Understanding LLMs' Cross-Lingual Context Retrieval: How Good It Is And Where It Comes From

Anonymous ACL submission

Abstract

The ability of cross-lingual context retrieval is a fundamental aspect of cross-lingual alignment of large language models (LLMs), where the model extracts context information in one language based on requests in another language. Despite its importance in real-life applications, this ability has not been adequately investigated for state-of-the-art models. In this paper, we evaluate the cross-lingual context retrieval ability of over 40 LLMs across 12 languages to understand the source of this ability, using cross-lingual machine reading comprehension (xMRC) as a representative scenario. Our results show that several small, post-trained open LLMs show strong cross-lingual context retrieval ability, comparable to closed-source LLMs such as GPT-4o. Our interpretability analysis shows that the cross-lingual context retrieval process can be divided into two main phases: question encoding and answer retrieval, which are formed in pre-training and post-training, respectively. Furthermore, the bottleneck of cross-lingual context retrieval lies at the last transformer layers in the second phase, where the effect of post-training can be evidently observed. Our results also indicate that larger LLMs need further multilingual post-training to fully unlock their cross-lingual context retrieval potential. [1]

1 Introduction

Since the rise of Large language models (LLMs), many models have demonstrated their strong capability in various NLP tasks (Chang et al., 2024), e.g. ChatGPT[2], Claude[3], Gemini (Gemini Team et al., 2024), LLaMA (Grattafiori et al., 2024), Qwen (Qwen et al., 2025), DeepSeek (DeepSeek-AI et al., 2024), etc. However, due to the dominance of English training data, most of these LLMs show their best performance in English (Lai et al., 2023b; Wang et al., 2023). To improve their performance and efficiency in non-English languages, cross-lingual alignment has become a major research topic for multilingual LLMs (Qi et al., 2023; Gao et al., 2024), which encourages LLMs to share capabilities across languages. For example, given requests with the same semantics but in different languages, LLMs should give consistent answers.

[1]Our code and data will be publicly available after the anonymous period.

[2]https://chatgpt.com/

[3]https://claude.ai/

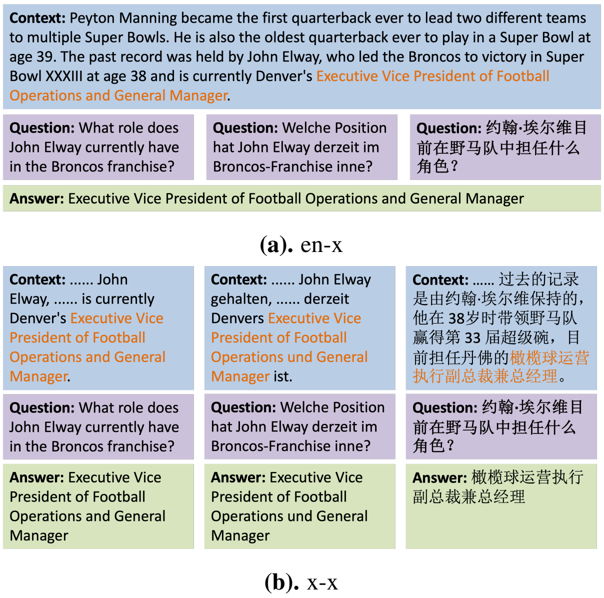


Figure 1: Examples of our en-x and x-x testing scenarios. The figures show examples in English (en), German (de), and Chinese (zh).

One fundamental aspect of such cross-lingual alignment is cross-lingual context retrieval, where the model needs to extract context information in the source language, e.g. English, to answer requests in the target language. However, there is yet no comprehensive evaluation and exploration of this ability.

