Do Retrieval Heads speak the same language?

Vishvesh Trivedi*
New York University
vt2369@nyu.edu

Shaswat Patel*
New York University
spp9399@nyu.edu

Yue Han*
New York University
yh5404@nyu.edu

Abstract

Demystifying Large Language Models (LLMs) is a crucial task to better scale the models during inference by optimizing pruning and KV caching strategies. The recent discovery of retrieval heads along with their use in optimizing KV caching has been profound. But limited works focus on multilingual extensions of the mechanistic explanations. We extend the recent work on retrieval heads to multilingual settings. Our initial observations suggest the following aspects of retrieval heads: (i) Not all retrieval heads are common across languages, with nearly 30-40% being language-specific.; (ii) he strength of retrieval heads is strongly correlated with their language-agnostic behavior with strongest retrieval heads common across all three languages and vice-versa; and (iii) Masking top-k language agnostic heads lead to negatively impacts models retrieving capabilities across all the languages. These observations further highlight the critical role of retrieval heads, particularly the strong retrieval heads. We believe these insights can inform the development of more efficient pruning and KV-caching strategies. Code and experiments are available at https://github.com/shaswatpatel123/Retrieval_Head.

1 Introduction

Recent advances in Large Language Models (LLMs) have shown impressive results in various tasks like reasoning, coding, general knowledge, etc. (DeepSeek-AI et al., 2025; Qwen et al., 2025) To better understand the role of different components in LLMs various mechanistic studies have been conducted (Olsson et al., 2022a; Wu et al., 2024; Fu et al., 2024). Recent work has uncovered the presence of *retrieval heads*, a small subset of attention heads that play a crucial role in retrieving relevant information from long context (Wu et al., 2024). Pruning these heads leads to failure in retrieving information and models start to hallucinate. Furthermore, these heads play a crucial role in Chain-of-thought (CoT) prompting strategies and pruning them leads to measurable drop in performance in CoT-based downstream tasks.

In this study, we propose to extend the retrieval heads functionality to multilingual settings. Mainly, we propose to answer the following questions:

1. Are retrieval heads language agnostic? What characteristics are shared by retrieval heads common across languages?

2. How does translation impact head activations?

3. What is the effect of masking retrieval heads associated with a certain language?

Our findings suggest that not all retrieval heads are language agnostic with nearly 30-40% being language-specific. The language agnostic heads are mainly the strongest retrieval heads while the language specific heads are mainly weaker retrieval heads. Lastly, pruning language specific heads lead to significantly large drop in downstreaming task when compared to pruning language agnostic heads. Overall, our finds further demystify the characteristic of retrieval heads by analyzing the activation patterns of attention in multilingual setting.

2 Related Work

Various studies have proposed techniques to mechanistically demystify attention heads and have discovered various functionalities performed by a certain set of attention heads (Zheng et al., 2024). McDougall et al. (2024) identified Copy Suppression behavior of layer 10 head 7 in GPT-2. Copy Suppression is the functionality of the attention head to reduce the prediction of a token that has appeared previously in the context. This can prevent the model from naively copying tokens, leading to improved calibration and more accurate predictions.

*All authors have contributed equally

Gould et al. (2024) identified Successor heads that perform incrementation of tokens in naturally ordered sequence. Olsson et al. (2022b) extend the research into Induction heads for in-context learning setting, and correlate the importance of induction heads for in-context learning. Mainly, they identify that emergence of Induction Heads coincides with sudden improvement in in-context learning abilities of the models. Wu et. al. identified Retrieval Heads which are responsible for extracting information from extended context.(Wu et al., 2024)

Similarity, prior works have discovered several characteristics of multilingual LLMs. Wendler et al. (2024a) identified language bias in Llama models where-in the Llama model's initial and final layers work on the original input language but middle-layers largely work on space closer to English. Zhang et al. (2025) recently investigated LLMs for multilingual settings and observed that LLMs share circuits for identical syntactic processes and employ distinct attention heads and feed-forward layers for language specific linguistic processes. Other works have tried to identify language-specific neurons (Tang et al., 2024) and studied how multilingual language models remember facts (Fierro et al., 2025). In similar vein, our work uncovers the characteristics of retrieval heads in multilingual language models.

