TOKEN ALIGNMENT HEADS: UNVEILING ATTENTION'S ROLE IN LLM MULTILINGUAL TRANSLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, large language models (LLMs) have made remarkable progress, with multilingual capability emerging as a core foundational strengths. However, the internal mechanisms by which these models perform translation remain incompletely understood. In this paper, we elucidate the relationship between the attention mechanism in LLMs and their translation abilities. We find that certain attention heads, which we term token alignment heads, are specifically responsible for mapping tokens from the source language to the target language during inference. Through a systematic investigation across various models, we confirm that these token alignment heads exhibit several key characteristics: (1) Universality: They are present in all LLMs we studied. (2) Sparsity: They constitute only a small fraction of all attention heads. (3) Consistency: The set of token alignment heads activated by the model shows strong consistency across different language pairs. (4) Causality: Interventionally removing these heads leads to a sharp decline in the model's translation performance, while randomly removing non-token alignment heads has little impact on translation ability. (5) Functional Specificity: Ablating token alignment heads disproportionately harms translation but has a varied impact on other multilingual tasks. We also traced the formation of token alignment heads during pre-training, revealing an evolutionary path of rapid proliferation, stabilization, and eventual pruning. Furthermore we leverage these token alignment heads to filter multilingual training data, and our experiments show that these data could enhance translation capabilities of the models.

1 INTRODUCTION

Recently released Large Language Models (LLMs) (Comanici et al., 2025; OpenAI, 2025; Anthropic, 2025; Liu et al., 2024; Yang et al., 2025) have demonstrated remarkable multilingual capabilities, showing significant improvements in both the complexity of multilingual tasks they can handle and the range of languages they support. Multilingual proficiency has now become an essential foundational ability for state-of-the-art LLMs. Among these capabilities, translation is particularly crucial, as it not only represents a key application but also underpins the overall multilingual performance of these models. A deeper understanding of the underlying mechanisms of translation in LLMs is therefore vital, not only for a comprehensive view of their inner workings but also for providing valuable insights to guide the development of superior multilingual training strategies, in terms of both model architecture and data selection.

A growing body of research has begun to demystify how LLMs process multilingual information (Artetxe et al., 2020; Lindsey et al., 2025; Datta et al., 2020; Chang et al., 2022). Several recent works have explored the internal mechanics of multilingualism in LLMs. For instance, Zhao et al. (2024) found that LLMs often initially process queries by converting multilingual inputs into an English-centric representation before solving tasks. Similarly, Schut et al. (2025) revealed that these models tend to make decisions and reason within an English-dominated semantic space. These findings underscore the hypothesis that an internal translation process is a core component of LLM multilingualism, highlighting the importance of understanding these translation mechanisms to further elucidate their broader multilingual capabilities.

Prior research (Michel et al., 2019; Vig & Belinkov, 2019; Finlayson et al., 2021; Elhage et al., 2021a) into the attention mechanism has established the functional specialization of individual

Response By Llama–3.1–8B:
Translate this sentence from English to Simplified Chinese
Sentence: "We now have 4–month–old mice that are non–diabetic that used to be diabetic," he added.
Answer: **"我们现在有4个月大的老鼠，它们以前是糖尿病的，现在不再是糖尿病了。" 他补充道**

Response By **top 30 token alignment heads pruned** Llama–3.1–8B:
Translate this sentence from English to Simplified Chinese
Sentence: "We now have 4–month–old mice that are non–diabetic that used to be diabetic," he added.
Answer: **"We now have 4–month–old mice that are non–diabetic that used to be diabetic," he added.**

Response By **30 random heads pruned** Llama–3.1–8B:
Translate this sentence from English to Simplified Chinese
Sentence: "We now have 4–month–old mice that are non–diabetic that used to be diabetic," he added.
Answer: **"我们现在有4个月大的老鼠，它们以前是糖尿病的，现在不再是糖尿病了。" 他补充道**

Figure 1: An example of token alignment head pruning in Llama-3.1-8B. The top panel shows the correct translation generated by the original model. The middle panel demonstrates that after the top 30 token alignment heads are pruned, the model loses its ability to translate and reverts to copying the English source text, while its basic copy-paste functionality is preserved. In contrast, the bottom panel shows that pruning 30 random non-token alignment heads (control) has no impact on the translation output.

heads. Early work on Transformer (Vaswani et al., 2023) models for machine translation showed that many attention heads were redundant and could be pruned with minimal impact on performance (Voita et al., 2019; Kovaleva et al., 2019). Subsequent studies, such as Kim et al. (2021); Ma et al. (2021); Zhang et al. (2025), have investigated the role of attention heads in translation by measuring their impact on translation metrics. These studies identified certain attention heads as critical for translation and noted that the sets of important heads are highly similar across different language pairs. However, these approaches often have limitations. They frequently rely on task-specific evaluation metrics, are typically conducted on smaller or single models, and their methods for identifying important heads can be opaque. Crucially, they often stop at identifying which heads are important, without fully explaining how these heads mechanistically contribute to the translation process.

Unlike previous work that measures head importance based on downstream benchmark performance, our approach shifts the focus to identifying the underlying mechanism. Inspired by the discovery of other functionally specialized circuits in LLMs, such as "induction heads" that implement in-context learning (Olsson et al., 2022) and "retrieval heads' designed for knowledge retrieval (Wu et al., 2025a), we hypothesized that a similar specialization must exist for translation. We posited that beyond attention heads performing generic copy-paste behaviors, there must be a set of heads specifically responsible for the core translation task: mapping tokens from a source language to their corresponding tokens in a target language. We term these attention heads, characterized by their direct cross-lingual token alignment behavior which can be regarded as a form of word alignment(Brown et al., 1993; Och & Ney, 2003), "token alignment heads". Figure 1 provides an illustration of this functional specialization: when we ablate the top 30 token alignment heads from a Llama-3.1-8B model (Grattafiori et al., 2024), its translation capability collapses. Crucially, the model does not simply fail; it reverts to a more basic copy-paste behavior, reproducing the English input verbatim. This demonstrates that the model's general ability to copy tokens remains intact, and that the ablated heads perform a specific, non-copying function of cross-lingual mapping.

In this work, we conduct a systematic investigation across a series of LLMs to validate the existence and properties of token alignment heads. We reveal that token alignment heads play a pivotal role in the translation capabilities of LLMs, thereby uncovering the relationship between translation mechanisms and attention. We further identify and validate several key characteristics of these token alignment heads: (1) Universality: All large models we studied possess such token alignment heads; (2) Sparsity: Only a small subset of attention heads function as token alignment heads; (3) Consistency: The token alignment heads activated by the model when translating different language pairs exhibit strong consistency. (4) Causality: Ablating these token alignment heads through causal intervention leads to a significant drop in translation performance, whereas removing a random equivalent number of non-token alignment heads has little impact. (5) Functional Specificity: Ablating token alignment heads disproportionately harms translation but has a varied impact on other multilingual tasks, suggesting the funcional specificity of token alignment heads. Furthermore, we investigate the formation process of token alignment heads by analyzing the entire pre-training life-cycle of a model. Our analysis reveals a distinct developmental trajectory in three phases: an initial rapid proliferation of token alignment heads that coincides with the acquisition of translation ability, followed by a period where the core set of heads stabilizes, and finally a long phase of consolidation and pruning. This discovery provides insight into how specialized circuits emerge and are optimized during large-scale training.

We further substantiate these findings through a practical application. We introduce TRater, an algorithm that leverages token alignment heads to score multilingual training data based on its importance to the translation mechanism. Our experiments reveal that a small fraction of data identified by TRater is responsible for the model's final translation proficiency. This result provides evidence for the causal role of token alignment heads. The discovery that a core capability like translation is governed by a sparse and functionally specialized circuit provides a concrete target for future research. Ultimately, these insights pave the way for more efficient and robust multilingual systems, enabling targeted architectural innovations, data curation strategies guided by mechanistic understanding.

2 Detecting Token Alignment Head

In this section, we introduce the algorithm for detecting token alignment heads. Since the model involves cross-lingual token alignment during the translation process, we first need to identify the mapping relationship between tokens from the source language to the target language. Then, we define a metric called the translation score to recognize attention heads that implement the model's translation mechanism. The translation score measures the frequency with which an attention head maps tokens from the source language to the target language. If an attention head exhibits a relatively high translation score, it indicates that this attention head frequently performs cross-lingual token alignment when processing different translation texts. Such attention heads are what we refer to as token alignment heads.

2.1 Token Alignment Annotation

Since existing token alignment tools do not cover all languages, we utilize OpenAI's GPT-4.1 model to annotate token alignments in translation texts. Specifically, we require the large language model to identify the corresponding source language token for each target language token and provide a confidence score for each token alignment. To ensure the accuracy of the annotations, we only retain token alignments with a confidence score greater than 0.9. If no corresponding source token exists, it is marked as None. Figure 2 shows an example of token alignment results from English to Chinese.

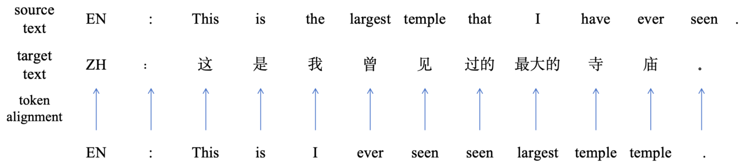

source text
EN : This is the largest temple that I have ever seen .
target text
ZH ： 这 是 我 曾 见 过的 最大的 寺 庙 。
token alignment
EN : This is I ever seen seen largest temple temple .

Figure 2: A token alignment example from English text to Chinese text

2.2 Translation Score

In the decoding process of translation, we define the translation score as the frequency of valid token alignments by attention heads. Specifically, during the greedy decoding process, let the currently generated token be t, and let the attention score of the attention head be denoted as $\mathbf{w} \in \mathbb{R}^{|\mathbf{x}|}$. If token t has a corresponding source token s with position idx as s_{idx}, and the attention head assigns the highest attention score probability to the source token s, then we consider that the attention head has successfully completed a language-pair token alignment. Formally, we have: $\mathbf{w}_{s_{\text{idx}}} = \max(\mathbf{w})$.

Let g_h denote the number of valid language-pair token alignments performed by attention head h, and let m be the total number of target tokens that have a corresponding valid source token. Then, the translation score of attention head h is defined as:

$$\mathrm{TS}_h = \frac{g_h}{m} \tag{1}$$

2.3 Token Alignment Head Detection

To empirically identify token alignment heads, we compute the Translation Score for every attention head in the model using the dev split of the FLORES-101 dataset (Goyal et al., 2021). For each of the approximately 900 source-target sentence pairs in a given language direction, we calculate the TS for all heads. The final score for each head is the average TS computed across all examples in that language pair. An attention head is then classified as a token alignment head if its final Translation Score exceeds a predefined threshold of 0.1.

3 BASIC PROPERTIES OF Token Alignment HEADS

In this section, we characterize the fundamental properties of the identified token alignment heads. To ensure the robustness and generalizability of our findings, our analysis spans a diverse set of open-source models. This selection covers a range of parameter sizes (1.7B to 30B), architectures (dense and Mixture-of-Experts (Jacobs et al., 1991)), and training stages (pre-trained and instruction-tuned). The models include Llama-3.1-8B, Mistral-7B-Instruct-v0.3[1], Mistral-7B-v0.3 (Jiang et al., 2023), Qwen2.5-7B (Qwen et al., 2025), Qwen3-1.7B, and Qwen3-30B (Yang et al., 2025).

3.1 Universality and Sparsity

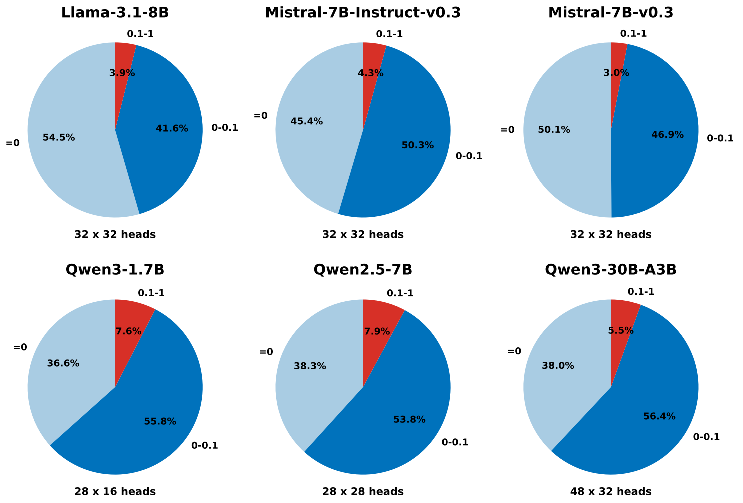


Figure 3: Translation score distribution for different models. According to the different translation scores, the model's attention heads are divided into three categories: token alignment heads (red), infrequently activated heads (blue), and heads with near-zero activation (light blue). All models studied have token alignment heads, and the proportion of token alignment heads is relatively small. Most of the heads are either not activated or are activated at a low frequency.

