

Bridging Latent Reasoning and Target-Language Generation via Retrieval-Transition Heads

Anonymous ACL submission

Abstract

Recent work has identified retrieval heads, specialized attention heads in Transformer largely responsible for retrieving information from the context. In this work, we investigate retrieval heads in multilingual contexts and then ask whether these are indeed the most important attention heads for multilingual reasoning. In multilingual language models, we find that retrieval heads are often shared across multiple languages. We further identify *Retrieval-Transition heads (RTH)*, which govern the shift from language-agnostic latent reasoning to target-language output. We quantify the semantic mapping between the concept space and target language tokens to identify RTHs. Unlike standard retrieval heads that copy from in-context, RTHs facilitate latent-to-language semantic mapping. Our experiments reveal that RTHs are distinct from retrieval heads and more vital for Chain-of-Thought reasoning in multilingual LLMs. For Qwen-2.5 7B Instruct, masking the 25 most influential RTHs triggers a 22.7-point average drop in MMLU-ProX accuracy across four languages, higher than the 14.3-point average decline observed when masking an equal number of standard retrieval heads. These results prove that RTHs are functionally unique and essential for maintaining both reasoning and coherence in multilingual models. Our work advances understanding of multilingual LMs by isolating the attention heads responsible for latent-to-language mapping and demonstrating its role in multilingual reasoning performance.

1 Introduction

Large Language Models (LLMs) have achieved state-of-the-art performance in reasoning and knowledge-intensive tasks (DeepSeek-AI et al., 2025; Qwen et al., 2025). As LLMs are deployed globally, their ability to retrieve, reason, and respond across multiple languages became crucial. Recent research has linked specific behaviors to

internal circuits (Zhou et al., 2024; Ferrando et al., 2024), notably 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 investigate multilingual retrieval by extending standard Needle-In-A-Haystack (NIAH) analysis to multiple languages. Evaluating three model families across English, German, Chinese, and Swahili, we find a consistent mixture of language-shared and language-exclusive retrieval heads. For instance, Qwen2.5-7B exhibits 59% overlap across all four languages, while Phi-3.5 shows significantly higher language-specific specialization with 21% en-only heads (Figure 2). We observe a rough correlation between model pre-training corpus and the "shared" linguistic nature of its retrieval head, similar to analogous findings for MLP heads (Chen et al., 2025).

We argue that multilingual proficiency requires more than "copy-paste" retrieval; models must translate retrieved evidence into syntactically and semantically grounded target-language output. Models often utilize a language-agnostic latent space, before mapping intermediate "thoughts" into the final target language (Wendler et al., 2024a; Wang et al., 2025). This raises a critical question: *what enables the transition from language-agnostic reasoning to language-specific token generation?* We define Retrieval-Transition Heads (RTH) that facilitate this step. To quantify this, we introduce the Retrieval-Transition Score (RTS), which measures whether a head's attention aligns target-language output tokens with their corresponding source-language "needle" counterparts (Figure 3).

Finally, we demonstrate that RTHs are causally necessary for reasoning in multilingual LLMs, *even for English language tasks*. On the MMLU-ProX benchmark, masking RTHs induces a severe collapse in Chain-of-Thought (CoT) performance. For example, English accuracy drops 57.4% to

32.2%, and German accuracy drops from 47.3% to 7.07%—a sharper drop than caused by masking standard RHs (Figure 6).

Our main 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 propose 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 MMLU-ProX and qualitative analysis of failure modes, we causally prove that RTHs are essential for CoT reasoning and linguistic coherence in multilingual LLMs.

2 Related Work

Role of attention heads Various studies have proposed techniques to explain 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. They observe that emergence of Induction Heads coincides with sudden improvement in in-context learning abilities of the models. Most relevant to our work, Wu et al. (2024) identify *Retrieval Heads* as the primary drivers of long-form 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 discovery has enabled practical applications, 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. (2024a) 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 but 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 behaviour and develop a core language-agnostic parameter space composed of MLP heads termed as "shared neurons". Further, Zhang et al. (2025) 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, the backbone of complex reasoning, 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 standard 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 identify retrieval heads for different languages using the Needle-In-A-Haystack (NIAH) task

(Kamradt, 2023) following the same experimental setup as Wu et al. (2024). For completeness, we describe the process 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 about the answer and has to retrieve k to answer the question. NIAH showcases a model’s ability to accurately answer questions based on a long context. In our study, we further extend NIAH to various languages.

Retrieval score (RS): Following prior work Wu et al. (2024), we calculate an attention head’s retrieval score by considering the proportion of decoded tokens part of the needle sentence where the head assigned the most attention weight. Concretely, we define the score for an attention head per example S_h as:

$$S_h = \sum_{t=1}^N \mathcal{N}^t,$$

$$\text{where, } \mathcal{N}^t = \begin{cases} \frac{1}{N}, & \text{if } \arg \max(a_h^t) \in k \\ 0, & \text{otherwise.} \end{cases}$$

Here, N is needle length, a_h^t is the attention assigned to each input token by head h at decoding timestep t , and k is the output token that is retrieved from the needle sentence.

Using retrieval scores associated with attention head we can quantify which heads illustrate copy-paste behavior for different languages.

3.2 Multilingual Experimental Setting

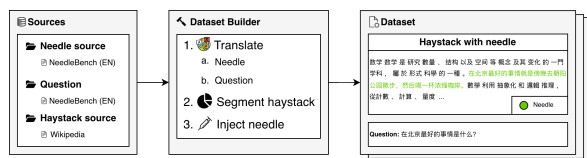


Figure 1: Multilingual NIAH dataset construction pipeline. We sample needles, questions and haystacks, translate the needle and question into the target language, segment the haystack, and inject the translated needle to produce the final evaluation instances.

Multilingual NIAH data construction We chose to study English (en), German (de), Chinese (zh),

and Swahili (sw). These languages have been selected to cover a wide spectrum of known language families (Lauscher et al., 2020; Philippy et al., 2023) and pretraining resources. English, and German belong to the Germanic branch of the Indo-European language family and share linguistic similarities, while Chinese, part of the Sino-Tibetan family, presents a distinct linguistic profile (De Gregorio et al., 2024). In contrast to above languages, Swahili represents a low-resource language to discover differences in model behaviour due to scarcity of pretraining tokens. We also report the effect of including more number of diverse and low-resource languages i.e. Arabic (ar), Dutch (nl), and Hindi (hi) to the retrieval head language overlap analysis in Appendix A.3.

As seen in Figure 1, to extend the NIAH task for different languages, we first create a haystack using Wikipedia dumps of each language.¹ To create the needle, we use Google Translate² to translate the same needles as used by Wu et al. (2024) to ensure consistency of evaluation. 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 20 context-length settings and 10 needle-depth percentages, and report each head’s mean score averaged over all configurations.

Models We evaluate on three open-source model families: Qwen2.5 7B Instruct (Qwen et al., 2025), Phi3.5 3B MiniInstruct (Abdin et al., 2024), and LLaMa-3.1 8B Instruct (Grattafiori et al., 2024).³ We choose these models to span widely-used open-source multilingual LLMs with diverse architectures and training recipes.

