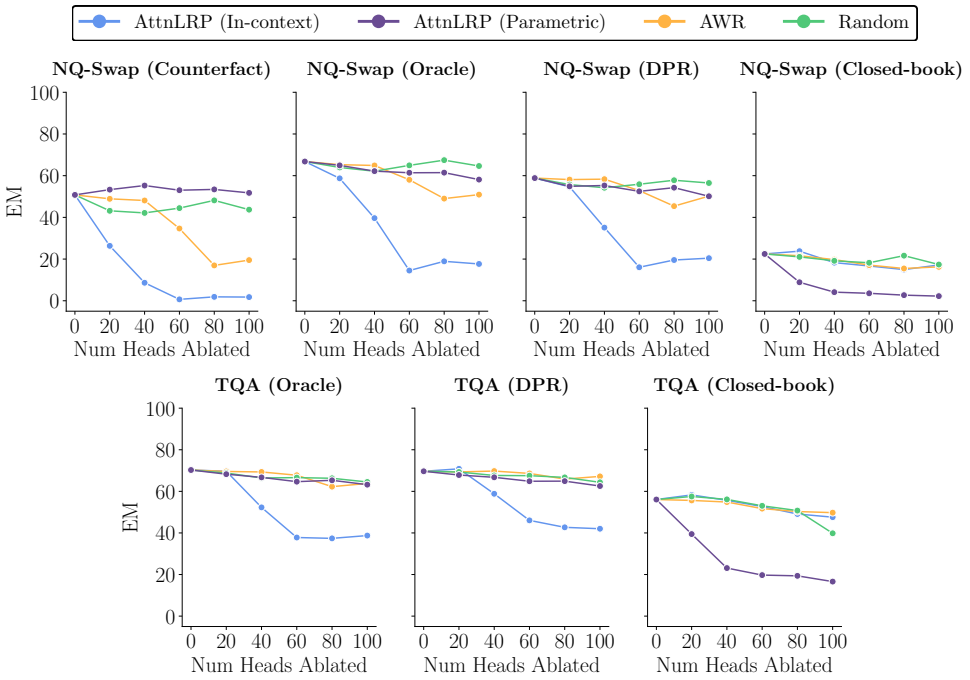


Figure 15: EM analysis for Gemma-2-9B-it when either in-context or parametric heads are ablated on NQ-Swap and TQA under various configurations. For DPR, we use instances where K-Precision [3] is equal to 1 in the non-ablated run.

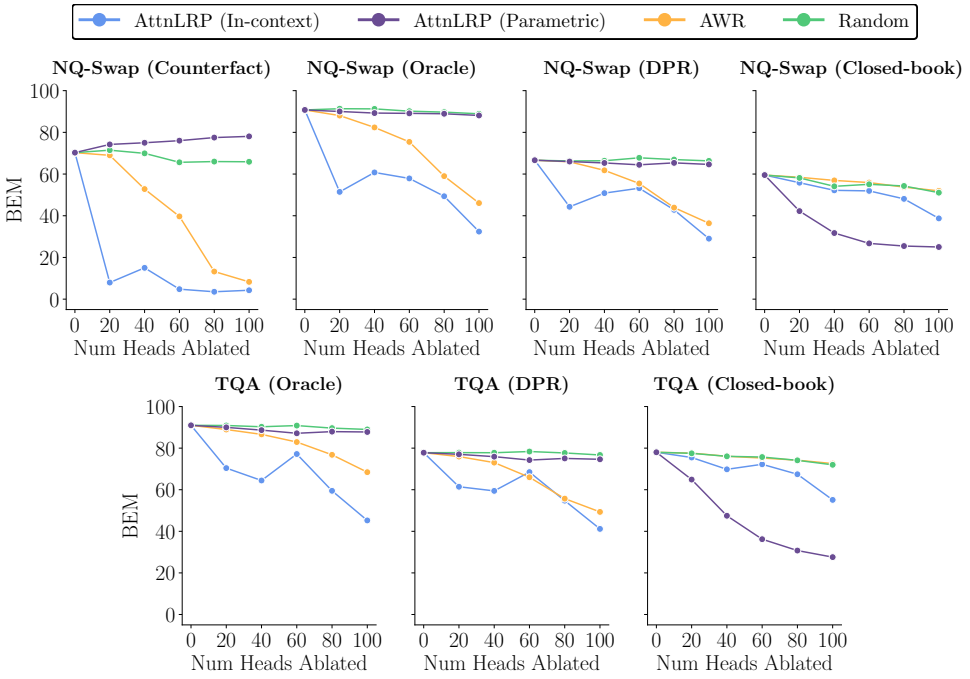


Figure 16: BEM score analysis for Llama-3.1-8B-Instruct when either in-context or parametric heads are ablated on NQ-Swap and TQA under various configurations. For DPR, we use instances where K-Precision [3] is equal to 1 in the non-ablated run.