Our analysis first reveals two fundamental properties: universality and sparsity. As illustrated in Figure 3, token alignment heads (defined as having a translation score > 0.1) are present in every model we examined, irrespective of its size, architecture, or training stage. This confirms that they are a universal, emergent feature of multilingual LLMs.

Concurrently, token alignment heads are exceptionally sparse. They constitute less than 8% of the total attention heads in all models, and as few as 3% in Mistral-7B-v0.3. Additionally, infrequently

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

activated heads during the translation process account for nearly 50%. The remaining attention heads, which are almost never activated, account for between 36% and 55%.

We next analyze the positional distribution of these heads within the model architecture. Figure 4 shows a consistent pattern: token alignment heads are predominantly concentrated in the middle layers of the models. In contrast, the earliest and latest layers contain very few token alignment heads. This aligns with the broader understanding of Transformer architectures, where initial layers are thought to handle surface-level feature extraction and final layers are responsible for structuring the output.

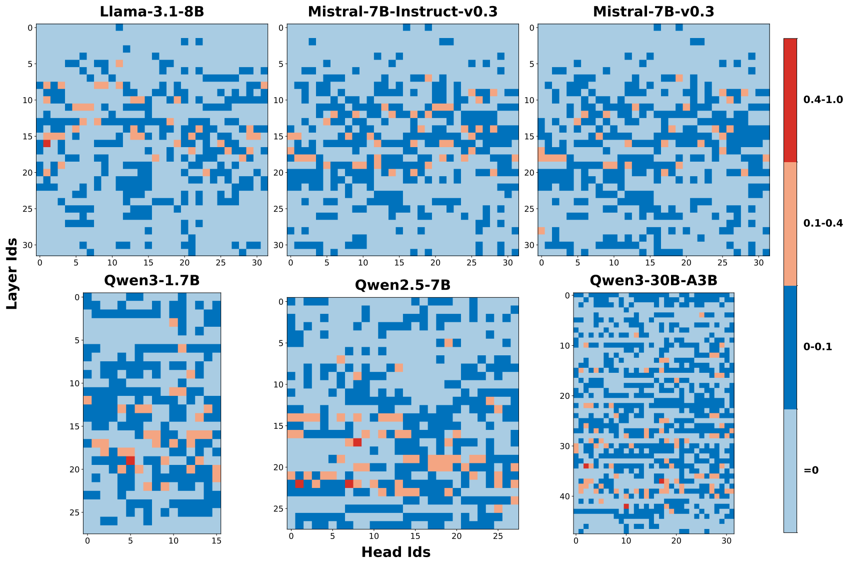
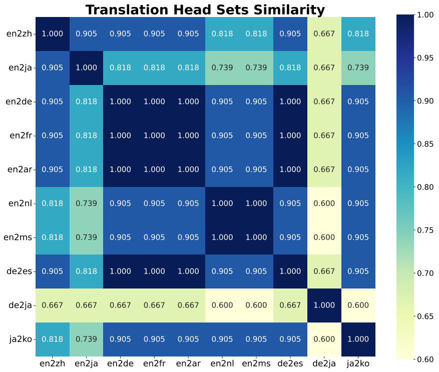


Figure 4: Positional distribution of translation scores in different models. Each heatmap visualizes the translation score distribution. The color intensity corresponds to the translation score, with warmer colors (red/orange) indicating higher scores and cooler colors (blue) indicating lower scores.

3.2 Language Consistency



	en2zh	en2ja	en2de	en2fr	en2ar	en2nl	en2ms	de2es	de2ja	ja2ko
en2zh	1.000	0.905	0.905	0.905	0.905	0.818	0.818	0.905	0.667	0.818
en2ja	0.905	1.000	0.818	0.818	0.818	0.739	0.739	0.818	0.667	0.739
en2de	0.905	0.818	1.000	1.000	1.000	0.905	0.905	1.000	0.667	0.905
en2fr	0.905	0.818	1.000	1.000	1.000	0.905	0.905	1.000	0.667	0.905
en2ar	0.905	0.818	1.000	1.000	1.000	0.905	0.905	1.000	0.667	0.905
en2nl	0.818	0.739	0.905	0.905	0.905	1.000	1.000	0.905	0.600	0.905
en2ms	0.818	0.739	0.905	0.905	0.905	1.000	1.000	0.905	0.600	0.905
de2es	0.905	0.818	1.000	1.000	1.000	0.905	0.905	1.000	0.667	0.905
de2ja	0.667	0.667	0.667	0.667	0.667	0.600	0.600	0.667	1.000	0.600
ja2ko	0.818	0.739	0.905	0.905	0.905	0.905	0.905	0.905	0.600	1.000

Figure 5: Jaccard similarity matrix of token alignment head sets across various language pairs. The axes list ten different language pairs. The token alignment heads between different language-pairs are very similar, most of the similarity scores are above 0.9.

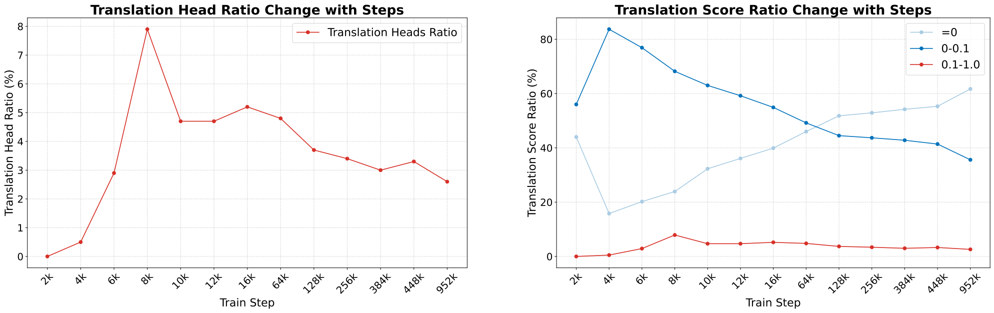
To investigate the consistency of token alignment heads in the model across different language pairs, we selected the following language pairs: English-Chinese (en2zh), English-Japanese (en2ja), English-German (en2de), English-French (en2fr), English-Arabic (en2ar), English-Dutch (en2nl), English-Malay (en2ms), German-Spanish (de2es), German-Japanese (de2ja), and Japanese-Korean (ja2ko). These pairs cover a variety of linguistic families. Without loss of generality, we focus this analysis on the Llama-3.1-8B model. For each language pair, we selected the top 20 token alignment heads to form its translation set. Then we compute the pairwise similarity between these sets using the Jaccard index:

$$\mathrm{Sim}_{S,T} = \frac{|S \cap T|}{|S \cup T|} \tag{2}$$

As shown in Figure 5, the similarities between all language pairs are relatively high, with Jaccard similarity scores consistently exceeding 0.8 for most pairs, and never dropping below 0.6. This indicates that a largely invariant set of attention heads is responsible for translation across diverse linguistic families, demonstrating the strong cross-lingual generalizability of token alignment heads.

4 Formation Process of Token Alignment Heads

This section investigates the formation process of token alignment heads during the model's pre-training lifecycle. To trace their development, we trained an 8B parameter model, architecturally identical to Llama-2 model (Touvron et al., 2023), from scratch. The model was trained for a total of 15 trillion tokens. The specific composition of the dataset and the hyperparameters for training are detailed in Appendix A.1. To understand the formation process of token alignment heads, we analyzed model checkpoints at multiple intervals throughout training to map the evolutionary trajectory of token alignment heads.



(a) The ratio of token alignment heads rises sharply in the early training stages, reaching a peak of approximately 8% at 8k step. Following this peak, the proportion drops and stabilizes around 5% between 10k and 64k steps. Subsequently, it enters a long phase of gradual decline, settling at 2.6% by the end of training.

(b) The distribution of all attention heads by activity level. The proportion of inactive heads (TS = 0, light blue line) starts at a low point but steadily increases throughout training, reaching over 60% by the final step. Conversely, the proportion of low-activity heads (dark blue line) begins as the dominant group but consistently decreases over time.

Figure 6: The evolutionary trajectory of token alignment heads during training

Phase 1: Rapid Proliferation (Early Training Stage, 0-8k steps). In the initial stages of training, as shown in Figure 6(a), the proportion of token alignment heads experiences a rapid proliferation, growing from zero to its peak. This period of rapid circuit formation coincides directly with the steepest gains in the model's translation performance, where the FLORES chrF++ metric surged from 12.58 to 45.77. This suggests that the initial acquisition of translation ability is contingent on the rapid emergence of these specialized heads.

Phase 2: Set Stabilization (Early-to-Mid Training Stage, 10k-64k steps). From 10k to 64k steps, the proportion of token alignment heads stabilized around 5%, and the core set of these heads becomes remarkably stable. To quantify this, we define a conditional overlap ratio metrics which measures the overlap between the token alignment head set at any given step (A) and the final reference set at the end of training (B):

$$\frac{|A \cap B|}{|B|} \tag{3}$$

As shown in Figure 7, from approximately 8k steps onward, conditional overlap ratio remains consistently high. This indicates that the token alignment heads formed rapidly in the early stage of training are largely maintained throughout the subsequent training process.

Phase 3: Consolidation and Pruning (Mid-to-Late Training Stage, 64k-952k steps). In this longest phase of training, we observe a gradual decline in the overall proportion of token alignment heads, which settles at 2.6% (Figure 6(a)). Given that the core set of heads remains stable (Phase 2), this decline implies that heads with weaker or more redundant translation capabilities are being "pruned"—their translation scores fall below the threshold as the network refines its functions.

We hypothesize this pruning is part of a broader network-wide optimization towards increased sparsity and computational efficiency. This is corroborated by the shifting distribution of head activity shown in Figure 6(b). As training progresses, the proportion of completely inactive heads steadily increases, reaching 61.7% by the end. This happens at the expense of low-to-moderately active token alignment heads. In essence, the model learns to solve the translation task not by using more heads, but by relying more heavily on a smaller, more efficient, and highly specialized set of circuits, while deactivating others. This process of over-producing and then refining specialized circuits appears to be a key mechanism in the development of efficient neural networks.

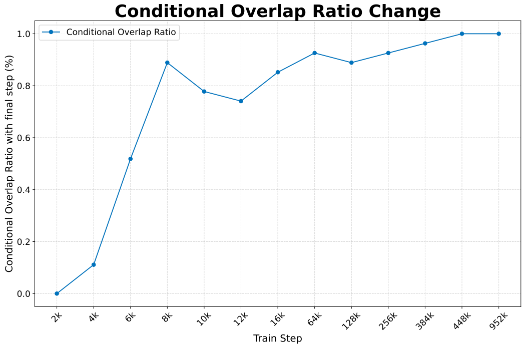


Figure 7: Stability of the token alignment head set over time. The conditional overlap ratio exhibits a steep and rapid increase during the initial training phase, rising from zero to nearly 0.9 at 8k step. From this point onward, the overlap remains consistently high, fluctuating but generally staying above 0.8 and approaching 1.0 by the end of training.

5 INFLUENCE ON DOWNSTREAM TASKS

In this section, we investigate the impact of token alignment heads on downstream benchmarks. First, we analyze the influence of token alignment heads on the model's translation performance to demonstrate their causality. Here, we select the FLORES101 benchmark to evaluate the model's translation performance. Next, we examine the impact of token alignment heads on the model's general multilingual capabilities. We evaluate the following benchmarks: translated Hellaswag (Zellers et al., 2019), ARC-Easy and ARC-Challenge (Clark et al., 2018) which are detailed in Appendix A.2, XWinograd (Tikhonov & Ryabinin, 2021), XStoryCloze (Mostafazadeh et al., 2016), XNLI (Conneau et al., 2018), XCOPA (Ponti et al., 2020), and a localized multilingual variant of the MMMLU[2] test set denoted as XMMLU which includes JMMLU[3], CMMLU (Li et al., 2024), AMMLU[4], IndoMMLU (Koto et al., 2023), and VMLU[5]. We study the influence of token alignment heads by mask those token alignment heads.

