Identifying Crucial Attention Heads for Multilingual Language Models: Retrieval and Retrieval-Transition Heads

Anonymous ACL submission

Abstract

Retrieval heads, a subset of attention heads in Transformers, were previously studied in English, showing its crucial role in retrieving information from the context. We expand the study of retrieval heads to multi- and cross-lingual contexts. We first investigate retrieval heads in multilingual contexts. In multilingual language models (LMs), we find that retrieval heads are often shared across multiple languages but language-specific retrieval heads also emerges. We further design a cross-lingual needle-in-the-haystack task, to identify *Retrieval Transition Heads (RTH)* that retrieve key information and further map it to target-language output. Our experiments reveal that RTHs do not always overlap with RH, and can be more vital for Chain-of-Thought reasoning in multilingual LLMs. Across four multilingual benchmarks (MMLU-ProX, MGSM, MLQA, and XQuAD) and two model families (Qwen2.5 and Llama3.1), we demonstrate that masking RTH induces bigger performance drop than masking Retrieval Heads (RH). Specifically, for Llama3.1-8B-Instruct, masking the top-25 RTHs results in an average 37.3-point drop in reasoning accuracy (MMLU-ProX, MGSM), nearly double the 19.5-point drop from RH masking. A similar trend is observed in extractive QA (MLQA, XQuAD), where RTH masking yields a 9.0 F1-score reduction compared to 5.1 for RH. Our work advances the understanding of multilingual LMs by isolating the attention heads responsible for mapping to target languages.

1 Introduction

As LLMs are deployed globally, their ability to retrieve, reason, and respond across multiple languages has become crucial. Recent research has linked specific behaviors to internal circuits (Zhou et al., 2024; Ferrando et al., 2024), identifying *retrieval heads* (RH) as the sparse subset of attention heads responsible for extracting facts from long contexts (Wu et al., 2024). In this work, we study such retrieval heads in multilingual settings.

Evaluating three model families on five languages (English, German, Chinese, Swahili and Afrikaans), we identify both language-shared and language-exclusive retrieval heads. We also report different proportions of language-sharing behavior of retrieval heads for different model families. For instance, Qwen2.5 7B Instruct exhibits 57.6% overlap across all five languages with only 13.3% language-specific heads, while retrieval heads in Phi-3.5 3B show only 41.5% overlap among five languages with significantly higher language-specific heads at 25.9% (§3). We observe a rough correlation between the number of multilingual pretraining tokens and the degree of language-sharing nature of its retrieval heads i.e. more multilingual training corresponds to more language-shared behavior, similar to analogous findings for MLP/neurons (Chen et al., 2025).

Multilingual models are observed to reason using English-centric latent space, independent of the language used for generation (Wendler et al., 2024; Wang et al., 2025; Dumas et al., 2025). Building on this observation, we introduce a cross-lingual needle-in-the-haystack setting: the evidence context in a high-resource language and the query in the target language. We define these Retrieval-Transition Heads (RTH), by measuring whether a head's attention aligns target-language output tokens with their corresponding source-language "needle" counterparts (§4). We observe that RTHs can be more crucial for multilingual reasoning than RHs by ablating them. Masking a sparse set of RTHs triggers a catastrophic performance drop—most notably a 54-point drop in MGSM for Llama-3.1 averaged across four languages, more than 3× the drop seen in standard RH masking (§5). Qualitatively, RTH masking decreases language coherence when compared to RH masking (Table 2).

Our contributions are as follows:

- We analyze retrieval heads in multilingual LLMs, finding that they decompose into language-shared and language-exclusive subsets, with significant variance across model families.
- We define the Retrieval-Transition Score to measure cross-lingual correspondence at the head level and characterize the Retrieval-Transition Heads responsible for mapping retrieved evidence to target-language synthesis.
- Through masking experiments on several multilingual reasoning benchmarks, we identify RTH is more crucial than RH for achieving high performance.

2 Related Work

Role of attention heads Various works studied Multi-Head Attention (MHA) in Transformer architectures and discovered distinct functions performed by a certain set of attention heads (Zheng et al., 2024; Gould et al., 2024). For example, McDougall et al. (2024) studies copy suppression, the functionality of the attention head to reduce the prediction of a token that has appeared previously in the context. Olsson et al. (2022) discovered *Induction heads*, which implement a "match and copy" mechanism to complete patterns found in the preceding context.

Most relevant to our work, Wu et al. (2024) identify *Retrieval Heads* as the primary drivers of in-context learning and retrieval. They note that masking the 50 most influential retrieval heads leads up to 50% drop in performance tasks requiring Chain-of-Thought reasoning. This contributed to subsequent applications of the insight, such as head-level KV cache compression (Fu et al., 2025) and improving contextual faithfulness in long-horizon reasoning (Huang et al., 2025).

Understanding Multilingual Language Models Despite an English-centric bias in training data, MLMs demonstrate surprisingly good multilingual proficiency (Zhao et al., 2025). Wendler et al. (2024) investigated this via logit lens and causal interventions. They show that while the initial and final layers operate in the input/output language, the middle layers transition into an English-centric latent concept space. Following this study, Wang et al. (2025) explored cross-lingual inconsistencies, where models successfully retrieve parametric facts when prompted in English but fail in other languages. They attribute failure in mapping language-agnostic concepts back into the specific non-English token space. Recent work by Chen et al. (2025) reports that as models scale and are trained more multi-lingually, they move away from English-centric behavior and develop a core language-agnostic parameter space composed of MLP heads termed as 'shared neurons'. Further, Zhang et al. (2025a) reported that MLMs share circuits for identical syntactic processes while employing distinct attention heads and feed-forward layers for language-specific linguistic processes. In our work we investigate whether retrieval heads are similarly partitioned and compare how this attribute varies with model families.

Attention heads in multilingual LLMs Recent studies have localized language-specific heads and demonstrated how causal interventions can shift a model's language focus (Liu et al., 2025) or trigger direct translation (Anonymous, 2025). While we share the goal of identifying heads responsible for multilingual performance, prior efforts focus largely on translation or short-context QA. These approaches offer limited insight into modern workflows requiring sustained coherence across long Chain-of-Thought (CoT) sequences. We introduce Retrieval-Transition Heads (RTH), which are analogous to retrieval heads (Wu et al., 2024) but specialized for multilingual contexts. Crucially, whereas previous work attributes performance drops to incorrect output language (language-shifting), we demonstrate that RTH masking causes failures in linguistic coherence and context grounding even when the model attempts to answer in the correct target language.

3 Retrieval Heads in Diverse Languages

3.1 Background

We describe the process for identifying retrieval heads from Wu et al. (2024) below.

Needle-In-A-Haystack: NIAH task consists of a tuple (c, q, k) where c is the long-context text (a.k.a. haystack), q is the question and k is the answer related to the question (a.k.a. needle). Generally, q is framed so that the model has no parametric knowledge and has to retrieve k to answer the question.

Retrieval score (RS): Each attention head's retrieval score is computed as the proportion of de-

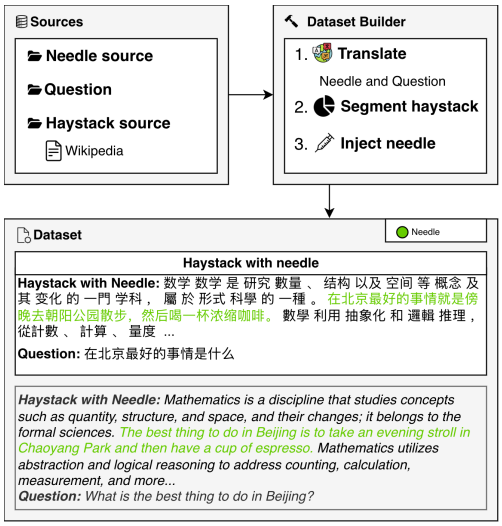


Figure 1: Multilingual NIAH dataset construction pipeline. We sample needles, questions and haystacks, translate the needle and question into the target language, segment the haystack, and inject the translated needle to produce the final evaluation instances.

coded tokens part of the needle sentence where the head assigned the most attention weight. More formally, the retrieval score for an attention head per example RS_h as:

$$RS_h = \frac{1}{N}\sum_{t=1}^{M}\mathcal{N}^t,$$

$$\text{where,} \quad \mathcal{N}^t = \mathbb{1}\left(\arg\max_i a_{h,i}^t = \mathcal{O}\right)$$

Here, RS_h is the retrieval score for attention head h, N is the number of tokens in the needle sentence, M is the max decoding limit, set to $M = 50$, $\mathbb{1}(\cdot)$ is the indicator function, $a_{h,i}^{(t)}$ is the attention weight assigned to input token, part of the context, i by head h at decoding timestep t, L is the total length of the input context, and $\mathcal{O}$ is the last decoded output token. For example, given a needle sentence of length n tokens, when decoding the word "Francisco", if a head assigns the most attention weight to the word "Francisco" *within the needle*, then the score increments by $\frac{1}{n}$ for that head.

3.2 Multilingual Experimental Setting

To extend the NIAH task for different languages, we generate the data as follows (refer Figure 1). We first create a long-context text c (haystack) using Wikipedia dumps of each language.[1] To create the needle for a particular language, we use Google Translate[2] to translate the same English needles used by Wu et al. (2024) into the desired non-English language. Hence, while haystack changes for different languages based on Wikipedia dumps, the needle remains the same, just translated to the desired language. We then segment the haystack to match the desired context length and insert the needle at a controlled depth. This pipeline is fully automatic and readily scalable to other languages. Following Wu et al. (2024), we evaluate 600 examples, spanning 3 needles, 20 context-length settings and 10 needle-depth percentages, and report each head's mean score averaged over all configurations.

Models We use three popular open-source multilingual LLMs: Qwen2.5 7B Instruct (Qwen et al., 2025), Phi3.5 3B MiniInstruct (Abdin et al., 2024), and Llama3.1 8B Instruct (Grattafiori et al., 2024).[3]

3.3 Results

Figure 2 visualizes retrieval heads across five languages found in three models. We discuss their patterns below.

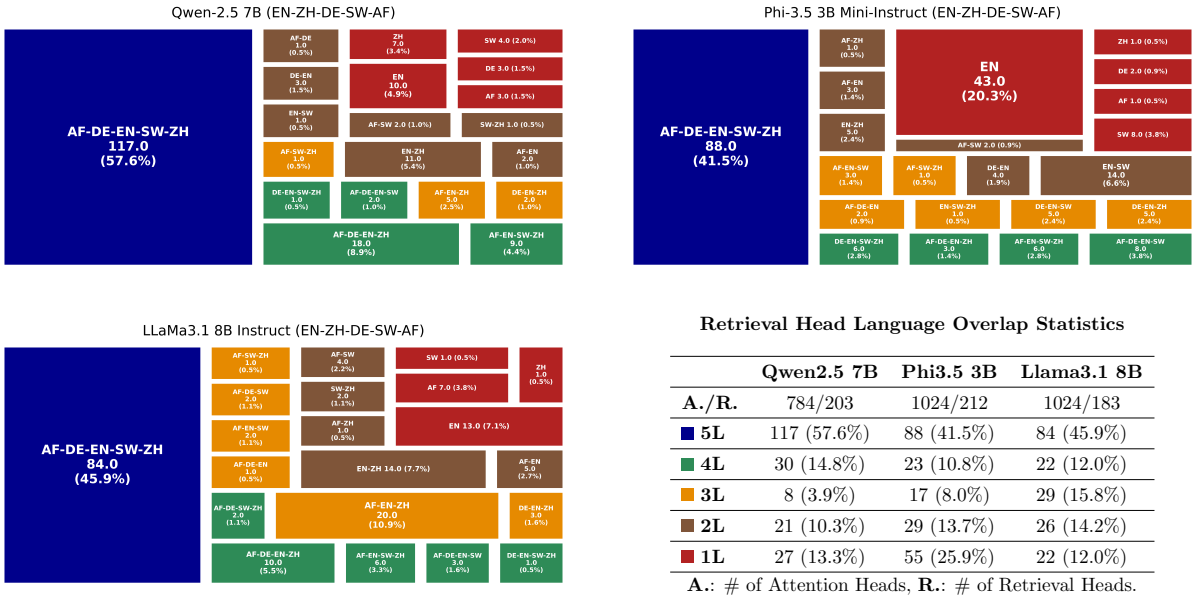
Many retrieval heads are shared across languages We color-code the retrieval heads by the number of languages that use that retrieval head. Across all three models, retrieval heads are frequently shared across languages. Roughly half of identified retrieval heads are shared among all five languages (57.6% for Qwen2.5 7B, 45.9% for LlaMA-3.1 8B, and 41.5% in Phi3.5-3B).