In this paper, we evaluate the cross-lingual context retrieval ability of SOTA multilingual LLMs, and analyze the source and bottleneck of such capability. We use cross-lingual machine reading comprehension (xMRC) (Cui et al., 2019) as a simplified but representative scenario, where the target knowledge to be retrieved is literally included in the context, and the model only has to copy part of the context as the answer. This ablates the need for knowledge storage and multilingual text generation, thus better focusing the research on cross-lingual context retrieval. Our main findings are:

- Several SOTA post-trained open-source LLMs, especially in their 7-9B versions show strong xMRC ability, catching up with closed-source models.
- Post-training significantly improves the estimated oracle performance to a very high level in all tested languages, setting space for improvement of real performance.
- The xMRC process can be divided into two phases within the model: question encoding and answer retrieval.
- The question encoding phase forms in pre-training, and its stability correlates with the base model capability; the answer retrieval phase forms in post-training, and serves as a bottleneck for xMRC.

2 Methodology

2.1 Evaluation methods

2.1.1 Scenarios

We use English to non-English (en-x) as a typical and common cross-lingual scenario, where the context and answer are in English, and the question is in non-English. Figure 1a shows a testing sample of en-x task across 3 example languages.

Meanwhile, we use non-English monolingual (x-x) as a comparative to the en-x scenario, in order to distinguish xMRC ability with monolingual MRC ability. Figure 1b shows a testing sample of x-x task in the same 3 example languages.

2.1.2 Metrics

Performance metrics. We use the F1 score and exact match (EM) to evaluate xMRC performance, as adopted by XQuAD. Because the models tend to include unnecessary source text in their correct answers, EM will under-estimate the correctness of the models' response. As a result, we use F1 as the main performance metric.

Cross-lingual performance metrics. To measure the level of the models' performance alignment between English and non-English languages, we calculate the average performance on all the non-English languages, as well as the ratio of non-English performances over English ones.

2.1.3 Performance bottleneck analysis

Error type ablation. We distinguish content error in the model predictions from other less important types of error. Based on preliminary observations, we choose several major types of distracting errors:

- **Language**: answering in x instead of en;
- **Gibberish**: giving irrelevant, non-sense or hallucination output;
- **Refusal**: refusing to answer the question;
- **Blank**: providing no answer at all.

The latter 3 error types can be aggregated as generation failure errors. We measure the proportion of these errors in the models' incorrect responses. For example, the language error rate for test setting en-x is calculated as:

$$E_{\text{lang}}(\text{en}, \text{x}) = \sum_{(r,a)\in W_{\text{en,x}}} \frac{\mathbb{I}(\text{Lang}(r) = \text{en})\mathbb{I}(\text{Lang}(a) = \text{x})}{|W_{\text{en,x}}|}$$

where r is the reference answer, a is the model prediction, and $\text{Lang}(\cdot)$ is a language detector for given text. Meanwhile, a generation failure error rate, e.g. gibberish, will be:

$$E_{\text{gib}}(\text{en}, \text{x}) = \sum_{(r,a)\in W_{\text{en,x}}} \frac{\mathbb{I}(\text{Type}(r) = \text{Gibberish})}{|W_{\text{en,x}}|}$$

where $\text{Type}(\cdot)$ is an LLM-based error type classifier for generation failure samples.

Oracle performance estimation Because the xMRC score can be affected by errors in the generation process, it may not reflect the full potential of the model's cross-lingual context retrieval ability. To ablate this effect, we estimate the oracle retrieval performance of the models by perturbation-based attribution[4] on contextual sentences or spans, which calculates the importance of each of them to the output. If the sentence/span with the correct answer in it receives the largest importance,

[4]We use Captum to perform the attribution (https://github.com/pytorch/captum)

we consider the model's oracle retrieval on this testing sample is correct. Then, we calculate the accuracy of such selection respectively with one generation step or the whole generation process as the attributed target.

2.2 Analysis methods

2.2.1 Layer-wise attribution to reflect forward process

To better understand the forward calculation process of models performing xMRC retrieval, we need layer-wise attribution in order to see from which input part the LLM draws information into the residual stream (Voita et al., 2024) at each layer.