[2] https://huggingface.co/datasets/openai/MMMLU

[3] https://huggingface.co/datasets/nlp-waseda/JMMLU

[4] https://huggingface.co/datasets/Hennara/ammlu

[5] https://vmlu.ai

5.1 Translation Capacity

In this subsection, we investigate the impact of token alignment heads on the translation performance of Llama-3.1-8B, Mistral-7B-v0.3, Qwen2.5-7B, Qwen3-1.7B, and Qwen3-30B using the FLORES-101 benchmark. We compare translation metrics before and after masking token alignment heads, as well as after masking random non-token alignment heads. To clearly illustrate the effects, we report the difference between the metrics of the masked models and those of the baseline models (without masking). Figure 12 presents the changes in BLEU and chrF++ scores. Masking token alignment heads leads to substantial declines in both metrics, with the largest drops exceeding 17 points for BLEU and 25 points for chrF++. In contrast, masking random heads has only a minimal effect. These results demonstrate that token alignment heads have a direct and significant influence on the models' translation capabilities, a property we refer to as the causality of token alignment heads.

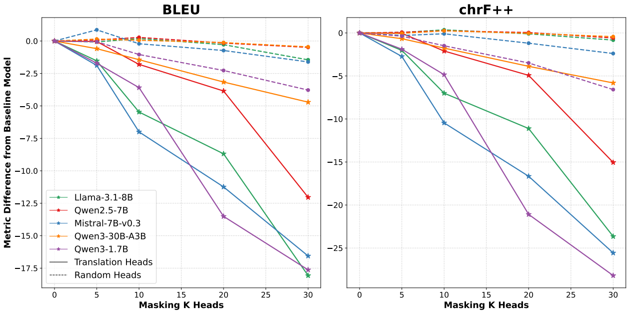


Figure 8: Impact of masking token alignment heads versus random heads on FLORES benchmark scores. Masking token alignment heads leads to a significantly larger performance drop compared to masking random heads

5.2 Multilingual Capacity

We then investigate the impact of ablating token alignment heads on a broader suite of multilingual benchmarks. From the results presented in Figure 9, we observe a clear hierarchy of dependency on token alignment heads. Benchmarks such as Hellaswag_ML, ARC_C_ML, and ARC_E_ML exhibit a significant performance drop (up to 10 points), which is consistently larger than the drop from ablating random heads. This suggests that these tasks, while not pure translation, partially rely on the cross-lingual mapping capabilities provided by token alignment heads. This functional overlap may stem from translation artifacts in their data creation process or a genuine need for cross-lingual conceptual alignment to solve the tasks.

In contrast, for other benchmarks like XNLI and XCOPA, token alignment heads demonstrate weak causality, as their ablation often results in a smaller performance drop than that of random ablation. This indicates that these tasks depend on different multilingual mechanisms within the model, likely operating at a higher semantic level that does not require the token-level mapping performed by token alignment heads. These findings suggest that token alignment heads provide a foundational cross-lingual alignment capability that various downstream tasks leverage to different degrees.

5.3 Token Alignment head as Data Rater

To further probe the relationship between token alignment heads and multilingual data, we introduce TRater, a data-filtering algorithm. TRater leverages token alignment heads to score data samples based on their importance to the translation mechanism. We compute the score of sample x as follows:

$$\text{score}(x) = \frac{1}{m} \sum_{i} \left(L(\theta_{\text{mask}}, x_i) - L(\theta, x_i) \right) \tag{4}$$

where L denotes the token level cross entropy loss, θ denotes the original model parameters, θ_{mask} represents the model parameters after masking the top 20 token alignment heads, i is the token

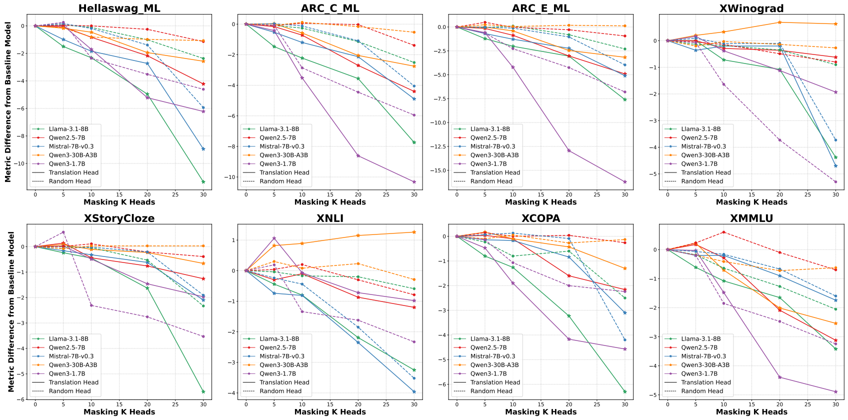


Figure 9: Performance (Accuracy) change across multilingual benchmarks when masking token alignment heads versus random heads. Tasks like Hellaswag_ML and ARC_ML show substantial drops when token alignment heads are ablated, while others such as XNLI and XCOPA are less affected, reflecting varying degrees of reliance on cross-lingual alignment.

index in x, and m is the total number of tokens. This score quantifies the performance degradation on a sample when token alignment heads are removed, with higher scores indicating greater reliance on these heads.We conduct experiments on the 1.5B model, training on a total of 1T tokens. The dataset comprises 700B tokens of English web data and 300B tokens of multilingual web data. Using the TRater algorithm, we score the 300B multilingual data, and we select the top 1.3% for each language. We design the following two experiments to validate the impact of the selected data:

Remove: From the baseline multilingual datas, we exclude the selected data. To maintain the data proportions unchanged, we increase the volume of the remaining data per language. And we ensure no additional duplicates compared to the baseline.

Enhance: The selected data is triplicated, while the remaining data is proportionally down-sampled to preserve the overall language distribution.

Table 1: Performance of baseline, remove, and enhance settings across multiple metrics.

Model	flores_chrF++	MMMLU	Hellaswag_ML	ARC_C_ML	ARC_E_ML	XWinograd	XStoryCloze	XNLI	XCOPA	XMMLU
baseline	43.87	**26.89**	44.88	31.40	53.97	**75.70**	58.40	41.94	62.88	30.71
remove	41.33	26.58	44.69	31.37	54.54	73.63	58.15	**42.10**	63.20	30.84
enhance	**46.68**	26.71	**44.95**	**31.54**	**54.94**	74.33	**58.44**	41.72	**63.70**	**30.88**

The experimental results are presented in Table 1. From the table, we observe that the data filtered by token alignment heads is crucial for the model's translation capabilities: the **Remove** setup exhibits a noticeable decline in translation performance compared to the baseline, whereas the **Enhance** setup shows a observable improvement.

However, the impact of this selected data on other multilingual benchmarks is less pronounced than its effect on FLORES. This can be attributed to two main factors. Firstly, the performance change observed in FLORES due to the selected data (about 2-3 points) is substantially smaller than the drops seen when ablating token alignment heads (over 10 points). Consequently, any impact on other multilingual benchmarks becomes less perceptible. Secondly, while the function of token alignment heads is partially leveraged by some non-translation tasks, these benchmarks only require the model to possess a foundational translation capability. Once this sufficient baseline is established, further increasing the proportion of translation-centric data yields diminishing returns for the model's general multilingual performance. Indeed, a qualitative analysis confirms the selected data is highly translation-specific, consisting predominantly of bilingual corpora. Several representative examples are detailed in Appendix A.3.

6 Related Work

A substantial body of work exists on understanding the internal mechanisms of large language models. Based on the granularity of analysis, we broadly categorize these studies into three main areas: semantic space, neuron-level mechanisms, and head-level mechanisms.

6.1 Semantic Space

Prior works (Wendler et al., 2024; Schut et al., 2025; Zhao et al., 2024; Wu et al., 2025b; Harrasse et al., 2025) study the geometry of multilingual representations and often concludes that models "think in English" or in a shared latent semantic space in middle layers. These works explain where multilingual information lives and how information exists (English or Language-Agnostic Space). Our results are complementary: we identify token alignment heads concentrated in similar middle layers and show that they implement token-level cross-lingual alignment, routing the aligned source token's representation into the target position.

6.2 Neuron-level mechanisms

Utilizing causal mediation analysis across a diverse range of in-context learning (ICL) tasks, Todd et al. (2024) identified a key mechanism termed "function vectors", which trigger the model to execute specific procedural tasks. Similarly, Wang et al. (2024) employed causal mediation procedures to locate attention heads pivotal for machine translation, leveraging these heads to construct translation vectors that mitigate language mismatch errors. In contrast, our work shifts focus from the task-triggering level to the token-execution level. We find that token alignment heads facilitate the actual cross-lingual alignment. Analogously, if the function vector acts as the "master switch" activating the translation mode, the token alignment heads are the vital machinery carrying out the translation itself.

From the perspective of languages, some studies (Liu et al., 2025; Zhao et al., 2024) pinpoint neurons that are specialized for encoding language identity and language-specific features and shows that ablating or fine-tuning them selectively affects particular languages. This explains which subcircuits are responsible for "being in language X". In contrast, we operate at the head level and focus on the cross-lingual alignment step, i.e., how information moves between languages during translation.

6.3 Head-level mechanisms

Works on induction heads, retrieval heads, and circuits (Elhage et al., 2021b; Olsson et al., 2022; Wu et al., 2025a; Bricken et al., 2024; Zhang et al., 2024) shows that a small number of specialized heads can explain non-trivial capabilities such as in-context learning or long-context retrieval. Recent work (Liu et al., 2025; Zhang et al., 2025) identifies language heads or translation-related heads by ranking heads via their impact on downstream loss, perplexity, or logits on specific benchmarks, sometimes using path patching. Our approach is closely related but uses a different identification signal: we define Token Aligenement Heads (TAH) using alignment-based translation score—heads are selected because they consistently link target tokens to their externally aligned source tokens, independent of any particular evaluation task. Masking experiments are then used only as a causal validation step. This makes our notion of specialization explicitly lexical and cross-lingual.

7 Conclusion

In this paper, we identified a special class of attention heads responsible for mapping source language tokens to target language tokens during translation. We experimentally confirmed that these heads are universal, consistent, and have a direct causal effect on the model's translation capabilities. We also uncovered their evolutionary process during pre-training, which involves rapid formation, stabilization, and pruning. More importantly, We found that a tiny fraction of critical data filterd by token alignment heads, proves decisive for translation performance but its impact on other multilingual tasks is less pronounced. This finding suggests that translation operates as a separable module within LLMs. Our work pave the way for more efficient and robust multilingual systems, enabling targeted architectural innovations, data curation strategies guided by mechanistic understanding.

References

Anthropic. Claude 4. Large language model, 2025. URL https://www.anthropic.com/news/claude-4.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.421. URL http://dx.doi.org/10.18653/v1/2020.acl-main.421.

Trenton Bricken, Adly Templeton, Alexander G. M. G. Anderson, Liane Lovitt, Jerry Chen, Kamal Ndousse, William Saunders, Christopher Olah, and Tom Henighan. Scaling monosemanticity: The sparse manifold hypothesis. *Transformer Circuits*, May 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html. Accessed: 2025-11-21.

Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, and Robert L Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2): 263–311, 1993.

Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. The geometry of multilingual language model representations, 2022. URL https://arxiv.org/abs/2205.10964.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations, 2018. URL https://arxiv.org/abs/1809.05053.

Arindrima Datta, Bhuvana Ramabhadran, Jesse Emond, Anjuli Kannan, and Brian Roark. Language-agnostic multilingual modeling. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8239–8243. IEEE, 2020.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021a. https://transformer-circuits.pub/2021/framework/index.html.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Roger Grosse, Zac Hatfield-Dodds, Danny Hernandez, Tristan Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Sam Ringer, Sam Johnston, Scott Thompson, Tom B. Brown, Jared Kaplan, Chris Olah, and Dario Amodei. A mathematical framework for transformer circuits. *Transformer Circuits*, Jul 2021b. URL https://transformer-circuits.pub/2021/framework/index.html. Accessed: 2025-11-21.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. Causal analysis of syntactic agreement mechanisms in neural language models, 2021. URL https://arxiv.org/abs/2106.06087.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation, 2021. URL https://arxiv.org/abs/2106.03193.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Ping Guo, Yubing Ren, Binbin Liu, Fengze Liu, Haobin Lin, Yifan Zhang, Bingni Zhang, Taifeng Wang, and Yin Zheng. Exploring polyglot harmony: On multilingual data allocation for large language models pretraining, 2025. URL https://arxiv.org/abs/2509.15556.

Wenhan Han, Yifan Zhang, Zhixun Chen, Binbin Liu, Haobin Lin, Bingni Zhang, Taifeng Wang, Mykola Pechenizkiy, Meng Fang, and Yin Zheng. Mubench: Assessment of multilingual capabilities of large language models across 61 languages, 2025. URL https://arxiv.org/abs/2506.19468.