Chua et al. conducted a comprehensive analysis investigating how well multilingual large language models (MLLMs) perform across different languages. Their findings show that MLLMs tend to perform best when both the questions and documents are in English, while performance significantly degrades when the relevant facts are unavailable in the target language. This suggests that knowledge within these models is heavily concentrated in English. Furthermore, translating non-English questions into English improves retrieval performance, indicating that the primary bottleneck is access to knowledge rather than linguistic fluency. To support their conclusions, the authors constructed a crosslingual open-retrieval question answering benchmark spanning 26 languages and derived insights through controlled experimental comparisons. However, the internal mechanisms within MLLMs responsible for such performance disparities remain unexplored. Therefore, our work aims to extend retrieval head analysis to multilingual settings in order to uncover the internal factors contributing to these performance differences. (Chua et al., 2025)

3 Experimental Setup

3.1 Detecting Retrieval Head

We identify retrieval heads in multilingual settings using the Needle-In-A-Haystack (NIAH) task (Kamradt, 2023). We derive inspiration from Wu et al. (2024) and follow a similar experimental setup with few changes to accommodate discrepancies induced by multilingual settings.

Needle-In-A-Haystack: NIAH task consists of a tuple (c, q, k) where c is long-context text(also known as haystack), q is the question and k is the answer related to the question(also known as needle). Generally, the q is such that the model has no parametric knowledge regarding the answer and has to retrieve the 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: Following prior work Wu et al. (2024) we calculate retrieval score by considering the ratio between the intersection of the needle tokens and decoding tokens to the number of decoding tokens.

$$\textit{Retrieval score} = \frac{|g_h \cap k|}{|k|}$$

where g_h is the token copy and pasted by a given attention head h, k is the needle inserted into the long context. Hence, the retrieval score ranges from 0 to 1 and represents each attention head's role in retrieval.

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

3.2 Extending Needle In A Haystack to the multilingual setting

To extend the NIAH task for different languages, we first create a haystack using Wikipedia dumps of each language.[1] For needles, we use Google Translate[2] to translate existing needles from NeedleBench (Li et al., 2024) by choosing three different sets of needle, question and answer tuples.

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

[2] https://translate.google.com/?sl=hi&tl=mr&op=translate

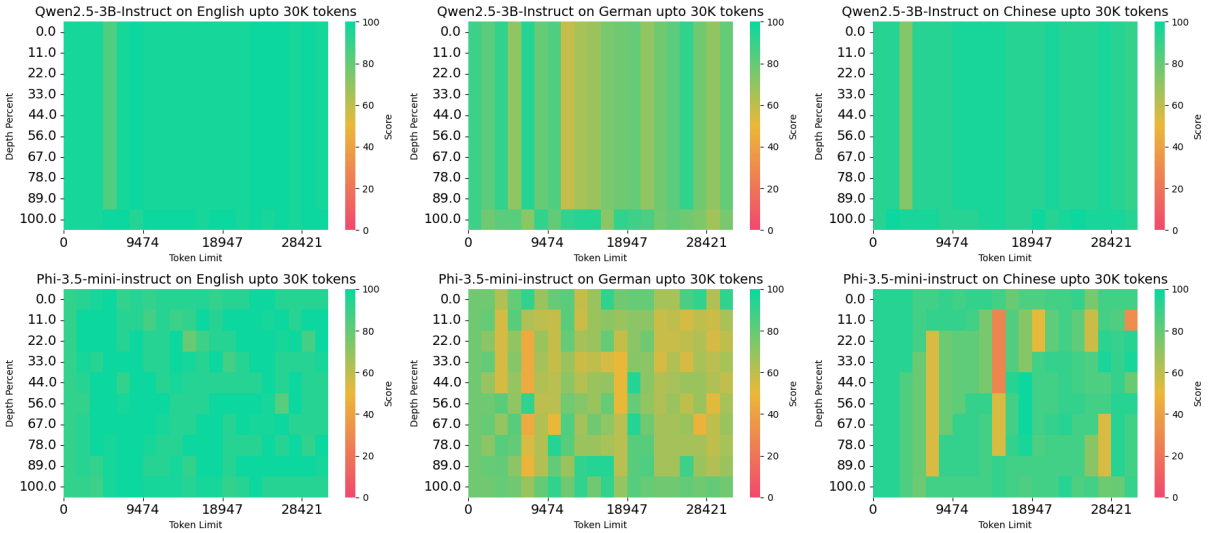

Figure 1: Needle-in-a-haystack results across three different languages. From left to right - English, German, Chinese. From top to bottom - Qwen-2.5-3B Instruct, Phi3.5 MiniInstruct. Depth Percent refers to the % of depth in the haystack where the needle is inserted. Most languages perform well across both models except German as certain noun is not faithfully translated.