Previous studies have proposed multiple layer-wise attribution methods, including attention-based (Hao et al., 2021; Ferrando et al., 2022), backpropagation-based (Voita et al., 2021), and decomposition-based (Yang et al., 2023a) etc. Here we choose AttentionLRP (Achtibat et al., 2024), because it can attribute latent representations in each model layer and is especially suitable for transformer models.

Based on this, we define the **Major Relevance Depth (MRD)** to estimate the maximum depth to which a token representation x needs to be encoded, by calculating the layer number corresponding to the 95th percentile of its attributed relevance to model m's output:

$$\mathrm{MRD}(m,x) = \min_{1\leqslant n\leqslant N} n$$

$$\text{s.t.} \sum_{i=1}^{n} r_{\text{out}}(x,i) \geqslant 0.95 \sum_{i=1}^{N} r_{\text{out}}(x,i)$$

where $r_{\text{out}}(x,i)$ refers to the normalized relevance score of token representation x in layer i to the final output. Note that N is one less than the total layer number (e.g. 31 for LLaMA-3.1-8B) because we attribute the output of the last layer to its precedent layers. If the MRD of an input token is $\hat{n}$, then we estimate that the token's information participates in the context retrieval process only in the first $\hat{n}$-th layers. Then, for parts of the input, i.e., task description, demonstrations, question and context, we take the maximum MRD of tokens in each part to represent them, and calculate the mean MRD for each part on a group of testing samples to reflect the general distribution.

Name	Modes	Sizes
LLaMA 2	Chat	7B
LLaMA 3.1	Base / Instruct	8B / 70B
Mistral V0.3	Base / Instruct	7B
Qwen 2.5	Base / Instruct	7B / 72B
DeepSeek V2	Base / Chat	Lite (16B)
Gemma 2	Base / IT	9B
GPT-3.5-Turbo-0125	-	-
GPT-4o	-	-

Table 1: Selected models for main evaluation and analysis.

2.2.2 Hidden state similarity to measure cross-lingual alignment

To observe the cross-linguality of model internal representations during the context retrieval process, we collect the hidden states of the input sequence across all model layers, and calculate their cross-lingual similarity. Because our testing samples are parallel across all the testing languages, we can use the ratio of the mean of sample-wise over language-wise cosine similarities to define a cross-lingual similarity ratio $S(\text{en,x})$ between the model calculation processes in English and language x:

$$\begin{aligned} S(\text{en,x}) &= \overline{\text{Sim}}(E,X)/\overline{\text{Sim}}(X,X) \\ &= \frac{(K-1)\sum_{e_k\in E, x_k\in X}\text{Sim}(e_k,x_k)}{2\sum_{x_i,x_j\in X}\text{Sim}(x_i,x_j)} \end{aligned}$$

where $\overline{\text{Sim}}$ denotes the mean cosine similarity, E denotes the hidden states from English MRC tasks, and X denotes the hidden states from en-x xMRC tasks. K is the total number of testing samples, e_k and x_k are the hidden states from the k-th parallel sample pair between English and language x, x_i, x_j are the hidden states from every two different samples in language x, and the cosine similarity $\text{Sim}(x,y) = x \cdot y/|x||y|$.

3 Experiment Settings

3.1 Dataset

We use the XQuAD dataset to measure the xMRC performance of LLMs, because its testing samples are parallel in all the 12 included languages and thus suitable for cross-lingual transforming. The XQuAD dataset is composed of 1190 parallel data points for each of its 12 languages, ensuring a consistent evaluation setup across languages. With an average context length of 702.50, the dataset provides sufficiently rich contexts for machine reading comprehension. Furthermore, the 12 languages cover diverse language families, scripts, and resource levels, making our evaluation more trustworthy and representative.