Abir Harrasse, Florent Draye, Zhijing Jin, and Bernhard Schölkopf. Tracing multilingual representations in llms with cross-layer transcoders, 2025. URL https://arxiv.org/abs/2511.10840.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Dongsheng Jiang, Yuchen Liu, Songlin Liu, Jin'e Zhao, Hao Zhang, Zhen Gao, Xiaopeng Zhang, Jin Li, and Hongkai Xiong. From clip to dino: Visual encoders shout in multi-modal large language models. *arXiv preprint arXiv:2310.08825*, 2023.

Zae Myung Kim, Laurent Besacier, Vassilina Nikoulina, and Didier Schwab. Do multilingual neural machine translation models contain language pair specific attention heads? *arXiv preprint arXiv:2105.14940*, 2021.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. Large language models only pass primary school exams in indonesia: A comprehensive test on indommlu, 2023. URL https://arxiv.org/abs/2310.04928.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the dark secrets of bert, 2019. URL https://arxiv.org/abs/1908.08593.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. Cmmlu: Measuring massive multitask language understanding in chinese, 2024. URL https://arxiv.org/abs/2306.09212.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the biology of a large language model. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Xin Liu, Qiyang Song, Qihang Zhou, Haichao Du, Shaowen Xu, Wenbo Jiang, Weijuan Zhang, and Xiaoqi Jia. Focusing on language: Revealing and exploiting language attention heads in multilingual large language models, 2025. URL https://arxiv.org/abs/2511.07498.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Weicheng Ma, Kai Zhang, Renze Lou, Lili Wang, and Soroush Vosoughi. Contributions of transformer attention heads in multi-and cross-lingual tasks. *arXiv preprint arXiv:2108.08375*, 2021.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one?, 2019. URL https://arxiv.org/abs/1905.10650.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and evaluation framework for deeper understanding of commonsense stories, 2016. URL https://arxiv.org/abs/1604.01696.

Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003. doi: 10.1162/089120103321337421. URL https://aclanthology.org/J03-1002/.

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. In-context learning and induction heads, 2022. URL https://arxiv.org/abs/2209.11895.

OpenAI. Gpt-5. Large language model, 2025. URL https://openai.com/index/introducing-gpt-5/.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. Xcopa: A multilingual dataset for causal commonsense reasoning, 2020. URL https://arxiv.org/abs/2005.00333.

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. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

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

Alexey Tikhonov and Max Ryabinin. It's all in the heads: Using attention heads as a baseline for cross-lingual transfer in commonsense reasoning, 2021. URL https://arxiv.org/abs/2106.12066.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models, 2024. URL https://arxiv.org/abs/2310.15213.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model, 2019. URL https://arxiv.org/abs/1906.04284.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned, 2019. URL https://arxiv.org/abs/1905.09418.

Weichuan Wang, Zhaoyi Li, Defu Lian, Chen Ma, Linqi Song, and Ying Wei. Mitigating the language mismatch and repetition issues in llm-based machine translation via model editing, 2024. URL https://arxiv.org/abs/2410.07054.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in English? on the latent language of multilingual transformers. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15366–15394, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.820. URL https://aclanthology.org/2024.acl-long.820/.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=EytBpUGB1Z.

Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. The semantic hub hypothesis: Language models share semantic representations across languages and modalities, 2025b. URL https://arxiv.org/abs/2411.04986.

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, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger

Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL `https://arxiv.org/abs/2505.09388`.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence?, 2019. URL `https://arxiv.org/abs/1905.07830`.

Hongbin Zhang, Kehai Chen, Xuefeng Bai, Xiucheng Li, Yang Xiang, and Min Zhang. Exploring translation mechanism of large language models. *arXiv preprint arXiv:2502.11806*, 2025.

Ruochen Zhang, Qinan Yu, Matianyu Zang, Carsten Eickhoff, and Ellie Pavlick. The same but different: Structural similarities and differences in multilingual language modeling, 2024. URL `https://arxiv.org/abs/2410.09223`.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. How do large language models handle multilingualism? *Advances in Neural Information Processing Systems*, 37:15296–15319, 2024.

A Appendix

A.1 Training Setting

In Section 4, We use AdamW (Loshchilov & Hutter, 2019) optimizer to train a 8B dense model with a structure identical to Llama-2 model Touvron et al. (2023). The training dataset, totaling 15 trillion tokens, comprises English and multilingual data sourced from cleaned and filtered ccwarc, along with open-source resources including Wikipedia, books, academic papers, mathematics, code, and parallel corpora. The hyperparameters were set as follows: learning rate (lr) = 3.6×10^{-4}, global batch size (gbs) = 4096, sequence length = 4096, weight decay = 0.1. We employed a cosine learning rate scheduler that decayed to 0.1 of the peak learning rate. For our analysis, we selected model checkpoints at 2,000, 4,000, 6,000, 8,000, 12,000, 16,000, 64,000, 128,000, 256,000, 384,000, 448,000, and 952,000 steps (near the end of training).

In Section 5.3, we trained the 1.5B model using the AdamW optimizer with the following hyperparameters: global batch size (gbs) = 4096, sequence length = 4096, weight decay = 0.1, learning rate (lr) = 5.0×10^{-4}, and with lr cosine decay to 5.0×10^{-5}, the multilingual web data includes a total of 17 languages, specifically including German, Spanish, French, Indonesian, Thai, Korean, Vietnamese, Arabic, Turkish, Italian, Malay, Chinese, Portuguese, Japanese, Dutch, Russian and Filipino. The proportion of different languages is determined by the method in Guo et al. (2025).

A.2 Translated Benchmark

For evaluating Hellaswag_ML, ARC_C_ML and ARC_E_ML, we use the MuBench dataset (Han et al., 2025). The evaluation covers 18 languages represented in our training data, namely: English, German, Spanish, French, Indonesian, Thai, Korean, Vietnamese, Arabic, Turkish, Italian, Malay, Chinese, Portuguese, Japanese, Dutch, Russian and Filipino.

A.3 Text Cases for TRater

Table 2: Text cases filtered by TRater for German, French and Spanish.

Language	Example
German	Zwischen zwei Seen, die unterschiedlicher kaum sein können, liegt Wandlitz. Der Wandlitzsee, bebaut, kaum zugänglich mit unzähligen Wassergrundstücken, der Liebnitzsee, frei zugänglich, Badestelle, Fähre zur Insel und Naherholungsgebiet im Buchenwald... Wandlitz is located between two lakes that could hardly be more different. The Wandlitzsee, built-up, hardly accessible with countless water properties, the Liebnitzsee, freely accessible, bathing area, ferry to the island and local recreation area in the beech forest...
French	Bruno Houssin, designer français et professeur à l'école de design de Nantes. Diplômé de l'école Boulle de Paris, en Architecture intérieure et Design en 1986... Bruno Houssin, French designer and teacher at the Nantes School of Design. Graduated from the Boulle school of Paris, in Interior Architecture and Design in 1986...
Spanish	Las placas tectónicas son como grandes balsas que se reparten por toda la corteza del planeta. Unas son de carácter continental, otras de carácter oceánico, contando las primeras con un espesor mayor que el de las segundas... Le tremblement de terre en Haïti est partie de l'ensemble de la libération des tensions accumulées à l'occasion du mouvement des plaques tectoniques dans les Caraïbes et en Amérique du Nord...

Table 3: Text cases filtered by TRater for Italian, Portuguese, Chinese and Dutch.

Language	Example
Italian	Dorothy Bhawl è un artista autodidatta interessato al mondo contemporaneo, soprattutto alla realtà che appartiene e avvolge questa epoca: quello della comunicazione, social network, spiritualità e grottesco con un sentimento di odi et amo... Dorothy Bhawl is a self-taught artist interested in the contemporary world, especially in the reality that belongs and envelops this era: that of communication, social networks, spirituality and grotesque with a feeling of hatred and love...
Portuguese	André Rigatti Centro Universitário Maria Antonia USP Sempre próximas a suas bordas, as pinturas de André Rigatti possuem pequenas aberturas, por onde se deixa ver o processo que dá origem a trabalhos de textura matérica mais ou menos acentuada, resultados da sobreposição de diversas camadas de tinta, aplicadas cada uma seguindo uma direção diferente do pincel... Always close to its edges, André Rigatti's paintings have small openings, where you can see the process that gives rise to more or less accentuated texture work, results of the overlapping of several layers of paint, applied each following a direction...
Chinese	原文: 版印书籍，唐人尚未盛为之。自冯瀛王始印五经，已后典记，皆为版本。庆历中，有布衣毕升，又为活版。其法用胶泥刻字，薄如钱唇，每字为一印，火烧令坚。先设一铁板，其上以松脂腊和纸灰之类冒之。欲印则以一铁范置铁板上，乃密布字印。满铁范为一板，持就火炀之，药稍熔，则以一平板按其面，则字平如砥... 译文： 用刻板印刷书籍，唐朝人还没有大规模采用它。五代始才开始印刷五经，以后的各种图书都是雕板印刷本。庆历年间，有位平民毕升，又创造了活板。他的方法是用胶泥刻成字，字薄得像铜钱的边缘，每个字制成一个字模，用火来烧使它坚硬。先设置一块铁板，它的上面用松纸、蜡混合纸灰这一灰东西覆盖它...
Dutch	Kees Blom (Apeldoorn, 1968) komt uit een artistieke familie. Zijn vader had al een passie voor de schilderkunst maar pas zoon Kees lukt het om de stap naar zelfstandig kunstschilder te zetten...Kees Blom (Apeldoorn, 1968) comes from an artistic family. While his father already had a passion for painting, son Kees succeeds in taking the step to become an independent painter...

A.4 Case Study For Token Alignment Heads Pruning

To clarify the role of Token Alignment Heads (TAHs), we conducted a systematic and detailed analysis of the model's performance on translation tasks when TAHs are masked (specifically focusing on cases where the model translated correctly before masking TAHs but failed to do so afterwards). We observed that the failure modes resulting from masking THs can be broadly categorized into three types:

1. No Translation (46%) The model fails to generate the target language and merely repeats the source text content. As illustrated in Figure 10, masking the Token Alignment Heads completely disables the model's word alignment capability for certain queries, preventing the generation of target language output.

2. Missing Details (36%) The translated output lacks specific details found in the source text. In these instances, the model's word alignment capability is partially compromised. While the model can still generate text in the target language, it fails to correctly map certain source information to the target output. This results in the omission of key details, leading to incomplete or imprecise

Translate this sentence from English to Simplified Chinese
Sentence: In late 2017, Siminoff appeared on shopping television channel QVC.
[**No Heads Pruned**] Answer: 2017年末，辛明诺夫出现在购物电视频道QVC上。
[**30 Token Alignment Heads Pruned**] Answer: In late 2017, Siminoff appeared on shopping television channel QVC.
[**30 Random Heads Pruned**] Answer: 在2017年晚些时候，Siminoff出现在购物电视频道QVC上。

Sentence: One antibody cocktail, ZMapp, initially showed promise in the field, but formal studies indicated it had less benefit than sought in preventing death.
[**No Heads Pruned**] Answer: 一种抗体混合物，ZMapp，最初在野外显示出有希望，但正式研究表明它比预期的更少有益于预防死亡。
[**30 Token Alignment Heads Pruned**] Answer: One antibody cocktail, ZMapp, showed promise in the field, but formal studies indicated it had less benefit than sought in preventing death.
[**30 Random Heads Pruned**] Answer: 一种抗体混合物，ZMapp，最初在野外表现出希望，但正式研究表明，它比预期预防死亡的作用要小。

Sentence: Scientists say this animal's plumage was chestnut-brown on top with a pale or carotenoid-colored underside.
[**No Heads Pruned**] Answer: 科学家说，这种动物的羽毛上面是棕色，下面是淡黄色或是橙色。
[**30 Token Alignment Heads Pruned**] Answer: Scientists say this animal's plumage was chestnut-brown on top with a pale or carotenoid-colored underside.
[**30 Random Heads Pruned**] Answer: 科学家们说，这种动物的羽毛是棕色顶部和淡色或橘色底部。

Figure 10: No Translation cases for masking token alignment heads. For the queries shown in the figure, after masking the token alignment heads, the model no longer outputs content related to the target language, but simply repeats the source text.

translations. Figure 11 presents three such cases, where the text highlighted in green represents the details omitted by the model.