3.3 Analyzing distribution of retrieval heads across different languages

We find that retrieval heads also exhibit a mixture of language-shared and language-exclusive heads, consistent with patterns discovered in recent multilingual studies (Liu et al., 2025; Chen et al., 2025). However, from Figure 2 we report a considerable variance of retrieval head intersection across four languages - English, German, Chinese,

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

²<https://translate.google.com/>

³Given all are Instruct models, we omit the use of words ‘Instruct’ & ‘MiniInstruct’ for better readability.

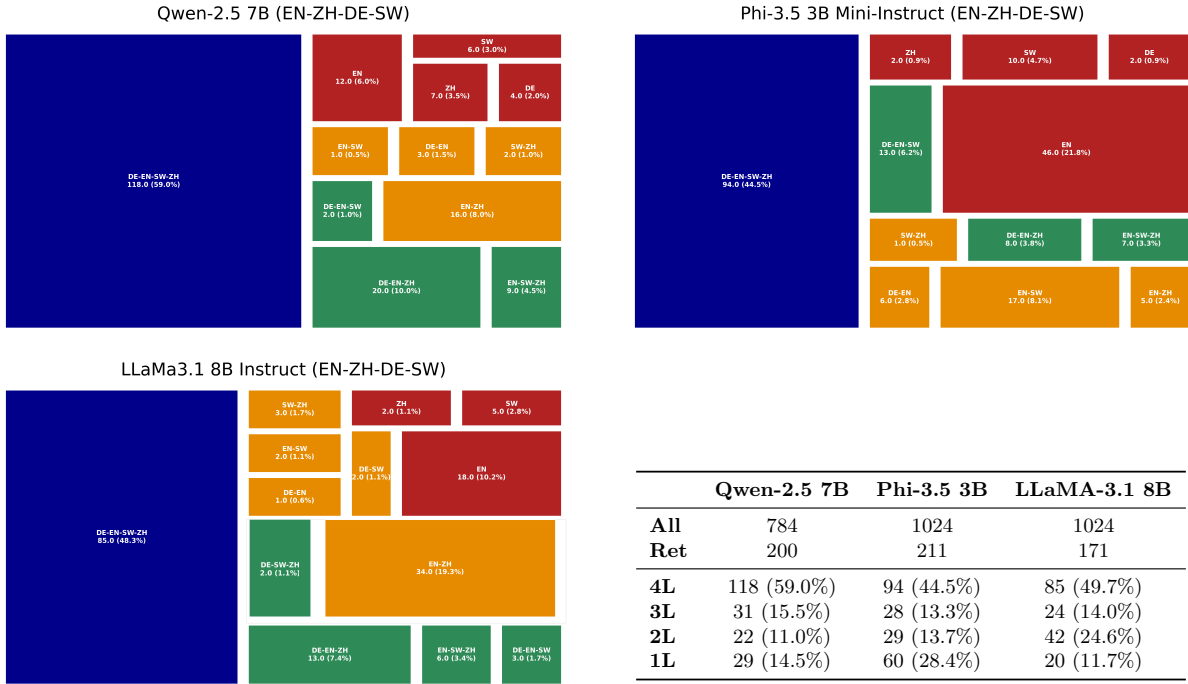


Figure 2: Language-specific (red) and shared (orange, green, blue) retrieval heads in three multilingual LMs for four languages - English (en), Chinese (zh), German (de), Swahili (sw). The size of each block is proportional to the number of retrieval heads that are language-exclusive (red), bilingual (orange), trilingual (green), or language-shared (blue). Qwen-2.5 7B exhibits strong multilingual head overlap (59% 4L), LLaMa-3.1 8B has high bilingual (mainly en-zh) head overlap (24.6% 2L), and Phi-3.5 3B shows a measurably higher language-specific (mainly en-specific) behavior (28.4% 1L) compared to the other two model families.

and Swahili for different model families. For instance, 118 of 200 (59.0%) heads in Qwen2.5-7B are shared among all four languages while 94 of 211 heads (44.5%) in Phi3.5-3B are shared. Moreover, the proportion of bilingual en-zh heads in LLaMa-3.1 8B accounts for 19% of all retrieval heads, more than twice the most abundant bilingual pairs in Qwen-2.5 7B (8.0% en-zh) and Phi-3.5 3B (8.1% en-sw). Finally, it is observed that Phi3.5-3B exhibits a disproportionately high number of en-specific retrieval heads (21.8%), while other models like Qwen-2.5 7B and LLaMa-3.1 8B have lower counts for en-specific heads (6.0% and 10.2% respectively). We attribute this English-inclined behaviour 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 varying architectures (refer Appendix A.1 and A.2 respectively) suggesting an invariance to these factors.

4 Retrieval-Transition heads

While retrieval heads facilitate reasoning by copying information from the previous context (Wu

et al., 2024), this mechanism alone is insufficient for LLMs trained on multiple languages. Because MLMs operate within a language-agnostic concept space (Wendler et al., 2024b; Wang et al., 2025; Chen et al., 2025), they must do more than just copy; they must translate latent, abstract thoughts into specific target-language tokens. Modifying experimental setup from standard retrieval head detection, we identify distinct heads, which we name as *Retrieval-Transition heads (RTH)* that perform this mapping behavior. Consequently, we show that masking these heads causes a more severe drop in CoT performance than standard retrieval heads (RT), as the model is impaired from its ability to verbalize latent thoughts.

4.1 Experimental Setup

Cross-Lingual NIAH Task: We use English as a pivot language to approximate the language-agnostic concept space as the latent token vocabulary is not directly accessible. However, we demonstrate that this choice is functionally arbitrary and depends solely on which language most closely aligns with the model’s internal representations (c.f. Figure 4 in Wang et al. (2025)). In models with