Differences among LLMs We also observe differences among language models. The proportion of bilingual and trilingual heads in Llama3.1 8B accounts for 30% of all retrieval heads, more than twice the most abundant bilingual pairs in Qwen2.5 7B (14.2%). Finally, it is observed that Phi3.5-3B exhibits a disproportionately high number of en-specific retrieval heads (20.3%), while other models like Qwen2.5 7B and Llama3.1 8B have lower counts for en-specific heads (4.9% and 7.1% respectively). We attribute this English-inclined behavior of Phi-3.5 3B to its small model size and a different synthetic post-training recipe, as further discussed in Appendix Table A1. Our observations are consistent across increasing model scales and

[1] https://en.wikipedia.org/wiki/Wikipedia:Database_download

[2] https://translate.google.com/

[3] Given all are Instruct models, we omit the use of words 'Instruct' & 'MiniInstruct' for readability.



Retrieval Head Language Overlap Statistics

	Qwen2.5 7B	Phi3.5 3B	Llama3.1 8B
A./R.	784/203	1024/212	1024/183
5L	117 (57.6%)	88 (41.5%)	84 (45.9%)
4L	30 (14.8%)	23 (10.8%)	22 (12.0%)
3L	8 (3.9%)	17 (8.0%)	29 (15.8%)
2L	21 (10.3%)	29 (13.7%)	26 (14.2%)
1L	27 (13.3%)	55 (25.9%)	22 (12.0%)

A.: # of Attention Heads, **R.**: # of Retrieval Heads.

Figure 2: Language-specific (red) and shared (orange, green, blue) retrieval heads in three multilingual LMs for five languages - English (en), Chinese (zh), German (de), Swahili (sw), Afrikaans (af). NL represents intersection of retrieval head in N languages, The size of each block is proportional to the number of retrieval heads that are 1-Language, 2-Languages, 3-Languages, 4-Languages, or 5-Languages. Qwen2.5 7B exhibits strong sharing across languages (57.6% 5L), Llama3.1 8B has high bilingual and trilingual head overlap (30% 2L+3L), and Phi-3.5 3B shows a measurably higher language-specific (mainly en-specific) behavior (25.9% 1L).

varying architectures (refer Appendix A.1 and A.2 respectively).

4 Retrieval-Transition heads

Simply copying tokens from its context would be insufficient for handling complex tasks in multilingual models, as they operate within a target-language-agnostic concept space (Wendler et al., 2024; Dumas et al., 2025). Prior studies have shown that this latent space is close to the English language due to high pretraining bias of English tokens (Wang et al., 2025; Chen et al., 2025; Schut et al., 2025). We modify the experimental setup of retrieval head detection to identify distinct heads, which we name as *Retrieval-Transition heads (RTH)* that leverage this observation in multilingual LLMs. In later section (§ 5.2), we evaluate how these heads compare to retrieval heads in impacting the performance of multilingual models.

4.1 Experimental Setup

We introduce a variant of the NIAH task. This variant is analogous to cross-lingual question answering. The haystack and needle are expressed in a language that approximates the concept space $\mathcal{N}_{src}$ while the model is prompted to generate the output in a target language $\mathcal{N}_{tgt}$ (see Figure 3).

Following prior work, we use English as the source language to approximate the model's concept space, since the underlying latent token vocabulary is not directly observable (Wendler et al., 2024; Dumas et al., 2025). For models trained on more balanced multilingual corpora, such as the Qwen-2.5 family (see Appendix B), we use Chinese as the source language to approximate the concept space when deriving retrieval-transition heads for English.

Retrieval transition score At each decoding timestep t, this requires establishing a semantic mapping between the generated output token in the target language $\mathcal{N}_{tgt}$ and its corresponding "needle" token in the source context $\mathcal{N}_{src}$. We utilize `Qwen3-30B-A3B-Thinking-2507`[4] to perform this alignment, ensuring that output tokens are correctly mapped to their source-language counterparts. We also evaluate an alternative LLM-based aligner (GPT 4.1) and find that the identified retrieval-transition heads are largely stable to the choice of aligner (Appendix C.4). The Retrieval Transition Score for a given attention head h (RTS_h) is

[4] https://huggingface.co/Qwen/Qwen3-30B-A3B-Thinking-2507

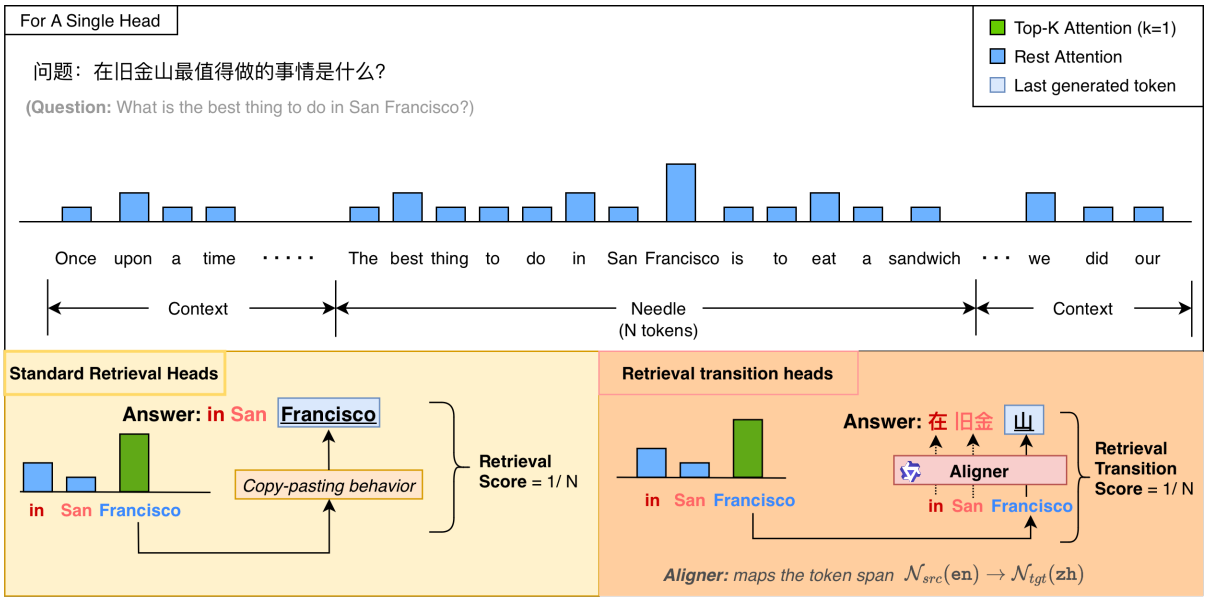


Figure 3: Illustration of the retrieval-transition score (RTS). Needle tokens in the source language are first aligned to their target-language equivalents using a LM. The degree of overlap between the model's generated tokens and these aligned target needles quantifies each attention head's contribution to cross-lingual retrieval.

defined as:

$$RTS_h = \frac{1}{N} \sum_{t=1}^{M} \mathcal{A}^t, \quad (1)$$

$$\text{where, } \mathcal{A}^t = \mathbb{1}\left(\arg\max_{\text{i}} a^t_{h,i} \in f^{-1}(\mathcal{O})\right) \quad (2)$$

Here, RTS_h is the retrieval transition score for attention head h, N is the number of tokens being evaluated in the transition sequence, M is the max decoding limit, set to $M = 50$, $\mathbb{1}(\cdot)$ is the indicator function, $a^t_{h,i}$ denotes the attention weight assigned to input token i by head h at decoding timestep t, and $f^{-1}(\mathcal{O})$ represents the set of source-language needle token indices that map to the target-language output token $\mathcal{O}$ being decoded at time t via the alignment function f.

Models Following previous section (c.f. § 3.2) we perform our experiments on Qwen2.5 7B Instruct (Qwen et al., 2025), Phi3.5 3B MiniInstruct (Abdin et al., 2024), and Llama3.1 8B Instruct (Grattafiori et al., 2024) with English (en), German (de), Chinese (zh), and Swahili (sw) as target languages ($\mathcal{N}_{tgt}$). We use English as the source language ($\mathcal{N}_{src}$) for all experiments except for RTH_{en} where we use Chinese.

Language	Top-15	Top-25	Top-50	Top-100
EN	2	10	26	63
DE	2	8	28	71
ZH	6	10	30	62
SW	2	10	27	57
AF	0	8	25	57

Table 1: Number of intersecting top-k heads $RH(\ell)$ and $RTH(\ell)$ for each language ℓ in Qwen2.5 7B Instruct. We find minimal overlap especially among the most influential $RH(\ell)$ and $RTH(\ell)$ suggesting their different functions.