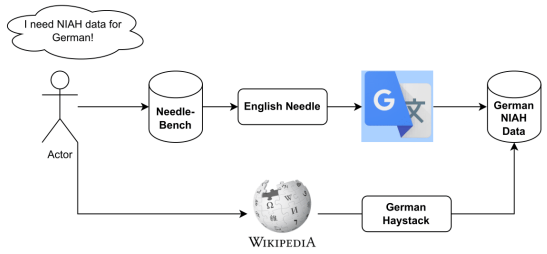

Figure 2: Pipeline to extend Needle-In-A-Haystack task to multiple languages

Therefore, while the haystack changes with the language, the content of the needle remains uniform for a fair comparison. We have so far extended the existing benchmark for Chinese, German, and Arabic languages and provide the results in Figure 1.

4 Results

4.1 Analyzing distribution of retrieval heads across different languages

Wu et al. (2024) characterized four properties for retrieval heads namely - Universality, Sparsity, Intrinsic Nature, and Dynamic Activation. The work extensively described experiments presenting evidence for each property. We extend this form of analysis in the multilingual setting. First, we determine the degree of intersection of retrieval heads across different languages. Later we examine the correlation of strength of each retrieval head to its corresponding intersecting class. Finally we show that the distribution of retrieval heads is also correlated with the corresponding language-distance such that retrieval heads are similar for similar languages and vice versa. We conduct our experiments on two open-source models Qwen2.5-3B Instruct(Qwen et al., 2025) and Phi3.5-Mini Instruct(Abdin et al., 2024). Our experiments provide the following findings:

Finding 1: Retrieval heads are composed of language-agnostic and language-dependent attention heads

We investigate the degree of intersection of retrieval heads for three languages—English, German, and Chinese—these three languages have been carefully picked based on their known language distances.(Lauscher et al., 2020; Philippy et al., 2023) English and German, both belonging to the Germanic branch of the Indo-European language family, share linguistic similarities, while Chinese, part of the Sino-Tibetan family, presents a distinct linguistic profile.(De Gregorio et al., 2024).

From Figure 4 inner-ring, we observe that 40-70% of retrieval heads are are common across all three languages (41 of 58 heads in Qwen-2.5_3B-Instruct, 28 of 66 in Phi-3.5_3B-Mini-Instruct). The rest of the retrieval heads that appear only in one or two languages are deemed specific to those languages. Other forms of correlation like

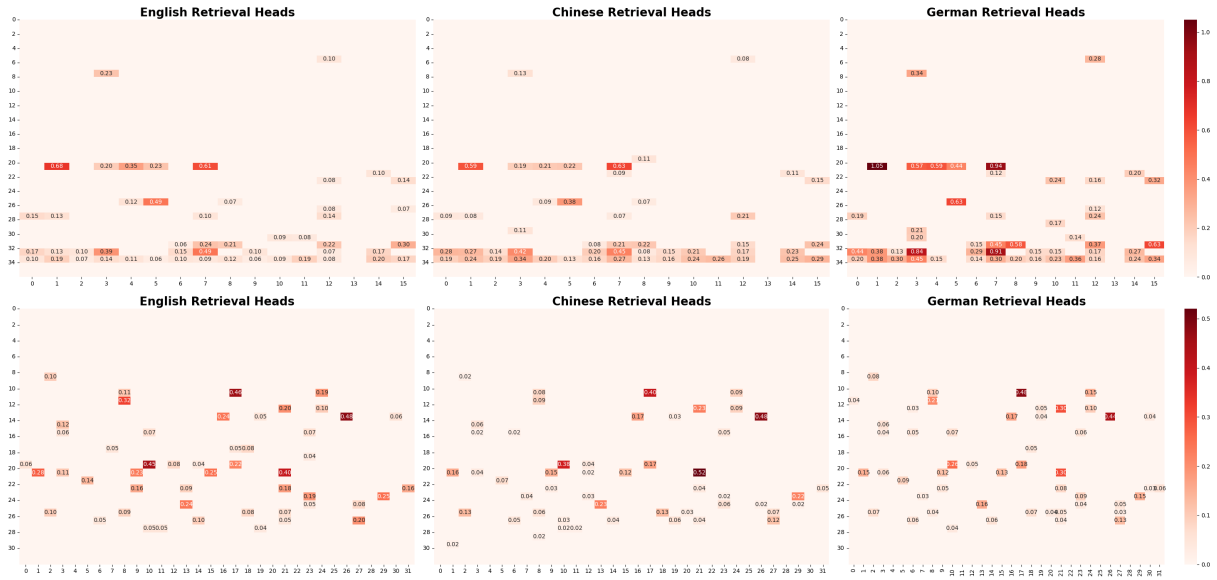

Figure 3: Visualizing retrieval heads for Qwen-2.5 and Phi-3.5 models across Transformer layers and head indexes for English, German, and Chinese languages. Top: Qwen-2.5-3B-Instruct; Bottom: Phi-3.5-3B-MiniInstruct