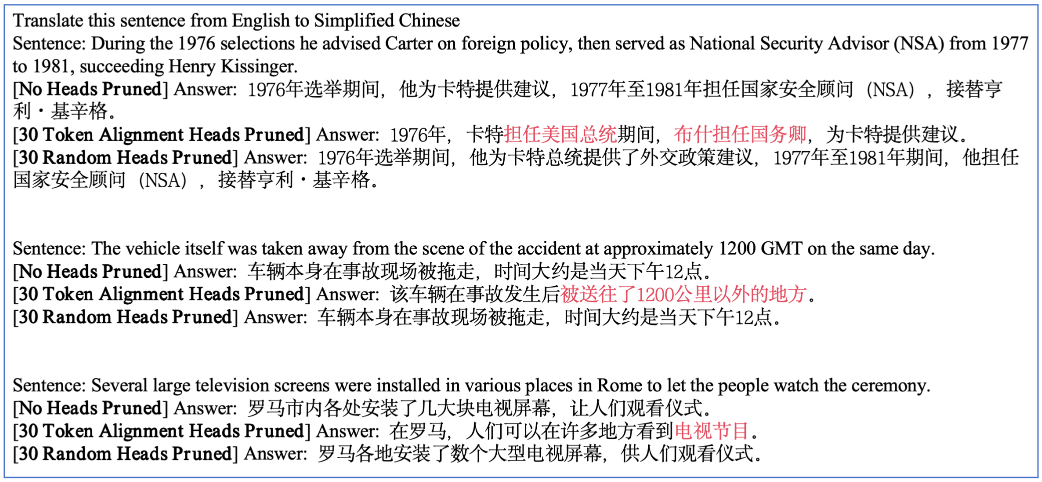
3. Hallucination (18%) The translated output contains content that is completely absent from the source text. This is likely because, without token alignment capabilities, the model cannot directly associate source tokens with target tokens. Consequently, it resorts to generating content in the target language space that is only loosely or tangentially related to the source tokens. Figure 12 displays three examples of this category, where the text highlighted in red indicates output generated by the model that is entirely unrelated to the source text.

Translate this sentence from English to Simplified Chinese
Sentence: Like some other experts, he is skeptical about whether diabetes can be cured, noting that these findings have no relevance to people who already have Type 1 diabetes.
[**No Heads Pruned**] Answer: 他和其他专家一样，对糖尿病能否治愈持怀疑态度，指出这些发现对已经患有1型糖尿病的人没有任何意义。
[**30 Token Alignment Heads Pruned**] Answer: 他认为这些发现对那些已经有了糖尿病的人没有什么帮助。
[**30 Random Heads Pruned**] Answer: 和其他专家一样，他对糖尿病能否治愈持怀疑态度，指出这些发现对已经患有1型糖尿病的人没有任何意义。

Sentence: This not only confirms that at least some dinosaurs had feathers, a theory already widespread, but provides details fossils generally cannot, such as color and three-dimensional arrangement.
[**No Heads Pruned**] Answer: 这不仅证实了至少有一些恐龙有羽毛，一个已经广泛流传的理论，而且提供了化石通常不能提供的细节，如颜色和三维排列。
[**30 Token Alignment Heads Pruned**] Answer: 这不仅证实了至少有一种恐龙有羽毛，而且还证实了这种羽毛的颜色和羽毛的三维结构。
[**30 Random Heads Pruned**] Answer: 这不仅证实了至少有些恐龙有羽毛——这一理论已经广为流传——而且还提供了化石通常无法提供的细节，比如颜色和三维排列。

Sentence: During his trip, Iwasaki ran into trouble on many occasions.
[**No Heads Pruned**] Answer: 在他旅行期间，伊瓦斯基遇到了很多麻烦。
[**30 Token Alignment Heads Pruned**] Answer: 在他旅行的过程中，他遇到了很多麻烦。
[**30 Random Heads Pruned**] Answer: 在旅行期间，伊瓦斯基遇到了很多麻烦。

Figure 11: Cases of Missing Details for masking token alignment heads. For the queries shown in the figure, after masking the token alignment heads, the model is still able to generate target language content. However, because some cross-lingual token alignment capabilities are missing, the model loses certain details present in the source text.



Translate this sentence from English to Simplified Chinese
Sentence: During the 1976 selections he advised Carter on foreign policy, then served as National Security Advisor (NSA) from 1977 to 1981, succeeding Henry Kissinger.
[**No Heads Pruned**] Answer: 1976年选举期间，他为卡特提供建议，1977年至1981年担任国家安全顾问（NSA），接替亨利・基辛格。
[**30 Token Alignment Heads Pruned**] Answer: 1976年，卡特担任美国总统期间，布什担任国务卿，为卡特提供建议。
[**30 Random Heads Pruned**] Answer: 1976年选举期间，他为卡特总统提供了外交政策建议，1977年至1981年期间，他担任国家安全顾问（NSA），接替亨利・基辛格。

Sentence: The vehicle itself was taken away from the scene of the accident at approximately 1200 GMT on the same day.
[**No Heads Pruned**] Answer: 车辆本身在事故现场被拖走，时间大约是当天下午12点。
[**30 Token Alignment Heads Pruned**] Answer: 该车辆在事故发生后被送往了1200公里以外的地方。
[**30 Random Heads Pruned**] Answer: 车辆本身在事故现场被拖走，时间大约是当天下午12点。

Sentence: Several large television screens were installed in various places in Rome to let the people watch the ceremony.
[**No Heads Pruned**] Answer: 罗马市内各处安装了几大块电视屏幕，让人们观看仪式。
[**30 Token Alignment Heads Pruned**] Answer: 在罗马，人们可以在许多地方看到电视节目。
[**30 Random Heads Pruned**] Answer: 罗马各地安装了数个大型电视屏幕，供人们观看仪式。

Figure 12: Hallucination cases for masking token alignment heads. For the queries shown in the figure, after masking the token alignment heads, the model exhibits hallucinations, specifically generating output that contains information not present in the source text.

A.5 Basic Properties of Token Alignment Heads Across Comprehensive Model Families

To provide a more thorough and comprehensive evaluation of Token Alignment Heads, we conducted experiments across four well-established multilingual LLM families: Llama3, Qwen3, Mistral, and Gemma2. Our experimental design ensures broad coverage by including models of varying scales within each family: small (1B/2B parameters), medium (8B/9B parameters), and large (over 13B parameters). Where applicable, we evaluate the instruction-tuned (Instruct) variants. Additionally, for families that feature Mixture-of-Experts (MoE) architectures, our study includes a specific analysis of these models.

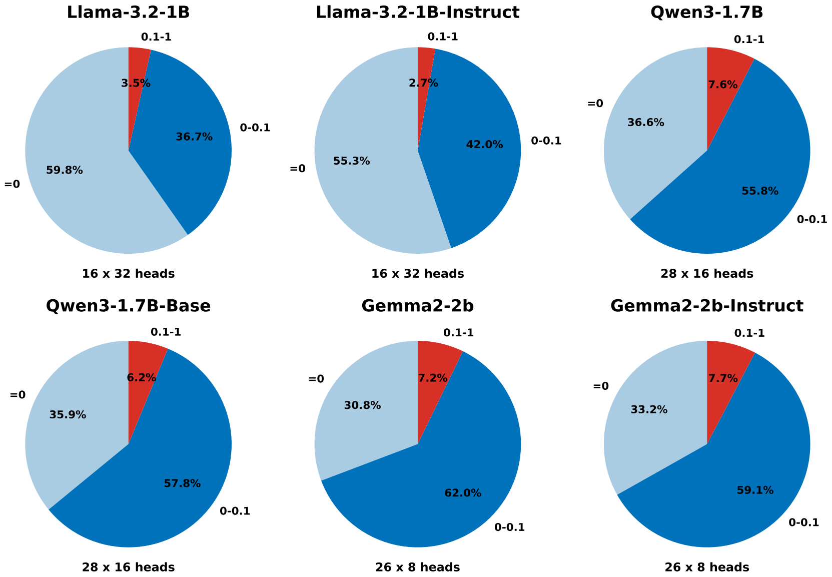


Figure 13: Translation score distribution for small size group models.

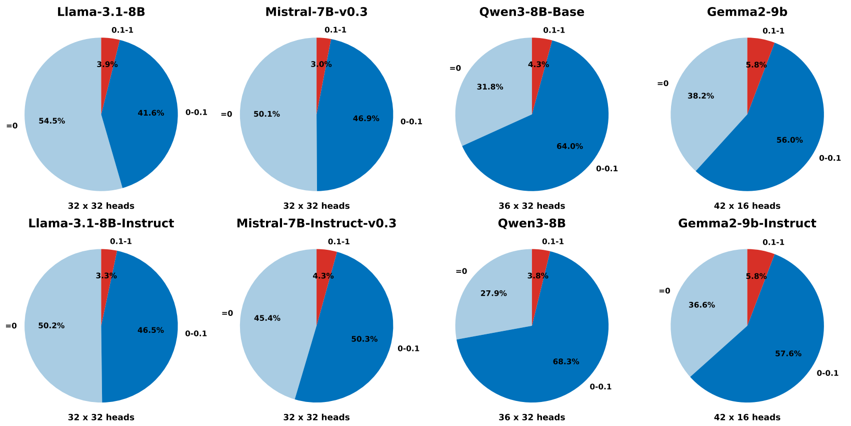


Figure 14: Translation score distribution for medium size group models.

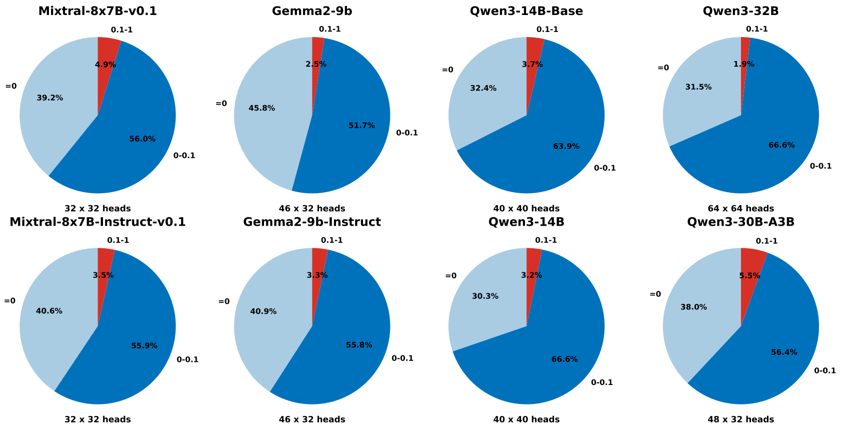


Figure 15: Translation score distribution for large size group models.

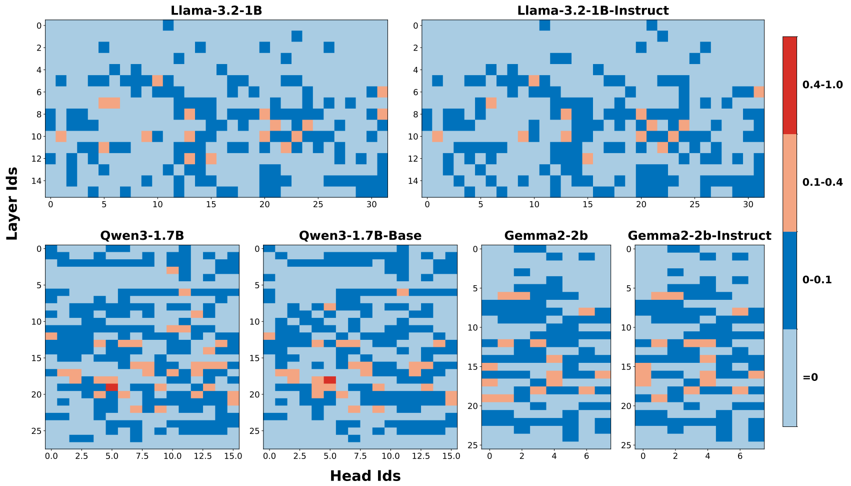


Figure 16: Positional distribution of translation scores in small size group models.

For clarity in presenting our experimental findings, we have organized the models into three distinct groups according to their scale: small, medium, and large. (The Mixtral-8x7B model is classified within the medium-scale group, as its quantity of attention heads is analogous to that of other models in this tier.) Our initial analysis focuses on demonstrating the basic characteristics of Token Alignment Heads. As illustrated in Figure 13, Figure 14 and Figure 15, Token Alignment Heads are a pervasive phenomenon, consistently identified across all models under investigation—irrespective of model scale, architecture (dense versus MoE), or training paradigm (Base versus Instruct). Crucially, these heads universally demonstrate the property of sparsity. Additionally, Figure 16. Figure 17 and Figure 18 illustrates the positional distribution of token alignment heads across these models. The general distribution pattern observed aligns with the description provided in the main body of the paper.