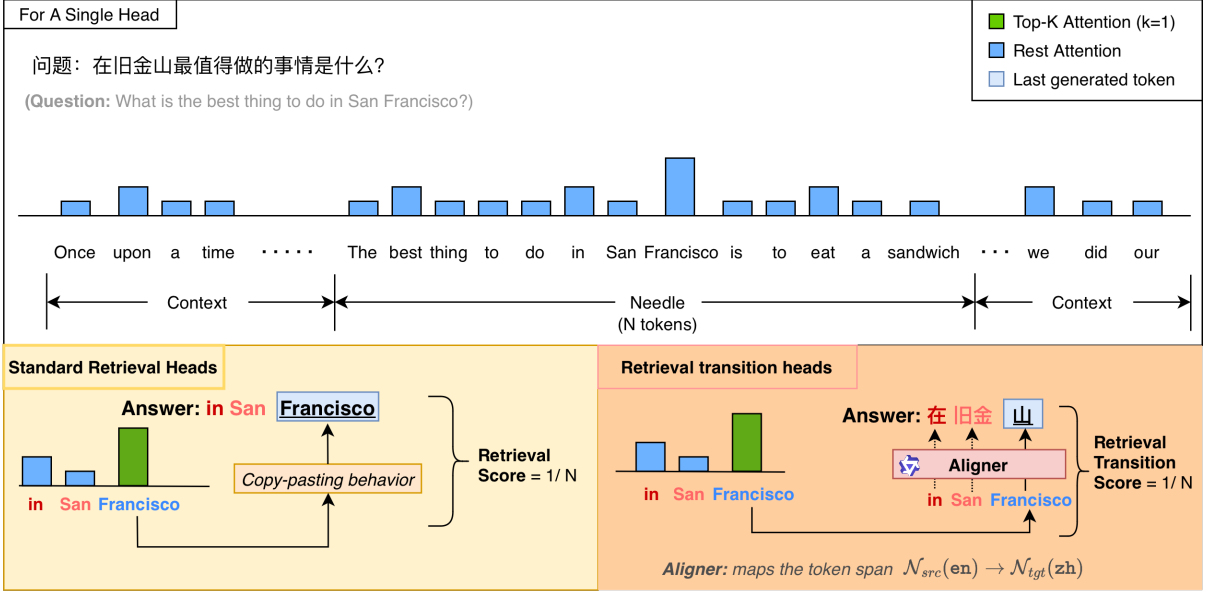


Figure 3: Illustration of the retrieval-transition score (RTS). Needle tokens in the source language are first aligned to their target-language equivalents using a LM. The degree of overlap between the model’s generated tokens and these aligned target needles quantifies each attention head’s contribution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

more balanced multilingual training (such as the Qwen-2.5 family) we utilize Chinese as the proxy for the concept space to derive retrieval-transition heads for English. Our experiments show that masking the identified heads leads to a significant performance collapse even in English-language CoT tasks, confirming that these heads are essential for the transition from the latent space to any specific target language. As illustrated in Figure 3, we introduce a modified variant of the NIAH task, where 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 given target language \mathcal{N}_{tgt} .

Retrieval transition score: To identify attention heads that facilitate retrieval across linguistic boundaries, we extend the standard retrieval score (RS) formulation to account for cross-token correspondence. 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-2507⁴ to perform this alignment, ensuring that output tokens are correctly mapped to their source-language counterparts. We also evaluate an alternative LLM-based aligner and find that the identified retrieval-transition heads

⁴<https://huggingface.co/Qwen/Qwen3-30B-A3B-Thinking-2507>

are largely stable to the choice of aligner (Appendix C.2). The Retrieval Transition Score for a given attention head h (RTS_h) is defined as:

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

$$\text{where } \mathcal{A}^t = \begin{cases} \frac{1}{N}, & \text{if } \operatorname{argmax}(a_h^t) \in f^{-1}(\text{token}_t) \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

Here, $f : \mathcal{N}_{src} \rightarrow \mathcal{N}_{tgt}$ is the alignment function mapping source-language needle tokens to their target-language equivalents, a_h^t denotes the attention distribution of head h at decoding timestep t , and $f^{-1}(\text{token}_t)$ represents the set of source tokens corresponding to the current target token being decoded.

Models and Languages evaluated: Same as Retrieval Head analysis (c.f. Section 3.2) we perform our experiments on Qwen2.5 7B Instruct (Qwen et al., 2025), Phi3.5 3B MiniInstruct (Abdin et al., 2024), and LLaMa-3.1 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.

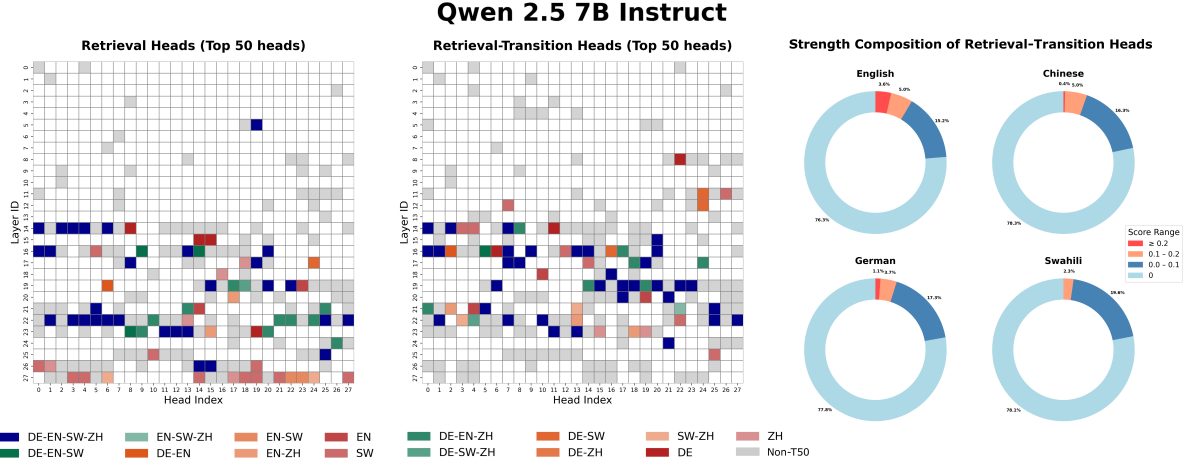


Figure 4: Layer-head distributions of Retrieval Heads (RH) and Retrieval-Transition Heads (RTH) across languages for Qwen-2.5 7B Instruct. Colored cells are part of Top-50 most prominent heads ranked by score. Among these colored heads, we find language-specific RH are dominant in the final layers while 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

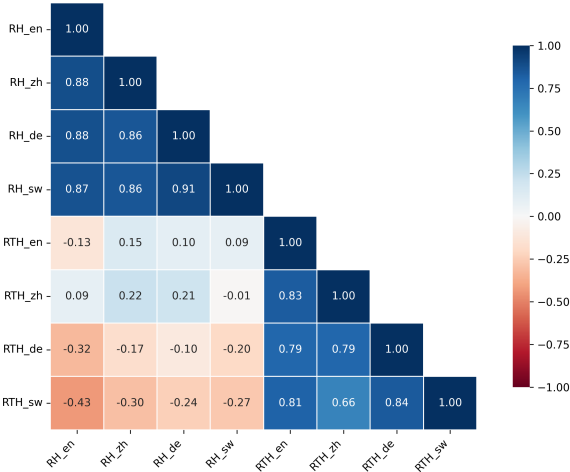


Figure 5: Spearman correlation score between top-50 most influential $RH(\ell)$ and $RTH(\ell)$ in Qwen-2.5 7B. $RTH(\ell)$ heads share little correlation with $RH(\ell)$ but high correlations among other $RTH(\ell_0)$ and vice-versa

