Do retrieval heads speak the same language?

NEW YORK UNIVERSITY

Vishvesh Trivedi, Shaswat Patel, Yue Han

NYU Courant Institute of Mathematical Sciences

Background & Motivation

Demystifying Large Language Models (LLMs) is a crucial task to better scale LLM during inference by optimizing pruning and KV caching strategies. Recent work shows that specific attention heads — termed retrieval heads — are crucial for retrieving relevant long-context information. Pruning these heads impairs model performance, especially in Chain-of-Thought (CoT) tasks.

In this study, we extend retrieval head analysis to multilingual settings. We systematically investigate:

1. Whether retrieval heads are common across languages or language-specific?
2. How does translation impact head activations?
3. What is the downstream effect of masking language-associated retrieval heads?

Our findings highlight important multilingual dynamics crucial for efficient LLM deployment.

Our contributions

1. Not all retrieval heads are common across languages, with nearly **30-40% being language-specific.**
2. The strength of retrieval heads is strongly correlated with their language-agnostic behavior with **strongest retrieval heads common across all three languages** and vice-versa.
3. Masking language agnostic heads have significant impact on model performance.

Method

We build on the methodology introduced by Wu et. al[2], with the corresponding algorithm outlined in Figure 2. To adapt the Needle-In-A-Haystack (NIAH) task to a multilingual setting, we synthetically generated needles and haystacks in Chinese and German. The full pipeline for this extension is shown in Figure 1.

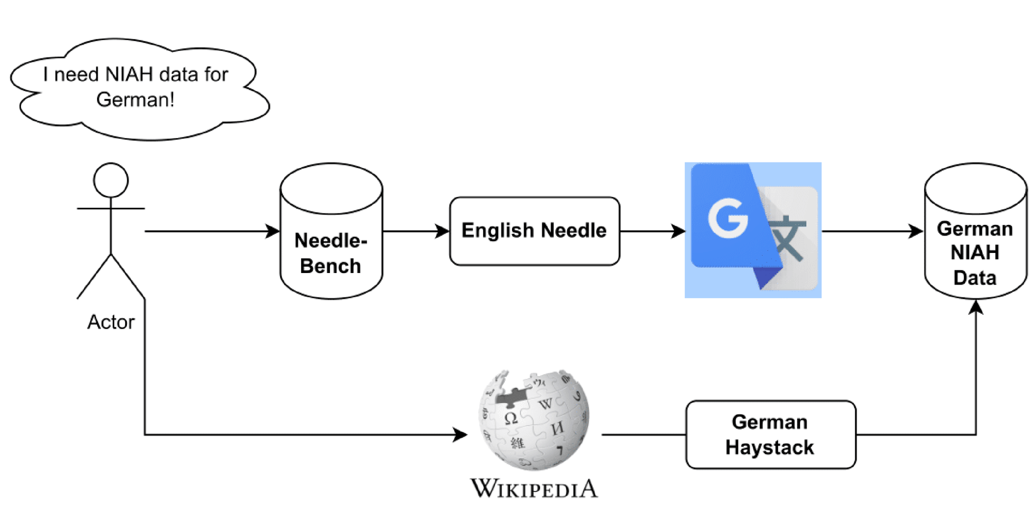

Figure 1. Pipeline to extend Needle-In-A-Haystack task to multiple languages

Algorithm 1 Decoding Procedure with Retrieval Score Calculation

Require: Model output with past key values *q_outputs*, input token *inp*, decoding length *decode_len*, optional block list *block_list*

Ensure: Decoded output tokens and retrieval scores

```
Initialize output ← [ ]
Initialize retrieval_score ← 3D list of size (layer_num × head_num) filled with (0)
for step_i ← 0 to decode_len − 1 do
    Reshape inp to shape (1, 1)
    outputs ← MODEL(inp, output_attentions=True)
    inp ← arg max(outputs)
    decoded_token ← ConvertIdsToTokens(inp)
    Append inp to output
    RetrievalCalculate(attentions, inp, decoded_token)
end for
return (output, retrieval_score)
```

Algorithm 2 Retrieval Score Calculation

Require: Attention matrix *attention_matrix*, retrieval score table *retrieval_score*, input token *inp*, decoded token *step_token*, top-k value *topk*

Ensure: Updated retrieval scores

```
for each layer index layer from 0 to layer_num −1 do
    for each head index head from 0 to head_num −1 do
        indices ← TopK(attention_matrix[layer][head])
        for each i in indices do
            if needle_start ≤ i < needle_end then
                Increment retrieval_score[layer][head] by r_score
                break
            end if
        end for
    end for
end for
```

$$r\_score = \frac{|g_h \cap k|}{|k|}$$

Figure 2. Algorithm to calculate retrieval scores

Related work

1. **Retrieval heads:** Attention heads responsible for retrieving tokens from in-context text.[1]
2. **Copy Suppression heads:** Attention heads that prevent models from naively copying tokens.[1]
3. **Successor heads:** Attention heads responsible for incrementation of tokens in naturally ordered sequence.[1]

Analyzing type of retrieval heads across different languages

We extend Wu et. al.[2]'s retrieval head analysis to the multilingual setting.