	F1 scores					error rates		
	en-en	mean en-x	mean x-x	en-x / en-en	x-x / en-en	mean language	mean generation	en-en generation
LLaMA-3.1-8B	75.97	49.01	70.28	0.64	0.93	0.32	8.88	5.60
LLaMA-3.1-70B	82.39	58.68	74.73	0.71	0.91	60.21	2.63	1.20
Mistral-V0.3-7B	79.57	58.74	64.92	0.74	0.82	21.24	14.25	0.49
Qwen-2.5-7B	62.42	57.51	66.11	0.92	1.06	0.96	3.29	1.51
Qwen-2.5-72B	86.03	78.92	81.16	0.92	0.94	10.90	2.53	0.00
DeepSeek-V2-Lite-16B	73.81	44.65	57.66	0.61	0.78	12.97	8.45	1.87
Gemma-2-9B	80.42	66.82	72.90	0.83	0.91	1.91	4.11	1.02
LLaMA-3.1-Instruct-8B	77.89	72.13	65.02	0.93	0.83	0.89	2.53	0.85
LLaMA-3.1-Instruct-70B	83.29	73.07	74.13	0.88	0.89	0.23	1.87	1.85
Mistral-V0.3-Instruct-7B	62.01	56.63	49.39	0.91	0.80	2.77	3.30	1.77
Qwen-2.5-Instruct-7B	81.83	76.43	71.61	0.93	0.88	0.67	3.21	2.75
Qwen-2.5-Instruct-72B	77.12	66.04	70.29	0.86	0.91	4.58	1.62	0.38
DeepSeek-V2-Chat-Lite-16B	70.30	54.03	49.95	0.77	0.71	2.36	5.92	0.58
Gemma-2-IT-9B	83.69	78.72	75.53	0.94	0.90	0.17	2.47	1.95
GPT-3.5-Turbo-0125	81.74	68.75	72.04	0.84	0.88	0.16	2.80	0.00
GPT-4o	83.29	78.76	75.68	0.95	0.91	0.10	1.40	0.00

Table 2: 2-shot F1 scores on en-x and x-x tasks, and 2-shot language error and generation failure error rates (%) on en-x tasks.

3.2 Models and tools

We adopt a variety of SOTA open and business LLMs, including LLaMA-3.1, Qwen-2.5, GPT-4o, etc., in smaller and larger sizes. Table 1 shows our selected model list in the main evaluation and analysis results. Also, Table 4 in Appendix A.1 shows a full list of all the tested models, where some of them are of older versions and alternative sizes compared with the main-list models.

We also adopt tools for error type identification. For language error detection, we use Lingua [5] with its high-accuracy mode, the accuracy of which is satisfactory in our tested languages. For generation failure errors detection, we use Qwen-2.5-72B-Instruct (prompt shown in Appendix A.2) to act like a classifier, distinguishing normal outputs and the three types of error.

3.3 Prompts

In our xMRC evaluation, we try our best to adopt the most suitable prompt templates for each tested model to minimize its limitation of performance. We try two prompting formats for each model and take the one with higher xMRC performance to represent that model. As shown in Appendix A.2, the v1 prompt format places the task description before the demonstrations, and the v2 prompt format places the task description after the demonstrations.

We also test each model in 2-shot and 0-shot settings. Because most of the tested LLMs perform better in the 2-shot setting, we focus our evaluation and analysis mainly on this setting. The 0-shot evaluation results are in Appendix A.3.

[5] https://github.com/pemistahl/lingua

4 Results

4.1 Evaluation results

The left part of Table 2 summarizes the F1 scores of our main-list models on both en-x cross-lingual MRC (xMRC) and x-x monolingual MRC tasks. Comparing the xMRC scores with the monolingual scores highlights the difference in model performance between cross-lingual and monolingual settings. Detailed F1 scores for each model, language, and task are further illustrated in Table 6 in Appendix A.3.

4.1.1 Cross-lingual performance

Generally, one can see the English MRC performance of most tested models are high (over 70 out of 100), but the xMRC scores ranges (from 45 to 78), **showing the performance gap in context retrieval with English and non-English queries, even with the same English context and answer**.

Down into individual models, while GPT-4o shows the highest cross-lingual performance and smallest language gap, several open-sourced post-trained model series, such as Gemma-2-IT, Qwen-2.5-Instruct and LLaMA-3.1-Instruct, also show comparable performance levels and small language gaps, indicating that newest training techniques contribute to higher xMRC performance.