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

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

5 Evaluation of Retrieval-Transition Heads

5.1 The distribution of retrieval-transition heads

In Figure 4 (left), we report the location of retrieval-transition heads in comparison to the location of retrieval heads. We observe that RT and RTH exhibit considerable degree of overlap. However, as seen in Table 1, the intersection between the two head groups is notably 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 as the search space expands to 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 Qwen 2.5 7B Instruct for RH and RTH across four languages: English (en), Chinese (zh), German (de), and Swahili (sw). As illustrated in the global correlation matrix (Figure 5), consistent clusters emerges within RH heads, and within RTH heads where coefficients remain high, typically ranging from 0.66 to 0.91. However, when comparing RTH with RH, even for the same language, we observe little to negative correlation. Moreover, most prominent RTHs appear predominantly in the middle layers (14-23 in Figure 4) while language-specific RHs appear mainly in the final layers (18 of top-50 heads in last 3 layers) in Qwen 2.5 7B Instruct. These results suggest that while the model has a unified way to retrieve data

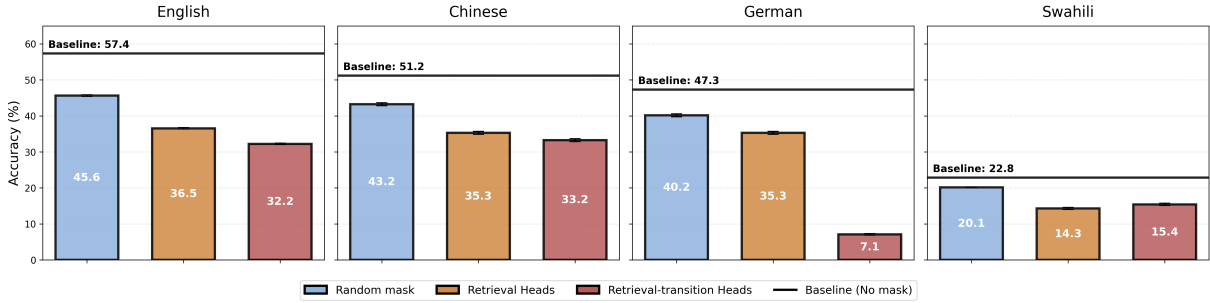


Figure 6: Impact of attention head masking (top- k where $k = 25$) on MMLU-ProX accuracy across languages for Qwen-2.5 7B Instruct. While removing standard retrieval heads (orange) leads to a performance decline, masking the retrieval-transition heads (RTH) (red) results in even more severe collapse in reasoning accuracy.

within a language, the transition from a latent concept space to a specific target language relies on a separate, specialized set of internal circuits.

5.2 The role of retrieval-transition heads in MLM performance

To probe the causal role of retrieval-transition heads, we mask both retrieval and retrieval-transition heads and measure the resulting change in MMLU-ProX accuracy (Xuan et al., 2025). MMLU-ProX is a multilingual, reasoning-centric multiple-choice QA benchmark that requires complex, multi-step reasoning and covers 29 languages. We evaluate each language under a monolingual protocol, where the question and answer options are presented exclusively in that language.

For each language (L), we evaluate three masking strategies: (i) masking the top- 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). We describe our results on Qwen-2.5 7B Instruct and additionally provide results by varying hyperparameter k in Appendix C.3. For $k > 25$, Figure A8 shows that retrieval failures—missing relevant information from the question—drive most of the degradation by disrupting the model’s chain-of-thought, which largely overwhelms the effect of weakening the transition from language-agnostic representations to language-specific decoding.

Figure 6 reports the impact of masking different head sets in the Qwen-2.5 7B Instruct model. Across languages, masking the top-25 retrieval heads causes a consistent and substantial reduction in MMLU-ProX accuracy, suggesting that these heads play a causal role in the model’s chain-of-thought (CoT) performance. In particular, accuracy

drops from 57.4 to (36.55 ± 0.43) in English, 51.2 to (35.28 ± 0.43) in Chinese, from 47.3 to (35.28 ± 0.43) in German, and from 22.84 to (14.28 ± 0.33) in Swahili. This aligns with prior findings that retrieval-related attention heads are important for reasoning-intensive tasks (Wu et al., 2024).

However, masking retrieval-transition heads leads to the largest degradation in all languages except Swahili: English drops from 57.4 to (32.24 ± 0.43) , Chinese drops from 51.2 to (33.25 ± 0.41) , German collapses from 47.3 to (7.07 ± 0.23) , and Swahili drops from 22.84 to (15.41 ± 0.33) . Overall, these results indicate that retrieval-transition heads are essential for multilingual CoT reasoning in MMLU-ProX, as masking them leads to poor performance across all languages, including English.

5.3 Qualitative Analysis

We perform a qualitative error analysis of failure cases induced by their removal. We identify two primary failure modes: *loss of coherence* and *worsened retrieval*. In the first mode, RTH masking impacts generation ability. The model often falls into nonsensical cycles and repetitive loops. In the other mode, the model makes meaningful sentences but often fails to retrieve crucial information from its context to reach to the correct answer. The information is present in the given context or its own CoT chain which the model fails to retrieve. To illustrate these effects, we examine two specific instances where the model succeeds under **No-Mask** and **RH-Mask** conditions but fails under the **RTH-Mask** condition.

Failure Mode 1: Lack of Coherence. Figure 7 shows the output comparison of Qwen-2.5 7B for a German MMLU-ProX Math problem. The

Effect of Retrieval-Transition Heads on language coherence	
<p>Question 1 : Antwort: Richtig oder falsch? Aussage 1: Jede Menge von zwei Vektoren in \mathbb{R}^2 ist linear unabhängig. Aussage 2: Wenn $V = \text{span}(v_1, \dots, v_k)$ und $\{v_1, \dots, v_k\}$ linear unabhängig sind, dann ist $\dim(V) = k$. <i>(Translated) Question 1</i> : Answer true or false. Statement 1: Any set of two vectors in \mathbb{R}^2 is linearly independent. Statement 2: If $V = \text{span}(v_1, \dots, v_k)$ and $\{v_1, \dots, v_k\}$ are linearly independent, then $\dim(V) = k$.</p>	
<p>(No attention heads masked) Model Output : Diese Aussage 1 ist falsch. Wenn man mehr als zwei Vektoren in \mathbb{R}^2 nimmt, sind sie notwendigerweise linear abhängig. Diese Aussage 2 ist wahr. Wenn ein Vektorraum V durch eine Basis $\{v_1, \dots, v_k\}$... Antwort : Falsch, Wahr</p> <p><i>(Translated)</i> This statement 1 is false. If there are more than two vectors in \mathbb{R}^2, they are necessarily linearly dependent. This statement 2 is true. If a vector space V is defined by a basis $\{v_1, \dots, v_k\}$,... Answer : False, True</p>	<p>(Top 25 Retrieval-Transition Heads masked) Model Output : Zuerst betrachten wir Aussage 1 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 ...</p> <p><i>(Translated)</i> 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 ...</p>

Figure 7: 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 to retrieve prior context
<p>Question 2 : Löse die Gleichung $14 = w + 23$. Optionen: A] 9 B] -9 C] -11 D] -14</p> <p><i>(Translated) Question 2</i>: Solve equation $14 = w + 23$ Options: A] 9 B] -9 C] -11 D] -14</p>
<p>(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</p> <p><i>(Translated)</i> 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:</p>

Figure 8: 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.