Figure 4: Distribution of Retrieval heads across different languages. en: English, de : German, zh : Chinese.

language pairwise intersections are less generalizable across models suggesting the dependence on the underlying architectures and training practices. Further experiments on a diverse range of model architectures and scales can help reinforce this observation.

Finding 2: Strong retrieval heads are generally language-agnostic while weaker heads are language exclusive.

Wu et al. (2024) classified attention heads based on their absolute retrieval scores into four categories - Strong ($\geq$ 0.5), Moderate $(0.1, 0.5)$, Weak $(0, 0.1)$, and Non-retrieval heads (0). Our initial experiments suggest that strength of retrieval heads is strongly co-related with its language-agnostic behaviour. As suggested by Figure 4 outer-ring for both the models most strong and moderate attention heads are shared across the three languages. Meanwhile, a majority of weak attention heads (100% for Qwen-2.5 and 73% for Phi-3.5) are exclusive to a single language or found in two languages. Additionally we find that the raw retrieval scores in Phi 3.5 are lower than Qwen

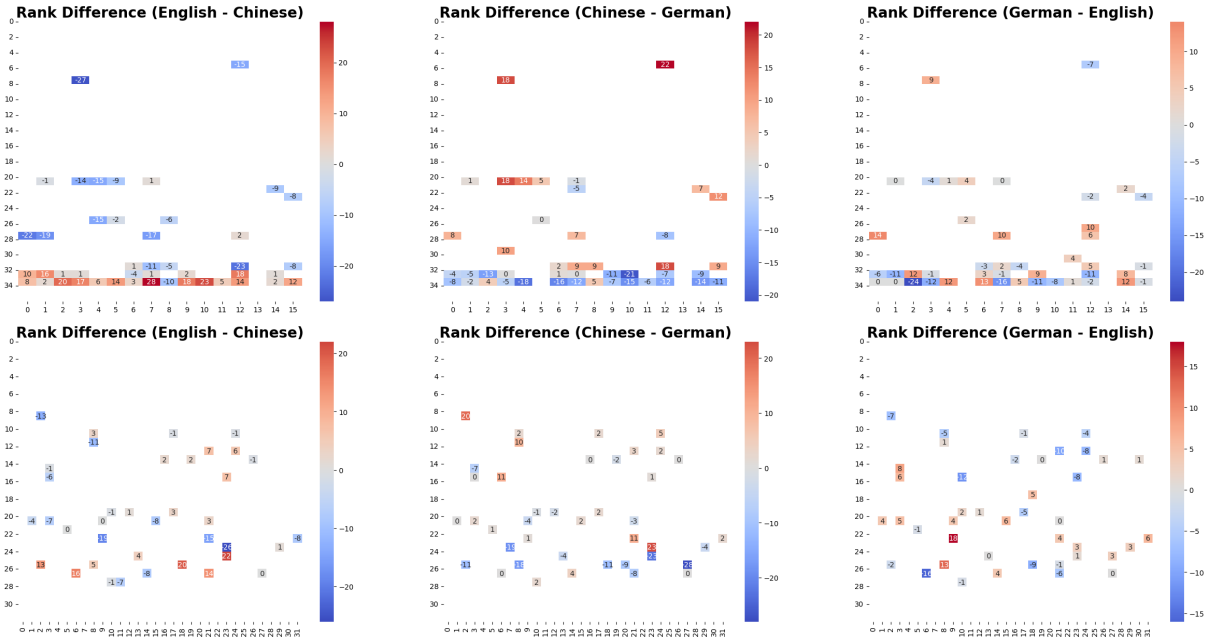

Figure 5: Difference in pair-wise ranks of retrieval heads across Transformer layers and head indexes.