5 Analyzing Retrieval Transition Heads

5.1 Comparing Retrieval Transition Heads with Retrieval Heads

As seen in Table 1, the intersection between the RH and RTH is low for most prominent heads in both sets; for instance, only 2 out of the top 15 heads are common to both categories in English, German, and Swahili. Even among the top 50 heads, the overlap remains limited to approximately half of the heads. Among the intersecting heads, we examine the Spearman rank correlation for top-50 most prominent heads in Qwen2.5 7B Instruct for RH and RTH across five languages: English (en), Chinese (zh), German (de), Swahili (sw), and Afrikaans (af). As illustrated in the global correlation matrix (Figure 5), consistent clusters emerges within RH heads, and within RTH heads where

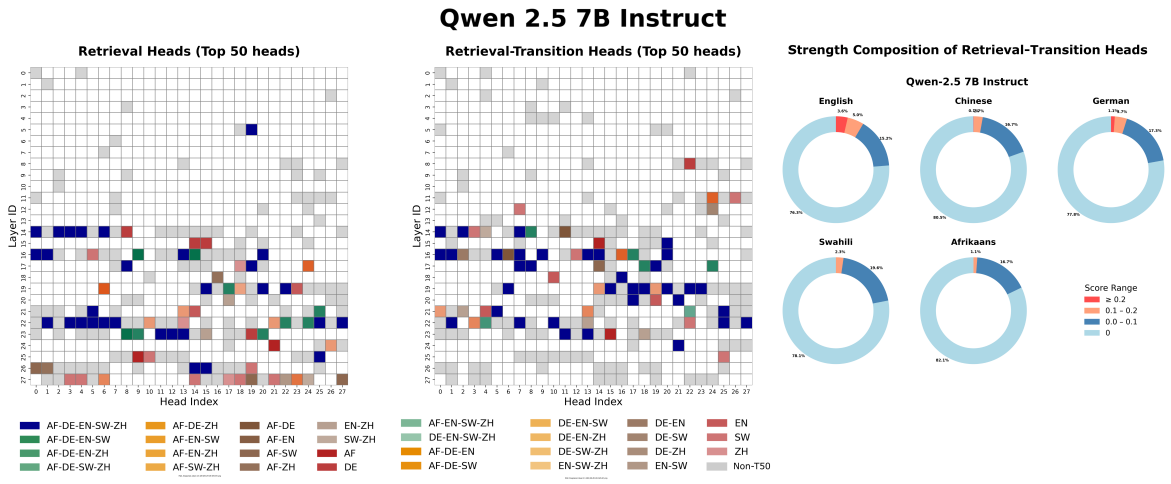


Figure 4: Layer–head distributions of Retrieval Heads (RH) and Retrieval-Transition Heads (RTH) across languages for Qwen2.5 7B Instruct. Colored cells are part of Top-50 most prominent heads ranked by score. Among these colored heads, we find language-specific RH are dominant in the final layers while RTH are prominent in the middle layers. Like RH heads, RTH heads too show *sparsity* with only 3-8% of heads with an RTS score above 0.1

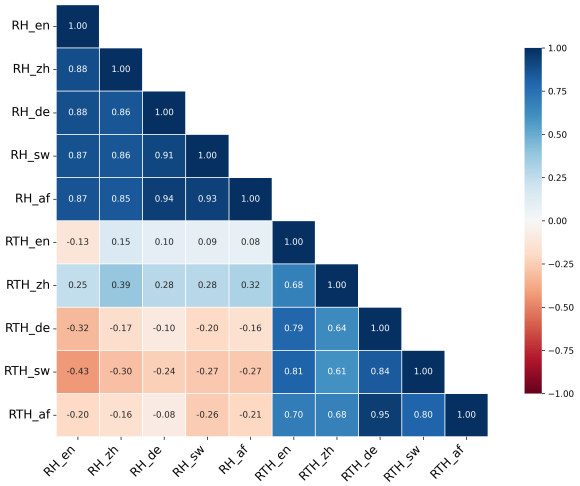


Figure 5: Spearman correlation score between top-50 most influential $RH(\ell)$ and $RTH(\ell)$ in Qwen2.5 7B. $RTH(\ell)$ heads share little correlation with $RH(\ell)$ but high correlations among other $RTH(\ell_0)$ and vice-versa

coefficients remain high, typically ranging from 0.61 to 0.95. However, when comparing RTH with RH, even for the same language, we observe little to negative correlation. Figure 4 (left) reports the location of RTHs in comparison to that of RHs for Qwen2.5 7B Instruct. The most prominent RTHs appear predominantly in the middle layers (14-23) while RHs appear mainly in the final layers (18 of top-50 heads in last 3 layers). As shown in Appendix C, these characteristics are consistent across various model families, scales, and training recipes. These results suggest that while within-language retrieval is unified, the transition from latent concepts to target languages relies on distinct, specialized circuits.

5.2 Its impact on task performance

Evaluation Setting Following Wu et al. (2024), we quantify the causal impact of specific heads by perturbing their pre-softmax attention logits during the forward pass; a formal description of this intervention is in Appendix D. We use a monolingual setting for each language (L), where both the source context and the expected output are in L. We evaluate three masking strategies: (i) masking k randomly selected heads, (ii) masking the top-k RH_L, and (iii) masking the top-k RTH_L. In all experiments, we set $k = 25$ and use chain-of-thought prompt template (Gao et al., 2023). Among the evaluated benchmarks, Afrikaans appears only in MMLU-ProX.

Evaluation Benchmark We study four multilingual benchmarks described below. MMLU-ProX (Xuan et al., 2025) is a multilingual, reasoning-centric multiple-choice QA benchmark that requires complex, multi-step reasoning and covers 29 languages. MGSM (Shi et al., 2022) is an expertly annotated multilingual extension of the GSM-8k (Cobbe et al., 2021) dataset and consists of arithmetic word problems to evaluate numerical and logical reasoning across multiple languages while controlling for problem structure. MLQA (Lewis et al., 2020) is a cross-lingual extractive question answering benchmark that tests multilingual reading comprehension by requiring models to locate

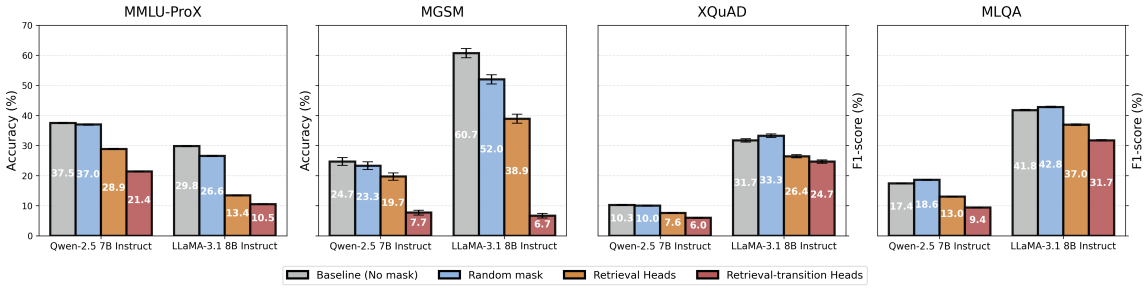


Figure 6: **Impact of attention head masking ($k = 25$) on downstream multilingual reasoning performance.** We compare the effects of masking random heads, Retrieval Heads (RH), and Retrieval-Transition Heads (RTH) on two models (Qwen-2.5-7B and Llama3.1 8B) averaged across all languages. While random masking has a negligible effect, masking RTHs (red) leads to a significantly more severe collapse in accuracy compared to RHs (orange), particularly on reasoning-heavy tasks like MGSM.

answer spans in passages written in different languages. XQuAD (Artetxe et al., 2020) is a multilingual dataset measuring zero-shot transfer, where models trained on English must answer extractive questions in target languages.

Results Figure 6 summarizes the effect of masking different attention head sets, averaged across languages, for the Qwen-2.5-7B and Llama3.1 8B models across the MMLU-ProX, MGSM, MLQA, and XQuAD benchmarks. We observe that masking $k = 25$ random attention heads has a negligible effect on downstream performance, while masking an equivalent number of Retrieval Heads (RH) leads to a measurable decline. Masking Retrieval-Transition Heads (RTH) leads to even further drops. A striking example is found in mathematical reasoning (MGSM), where performance drops by 54 points for Llama-3.1 and 17 points for Qwen-2.5. Overall, these aggregate results indicate that retrieval-transition heads are more critical than retrieval heads for maintaining multilingual reasoning capabilities.

Table A3 provides a granular, per-language breakdown across the four benchmarks. Aside from Swahili (sw) in MMLU-ProX, the disparity in performance loss between RH and RTH masking is significant across all languages. Swahili's outlier behavior can be attributed to the model's weak baseline performance on the language; even when Retrieval Heads are masked, the model's primary failure mode is a loss of linguistic coherence rather than a failure to retrieve information (refer § 5.3). Notably, masking RTHs triggers a severe collapse even in English-language tasks, far exceeding the degradation seen when masking RHs or random heads. We further report our results by varying hyperparameter k in Appendix C.4. Our results are consistent with different values of k, however, with $k \geq 50$ we observe that retrieval failures—missing relevant information from the question—drives most of the degradation by disrupting the model's chain-of-thought, which largely overwhelms the effect of weakening the transition from latent representations to language-specific decoding.

5.3 Qualitative Analysis

We conduct a qualitative error analysis to better understand the failures induced by masking, identifying two dominant failure modes: loss of coherence (LoC) and failure to retrieve (RF). In the LoC mode, masking RTHs fundamentally disrupts generation, causing the model to frequently collapse into nonsensical outputs, repetition, or looping behaviors. In contrast, in the FR mode, the model maintains fluency but does not retrieve the key information required to answer correctly—even when that information resides within the provided context or the model's own chain-of-thought.

To quantify these failure modes, we evaluate Qwen-2.5-7B at $k = 25$ using Qwen3-235B-A22B (Yang et al., 2025) as an LLM-as-a-judge (see Figure A12 for the evaluation prompt). We analyze "flipped" samples, which we define as instances answered correctly under No-Mask but incorrectly under both RH-Mask and RTH-Mask.

Table 2 shows the breakdown of different failure modes for two masking condition. The proportion of LoC failures is notably higher when masking RTHs compared to masking RHs. Furthermore, we observe distinct cross-lingual and cross-dataset variations. On MMLU-ProX, Swahili exhibits a unique error profile under RH-Mask compared to

Dataset	Setting	# Samples	RH Failure Modes			RTH Failure Modes		
			LoC	RF	Other	LoC	RF	Other
MMLU-ProX	en	1454	150 (10.32)	1269 (87.28)	34 (2.34)	446 (30.70)	997 (68.62)	10 (0.69)
	zh	1311	89 (6.79)	1202 (91.69)	18 (1.37)	303 (23.15)	1003 (76.62)	3 (0.23)
	de	1738	230 (13.23)	1494 (85.96)	10 (0.58)	1116 (64.29)	618 (35.60)	2 (0.12)
	sw	890	749 (84.16)	141 (15.84)	0 (0.00)	824 (92.58)	66 (7.42)	0 (0.00)
	af	1707	454 (26.60)	1236 (72.41)	17 (1.00)	1415 (82.89)	291 (17.05)	1 (0.06)
MGSM	en	52	0 (0.00)	52 (100.00)	0 (0.00)	6 (11.54)	45 (86.54)	1 (1.92)
	zh	62	3 (4.84)	54 (87.10)	5 (8.06)	6 (9.68)	56 (90.32)	0 (0.00)
	de	61	7 (11.48)	54 (88.52)	0 (0.00)	18 (29.51)	42 (68.85)	1 (1.64)
	sw	22	10 (45.45)	12 (54.55)	0 (0.00)	20 (90.91)	2 (9.09)	0 (0.00)

Table 2: **Quantitative breakdown of failure modes**—Loss of Coherence (LoC), Failure to Retrieve (RF), and Other—across MMLU-ProX and MGSM datasets for English (en), Chinese (zh), German (de), Swahili (sw) and Afrikaan (af) on Qwen2.5 7B Instruct. Each cell reports # (count) (% (percentage)). The table compares failure modes under RH vs. RTH masking on samples that flip in both settings (No-mask→RH and No-mask→RTH). Notably, masking RTHs consistently yields a higher proportion of LoC failures compared to RH masking across most languages, with German exhibiting the highest spike in loss of coherence for MMLU-ProX.

other languages, with generation failures occurring substantially more frequently than retrieval failures. Under RTH-Mask, however, lack of coherence failures mode increases substantially across all evaluated languages. We observe a similar trend on MGSM, where masking RTHs consistently induces higher rates of LoC as compared to masking RH. The primary cross-dataset discrepancy lies in how Swahili responds to RH-Mask: while RH masking triggers a high volume of generation failures on MMLU-ProX, this degradation is significantly less pronounced on MGSM. This suggests that when evaluating low-resource languages on tasks requiring generative synthesis with minimal fact retrieval (such as MGSM), RTHs are likely more critical than RHs, especially in comparison to tasks that demand precise information recovery from either the local context or parametric memory (such as MMLU-ProX). In Appendix E.3, we present several illustrative MMLU-ProX examples that highlight these failure modes: the model answers correctly under No-Mask and RH-Mask, but breaks down under RTH-Mask.