An interesting observation is that, for LLaMA-3.1-8B and Gemma-2-9B, the English performances after post-training are close to those of the base versions, but the cross-lingual performances greatly improve. Also, this phenomenon is more prominent in smaller (7-9B) than larger models, bring the former a smaller performance gap between English and non-English. This shows post-training, especially on smaller LLMs, is important to the xMRC task.

4.1.2 Comparison with Monolingual performance

In general, the performance gaps between English and non-English monolingual MRC are much smaller for most of the tested LLMs than xMRC, and the Qwen models even show higher non-English performance than English. This suggests that non-English language fluency is not the main cause of the ranging xMRC performance.

Also, for base models and post-trained large models (∼70B), the monolingual performances in non-English languages are always higher than cross-lingual performance. A possible explanation to this may be the cross-lingual task is more difficult and less frequent in training, and it requires cross-lingual understanding.

However, for post-trained, smaller models (7-9B), the pattern flips, where the monolingual performances become consistently lower than cross-lingual ones. This result suggests that these models become capable of **using their English ability to assist context retrieval with non-English queries**, overcoming the difficulty and low-frequency of the cross-lingual task. Also, since we should expect larger LLMs having better instruction following and understanding in general, this further highlights that post-training better elicits the cross-lingual context retrieval ability on smaller LLMs.

4.2 Performance bottleneck analysis

4.2.1 Error type ablation

Language error. The right part of Table 2 shows the average percentage of language errors among all languages. One can see an expected advantage of post-trained models against base models, because the former are better in following the cross-lingual task format. However, since the error rate is low for all the post-trained models, it cannot be viewed as a bottleneck for the xMRC task.

Generation failure errors. The right part of Table 2 also shows the average summed percentages of gibberish, refusal, and blank error in samples with wrong model answers (detailed percentages in Table 7-9c in Appendix A.4). One can see the generation failure rates are minor for most models, regardless of size and post-training, marking it not the bottleneck of xMRC either.

Model	Step		Sequence	
Language	en-en	en-x	en-en	en-x
LLaMA-3.1-8B	66.55	67.59	35.14	36.39
LLaMA-3.1-70B	47.47	54.92	47.89	56.97
LLaMA-3.1-Instruct-8B	89.86	83.42	93.75	86.66
LLaMA-3.1-Tuned-8B	86.49	81.04	89.53	83.07
LLaMA-3.1-Instruct-70B	92.82	90.67	95.44	93.54

Table 3: Oracle performance estimated (F1) for LLaMA models in en-en and en-x (average) scenarios. The estimation is performed with one generation step (left) and with the whole generated sequence, respectively.

4.2.2 Estimated oracle performance

As described in §2.1.3, the oracle performances of the LLaMA models estimated with one generation step and the whole sequence are shown in Table 3, where two phenomena stand out:

First, the estimated oracle performances of post-trained models are significantly higher than the base models, both on en-en and en-x settings, suggesting post-training is crucial to the xMRC task.

Second, the estimated oracle performance for en-x is close to en-en for all the LLaMA models, and the oracle for post-trained models are basically over 90%. This is quite different from the actual performance, where en-en is better than en-x, and the performances are not that high. This suggests the models have actually obtained the ability to locate the correct answer in the xMRC task, but the ability needs to be further elicited.

5 Two-phased mechanism of xMRC

The above results show that post-training on smaller LLM elicits the cross-lingual transfer of contextual retrieval ability. However, the mechanism that enables such transfer is unknown. Considering the model forward process from the input prompt to the output answer, we come up with a two-phase hypothesis of the xMRC process (taking en-x as an example):

1. **Question encoding.** The queries (mainly the non-English question) will be encoded into a shared semantic space, where queries in different languages are aligned and understood in a language-neutral way;

2. **Answer retrieval.** The encoded queries will be used to match the answer in the English context according to the task description and format, then the answer is generated by copying from the original context.