2.5, particularly in German. As stated earlier, this is due to poor retrieval accuracy in that language.[3]. Figure 3 visualizes the presence of retrieval heads across various transformer layers and head indexes. We observe that while retrieval heads discovered in Phi-3.5 do not show any localization behavior, retrieval heads in Qwen-2.5 are particularly abundant in the later layers of the transformer. This observation is in line with previous research regarding LLMs handling multilingual prompts. The LLM translates text to English in the initial layers while processing the request in the middle layers. Later, the English answer is translated back to the language of conversation in the final layers of the Transformer block. (Wendler et al., 2024b). We find that language-specific weaker heads are more localized in the final layers of the transformer, eliciting translation behavior (Figure 3)

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

Table 1: Comparison of language pair scores across Qwen-2.5 and Phi-3.5 models

[3]Hence in Phi-3.5 we do not find any heads that have a score greater than 0.5 for all three languages given the highest score in German is 0.45

Finding 3: The pairwise correlation between retrieval heads from different languages is associated with their known language-distance.

Language distance is a widely studied topic in Linguistics, where languages are grouped as being similar or dissimilar based on their origin, written scripture, indentation, and other factors. We find that the retrieval heads also exhibit a similar property where heads from linguistically closer languages are highly correlated as compared to heads from linguistically disparate languages. We utilize relative ranking of heads to measure retrieval strength instead of raw scores to normalize across various languages. Hence the head with the highest retrieval score is ranked 1 and accordingly for other heads. We then compare pairwise ranks between two languages where high correlation factor signifies more similarity between retrieval heads of those languages. Table 1 shows the pairwise rank correlation progressively increases with decrease in language distance *(from English and Chinese to English and German)*. Figure 5 visualizes the difference in pair-wise ranks between languages across various layers and head indexes.

4.2 Retrieval-translation heads evaluations

In this section, we investigate the effect of translation and attention head activation. our experimental setup follows the standard NIAH framework, we

Heads Masked	ROUGE (EN)	Drop (EN)	ROUGE (DE)	Drop (DE)	ROUGE (ZH)	Drop (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 2: Masking out language-agnostic top-k retrieval heads severely damage the retrieval capabilities across all languages. Causally proving that stronger retrieval heads are shared by all languages.**LA** refers to language-agnostic heads, while **LS** denotes language-specific heads.

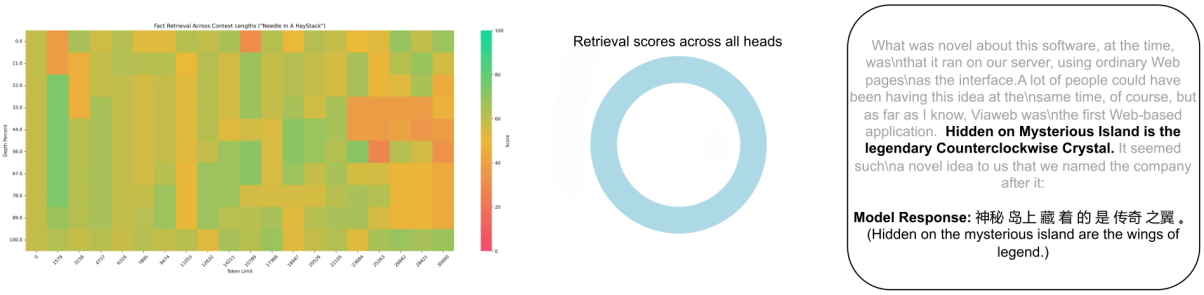

Figure 6: Multilingual evaluation on Qwen2.5 3B Instruct, where the haystack, needle, and prompt are in English. The model is expected to generate a response in Chinese.

introduce an additional constraint by prompting the model to respond in a specified target language while keeping the needle, haystack, and prompt in one specific language. This configuration is designed to identify attention heads involved in both retrieval and translation. Since the model must first retrieve the relevant information and then translate. This setup is similar to Fu et al. (2025) for identifying retrieval-reasoning attention heads. For this experiment, we use Qwen2.5-3B Instruct. The haystack, needle and prompt are in English while the model is prompted to generate the output in Chinese.

Finding 4: As shown in Figure 6, the model struggles in this setting, exhibiting lower ROUGE scores across different contexts and depths. We observe that the model straggles in this setting and observe that non of the attention heads have any retrieval scores(light blue indicates 0 retrieval scores). While some heads do exhibit non-zero retrieval scores, values below 0.01 are rounded to zero to avoid misclassifying noisy attention patterns as true retrieval behavior. This suggests that our current experimental setup does not generalize well to this translation scenario and a careful experiential design is required to identify retrieval-translation heads. We also provide examples for German language setting in Appendix Figure 1.