6 Conclusion and Future Work

This paper identifies Retrieval-Transition Heads (RTHs), a specialized sparse set of attention heads within multilingual LLMs responsible for bridging the gap between latent reasoning and target-language generation. Unlike retrieval heads that implement a copy-paste mechanism, RTHs bridge the target-language independent latent space to the final output. We demonstrate that RTHs are functionally indispensable for complex tasks: masking them triggers a catastrophic performance collapse on multilingual reasoning benchmarks. These findings have implications for efficiency methods. Recent frameworks for KV-cache compression (Fu et al., 2025; Tang et al., 2024), efficient inference (Xiao et al., 2024), and context re-ranking (Zhang et al., 2025b) rely on heuristics that prioritize retrieval behavior. Our results suggest these methods risk pruning the heads required for multilingual reasoning. We propose that future "transition-aware" strategies must incorporate RTH identification to preserve the specific heads responsible for translating latent context into linguistic fluency, thereby preventing the reasoning degradation that standard pruning would otherwise cause in multilingual settings.

Limitations

First, the Retrieval Transition Score (RTS) approximates the language-agnostic concept space using English as a proxy; while effective due to inherent model biases, this results in RTHs that are less sharply defined than standard retrieval heads. Another limitation of this study is its exclusive focus on attention heads. While heads are accessible units for causal masking, they represent only one part of the model's reasoning circuitry. Future work should explore the interplay between these heads, MLP 'shared neurons', and residual stream activations to provide a more holistic understanding of the latent-to-target language transition. Finally, as models are trained with more balanced multilingual data, the single-language approximation (e.g. English) becomes typologically distant from the model's latent space reducing the discriminative signal of the method. Moreover, future research should transcend this reliance on pivot lan-

guages by developing systems that learn to mimic true latent tokens. By generating a model-specific latent vocabulary, researchers can move toward identifying transition mechanisms that reflect the model's internal conceptual structures rather than cross-lingual projections.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Anonymous. 2025. Translation heads: Unveiling attention's role in LLM multilingual translation. In *Submitted to The Fourteenth International Conference on Learning Representations*. Under review.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4623–4637.

Yuxin Chen, Yiran Zhao, Yang Zhang, An Zhang, Kenji Kawaguchi, Shafiq Joty, Junnan Li, Tat-Seng Chua, Michael Qizhe Shieh, and Wenxuan Zhang. 2025. The emergence of abstract thought in large language models beyond any language. *arXiv preprint arXiv:2506.09890*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2025. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31822–31841. Association for Computational Linguistics.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. 2024. Do i know this entity? knowledge awareness and hallucinations in language models. *arXiv preprint arXiv:2411.14257*.

Yu Fu, Zefan Cai, Abedelkadir Asi, Wayne Xiong, Yue Dong, and Wen Xiao. 2025. Not all heads matter: A head-level KV cache compression method with integrated retrieval and reasoning. In *The Thirteenth International Conference on Learning Representations*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2023. A framework for few-shot language model evaluation.

Rhys Gould, Euan Ong, George Ogden, and Arthur Conmy. 2024. Successor heads: Recurring, interpretable attention heads in the wild. In *The Twelfth International Conference on Learning Representations*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuchun Fan, Xiachong Feng, Yangfan Ye, Weihong Zhong, Yuxuan Gu, Baoxin Wang, Dayong Wu, and 1 others. 2025. Improving contextual faithfulness of large language models via retrieval heads-induced optimization. *arXiv preprint arXiv:2501.13573*.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. Mlqa: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7315–7330.

Xin Liu, Qiyang Song, Qihang Zhou, Haichao Du, Shaowen Xu, Wenbo Jiang, Weijuan Zhang, and Xiaoqi Jia. 2025. Focusing on language: Revealing and exploiting language attention heads in multilingual large language models. *arXiv preprint arXiv:2511.07498*.

Callum Stuart McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. 2024. Copy suppression: Comprehensively understanding an attention head.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context learning and induction heads. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. Do multilingual llms think in english? *arXiv preprint arXiv:2502.15603*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, and 1 others. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Hanlin Tang, Yang Lin, Jing Lin, Qingsen Han, Shikuan Hong, Yiwu Yao, and Gongyi Wang. 2024. Razorattention: Efficient kv cache compression through retrieval heads. *arXiv preprint arXiv:2407.15891*.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schütze. 2025. Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models. *arXiv preprint arXiv:2504.04264*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2024. Retrieval head mechanistically explains long-context factuality. *arXiv preprint arXiv:2404.15574*.

Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. 2024. Duoattention: Efficient long-context llm inference with retrieval and streaming heads. *arXiv preprint arXiv:2410.10819*.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, and 13 others. 2025. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Suzhou, China. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Ruochen Zhang, Qinan Yu, Matianyu Zang, Carsten Eickhoff, and Ellie Pavlick. 2025a. The same but different: Structural similarities and differences in multilingual language modeling. In *The Thirteenth International Conference on Learning Representations*.

Wuwei Zhang, Fangcong Yin, Howard Yen, Danqi Chen, and Xi Ye. 2025b. Query-focused retrieval heads improve long-context reasoning and re-ranking. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23802–23816.

Yiran Zhao, Chaoqun Liu, Yue Deng, Jiahao Ying, Mahani Aljunied, Zhaodonghui Li, Lidong Bing, Hou Pong Chan, Yu Rong, Deli Zhao, and 1 others. 2025. Babel: Open multilingual large language models serving over 90% of global speakers. *arXiv preprint arXiv:2503.00865*.

Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Mingchuan Yang, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2024. Attention heads of large language models: A survey. *arXiv preprint arXiv:2409.03752*.

Tianyi Zhou, Deqing Fu, Vatsal Sharan, and Robin Jia. 2024. Pre-trained large language models use fourier features to compute addition. *Advances in Neural Information Processing Systems*, 37:25120–25151.

A Effect of model sizes, architecture, and low-resource languages to retrieval head analysis.

A.1 Effect of scaling model size on Qwen-2.5

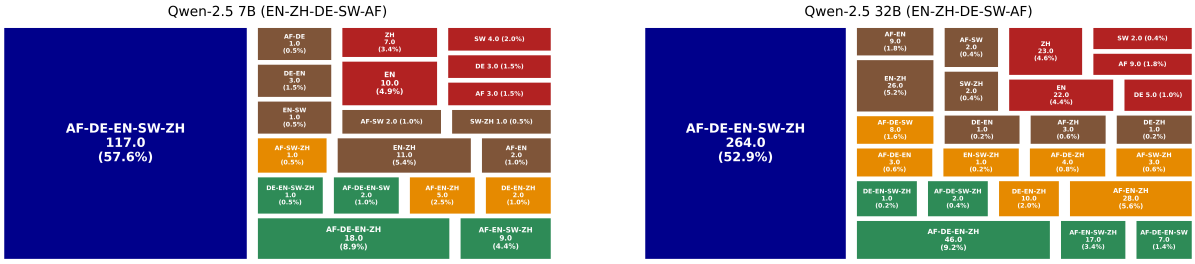


Figure A1: To assess the effect of model scaling, we evaluate Qwen2.5-14B Instruct and Qwen2.5-32B Instruct. Despite the larger parameter count, the proportions of shared and language-exclusive attention heads remain largely consistent.

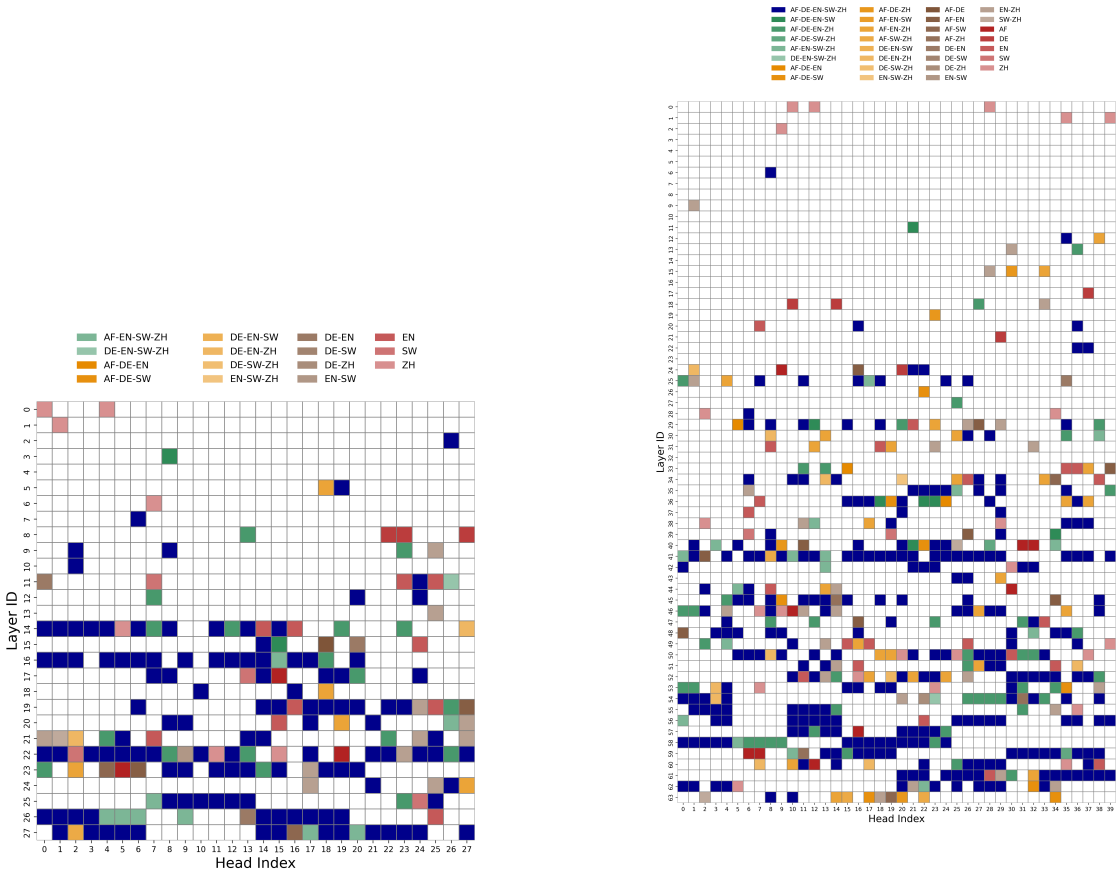


Figure A2: To assess how scaling affects the placement of shared and language-exclusive attention heads, we evaluate Qwen2.5-14B Instruct and Qwen2.5-32B Instruct. The spatial distribution of these head types remains largely consistent across scales.