Figure 2 shows an illustration of the hypothesized process. We test our hypothesis from at-

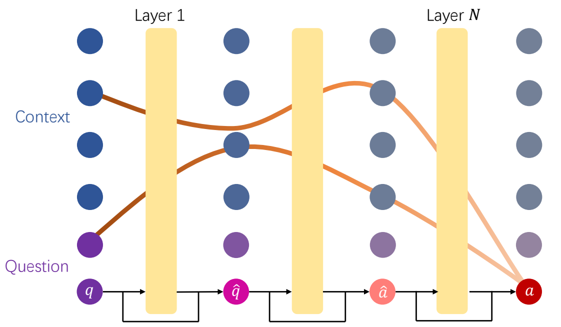
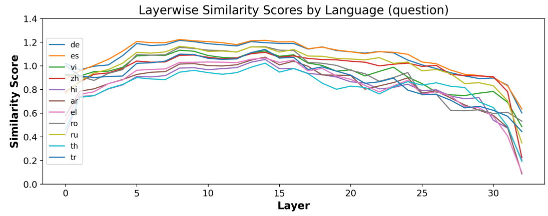
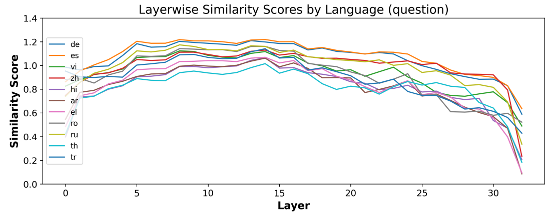


Figure 2: Illustration of the hypothesized two-phased xMRC process. As layer depths increases, the last question token will be gradually transferred to the first answer token.



(a). balanced samples



(b). en-superior samples

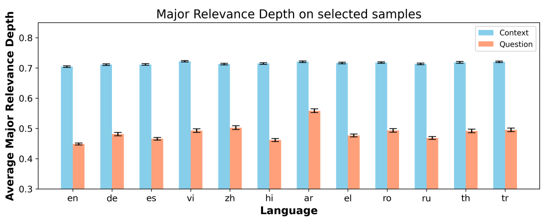
Figure 4: Question hidden state similarity between English and other languages in each layer of the LLaMA-3.1-Instruct-8B model.

tribution and hidden-state views. Since LLaMA-3.1-Instruct-8B shows high performance alignment across languages and is widely used in the community, we take its family to perform our analysis.

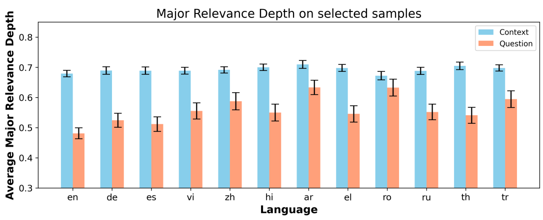
To further ensure the effect of post-training, we also conduct finetuning and compare the model behavior before and after it.

5.1 Attribution view

We use layer-wise attribution (§2.2.1) to identify the question encoding and answer retrieval phases on the LLaMA models. Figure 3 shows the mean MRD of the contexts and the questions for the LLaMA-3.1-Instruct-8B on testing samples that are identified as either "balanced" or "en-superior" in all tested languages. [6]



(a). balanced samples



(b). en-superior samples

Figure 3: Mean MRD of the context and question parts for LLaMA-3.1-Instruct-8B.

One can see that the mean question MRD is significantly and substantially lower than the mean context MRD in all tested languages and across the LLaMA models, revealing a clear phased behavior.

Also, by comparing the MRDs of "balanced" (Figures 3a) and "en-superior" samples (Figures 3b), we can see that the MRDs of "balanced" samples are more stable than those of "en-superior" samples, suggesting correlation between the high cross-lingual context retrieval ability and a clear two-phase behavior.