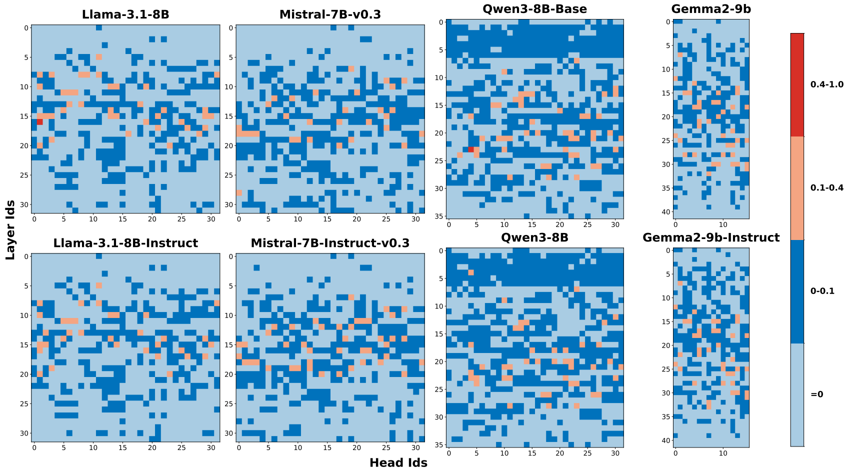


Figure 17: Positional distribution of translation scores in medium size group models.

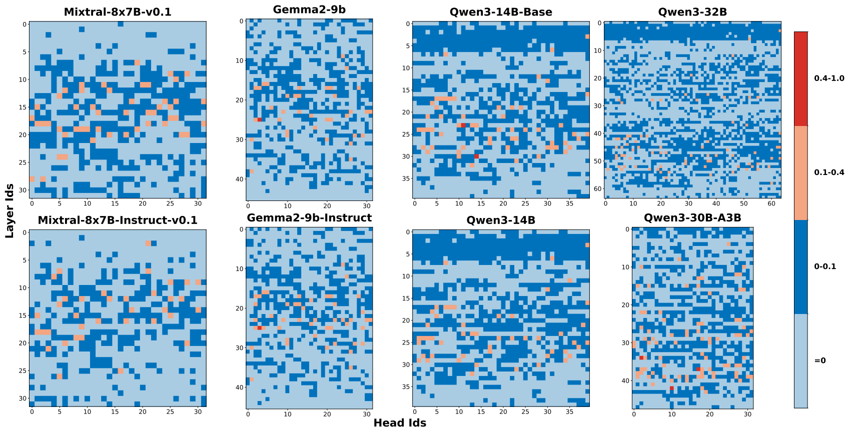


Figure 18: Positional distribution of translation scores in large size group models.

A.6 DOWNSTREAM TASKS INFLUENCE ACROSS COMPREHENSIVE MODEL FAMILIES

In this section, we present the impact of token alignment heads on downstream task performance, analyzed on a group-by-group basis. To better evaluate translation performance, we supplement the BLEU and chrF++ metrics from the main text with two additional metrics: BLEURT and COMET.

As can be seen, the experimental conclusions are consistent with those in the main text. As illustrated in Figure 19, Figure 20 and Figure 21, for BLEURT and COMET, masking token alignment heads leads to a significant drop in scores, whereas masking random heads results in only a minimal decrease. This demonstrates the causal role of token alignment heads in translation capability. For Hellaswag_ML, ARC_C_ML, and ARC_E_ML, masking token alignment heads causes a larger performance drop than masking random heads, but the overall magnitude of the decrease is far less pronounced than that for the translation metrics. This suggests that these metrics rely to some extent on token alignment capabilities. In contrast, for metrics like XWinograd and XNLI, performance after masking token alignment heads can be better than after masking random heads, indicating that these metrics prioritize other model abilities, such as reasoning, over translation capability.

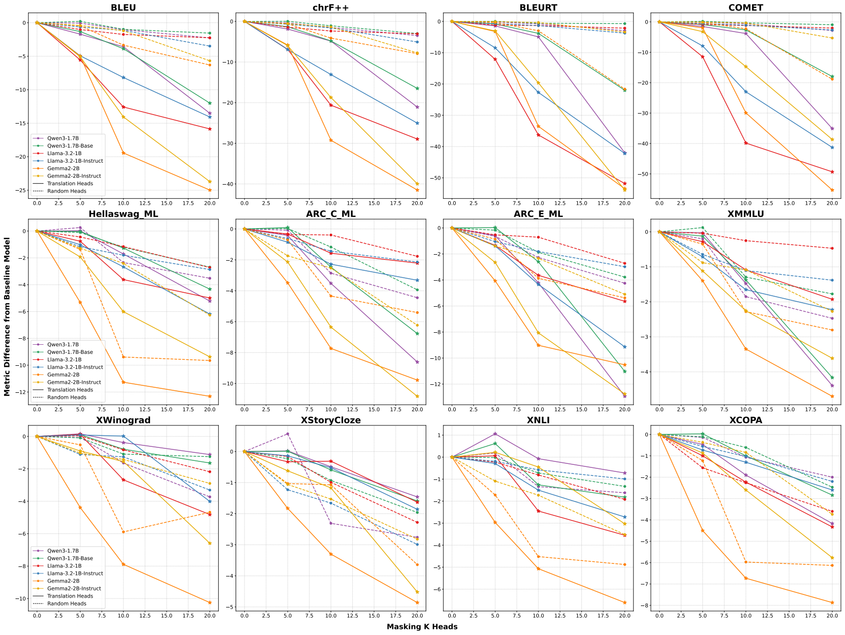


Figure 19: Performance change across downstream benchmarks for small size group models when masking token alignment heads versus random heads.

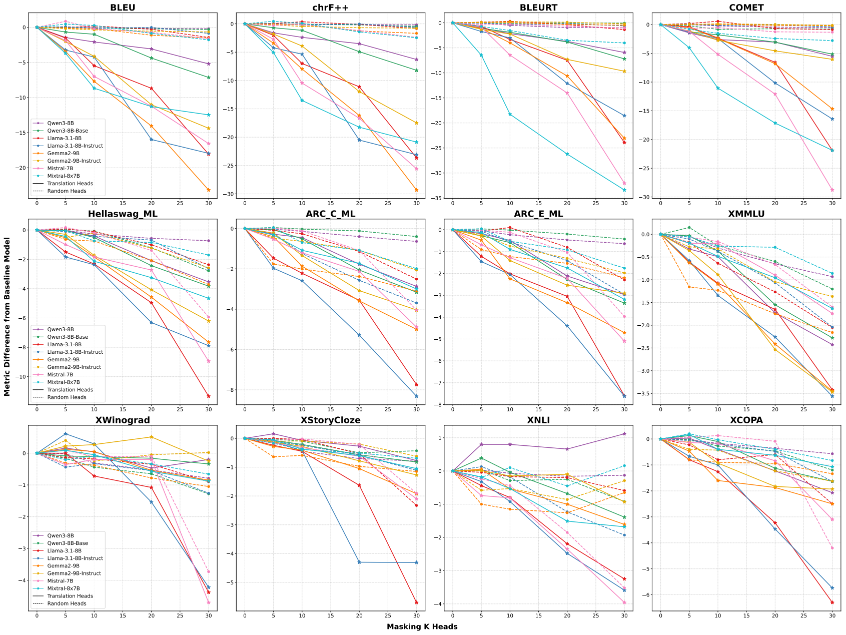


Figure 20: Performance change across downstream benchmarks for medium size group models when masking token alignment heads versus random heads.

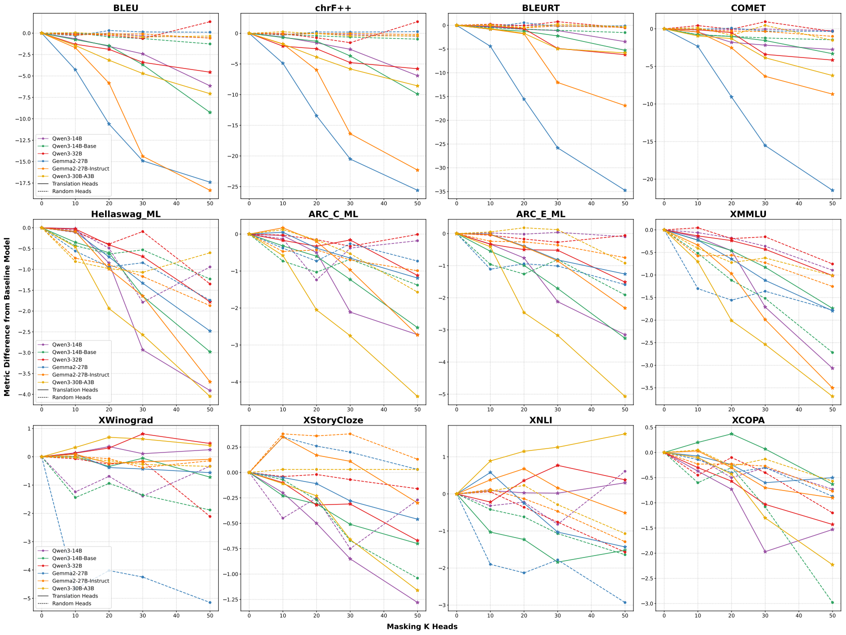


Figure 21: Performance change across downstream benchmarks for large size group models when masking token alignment heads versus random heads.

A.7 The Role of Token Alignment Heads on Multilingual Tasks

In this section, we showcase the utility of Token Alignment Heads on multilingual tasks through a series of case studies. To elucidate their performance characteristics under varying conditions, we analyze two distinct scenarios: the Hellaswag_ML task, which is moderately dependent on Token Alignment Heads, and the XNLI/XWinograd tasks, where the dependency is substantially weaker.

The Role of Token Alignment Heads in Hellaswag_FR

We need to choose the best ending for the context text.
context_text: Ceci est un **tutoriel** sur comment allumer un feu de camp. Il montre le **feu de camp** qui **brûle** sur le sol. il .
endings:
A: zeigt den Schornstein des Hauses.
B: zeigt, wie man ein paar drumherum legen kann, um etwas Licht zu haben.
C: beginnt mit den Dingen, die man für den Anfang braucht.
D: zeigt, wie man das Feuer hält und den Topf ein wenig darauf stellt, um das Propangas aufzufangen.

Key Word For Task: **tutoriel , feu de camp, brûle**

Qwen3-30B-A3B Attention Heads Translation Score: L40-H14 = 0.26   L16-H28 = 0.22   L26-H2 = 0.126   L28-H14 = 0.124

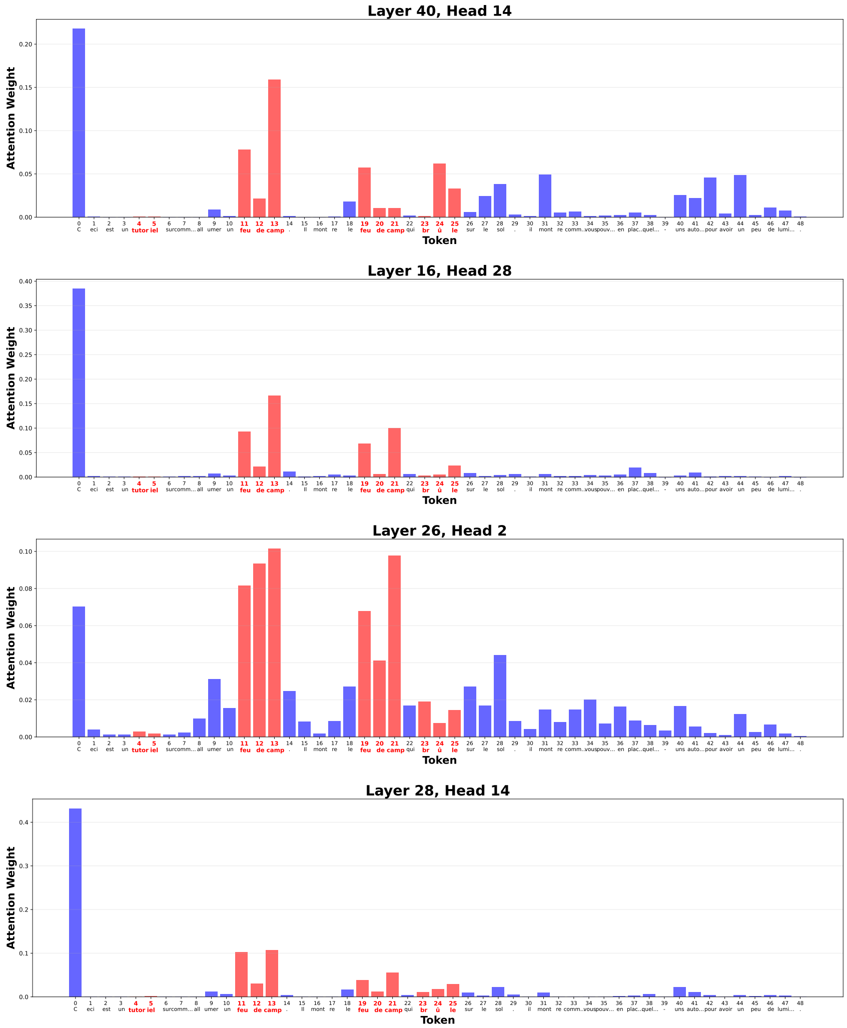


Figure 22: French Hellaswag case. The notation L40-H14 indicates Layer 40, Head 14, where ”L” stands for Layer and ”H” stands for Head. In this French Hellaswag case, L40-H14, L16-H28, L26-H2 and L28-H14 have translation scores greater than 0.1, identifying them as token alignment heads. We observe that all token alignment heads here can attend to the key tokens.