A.2 Comparing Mixture-of-Experts to fully activated model architecture on Phi-3.5

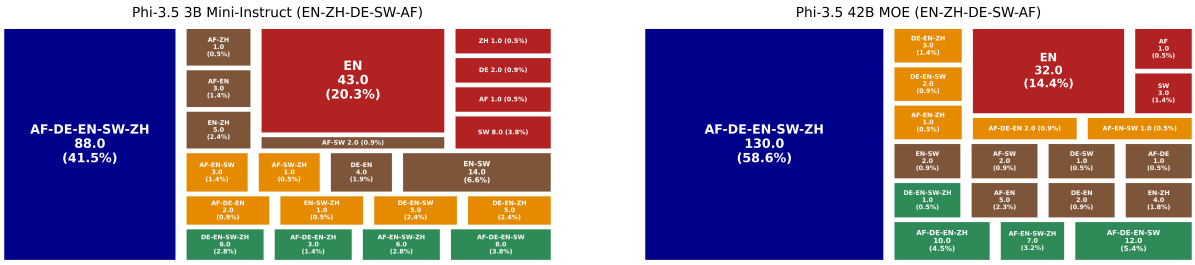


Figure A3: To compare architectures, we study Phi-3.5 3B Mini-Instruct and Phi-3.5 42B MoE (Mixture-of-Experts). For Phi-3.5 42B MoE architecture, we observe that the proportions of shared attention heads are higher compared to language-exclusive attention heads.

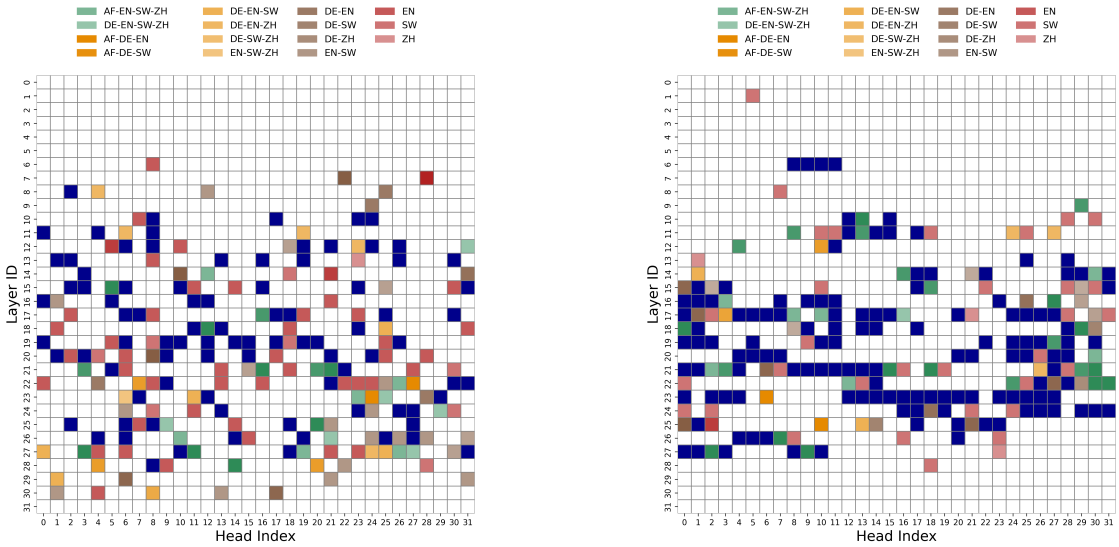


Figure A4: To assess how scaling affects the placement of shared and language-exclusive attention heads, Phi-3.5 3B Mini-Instruct and Phi-3.5 42B MoE (Mixture-of-Experts). The spatial distribution of these head types remains largely consistent across scales.

A.3 Considering the effect of including more low-resource languages

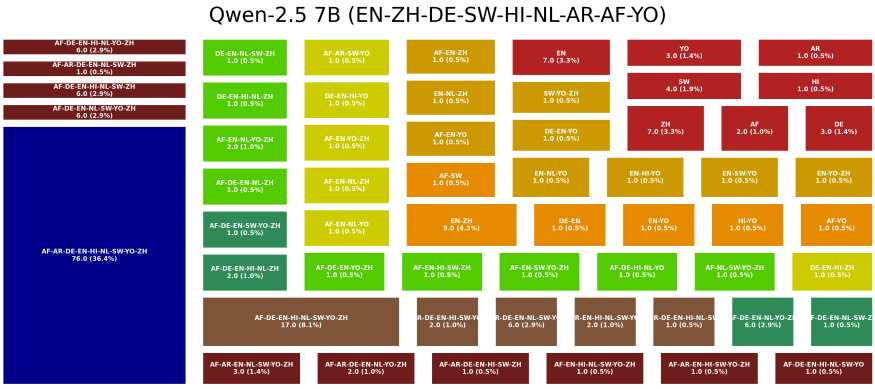


Figure A5: To study the effect of adding more low-resource languages, we evaluate Qwen2.5 7B Instruct on English, Chinese, German, Swahili, Hindi, Dutch, and Arabic. As the language set expands, the fraction of shared attention heads decreases, while language-exclusive heads become more prevalent.

B Language Specificity and Training Data Composition

The variance in retrieval head intersection observed in our analysis—ranging from the high degree of cross-lingual sharing in Qwen2.5 7B to the English-dominant specialization in Phi-3.5 3B—can be interpreted through the lens of the distinct data-scaling strategies and training recipes employed by each model family. To provide context for the variability in retrieval head behavior, we summarize the known multilingual and pre-training configurations for the three models in Table A1.

Table A1: Comparison of training data scale and multilingual distribution. Data compiled from official model cards and technical reports (Qwen et al. (2025); Abdin et al. (2024); Grattafiori et al. (2024)).

Feature	**Qwen2.5 7B**	**Llama3.1 8B**	**Phi-3.5 3B Mini**
Total Pre-training Tokens	18.0 Trillion	15.0 Trillion	3.4 Trillion
Primary Language Focus	Balanced (EN/CN)	English-Centric	English Reasoning
Non-English Proportion	High (∼40–50% Est.)	∼5% (∼750B tokens)	∼10% (Multilingual)
Post-training Strategy	Instruction/Reasoning	3.01% Multilingual SFT	Synthetic Textbooks
Multilingual Scope	29+ Official Languages	8+ Official (100+ total)	20+ Languages

We observe that the degree of cross-lingual head intersection is strongly correlated with pre-training data composition and scale. In Qwen2.5 7B, the massive 18T token corpus and high-density non-English data promote "universal" retrieval mechanisms, with 57.4% of heads shared across all five studied languages. Conversely, Phi-3.5 3B exhibits greater linguistic specialization, with the lowest 5-language intersection (41.5%) and a bias toward English-specific heads; this likely reflects its reliance on high-quality, primarily English synthetic data. These findings suggest that while scale drives language-agnostic representations, the specific data mixture determines whether a model develops universal or language-specific retrieval pathways.

C Characteristics of Retrieval Transition Heads

C.1 Are Retrieval Transition heads universally present in Multilingual Language Models?

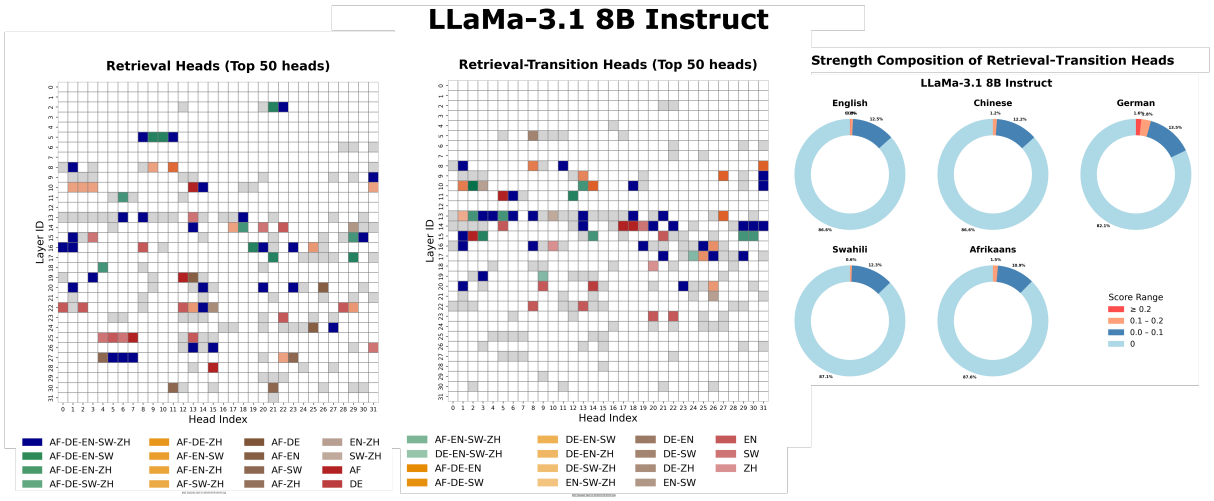


Figure A6: Layer–head distributions of Retrieval Heads (RH) and Retrieval-Transition Heads (RTH) across languages for Llama3.1 8B Instruct. Colored cells are part of Top-50 most prominent heads ranked by score. Among these colored heads, we find language-specific RH are dominant in the final layers while RTH are prominent in the middle layers. Like RH heads, RTH heads too show *sparsity* with only 0-5% of heads with an RTS score above 0.1

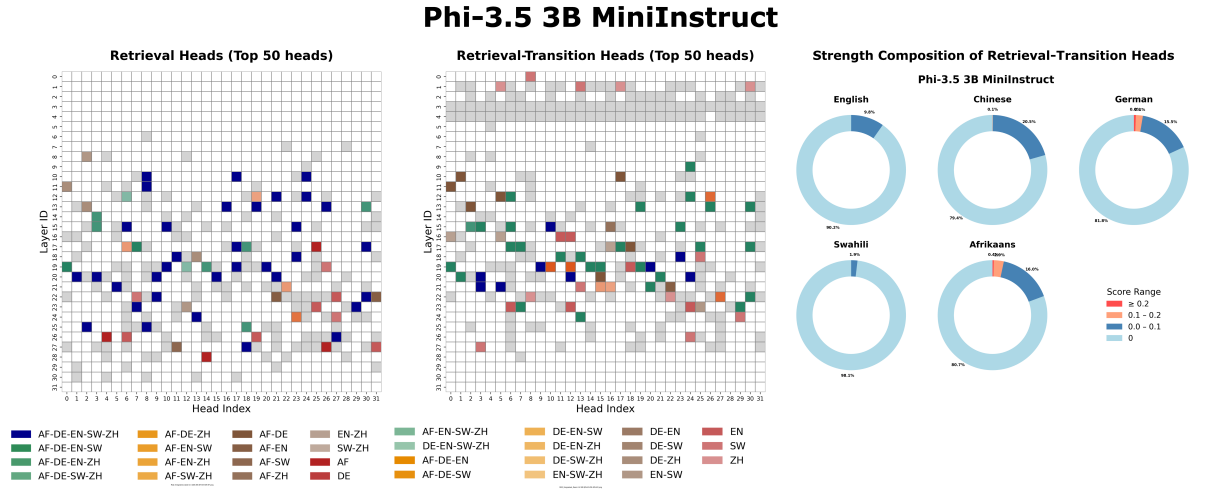


Figure A7: Layer–head distributions of Retrieval Heads (RH) and Retrieval-Transition Heads (RTH) across languages for Phi-3.5 3B MiniInstruct. Colored cells are part of Top-50 most prominent heads ranked by score. Among these colored heads, we find language-specific RH are dominant in the final layers while RTH are prominent in the middle layers. Like RH heads, RTH heads too show *sparsity* with only 0-3% of heads with an RTS score above 0.1

C.2 Effect of larger model size and MOE architecture on Retrieval Transition head distribution