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.

Failure Mode 2: Worsened Retrieval 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 8 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.

6 Conclusion and Future Work

We have identified Retrieval-Transition Heads as a distinct and essential circuit for multilingual intelligence. Our causal analysis via the Retrieval-Transition Score demonstrates that these heads are not merely participants in retrieval but are primary drivers of linguistic coherence and grounded reasoning. Future work could leverage these findings to develop language-aware KV cache compression strategies. While existing methods prioritize standard retrieval heads to maintain factuality, our results suggest that a "Transition-Aware" pruning strategy may be necessary to prevent the catastrophic loss of coherence observed in multilingual settings.

510 Limitations

511 In spite of strong causal results on downstream
512 tasks, our methodology faces two primary con-
513 straints. First, the Retrieval Transition Score (RTS)
514 approximates the language-agnostic concept space
515 using English as a proxy; while effective due to
516 inherent model biases, this results in RTHs that are
517 less sharply defined than standard retrieval heads.
518 Second, as source languages grow typologically
519 distant from the model’s latent pivot, the head dis-
520 tribution approaches uniformity, reducing the dis-
521 criminative signal of the RTH circuit. Future re-
522 search should transcend this reliance on pivot lan-
523 guages by developing systems that learn to mimic
524 true latent tokens. By generating a model-specific
525 latent vocabulary, researchers can move toward
526 identifying transition mechanisms that reflect the
527 model’s internal conceptual structures rather than
528 cross-lingual projections.

529 References

530 Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed
531 Awadallah, Ammar Ahmad Awan, Nguyen Bach,
532 Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat
533 Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck,
534 Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav
535 Chaudhary, Dong Chen, Dongdong Chen, and 110
536 others. 2024. [Phi-3 technical report: A highly capa-
537 ble language model locally on your phone](#). *Preprint*,
538 arXiv:2404.14219.

539 Anonymous. 2025. [Translation heads: Unveiling at-
540 tention’s role in LLM multilingual translation](#). In
541 *Submitted to The Fourteenth International Confer-
542 ence on Learning Representations*. Under review.

543 Yuxin Chen, Yiran Zhao, Yang Zhang, An Zhang,
544 Kenji Kawaguchi, Shafiq Joty, Junnan Li, Tat-Seng
545 Chua, Michael Qizhe Shieh, and Wenxuan Zhang.
546 2025. The emergence of abstract thought in large lan-
547 guage models beyond any language. *arXiv preprint*
548 [arXiv:2506.09890](#).

549 Juan De Gregorio, Raúl Toral, and David Sánchez. 2024.
550 [Exploring language relations through syntactic dis-
551 tances and geographic proximity](#). *EPJ Data Science*,
552 13(1).

553 DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan
554 Wang, Bochao Wu, Chengda Lu, Chenggang
555 Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan,
556 Damai Dai, Daya Guo, Dejian Yang, Deli Chen,
557 Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai,
558 and 181 others. 2025. [Deepseek-v3 technical report](#).
559 *Preprint*, arXiv:2412.19437.

560 Javier Ferrando, Oscar Obeso, Senthoooran Rajamanoha-
561 ran, and Neel Nanda. 2024. Do i know this entity?

knowledge awareness and hallucinations in language
562 models. *arXiv preprint arXiv:2411.14257*. 563

564 Yu Fu, Zefan Cai, Abedelkadir Asi, Wayne Xiong, Yue
565 Dong, and Wen Xiao. 2025. [Not all heads matter:
566 A head-level KV cache compression method with
567 integrated retrieval and reasoning](#). In *The Thirteenth
568 International Conference on Learning Representa-
569 tions*.

570 Leo Gao, Jonathan Tow, Baber Abbasi, Stella Bider-
571 man, Sid Black, Anthony DiPofi, Charles Foster,
572 Laurence Golding, Jeffrey Hsu, Alain Le Noac’h,
573 Haonan Li, Kyle McDonell, Niklas Muennighoff,
574 Chris Ociepa, Jason Phang, Laria Reynolds, Hailey
575 Schoelkopf, Aviya Skowron, Lintang Sutawika, and
576 5 others. 2023. [A framework for few-shot language
577 model evaluation](#).

578 Rhys Gould, Euan Ong, George Ogden, and Arthur
579 Conmy. 2024. [Successor heads: Recurring, inter-
580 pretable attention heads in the wild](#). In *The Twelfth
581 International Conference on Learning Representa-
582 tions*.

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

588 Lei Huang, Xiaocheng Feng, Weitao Ma, Yuchun Fan,
589 Xiachong Feng, Yangfan Ye, Weihong Zhong, Yuxuan
590 Gu, Baoxin Wang, Dayong Wu, and 1 others.
591 2025. Improving contextual faithfulness of large lan-
592 guage models via retrieval heads-induced optimiza-
593 tion. *arXiv preprint arXiv:2501.13573*.

594 Greg Kamradt. 2023. [Needle in a haystack - pressure
595 testing llms](#).

596 Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and
597 Goran Glavaš. 2020. [From zero to hero: On the
598 limitations of zero-shot language transfer with mul-
599 tilingual Transformers](#). In *Proceedings of the 2020
600 Conference on Empirical Methods in Natural Lan-
601 guage Processing (EMNLP)*, pages 4483–4499, On-
602 line. Association for Computational Linguistics.

603 Xin Liu, Qiyang Song, Qihang Zhou, Haichao Du,
604 Shaowen Xu, Wenbo Jiang, Weijuan Zhang, and
605 Xiaoqi Jia. 2025. Focusing on language: Revealing
606 and exploiting language attention heads in mul-
607 tilingual large language models. *arXiv preprint*
608 [arXiv:2511.07498](#).

609 Callum Stuart McDougall, Arthur Conmy, Cody Rush-
610 ing, Thomas McGrath, and Neel Nanda. 2024. [Copy
611 suppression: Comprehensively understanding an at-
612 tention head](#).