The Role of Token Alignment Heads in Hellaswag_DE

We need to choose the best ending for the context text.
context_text: Dies ist eine Anleitung, wie man ein Lagerfeuer entfacht. Sie zeigt das brennende Lagerfeuer auf dem Boden. sie.
endings:
A: zeigt den Schornstein des Hauses.
B: zeigt, wie man ein paar drumherum legen kann, um etwas Licht zu haben.
C: beginnt mit den Dingen, die man für den Anfang braucht.
D: zeigt, wie man das Feuer hält und den Topf ein wenig darauf stellt, um das Propangas aufzufangen.

Key Word For Task: **tutoriel , feu de camp, brûle**

Qwen3-30B-A3B Attention Heads Translation Score: L30-H4 = 0.175 L30-H11 = 0.127 L40-H14 = 0.262 L26-H2 = 0.126

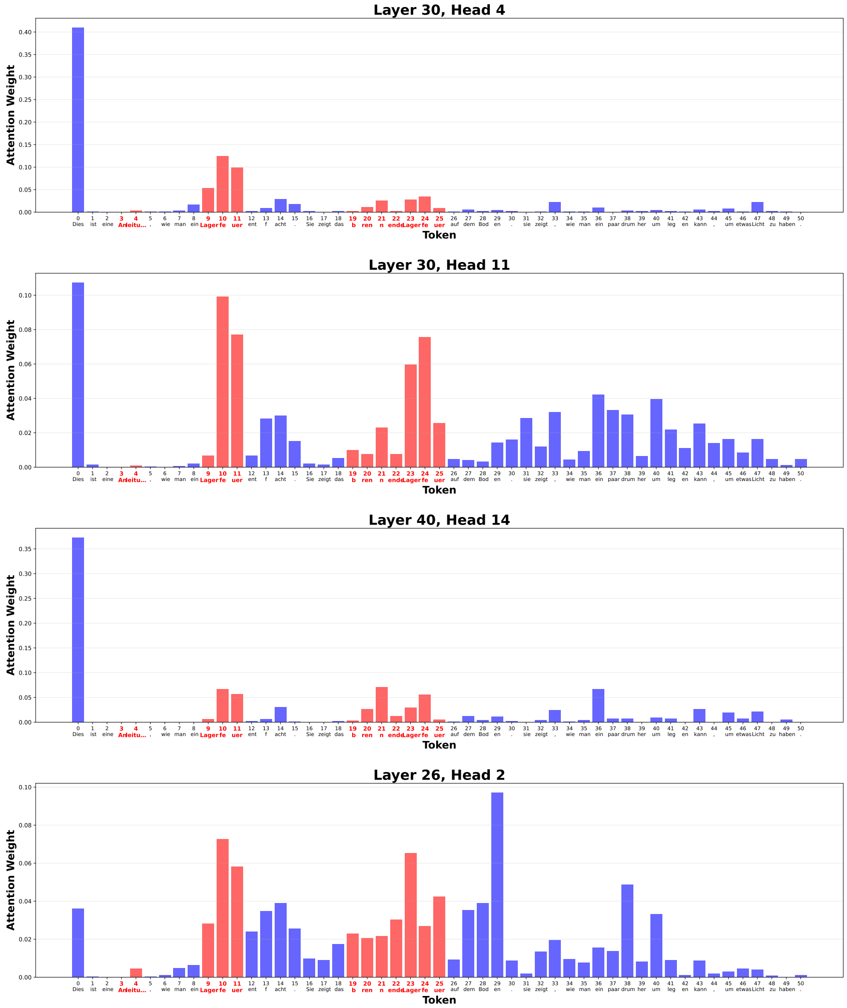


Figure 23: German Hellaswag case. The notation L30-H4 indicates Layer 30, Head 4, where ”L” stands for Layer and ”H” stands for Head. In this German Hellaswag case, L30-H4, L30-H1, L40-H14 and L26-H2 have translation scores greater than 0.1, identifying them as token alignment heads. We observe that all token alignment heads here can attend to the key tokens.

The Role of Token Alignment Heads in XNLI_FR

We need to decide whether A and B are entailment, contradiction or neutral.
A: Premièrement , nous avons **appliqué** trois fonctions alternatives (c-R) pour estimer l' incidence prématurée de la mortalité .
B: **Aucune** fonction cr n' a été **appliquée** à l' incidence .

Key Word For Task: **appliqué, Aucune , appliquée**

Qwen3-30B-A3B Attention Heads Translation Score: L15-H14 = 0.155 L10-H29 = 0.127 L24-H23 = 0 L29-H5 = 0

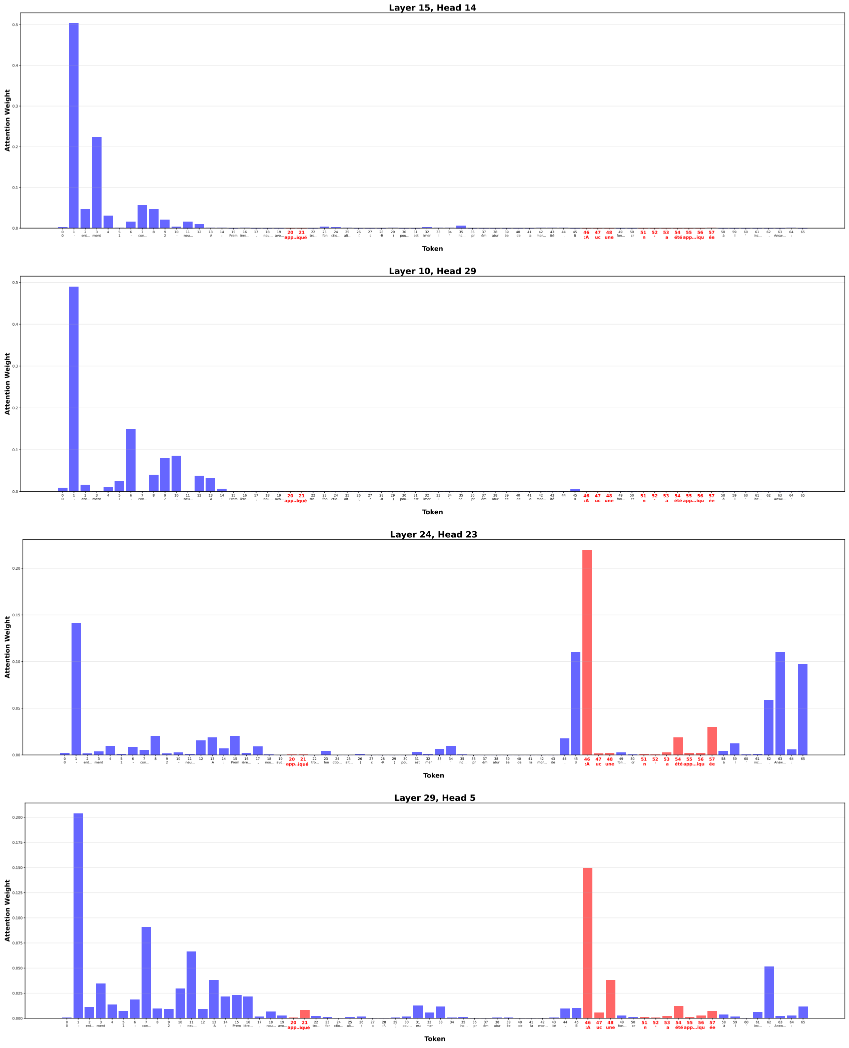


Figure 24: French XNLI case. The notation L15-H14 indicates Layer 15, Head 14, where ”L” stands for Layer and ”H” stands for Head. In this French XNLI case, Both L15-H14 and L10-H29 have translation scores greater than 0.1, identifying them as token alignment heads, whereas L24-H23 and L29-H5 are not. We observe that the token alignment heads tend to have attention weights close to zero on key tokens, as seen with L15-H14 and L10-H29 in the figure. In contrast, heads with relatively high attention weights on key tokens, such as L24-H23 and L29-H5, have very low translation scores.

The Role of Token Alignment Heads in XNLI_DE

We need to decide whether A and B are entailment, contradiction or neutral.
A: Sieh mal , das ist nicht so **schlimm** ein paar **Stunden** .
B: Ich **mag es nicht** , zu warten , nicht mal für 10 **Minuten** .

Key Word For Task: **schlimm, Stunden, mag es nicht, Minuten**

Qwen3-30B-A3B Attention Heads Translation Score: L13-H17 = 0.115   L10-H29 = 0.127   L0-H29 = 0.01   L1-H13 = 0.005

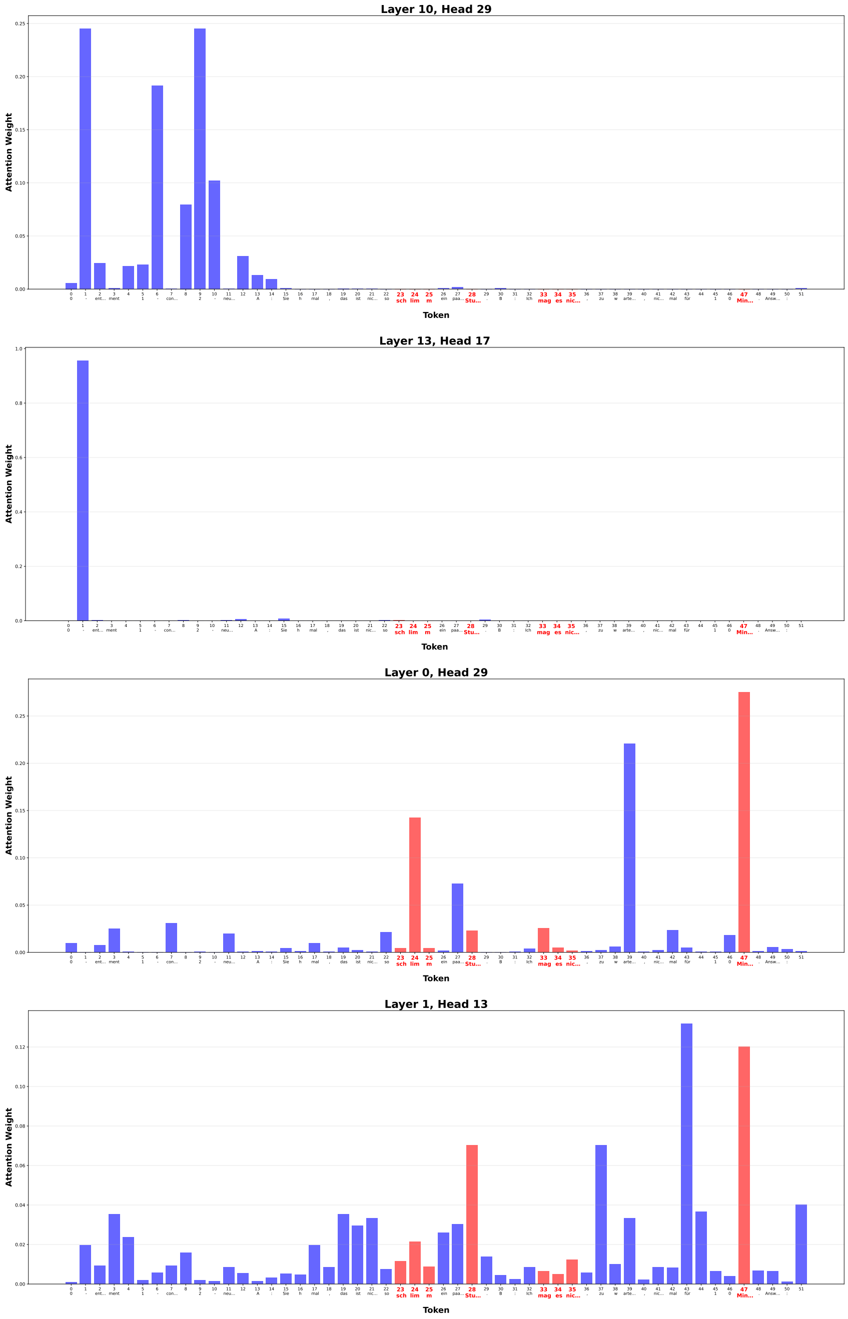


Figure 25: German XNLI case. The notation L13-H17 indicates Layer 13, Head 17, where ”L” stands for Layer and ”H” stands for Head. In this German XNLI case, Both L13-H17 and L10-H29 have translation scores greater than 0.1, identifying them as token alignment heads, whereas L0-H29 and L1-H13 are not. We observe that the token alignment heads tend to have attention weights close to zero on key tokens, as seen with L13-H17 and L10-H29 in the figure. In contrast, heads with relatively high attention weights on key tokens, such as L0-H29 and L1-H13, have very low translation scores.