Details and More results can be found in Appendix B.1. We found this pattern consistent for different LLaMA models and not affected by the prompt format, though it is weaker in LLaMA-2-Chat-7B, which has a smaller pre-trained capacity and weaker multilingual ability.

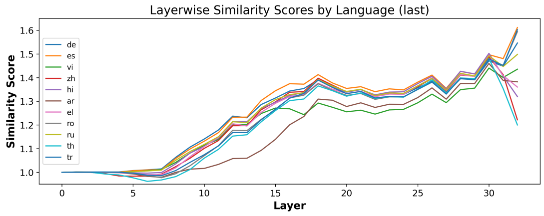
In summary, the attribution results supports the two-phase hypothesis, and indicates that the phased behavior is already formed after pre-training, regardless of model size and prompt format. The phasing strength correlates with the pre-training capacity and the xMRC task performance.

5.2 Hidden states view

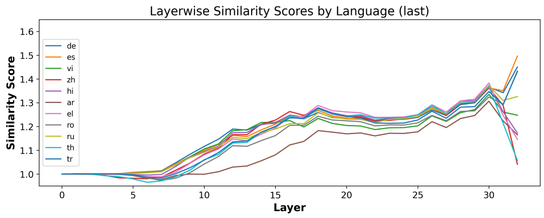
The hidden state similarity results also support our hypothesis. Figures 4, 5, and 6 show the en-x hidden state similarity of the question, the last input token (predicting the start of the answer) and the context parts for the LLaMA-3.1-Instruct-8B model. Results for other LLaMA models can be found in Appendix B.2. The observed trends are consistent:

- For question representations, they all show

[6] We identify a sample as "balanced" if the model F1 score on it is above 0.5 in all directions; and a sample as "en-supeiror" if the F1 score in English is higher than the average of other languages with a margin greater than 0.5.

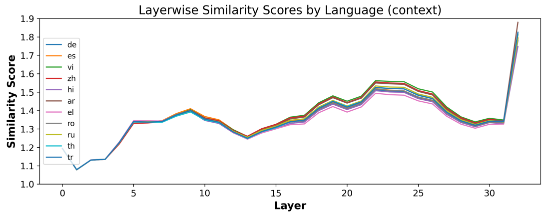


(a). balanced samples

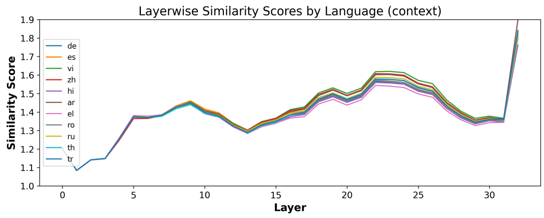


(b). en-superior samples

Figure 5: Last-input-token hidden state similarity between English and other languages in each layer of the LLaMA-3.1-Instruct-8B model.



(a). balanced samples



(b). en-superior samples

Figure 6: Context hidden state similarity between English and other languages in each layer of the LLaMA-3.1-Instruct-8B model.

a shared arc-shaped trend, where the highest similarity to English appears at the relative depth of around 1/3;

- For context representations, a consistent double-peak trend can be observed, with a "turning point" around the relative depth of 0.4 (matching the question MRD) and the second, higher peak at around 0.7 (matching the context MRD);

- For last input token representations, one can see a consistent "plateau" of similarity starting at around a relative depth of 0.5, which also matches the mean question MRD.

Again, for the less powerful model LLaMA-2-Chat-7B (in Appendix B.2 Figure 18), the trends are weaker: its question similarity to English varies much across languages, and the "plateau" of answer similarity to English starts later than other models, which is after the mean question and context MRDs.

These results suggest that the hidden states similarity through the xMRC process also undergo two main phases with evident distinction. This phased behavior already exists in pre-trained LLMs, and is preserved in post-training. Also, the phasing strength correlates with the model capability built during pre-training.

5.3 Pre-training vs. post-training

Although the phasing behavior is shaped during pre-training, other results show that post-training is also essential to complete the xMRC task.