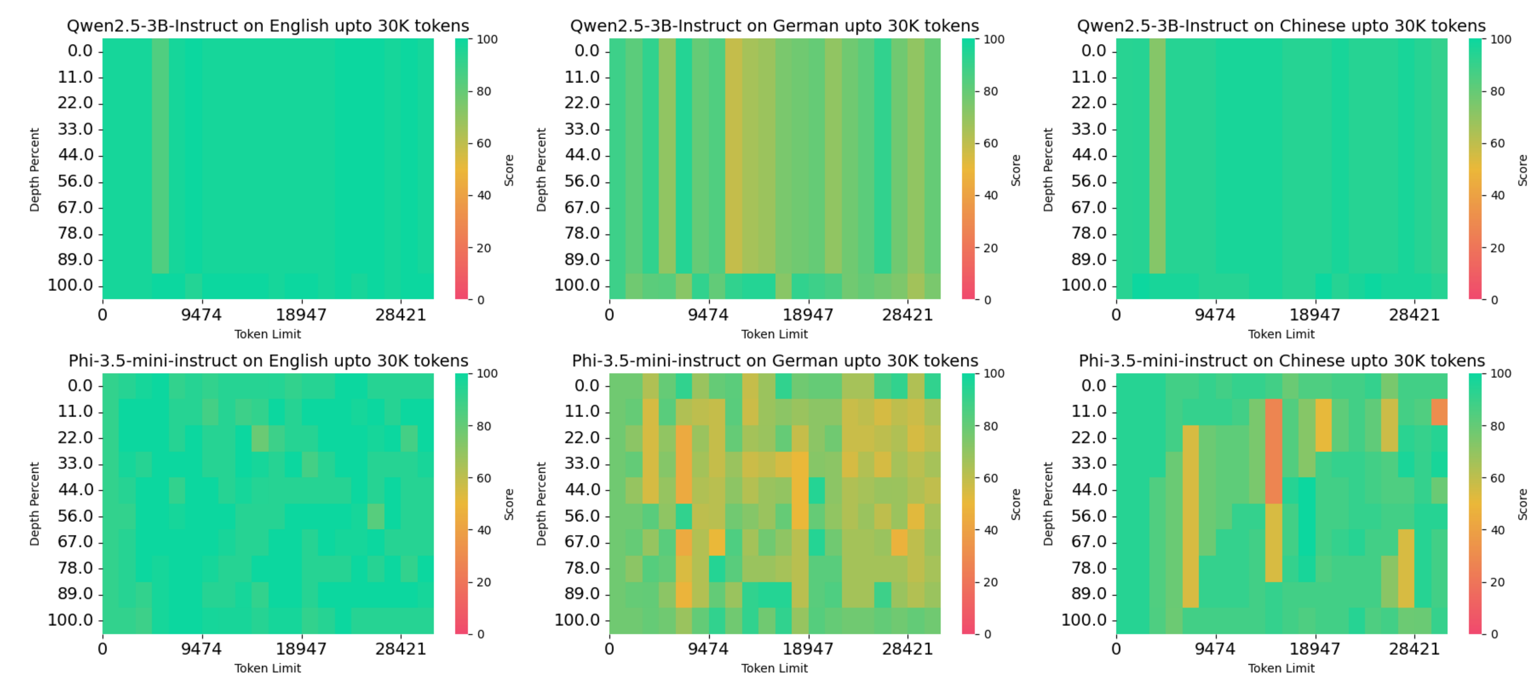

Figure 3. Needle-in-a-haystack results across three different languages. From left to right - English, German, Chinese. From top to bottom - Qwen-2.5-3B Instruct, Phi3.5 MiniInstruct. Depth Percent refers to the % of depth in the haystack where the needle is inserted. Most languages perform well across both models except German as certain noun is not faithfully translated.

Finding 1: Retrieval heads are a mix of language-agnostic and language-dependent attention heads. Nearly 50–70% of retrieval heads are shared across all three languages in Phi-3.5-3B-Mini-Instruct, and Qwen-2.5-3B-Instruct (Figure 4).