The Role of Token Alignment Heads in XWinograd_RU

Based on the context, we need to fill in the blank with the correct subject.
context: Кубок не помещается в коричневый чемодан, потому что _ слишком маленький.
1. Кубок.
2. чемодан.

Key Word For Task: чемодан
Qwen3-30B-A3B Attention Heads Translation Score: L27-H17 = 0.244 L22-H2 = 0.187 L16-H18 = 0 L16-H21 = 0

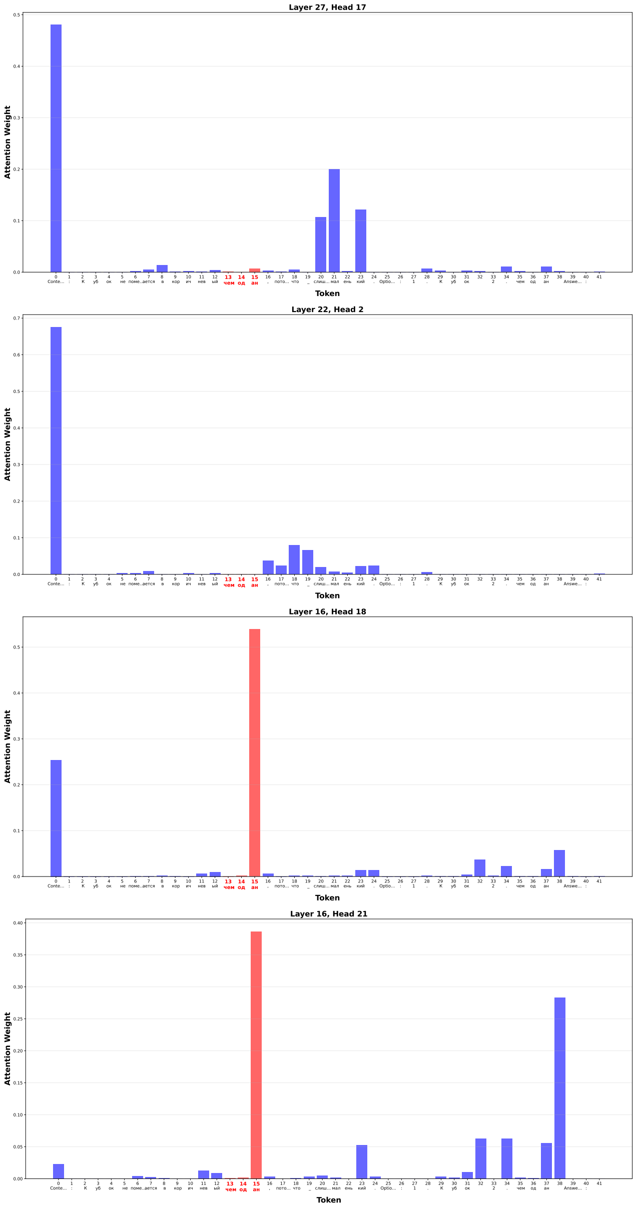


Figure 26: Russian XWinograd case. The notation L27-H17 indicates Layer 27, Head 17, where ”L” stands for Layer and ”H” stands for Head. In this Russian XWinograd case, Both L27-H17 and L22-H2 have translation scores greater than 0.1, identifying them as token alignment heads, whereas L16-H18 and L16-H21 are not. We observe that the token alignment heads tend to have attention weights close to zero on key tokens, as seen with L27-H17 and L22-H2 in the figure. In contrast, heads with relatively high attention weights on key tokens, such as L16-H18 and L16-H21, have very low translation scores.

The Role of Token Alignment Heads in XWinograd_FR

Based on the context, we need to fill in the blank with the correct subject.
context: Claire a frappé à la porte de **Sylvie**, mais _ n'a pas répondu.
1. Claire.
2. Sylvie.

Key Word For Task: **Sylvie**
Qwen3-30B-A3B Attention Heads Translation Score: L13-H18 = 0.24 L22-H2 = 0.187 L45-H29 = 0

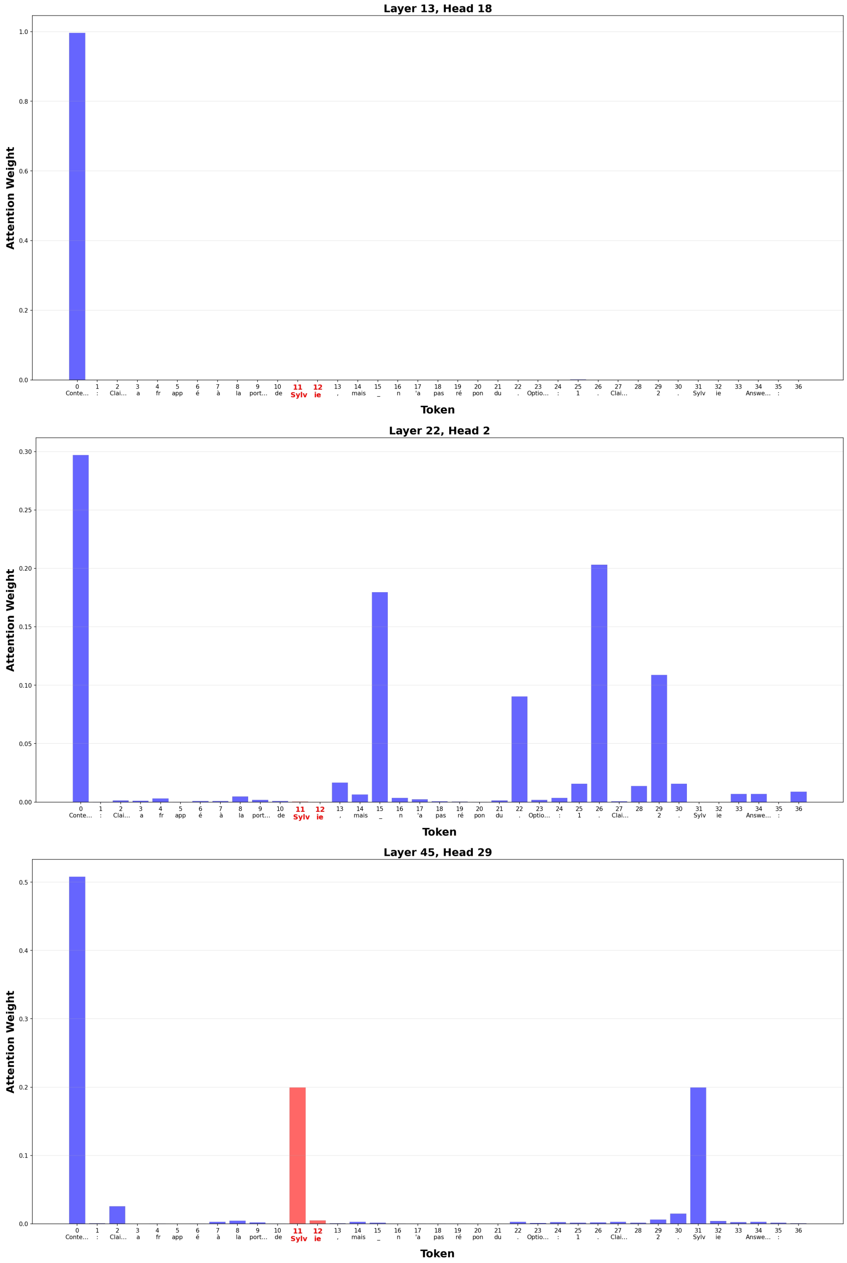


Figure 27: French XWinograd case. The notation L13-H18 indicates Layer 13, Head 18, where "L" stands for Layer and "H" stands for Head. In this French XWinograd case, Both L13-H18 and L22-H2 have translation scores greater than 0.1, identifying them as token alignment heads, whereas L45-H29 is not. We observe that the token alignment heads tend to have attention weights close to zero on key tokens, as seen with L13-H18 and L22-H2 in the figure. In contrast, heads with relatively high attention weights on key tokens, such as L40-H14, have very low translation scores.

A.8 Language Consistency Across 20 Language Pairs

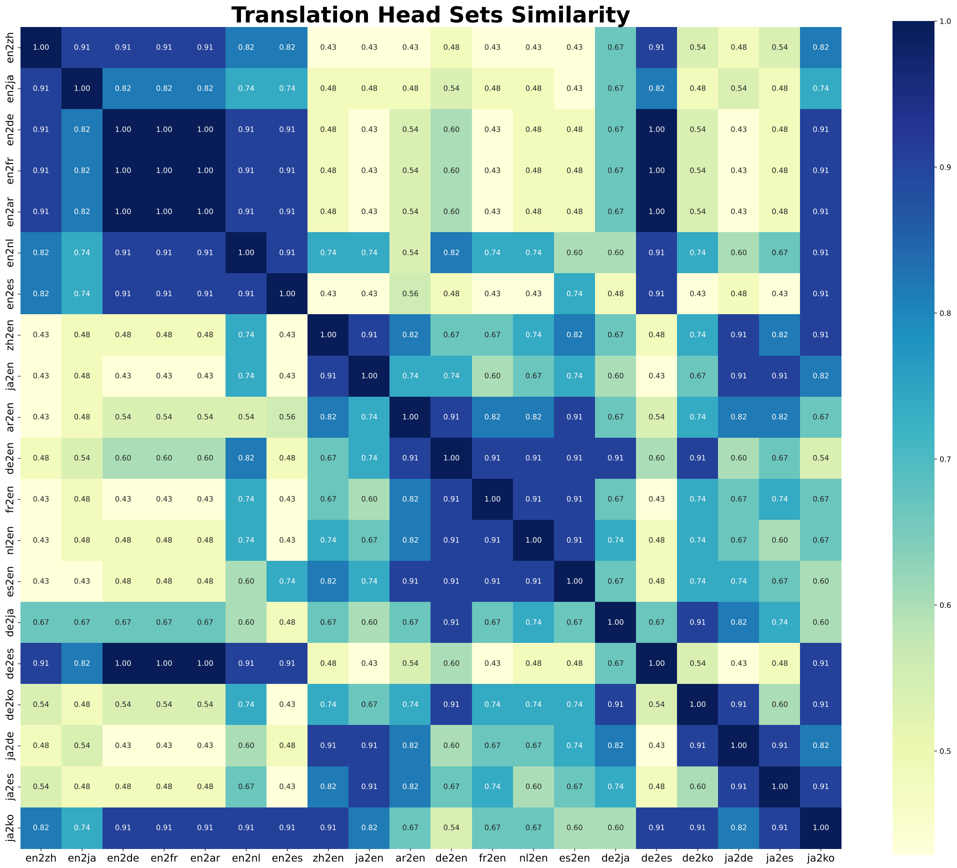


Figure 28: Jaccard similarity matrix of token alignment head sets across 20 language pairs. Here, we broadly categorize the selected language pairs into four groups: en2others (en2zh, en2ja, en2de, en2fr, en2ar, en2nl, en2es), others2en (zh2en, ja2en, ar2en, de2en, fr2en, nl2en, es2en), de2ja/es/ko, and ja2de/es/ko. It is observable that the similarity among the en2others pairs is high, whereas the others2en group and the other language pairs generally exhibit lower similarity.