4.3 Masking retrieval heads

In this section, we examines the effect of masking retrieval heads on NIAH task. Across the experiments we use Qwen2.5-3B Instruct model as it has shown near perfect ROUGE scores across for the NIAH task(Table 1). Wu et al. (2024) have previously observed that masking top retrieval heads negatively impact models capabilities across multiple tasks.

Finding 5: Masking language agnostic heads, following their importance rankings, causes performance degradation across all languages.

We extend masking experiments to multilingual NIAH setting by initially masking all the language-specific retrieval heads. Then, we progressively mask the top-k ranked retrieval heads, the top-k heads are ranked based on retrieval scores obtained via English experiments. In our experiments, we apply mask to top 8, and top 17 ranked retrieval heads.

As observed in Table 2, as the number of attention heads masked increases, the drop in ROUGE score increases. Furthermore, masking language specific head also decreases the ROUGE scores but the severity in drop is more profound for language agnostic heads.

5 Conclusion and Future Work

This paper further demystifies the retrieval heads and their characteristics in a multilingual setting. We conduct an in-depth analysis and illustrate that strong retrieval heads are language agnostic. The language agnostic heads strongly influence the downstreaming tasks compared to compared to language specific heads, which are generally weaker heads. We believe these insights can inform the development of more efficient pruning and KV-caching strategies.

Future work. We anticipate that our findings can serve as a foundation for numerous future research directions. Fu et al. (2025) build on Wu et al. (2024)'s work by allocating different KV caching budget based on retrieval reasoning scores. A similar extension is also possible for multilingual QA tasks based on our findings. As illustrated in our Multilingual Code Evaluation experiments, a careful design is required to identify retrieval-translation heads and can be a promising extension. Moreover, while our study centers on identifying retrieval heads in trained models, the dynamics of their emergence during the training process remain unexplored. The emergence of language-agnostic behavior and its relationship to the composition of the training data remains unexplored. These questions present promising avenues for future research, building upon the framework established in our study.

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, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone.

Lynn Chua, Badih Ghazi, Yangsibo Huang, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Amer Sinha, Chulin Xie, and Chiyuan Zhang. 2025. Crosslingual capabilities and knowledge barriers in multilingual large language models.

Juan De Gregorio, Raúl Toral, and David Sánchez. 2024. Exploring language relations through syntactic distances and geographic proximity. *EPJ Data Science*, 13(1).

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li,

Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2025. Deepseek-v3 technical report.

Constanza Fierro, Negar Foroutan, Desmond Elliott, and Anders Søgaard. 2025. How do multilingual language models remember facts? *arXiv preprint arXiv:2410.14387*.

Yu Fu, Zefan Cai, Abedelkadir Asi, Wayne Xiong, Yue Dong, and Wen Xiao. 2024. Not all heads matter: A head-level kv cache compression method with integrated retrieval and reasoning. *arXiv preprint arXiv:2410.19258*.

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*.

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*.

Greg Kamradt. 2023. Needle in a haystack - pressure testing llms.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Mo Li, Songyang Zhang, Yunxin Liu, and Kai Chen. 2024. Needlebench: Can llms do retrieval and reasoning in 1 million context window? *arXiv preprint arXiv:2407.11963*.

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, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022a. In-context learning and induction heads.

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, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022b. In-context learning and induction heads. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Identifying the correlation between language distance and cross-lingual transfer in a multilingual representation space. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, page 22–29. Association for Computational Linguistics.

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, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. *arXiv preprint arXiv:2402.16438*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024a. 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.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024b. 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.

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

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 *The Thirteenth International Conference on Learning Representations*.

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*.

A Appendix

Needle: The first band to play on the Moon was the Virtual Rocket Band.

Response: "Welches Band war das erste, das auf dem Mond spielte? English: According to the passage, the first band to play on the Moon was the Virtual Rocket Band. German: Das erste Band, das auf dem Mond spielte, war die"

Depth: 0 **Context:** 1579

Needle: Hidden on Mysterious Island is the legendary Counterclockwise Crystal.

Response: "Was ist der berühmte Gegenstand, der auf dem mysteriösen Inselgelände versteckt ist? English: The legendary item hidden on Mysterious Island is the Counterclockwise Crystal. German: Der berüh"

Depth: 0 **Context:** 22105

Figure 1: Retrieval-translation attention heads experiment failure cases where the haystack, needle and query is in English while the expected answer is in German.