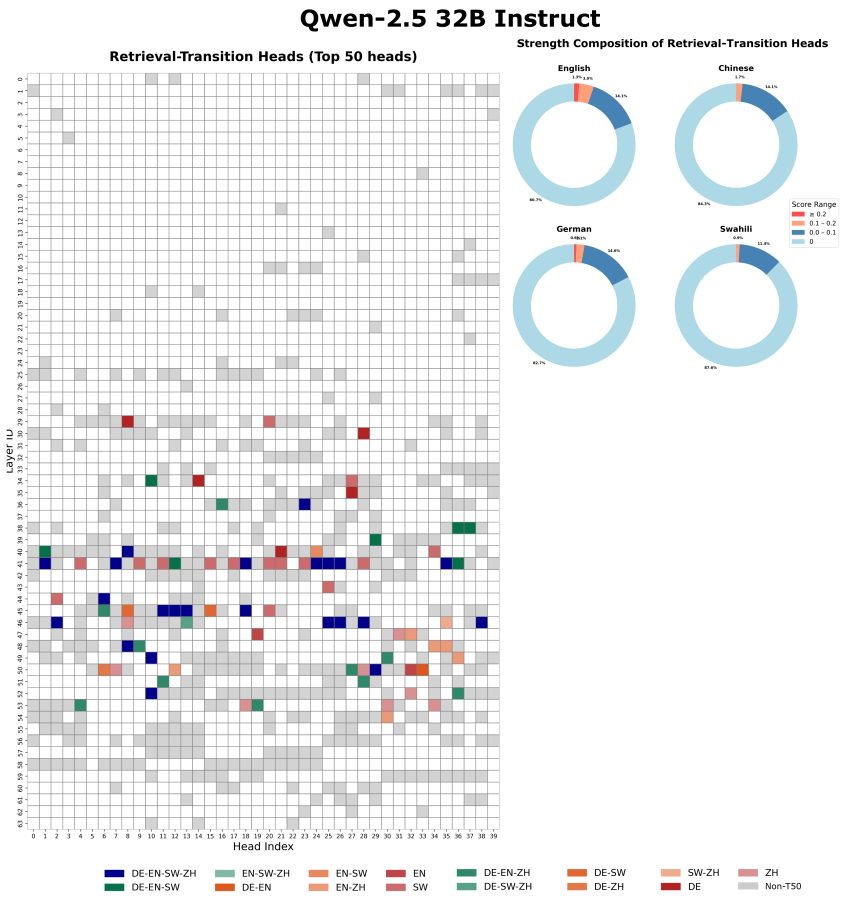


Figure A8: Layer–head distributions of Retrieval-Transition Heads (RTH) across languages for Qwen-2.5 32B Instruct. Colored cells are part of Top-50 most prominent heads ranked by score. We find that RTH heads consistently appear in larger model sizes as well.

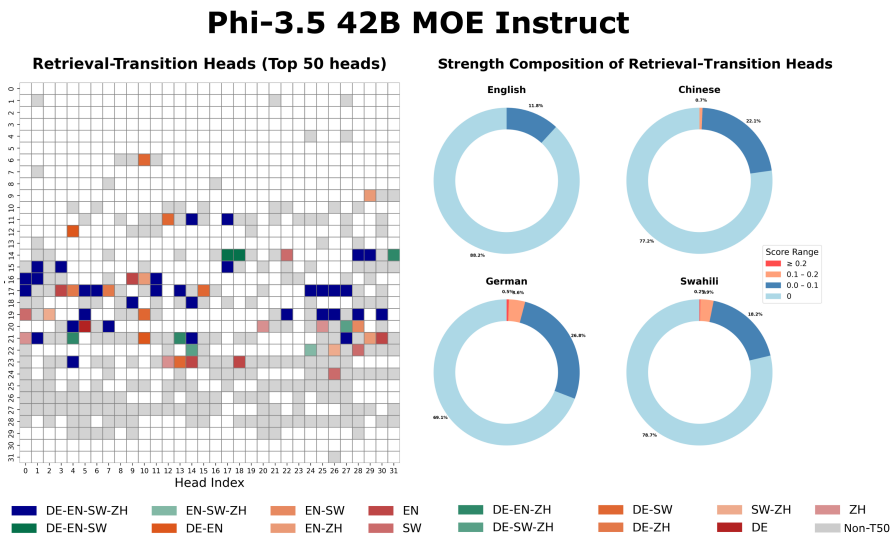


Figure A9: Layer–head distributions of Retrieval-Transition Heads (RTH) across languages for Phi-3.5 42B MOE Instruct. Colored cells are part of Top-50 most prominent heads ranked by score. We find that RTH heads also appear in different model architectures like Mixture-of-Experts.

C.3 Do Retrieval Transition heads also appear in pretrained-only models?

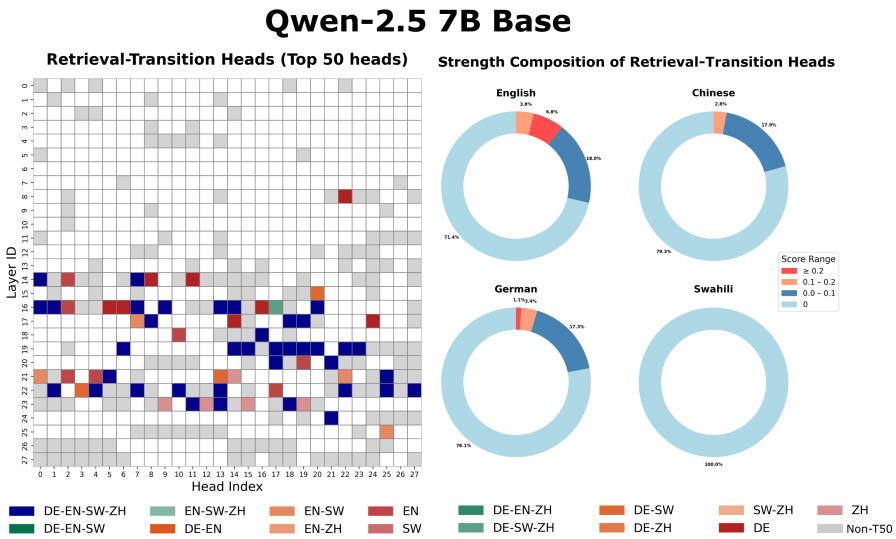


Figure A10: Layer–head distributions of Retrieval-Transition Heads (RTH) across languages for Qwen2.5 7B Base. Colored cells are part of Top-50 most prominent heads ranked by score. We find that RTH heads are also present in non-Instruct models, however are primarily English language dominated.

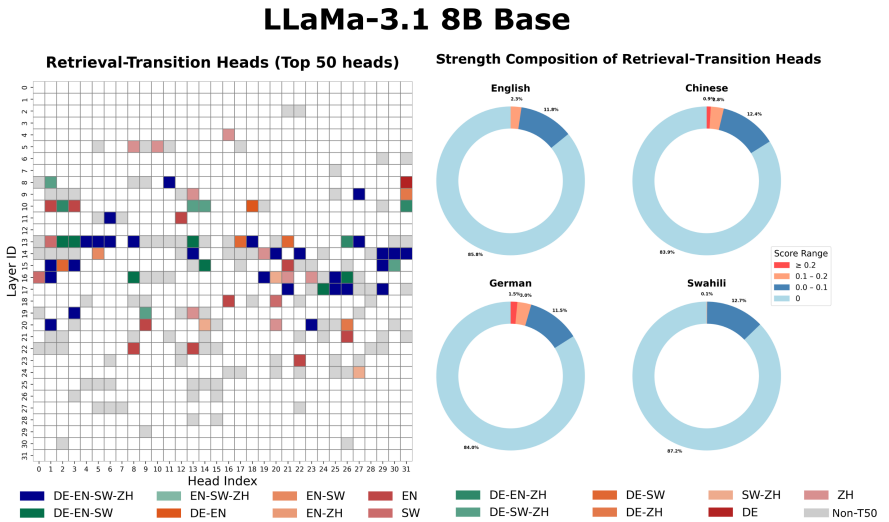


Figure A11: Layer–head distributions of Retrieval-Transition Heads (RTH) across languages for Llama3.1 8B Base. Colored cells are part of Top-50 most prominent heads ranked by score. We find that RTH heads are also present in non-Instruct models, however are primarily English language dominated.

C.4 Analyzing variability introduced by LLM-based aligner

Language	Top-15	Top-25	Top-50	Top-100
RTH_{en}	15	25	50	100
RTH_{zh}	13	23	48	98

Table A2: Intersection size of the top-k retrieval-transition heads identified using `Qwen3-30B-A3B-Thinking-2507` versus `GPT-4.1` as the LLM-based aligner. We run the analysis on `Qwen-2.5-7B Instruct` for RTH_{en} and RTH_{zh}. The top-k RTH sets largely overlap across aligners, indicating that head identification is robust to the choice of aligner.

D Causal Intervention through Attention Head Masking

Following (Wu et al., 2024), we quantify the causal impact of individual attention heads by performing targeted interventions during inference. Specifically, we perturb the attention logits of a predefined subset of heads prior to the softmax operation, effectively neutralizing their contribution to the subsequent layer representations.

Let $A^{(l,h)} \in \mathbb{R}^{q \times k}$ denote the pre-softmax attention logits for head h in layer l, computed as:

$$A^{(l,h)} = \frac{Q^{(l,h)} K^{(l,h)\top}}{\sqrt{d_h}} \tag{3}$$

where $Q^{(l,h)}$ and $K^{(l,h)}$ are the query and key projections, respectively. For a head designated for ablation, we replace its logit matrix with a large negative constant:

$$A^{(l,h)} \leftarrow -C \cdot \mathbf{1}_{q \times k} \tag{4}$$

where C is a sufficiently large scalar (we set $C = 10^9$). This intervention ensures that, upon softmax normalization, the resulting attention probabilities approach zero, effectively preventing the head from aggregating information from any key positions.

This masking is localized; all other heads and model parameters remain unmodified, and the residual stream is only affected by the absence of the ablated head's output. Operationally, we specify a set of $(\text{layer}, \text{head})$ tuples at inference time to selectively ablate different functional categories, such as retrieval or retrieval-transition heads, alongside randomly sampled heads as a control condition.

We evaluate the causal importance of these components by measuring the degradation in downstream benchmark performance relative to the original, unperturbed model. Significant performance loss under a specific mask is interpreted as evidence of that head's functional necessity for the task.