613 Catherine Olsson, Nelson Elhage, Neel Nanda,
614 Nicholas Joseph, Nova DasSarma, Tom Henighan,
615 Ben Mann, Amanda Askell, Yuntao Bai, Anna
616 Chen, Tom Conerly, Dawn Drain, Deep Gan-
617 guli, Zac Hatfield-Dodds, Danny Hernandez, Scott

618	Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context learning and induction heads. <i>Transformer Circuits Thread</i> . https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html .	675
619		676
620		677
621		678
622		679
623	Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Identifying the correlation between language distance and cross-lingual transfer in a multilingual representation space. In <i>Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP</i> , page 22–29. Association for Computational Linguistics.	680
624		681
625		682
626		683
627		684
628		685
629		686
630	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, Jiayi Yang, Jingren Zhou, and 25 others. 2025. <i>Qwen2.5 technical report</i> . Preprint, arXiv:2412.15115.	687
631		688
632		
633		
634		
635		
636		
637	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. <i>arXiv preprint arXiv:2504.04264</i> .	
638		
639		
640		
641		
642	Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024a. Do llamas work in English? on the latent language of multilingual transformers . In <i>Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.	
643		
644		
645		
646		
647		
648		
649	Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024b. Do llamas work in english? on the latent language of multilingual transformers. In <i>Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 15366–15394.	
650		
651		
652		
653		
654		
655	Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2024. Retrieval head mechanistically explains long-context factuality. <i>arXiv preprint arXiv:2404.15574</i> .	
656		
657		
658		
659	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 <i>Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing</i> , Suzhou, China. Association for Computational Linguistics.	
660		
661		
662		
663		
664		
665		
666		
667		
668		
669	Ruochen Zhang, Qinan Yu, Matianyu Zang, Carsten Eickhoff, and Ellie Pavlick. 2025. The same but different: Structural similarities and differences in multilingual language modeling . In <i>The Thirteenth International Conference on Learning Representations</i> .	
670		
671		
672		
673		
674		
	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. <i>arXiv preprint arXiv:2503.00865</i> .	
	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. <i>arXiv preprint arXiv:2409.03752</i> .	
	Tianyi Zhou, Deqing Fu, Vatsal Sharan, and Robin Jia. 2024. Pre-trained large language models use fourier features to compute addition. <i>Advances in Neural Information Processing Systems</i> , 37:25120–25151.	

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

689
690

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

691

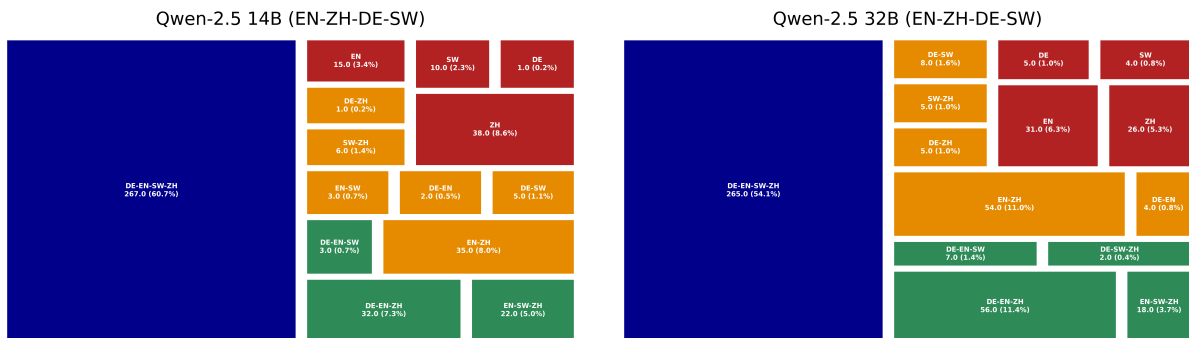


Figure A1: To assess the effect of model scaling, we evaluate Qwen2.5-14B Instruct and Qwen2.5-32B Instruct. Despite the larger parameter count, the proportions of shared and language-exclusive attention heads remain largely consistent.

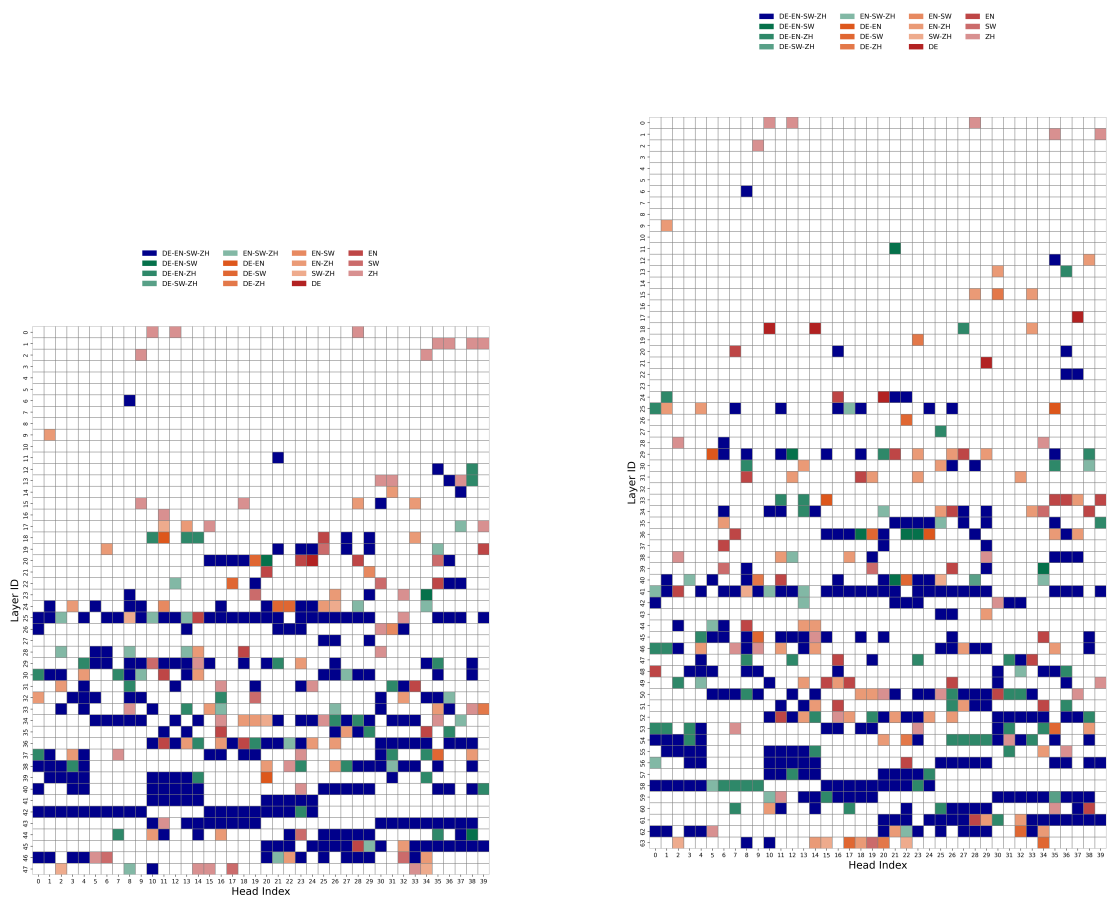


Figure A2: To assess how scaling affects the placement of shared and language-exclusive attention heads, we evaluate Qwen2.5-14B Instruct and Qwen2.5-32B Instruct. The spatial distribution of these head types remains largely consistent across scales.