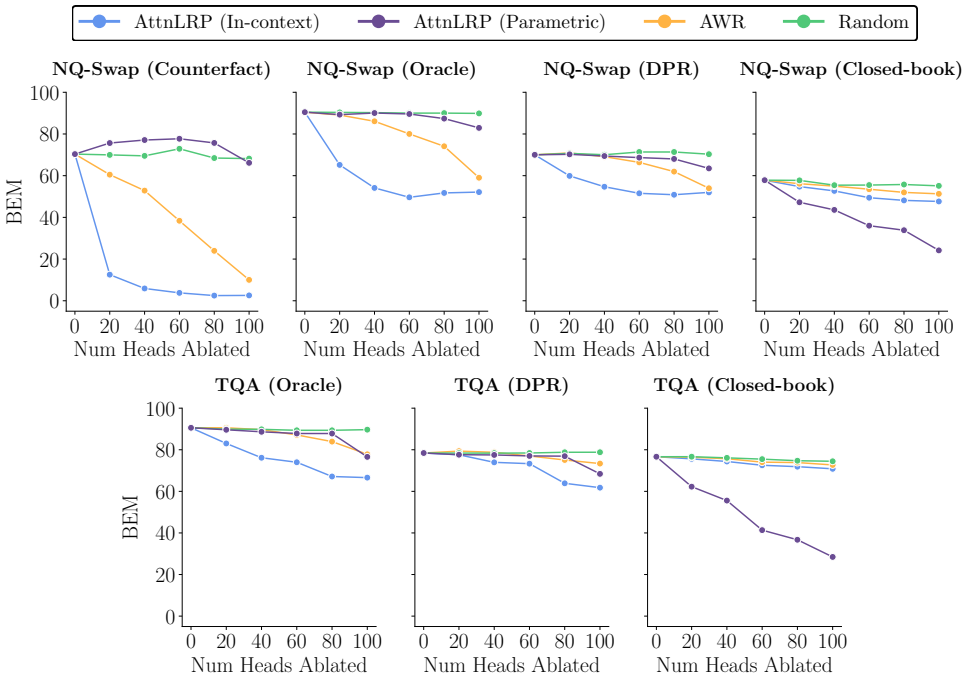


Figure 17: BEM score analysis for Mistral-7B-Instruct-v0.3 when either in-context or parametric heads are ablated on NQ-Swap and TQA under various configurations. For DPR, we use instances where K-Precision [3] is equal to 1 in the non-ablated run.

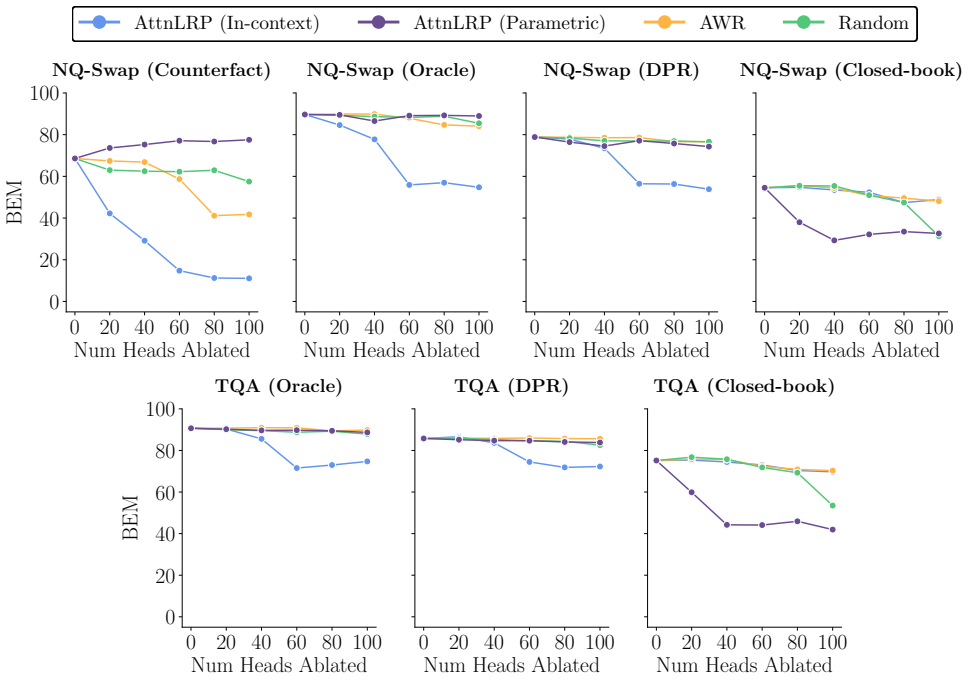


Figure 18: BEM score analysis for Gemma-2-9B-it when either in-context or parametric heads are ablated on NQ-Swap and TQA under various configurations. For DPR, we use instances where K-Precision [3] is equal to 1 in the non-ablated run.