E Additional experiments and Qualitative analysis of Retrieval-Transition Heads

E.1 Effect of hyperparameter k (number of heads masked)

	Heads	MMLU-ProX (Acc.)						XQuAD (F1-score)			
		en	zh	de	sw	af	**Avg**	en	zh	de	**Avg**
Qwen2.5 7B	None	42.56	44.22	40.57	22.59	32.98	36.58	15.72	0.82	14.28	10.27
	Random	$43.35^{+0.79}$	$42.66^{-1.56}$	$39.57^{-1.00}$	$22.32^{-0.27}$	$32.59^{-0.39}$	$36.10^{-0.49}$	$15.94^{+0.22}$	$0.93^{+0.11}$	$13.23^{-1.05}$	$10.03^{-0.24}$
	RH	$33.18^{-9.38}$	$33.42^{-10.80}$	$34.71^{-5.86}$	$14.18^{-8.41}$	$28.45^{-4.53}$	$28.79^{-7.80}$	$13.44^{-2.28}$	$0.38^{-0.44}$	$8.97^{-5.31}$	$7.60^{-2.68}$
	RTH	$30.61^{-11.95}$	$32.38^{-11.84}$	$7.07^{-33.50}$	$15.43^{-7.16}$	$3.92^{-29.06}$	$\mathbf{17.88}^{\mathbf{-18.70}}$	$11.80^{-3.92}$	$0.36^{-0.46}$	$5.78^{-8.50}$	$\mathbf{5.98}^{\mathbf{-4.29}}$
Llama3.1 8B	None	43.59	31.41	30.89	13.27	27.38	29.31	35.26	28.70	31.10	31.69
	Random	$41.70^{-1.89}$	$28.94^{-2.47}$	$22.87^{-8.02}$	$12.71^{-0.56}$	$21.39^{-5.99}$	$25.52^{-3.79}$	$37.18^{+1.92}$	$29.57^{+0.87}$	$33.03^{+1.93}$	$33.26^{+1.57}$
	RH	$17.03^{-26.56}$	$20.70^{-10.71}$	$12.65^{-18.24}$	$3.27^{-10.00}$	$6.81^{-20.57}$	$12.09^{-17.22}$	$30.35^{-4.91}$	$22.84^{-5.86}$	$26.12^{-4.98}$	$26.44^{-5.25}$
	RTH	$11.42^{-32.17}$	$24.76^{-6.65}$	$2.35^{-28.54}$	$3.50^{-9.77}$	$1.03^{-26.35}$	$\mathbf{8.61}^{\mathbf{-20.70}}$	$19.66^{-15.60}$	$24.99^{-3.71}$	$29.38^{-1.72}$	$\mathbf{24.68}^{\mathbf{-7.01}}$

	Heads	MGSM (Acc.)					MLQA (F1-score)			
		en	zh	de	sw	**Avg**	en	zh	de	**Avg**
Qwen2.5 7B	None	33.6	31.6	27.2	6.4	24.70	17.26	18.93	16.16	17.45
	Random	$32.0^{-1.6}$	$27.2^{-4.4}$	$30.0^{+2.8}$	$4.0^{-2.4}$	$23.30^{-1.40}$	$18.61^{+1.35}$	$19.88^{+0.95}$	$17.29^{+1.13}$	$18.59^{+1.14}$
	RH	$31.2^{-2.4}$	$22.8^{-8.8}$	$21.2^{-6.0}$	$3.6^{-2.8}$	$19.70^{-5.00}$	$15.05^{-2.21}$	$12.74^{-6.19}$	$11.20^{-4.96}$	$13.00^{-4.45}$
	RTH	$24.0^{-9.6}$	$2.8^{-28.8}$	$3.2^{-24.0}$	$0.8^{-5.6}$	$\mathbf{7.70}^{\mathbf{-17.00}}$	$13.27^{-3.99}$	$8.29^{-10.64}$	$6.54^{-9.62}$	$\mathbf{9.37}^{\mathbf{-8.08}}$
Llama3.1 8B	None	73.2	62.4	59.2	48.0	60.70	39.05	49.49	36.76	41.77
	Random	$69.2^{-4.0}$	$55.6^{-6.8}$	$56.0^{-3.2}$	$27.2^{-20.8}$	$52.00^{-8.70}$	$40.29^{+1.24}$	$49.49^{0.00}$	$38.61^{+1.85}$	$42.80^{+1.03}$
	RH	$56.0^{-17.2}$	$42.0^{-20.4}$	$38.8^{-20.4}$	$18.8^{-29.2}$	$38.90^{-21.80}$	$34.10^{-4.95}$	$44.25^{-5.24}$	$32.49^{-4.27}$	$36.95^{-4.82}$
	RTH	$3.6^{-69.6}$	$5.6^{-56.8}$	$15.2^{-44.0}$	$2.4^{-45.6}$	$\mathbf{6.70}^{\mathbf{-54.00}}$	$21.90^{-17.15}$	$38.92^{-10.57}$	$34.41^{-2.35}$	$\mathbf{31.74}^{\mathbf{-10.02}}$

Table A3: **Per-language performance degradation under top-k ($k = 25$) head masking.** Accuracy (MMLU-ProX, MGSM) and F1-scores (MLQA, XQuAD) are reported for Qwen-2.5-7B and Llama3.1 8B. The **Avg** column reports the mean across the evaluated languages, with the colored superscript denoting the absolute change relative to the no-masking baseline. Results highlight that RTH masking consistently induces the largest performance drops across high-resource (en, zh, de) and often in low-resource (sw) languages. Crucially, English-language tasks are also affected when masking RTH, suggesting that the latent reasoning mechanism in multilingual language models is invariant to the target language.

	Heads	MMLU-ProX (Acc.)						MLQA (F1-score)			
		en	zh	de	sw	af	Avg	en	zh	de	Avg
Qwen2.5 7B	None	42.56	44.22	40.57	22.59	32.98	36.58	17.26	18.93	16.16	17.45
	Random	42.85$^{+0.29}$	42.08$^{-2.14}$	29.94$^{-10.63}$	21.90$^{-0.69}$	29.65$^{-3.33}$	33.28$^{-3.30}$	16.71$^{-0.55}$	17.87$^{-1.06}$	14.26$^{-1.90}$	16.28$^{-1.17}$
	RH	39.85$^{-2.71}$	39.87$^{-4.35}$	37.08$^{-3.49}$	20.09$^{-2.50}$	29.82$^{-3.16}$	33.34$^{-3.24}$	15.82$^{-1.44}$	18.28$^{-0.65}$	15.11$^{-1.05}$	16.40$^{-1.05}$
	RTH	30.64$^{-11.92}$	35.37$^{-8.85}$	24.83$^{-15.74}$	15.58$^{-7.01}$	19.02$^{-13.96}$	**25.09$^{-11.50}$**	13.69$^{-3.57}$	15.17$^{-3.76}$	8.95$^{-7.21}$	**12.60$^{-4.85}$**
Llama3.1 8B	None	43.59	31.41	30.89	13.27	27.38	29.31	39.05	49.49	36.76	41.77
	Random	43.17$^{-0.42}$	31.47$^{+0.06}$	26.94$^{-3.95}$	14.60$^{+1.33}$	20.89$^{-6.49}$	27.41$^{-1.89}$	40.35$^{+1.30}$	50.11$^{+0.62}$	39.20$^{+2.44}$	43.22$^{+1.45}$
	RH	40.23$^{-3.36}$	26.97$^{-4.44}$	22.42$^{-8.47}$	1.70$^{-11.57}$	27.36$^{-0.02}$	23.74$^{-5.57}$	39.08$^{+0.03}$	46.82$^{-2.67}$	36.33$^{-0.43}$	40.74$^{-1.02}$
	RTH	9.57$^{-34.02}$	26.27$^{-5.14}$	3.48$^{-27.41}$	4.13$^{-9.14}$	0.94$^{-26.44}$	**8.88$^{-20.43}$**	30.88$^{-8.17}$	43.83$^{-5.66}$	36.02$^{-0.74}$	**36.91$^{-4.86}$**

	Heads	MGSM (Acc.)					XQuAD (F1-score)			
		en	zh	de	sw	Avg	en	zh	de	Avg
Qwen2.5 7B	None	33.6	31.6	27.2	6.4	24.70	15.72	0.82	14.28	10.27
	Random	29.6$^{-4.0}$	26.8$^{-4.8}$	18.0$^{-9.2}$	2.4$^{-4.0}$	19.20$^{-5.50}$	13.82$^{-1.90}$	0.88$^{+0.06}$	10.85$^{-3.43}$	8.52$^{-1.76}$
	RH	36.4$^{+2.8}$	22.4$^{-9.2}$	23.2$^{-4.0}$	3.6$^{-2.8}$	21.40$^{-3.30}$	13.89$^{-1.83}$	0.56$^{-0.26}$	12.64$^{-1.64}$	9.03$^{-1.24}$
	RTH	18.0$^{-15.6}$	23.6$^{-8.0}$	5.6$^{-21.6}$	2.0$^{-4.4}$	**12.30$^{-12.40}$**	12.04$^{-3.68}$	0.46$^{-0.36}$	7.34$^{-6.94}$	**6.61$^{-3.66}$**
Llama3.1 8B	None	73.2	62.4	59.2	48.0	60.70	35.26	28.70	31.10	31.69
	Random	72.4$^{-0.8}$	58.4$^{-4.0}$	57.6$^{-1.6}$	28.4$^{-19.6}$	54.20$^{-6.50}$	37.28$^{+2.02}$	28.81$^{+0.11}$	34.46$^{+3.36}$	33.52$^{+1.83}$
	RH	72.0$^{-1.2}$	62.4 $^{0.0}$	52.4$^{-6.8}$	48.0 $^{0.0}$	58.70$^{-2.00}$	36.68$^{+1.42}$	21.43$^{-7.27}$	29.52$^{-1.58}$	**29.21$^{-2.48}$**
	RTH	21.2$^{-52.0}$	24.8$^{-37.6}$	25.6$^{-33.6}$	19.6$^{-28.4}$	**22.80$^{-37.90}$**	28.48$^{-6.78}$	28.70 $^{0.00}$	30.88$^{-0.22}$	29.35$^{-2.33}$

Table A4: Impact of top-15 head masking on MMLU-ProX, MGSM, MLQA, and XQuAD using Qwen2.5 7B Instruct and Llama3.1 8B Instruct. We report accuracy for MMLU-ProX and MGSM, and F1 score for MLQA and XQuAD, computed using the lm-evaluation-harness. The **Avg** column reports the mean across the evaluated languages, with the colored superscript denoting the absolute change relative to the no-masking baseline.