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

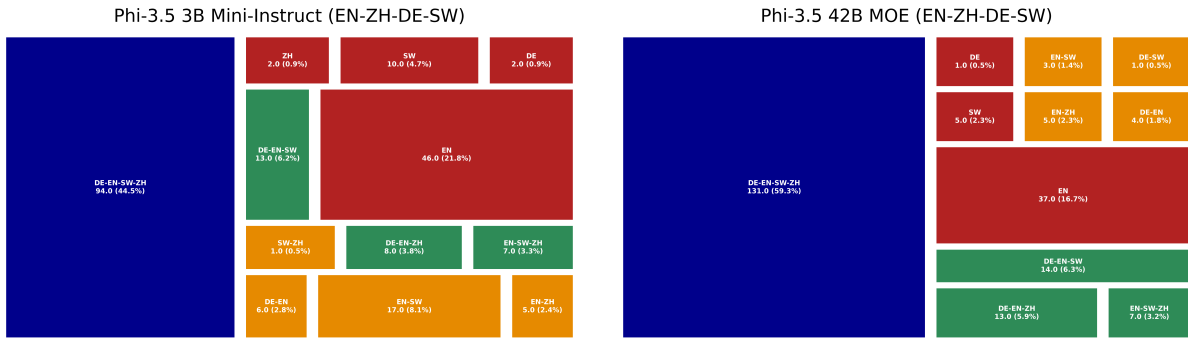


Figure A3: To compare architectures, we study Phi-3.5 3B Mini-Instruct and Phi-3.5 42B MoE (Mixture-of-Experts). For Phi-3.5 42B MoE architecture, we observe that the proportions of shared attention heads are higher compared to language-exclusive attention heads.

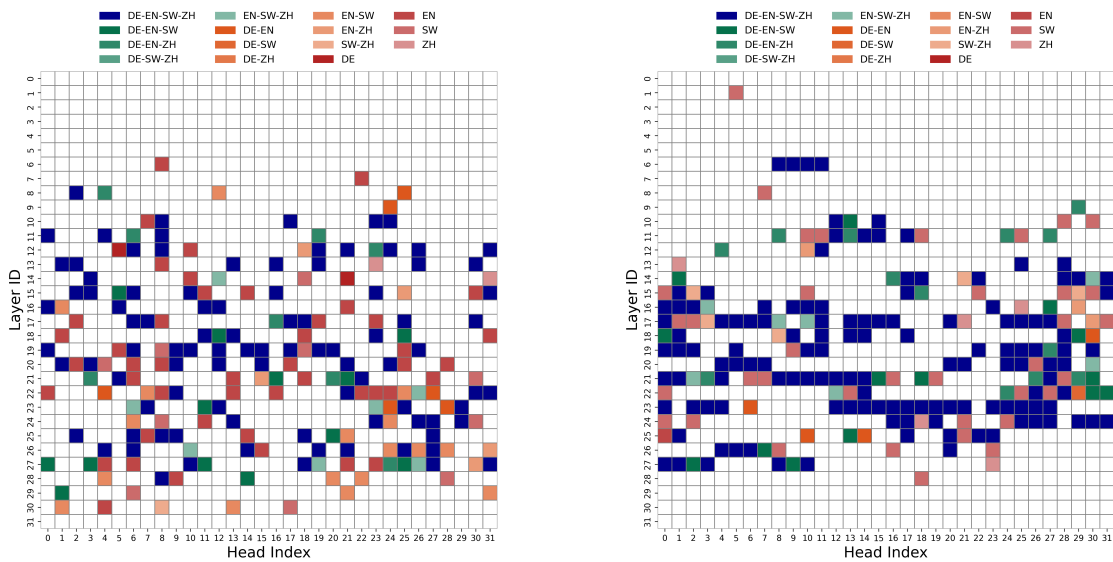


Figure A4: To assess how scaling affects the placement of shared and language-exclusive attention heads, Phi-3.5 3B Mini-Instruct and Phi-3.5 42B MoE (Mixture-of-Experts). The spatial distribution of these head types remains largely consistent across scales.

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

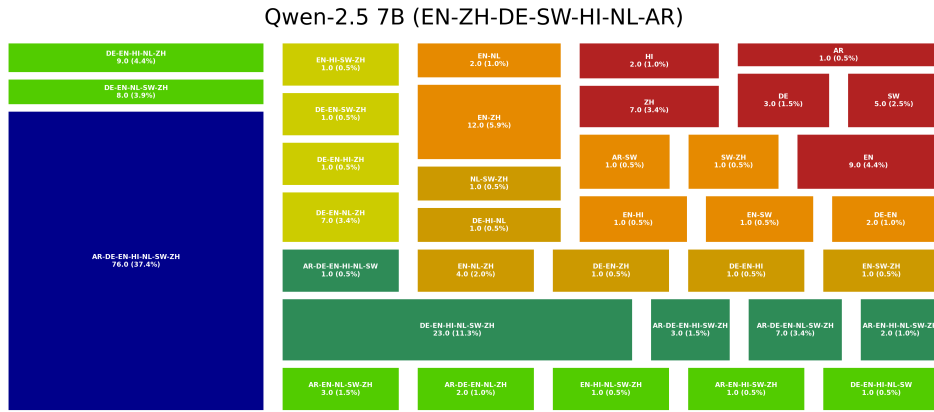


Figure A5: To study the effect of adding more low-resource languages, we evaluate Qwen2.5-7B Instruct on English, Chinese, German, Swahili, Hindi, Dutch, and Arabic. As the language set expands, the fraction of shared attention heads decreases, while language-exclusive heads become more prevalent.

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 Qwen-2.5 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	Qwen-2.5 7B	LLaMA-3.1 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

B.1 Correlating Data Composition with Retrieval Head Intersection

Our analysis of retrieval head distribution across English, German, Chinese, and Swahili reveals a strong correlation between the volume of multilingual training tokens and the degree of cross-lingual head intersection.

Broad Generality in Qwen-2.5 7B. As observed in our results, Qwen-2.5 7B exhibits the highest degree of intersection, with 59.0% (118/200) of its retrieval heads shared among all four analyzed languages. This architectural generality likely stems from its massive 18 trillion token corpus, which is the largest among the models studied. The inclusion of a significant, high-density non-English component (primarily Chinese at estimated 50%) during pre-training forces the model to develop shared internal representations for retrieval that transcend specific linguistic syntax, leading to a high proportion of "universal" retrieval heads.

Language-Specific Specialization in Phi-3.5 3B. In contrast, Phi-3.5 3B exhibits the lowest 4-language intersection (44.5%) and a disproportionately high count of English-specific heads (21.8%). This aligns with the "Phi-philosophy" of high-quality synthetic data. Since the majority of "textbook-quality" reasoning data used to train the Phi-3.5 series is generated in English, the model appears to allocate a significant portion of its retrieval capacity to English-specific patterns. Interestingly, the model still

716 maintains a notable en-sw bilingual pair (8.1%), likely reflecting the specific multilingual "boost"
717 introduced in the 3.5 version to improve its performance in lower-resource languages.

718 These results suggest that while total token count (scale) promotes language-agnostic retrieval mecha-
719 nisms (as seen in Qwen), the specific ratio of synthetic vs. natural data and the post-training alignment
720 strategy (as seen in Phi) significantly influences whether a model develops universal, or language-specific
721 retrieval pathways.