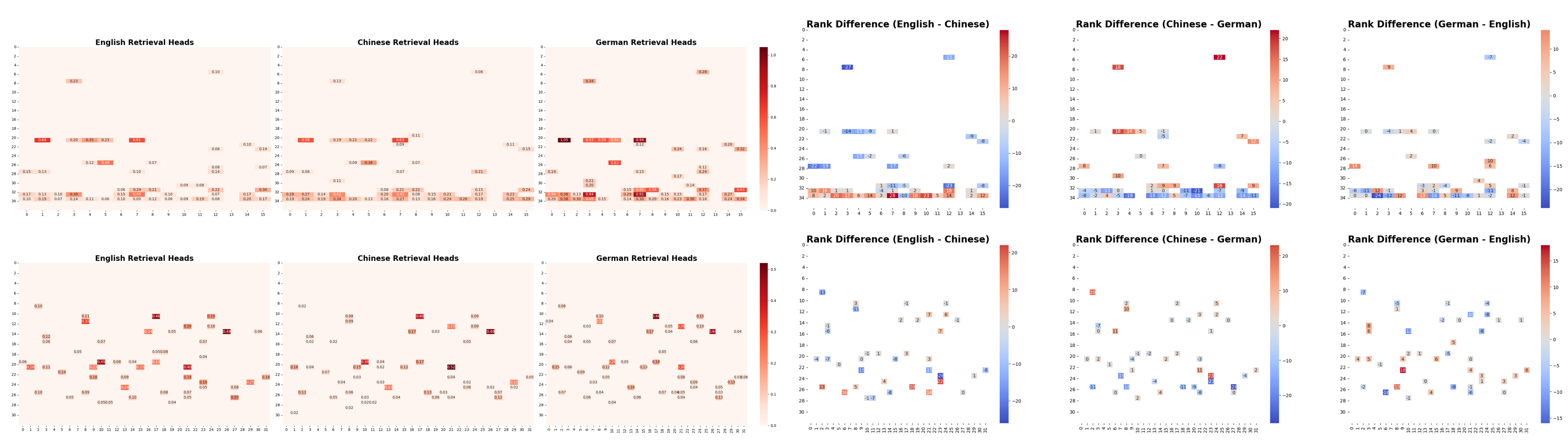

Figure 4. The distribution of retrieval heads for Qwen-2.5 and Phi-3.5 models across English, German, and Chinese languages. Top: Qwen-2.5-3B-Instruct; Bottom: Phi-3.5-3B-MiniInstruct. Left : Raw Retrieval Scores across languages; Right : Difference in pair-wise ranks of retrieval heads across languages

Finding 2 : Strength and ranking between retrieval heads is strongly associated with the underlying language Based on prior work of classifying retrieval heads (Strong (≥ 0.5), Moderate (0.1, 0.5), Weak (0, 0.1), and Non-retrieval heads (0)) we observe that all strong heads, and majority of moderate heads are shared across languages **Fig 5(a)**. Moreover, the rank correlation between language pairs is also closely related with their corresponding language distance **Table 1 (b)**

Figure 5. (a) Intersection of Retrieval heads across different languages. en: English, de : German, zh : Chinese.

Language/Model	Qwen-2.5	Phi-3.5
En-Zh	0.58	0.77
Zh-De	0.72	0.80
En-De	0.85	0.89

(b) Spearman rank correlations between retrieval head rankings across language pairs. Higher correlations are observed for linguistically closer languages (e.g., English-German) compared to distant pairs (e.g., English-Chinese), suggesting that retrieval head alignment reflects underlying language similarity.

Multilingual Evaluations

We assess the model's retrieval and translation performance in the NIAH framework, extending the original setup by prompting responses in Chinese (or German). **Finding 3: As shown in Figure 6, the model struggles in this setting, exhibiting lower ROUGE scores across different contexts and depths.**

Figure 6. Multilingual evaluation on Qwen2.5 3B Instruct, where the haystack, needle, and prompt are in English. The model is expected to generate a response in Chinese.

Causal Interventions through attention head masking

Finding 4: In Table 1, we demonstrate that masking these language agnostic heads, following their importance rankings, causes performance degradation across all languages.

Heads Masked	Acc (EN)	Drop/Head (EN)	Acc (DE)	Drop/Head (DE)	Acc (ZH)	Drop/Head (ZH)
0	0.976	–	0.786	–	0.939	–
17(LS)	0.925	5.22%	0.708	9.90%	0.877	6.60%
25(LA + LS)	0.853	12.6%	0.757	3.60%	0.787	16.18%
34(LA + LS)	0.790	19.1%	0.728	7.37%	0.858	8.63%

Table 1. Accuracy and drop per head masked across different masking configurations. **LA** refers to language-agnostic heads, while **LS** denotes language-specific heads.

References

[1] Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety – a review, 2024.

[2] Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality, 2024.