	Heads	MMLU-ProX (Acc.)						MLQA (F1-score)			
		en	zh	de	sw	af	Avg	en	zh	de	Avg
Qwen2.5 7B	None	42.56	44.22	40.57	22.59	32.98	36.58	17.26	18.93	16.16	17.45
	Random	40.95$^{-1.61}$	39.71$^{-4.51}$	35.75$^{-4.82}$	22.31$^{-0.28}$	29.84$^{-3.14}$	33.71$^{-2.87}$	17.33$^{+0.07}$	18.16$^{-0.77}$	16.27$^{+0.11}$	17.25$^{-0.20}$
	RH	19.48$^{-23.08}$	18.98$^{-25.24}$	14.84$^{-25.73}$	0.30$^{-22.29}$	8.10$^{-24.88}$	12.34$^{-24.24}$	6.23$^{-11.03}$	6.63$^{-12.30}$	5.05$^{-11.11}$	5.97$^{-11.48}$
	RTH	14.14$^{-28.42}$	24.13$^{-20.09}$	3.24$^{-37.33}$	11.05$^{-11.54}$	5.46$^{-27.52}$	**11.60$^{-24.98}$**	5.03$^{-12.23}$	5.39$^{-13.54}$	5.04$^{-11.12}$	**5.15$^{-12.30}$**
Llama3.1 8B	None	43.59	31.41	30.89	13.27	27.38	29.31	39.05	49.49	36.76	41.77
	Random	36.98$^{-6.61}$	27.28$^{-4.13}$	16.92$^{-13.97}$	5.60$^{-7.67}$	29.84$^{+2.46}$	23.32$^{-5.98}$	38.53$^{-0.52}$	46.12$^{-3.37}$	37.93$^{+1.17}$	40.86$^{-0.91}$
	RH	15.32$^{-28.27}$	10.83$^{-20.58}$	1.25$^{-29.64}$	0.80$^{-12.47}$	6.81$^{-20.57}$	7.00$^{-22.31}$	31.13$^{-7.92}$	43.17$^{-6.32}$	22.88$^{-13.88}$	32.39$^{-9.37}$
	RTH	5.92$^{-37.67}$	20.48$^{-10.93}$	1.40$^{-29.49}$	2.26$^{-11.01}$	1.11$^{-26.27}$	**6.23$^{-23.07}$**	22.12$^{-16.93}$	33.51$^{-15.98}$	26.56$^{-10.20}$	**27.40$^{-14.37}$**

	Heads	MGSM (Acc.)					XQuAD (F1-score)			
		en	zh	de	sw	Avg	en	zh	de	Avg
Qwen2.5 7B	None	33.6	31.6	27.2	6.4	24.70	15.72	0.82	14.28	10.27
	Random	16.4$^{-17.2}$	18.4$^{-13.2}$	13.2$^{-14.0}$	4.4$^{-2.0}$	13.10$^{-11.60}$	14.96$^{-0.76}$	0.67$^{-0.15}$	12.67$^{-1.61}$	9.43$^{-0.84}$
	RH	4.4$^{-29.2}$	5.6$^{-26.0}$	5.6$^{-21.6}$	1.2$^{-5.2}$	4.20$^{-20.50}$	5.57$^{-10.15}$	0.41$^{-0.41}$	4.29$^{-9.99}$	**3.42$^{-6.85}$**
	RTH	2.0$^{-31.6}$	0.8$^{-30.8}$	2.0$^{-25.2}$	0.4$^{-6.0}$	**1.30$^{-23.40}$**	4.86$^{-10.86}$	0.34$^{-0.48}$	5.19$^{-9.09}$	3.46$^{-6.81}$
Llama3.1 8B	None	73.2	62.4	59.2	48.0	60.70	35.26	28.70	31.10	31.69
	Random	67.2$^{-6.0}$	45.2$^{-17.2}$	41.6$^{-17.6}$	14.0$^{-34.0}$	42.00$^{-18.70}$	36.47$^{+1.21}$	20.81$^{-7.89}$	33.30$^{+2.20}$	30.19$^{-1.49}$
	RH	20.8$^{-52.4}$	25.6$^{-36.8}$	2.0$^{-57.2}$	4.0$^{-44.0}$	13.10$^{-47.60}$	29.16$^{-6.10}$	22.49$^{-6.21}$	18.84$^{-12.26}$	23.50$^{-8.19}$
	RTH	2.8$^{-70.4}$	1.6$^{-60.8}$	2.4$^{-56.8}$	3.2$^{-44.8}$	**2.50$^{-58.20}$**	18.04$^{-17.22}$	18.24$^{-10.46}$	19.87$^{-11.23}$	**18.72$^{-12.97}$**

Table A5: Impact of top-50 head masking on MMLU-ProX, MGSM, MLQA, and XQuAD using Qwen2.5 7B Instruct and Llama3.1 8B Instruct. We report accuracy for MMLU-ProX and MGSM, and F1 score for MLQA and XQuAD, computed using the lm-evaluation-harness. The **Avg** column reports the mean across the evaluated languages, with the colored superscript denoting the absolute change relative to the no-masking baseline.

E.2 LLM-as-a-judge prompt for failure-mode categorization

```
# LLM Judge: Failure Mode Classification

You are an expert evaluator tasked with analyzing why a language model failed to produce a correct response after
a causal intervention (e.g., attention masking).

## Your Task

You will be provided with three artifacts:
1. **Question**: The original question posed to the model
2. **Response 1 (Correct)**: The model's correct response without intervention (oracle reference)
3. **Response 2 (Incorrect)**: The model's incorrect response after intervention

Your goal is to classify the PRIMARY failure mode that caused Response 2 to be incorrect, using Response 1 as the
reference for what a correct response should look like.

---

## Failure Mode Categories

### A. Lack of Coherence

**Definition**: The model produces incoherent, ill-formed, or nonsensical output that fails to maintain logical
thought progression.

**Key Characteristics**:
- Ill-formed sentences or grammatically broken language
- Gibberish, word salad, or meaningless text
- Repetitive cycles of the same phrases or words
- Derails into unrelated tangents
- Cannot maintain a coherent train of thought
- May have correct facts but cannot organize them into a sensible response

**Example**: Model correctly retrieves facts but then produces: "The answer is... the answer... because the
statement is false true false... repeating cycle... cannot determine..."

---

### B. Lack of Retrieval Capabilities

**Definition**: The model produces coherent and fluent language but arrives at the wrong answer due to missing,
incorrect, or hallucinated information.

**Key Characteristics**:
- Language is fluent, grammatically correct, and well-structured
- Reasoning chain is coherent but based on wrong premises
- Missing key information from the context or its own reasoning
- Hallucinates facts not present in the context
- Fails to retrieve or recall relevant information from earlier in its chain-of-thought
- Makes logical inferences but from incorrect or incomplete information

**Example**:
**Context**: "Solve 14 = w + 23. Options: A) 9  B) -9  C) -11  D) -14"
**Model Output**: "To isolate w, we subtract 23 from both sides... This gives -9 = w. Therefore, the solution is
not directly included in the given options."
(Coherent and mathematically correct, but fails to retrieve "Option B" from the context and hallucinates that the
answer is missing).

---

### C. Other

**Definition**: The failure does not fit into the above two categories.

**When to use**:
- Model produces correct reasoning but outputs in the wrong language
- Model produces correct answer but in wrong format
- Technical errors (output truncation, encoding issues)
- Any other failure mode not captured by Lack of Coherence or Lack of Retrieval

**Important**: If you select "Other", you MUST provide a brief one-line description of the actual failure mode.

---

## Output Format

You must respond with ONLY a valid JSON object in the following format:

```json
{
  "failure_mode": "Lack of Coherence" | "Lack of Retrieval Capabilities" | "Other",
  "other_desc": "<one-line description if Other, otherwise empty string>"
}
```
```

Figure A12: LLM-as-a-judge prompt used to categorize flipped samples into failure modes (Loss of Coherence, Failure to retrieve, Other). The judge is given the original question, the correct answer chain-of-thought from the No-Mask condition, and the model outputs under different masking settings, and is instructed to assign a single failure label along with a brief justification.

E.3 Qualitative Examples of failure modes on masking Retrieval-Transition heads

Failure Mode 1: Lack of Coherence. Figure A13 shows the output comparison of Qwen2.5 7B for a German MMLU-ProX Math problem. The unmasked model outputs a concise and correct thought chain. However, when top-25 retrieval-transition heads for German (RTH_{de}) are masked, the model fails to reach the correct answer. This is even after the model retrieves the correct facts for both statements, correctly identifying the first as false and the second as true, but still unable to form a coherent thought. Instead, it devolves into repetitive, nonsensical cycles and fails to provide a final answer before hitting the decoding limit.

Effect of Retrieval-Transition Heads on language coherence

Question 1 : Antwort: Richtig oder falsch? Aussage 1: Jede Menge von zwei Vektoren in R^2 ist linear unabhängig.
Aussage 2: Wenn V = span(v1, ... , vk) und {v1, ... , vk} linear unabhängig sind, dann ist dim(V) = k.
(Translated) Question 1 : *Answer true or false. Statement 1: Any set of two vectors in R^2 is linearly independent. Statement 2: If V = span(v1, ... , vk) and {v1, ... , vk} are linearly independent, then dim(V) = k.*

(No attention heads masked)
Model Output : Diese Aussage 1 ist falsch.
Wenn man mehr als zwei Vektoren in R^2 nimmt, sind sie notwendigerweise linear abhängig.
Diese Aussage 2 ist wahr.
Wenn ein Vektorraum V durch eine Basis {v1, ... , vk} ... Antwort : Falsch, Wahr

(Translated)
This statement 1 is false. If there are more than two vectors in R^2, they are necessarily linearly dependent. This statement 2 is true.
If a vector space V is defined by a basis {v1, ... , vk},..
Answer : False, True

(Top 25 Retrieval-Transition Heads masked)
Model Output : Zuerst betrachten wir Aussage 1 in in Aussage 2 ... Wenn die Vektoren parallel sind, dann sind sie linear abhängig. Daher ist Aussage 1 in Aussage 2 falsch. Aussage 2 in Aussage 1: ... die Aussage 2 in Aussage 1 ist die Aussage 2 in Aussage 1 korrekt, da sie die Dimension der Spannung von Vektoren angibt ...

(Translated)
First, consider statement 1 in statement 2 ... If the vectors are parallel, then they are linearly dependent. Therefore, statement 1 in statement 2 is false. For statement 2 in statement 1.... statement 2 in statement 1, statement 2 in statement 1 is correct, since it specifies the dimension of the spans of vectors ...

Figure A13: Masking RTHs disrupts the mapping from latent reasoning to target-language syntax. While the model may grasp the underlying fact, it fails to synthesize grammatical output, resulting in structural gibberish or character-level hallucinations.

Failure Mode 2: Failure to Retrieve Capabilities. Even for fairly simple problems, requiring little reasoning, masking RTH heads causes hallucinations and incomplete retrieval, like RH heads caused in long-context reasoning (Wu et al., 2024). Figure A14 shows one such example using a simple Arithmetic in German. The model performs the arithmetic correctly and calculates the value of -9, but is unable to retrieve the corresponding option from the context due to top 25 most influential RTH heads being masked. It incorrectly concludes that the correct answer is not present in the options and ultimately starts hallucinating. Such failures are not observed when masking the top-25 RH for that language suggesting that RTH act as the essential bridge allowing an MLM to map internal reasoning back to specific target-language tokens and options.

Effect of Retrieval-Transition Heads on ability to retrieve information from given context

Question 2 : Löse die Gleichung 14 = w + 23.
Optionen: A] 9 B] -9 C] -11 D] -14

(Translated) Question 2: *Solve equation 14 = w + 23*
Options: A] 9 B] -9 C] -11 D] -14

(No attention heads masked)
Model Output : Wir suchen eine Zahl (x) so, dass (4 times x = 36).Um (x) zu finden, können wir beide Seiten der Gleichung durch 4 teilen:[x = frac{36}{4} = 9]Die Antwort ist (F).

(Translated)
We are looking for a number (x) such that (4 times x = 36). To find (x), we can divide both sides of the equation by 4: [x = frac{36}{4} = 9]. The answer is (F).

(Top 25 Retrieval Translation Heads masked)
Model Output : Die Gleichung lautet:\
n[14 = w + 23]. Um (w) zu isolieren, subtrahieren wir 23 von beiden Seiten der Gleichung:[14 - 23 = w]
Das ergibt:[-9 = w] Daher ist die Lösung nicht direkt in den gegebenen Optionen enthalten, aber wir können die Gleichung umstellen, um die korrekte Form zu erhalten

(Translated)
The equation is: n[14 = w + 23] To isolate (w), we subtract 23 from both sides of the equation: [14 - 23 = w]
This gives: [-9 = w] Therefore, the solution is not directly included in the given options, but we can rearrange the equation to obtain the correct form:

Figure A14: The model loses the ability to align internal reasoning with the prompt's constraints. Even if the logic remains correct, the model is unable look-back to the provided multiple-choice options, leading to answers that are ungrounded in the context.