C Additional experiments and analysis on Retrieval Transition Heads

722

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

723

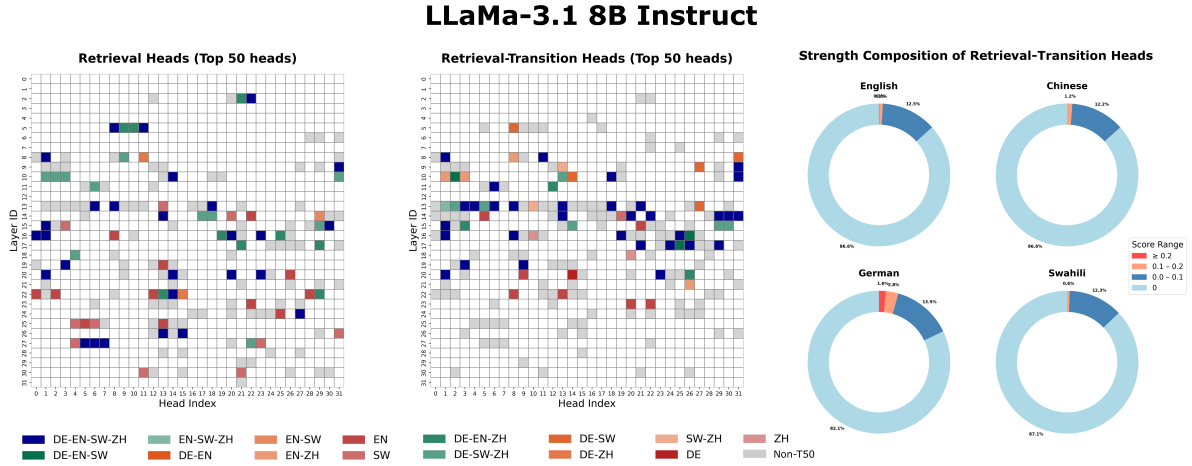


Figure A6: Layer-head distributions of Retrieval Heads (RH) and Retrieval-Transition Heads (RTH) across languages for LLaMa-3.1 8B Instruct. Colored cells are part of Top-50 most prominent heads ranked by score. Among these colored heads, we find language-specific RH are dominant in the final layers 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

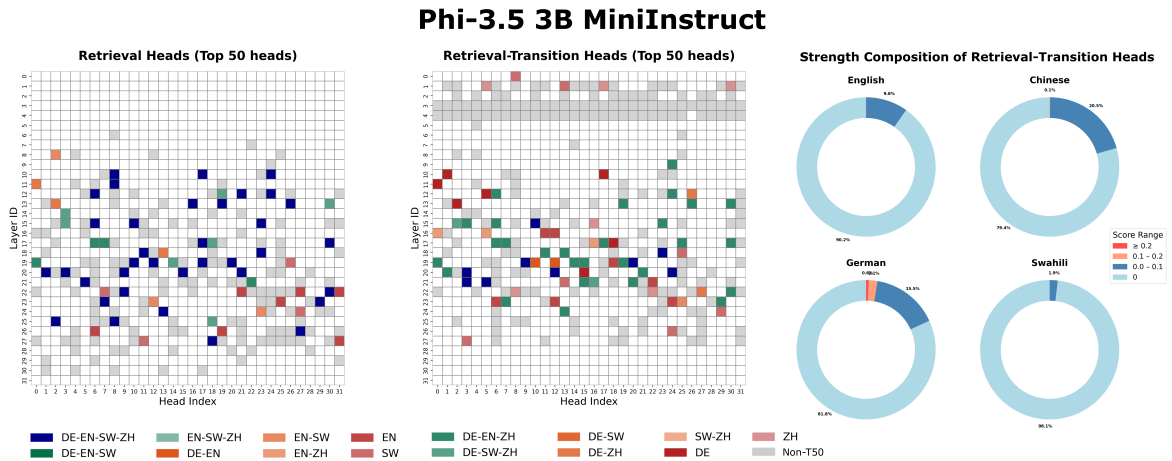


Figure A7: Layer-head distributions of Retrieval Heads (RH) and Retrieval-Transition Heads (RTH) across languages for Phi-3.5 3B MiniInstruct. Colored cells are part of Top-50 most prominent heads ranked by score. Among these colored heads, we find language-specific RH are dominant in the final layers 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 Analyzing variability introduced by LLM-based aligner

724

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.

C.3 Additional masking experiments on Qwen 2.5 7B Instruct: effect of parameter k

Impact of top 50 Head Masking on MMLU-ProX Accuracy (Qwen-2.5 7B Instruct)

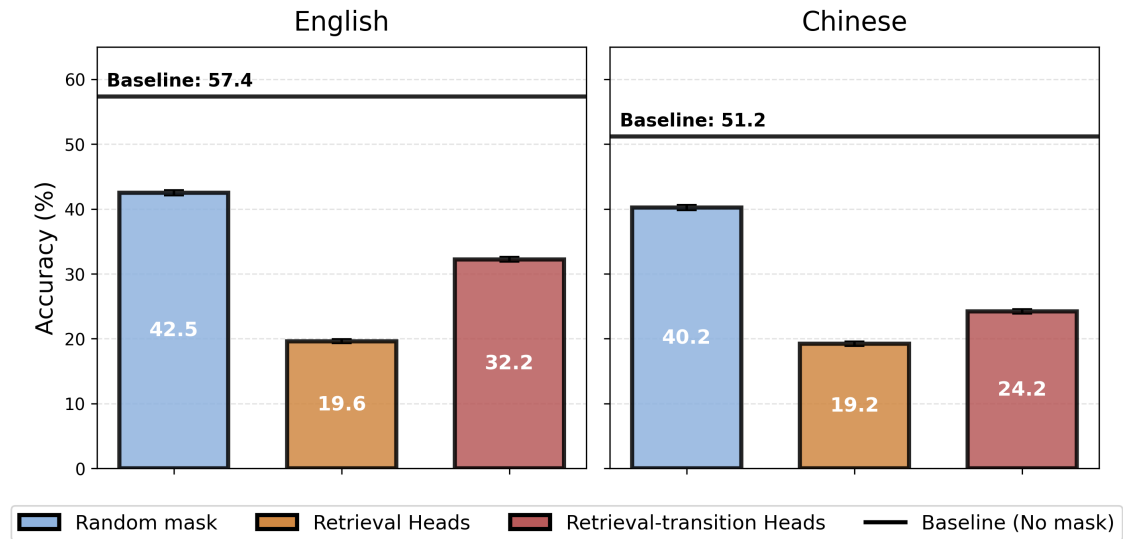


Figure A8: Effect of masking the top- k attention heads ($k = 50$) on MMLU-ProX accuracy for Qwen 2.5 7B Instruct. In contrast to the $k = 25$ setting, masking $k = 50$ retrieval heads (orange) causes a larger accuracy drop than masking $k = 50$ retrieval-transition heads (RTH; red). At this masking level, failures in retrieving relevant parts from the question and the resulting degradation of the model generated chain-of-thought dominate, thereby obscuring the comparatively smaller effect of impaired transition from a language-agnostic representation to language-specific decoding.