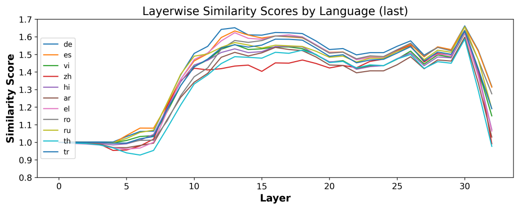
We observe that (in Appendix B.2 Figure 19) the last-input-token similarity of base models experience severe and consistent decline in the last few layers, across all non-English languages. Since in our xMRC task we expect the same English output for all tested languages, this drop in cross-lingual similarity can directly affect the performance and its cross-lingual alignment. Meanwhile, post-training significantly narrows this decline, especially on the 8B model. As a result, for the Instruct-8B model, languages with Latino alphabets reveal even higher similarity than the previous layers.

In this regard, a possible reason for the larger performance gap between English and non-English for larger post-trained models could be, their post-training is insufficient to form the answer retrieving phase. As shown in Figure 17, their last-input-token similarity drops in the final layers instead of rising, which is a feature of the base models.

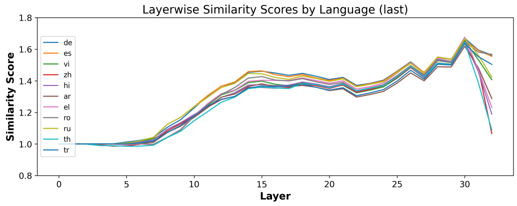
Another outcome of post-training is sensitivity to the demonstrations. One can see from Figure 6 and 17 (c) and (d), both post-trained models show divergence in context similarity, which is not seen in their base models. Since the context in the xMRC task is English-only, this divergence in similarity can only be the effect of cross-lingual demonstration preceding the context. Compared with the 8B model, the divergence of the 70B instruct model is too large, suggesting that it is too sensitive to the demonstrations, and may harm the cross-lingual performance.

5.4 Validation of post-training

To validate the effect of post-training, we tune the LLaMA-3.1-8B model use the TULU-v3 dataset (Lambert et al., 2025) into a model called LLaMA-3.1-Tuned-8B (Appendix B.3). Like the instruct model, the en-x xMRC performance of the tuned model become significantly higher than the base model (Table 10), and the decline of cross-lingual hidden state similarity in the last few layers also narrows (Figure 7). This is clear evidence for the essential role of post-training in completing the xMRC task.



(a). LLaMA-3.1-8B



(b). LLaMA-3.1-Tuned-8B

Figure 7: Change in Last-input-token hidden state similarity between English and other languages in each layer of LLaMA-3.1-8B after tuning on the same data samples as "balanced" of the base model.

6 Related Work

6.1 Cross-lingual alignment of LLMs

Previous studies have shown a misalignment of LLMs with English and other languages. With respect to performance, Lai et al. (2023b), Ahuja et al. (2024) and Wang et al. (2024) demonstrated that SOTA LLMs performed better in English, and showed inconsistency when dealing with non-English queries. Etxaniz et al. (2024) found that LLMs performed worse with non-English prompts than with self-translated English prompts. Beyond performance, Qi et al. (2023) demonstrated the low cross-lingual consistency of factual knowledge does of LLMs, and Gao et al. (2024) showed that multilingual pre-training and instruction tuning could only enhance superficial levels of cross-lingual alignment. However, the previous work mostly focused on queries with a consistent language, instead of cross-lingual queries.

There have also been many techniques to enhance LLMs' cross-lingual alignment. For example, adding parallel data in the pre-training stage (Lample and Conneau, 2019; Jiang et al., 2022; Wei et al., 2023; Lu et al., 2024); and the post-training stage, including instruction tuning (Li et al., 2023b; Zhang et al., 2025; Li et al., 2023a; Cahyawijaya et al., 2023; Chai et al., 2024; Kuulmets et al., 2024; Shaham et al., 2024; Kew et al., 2024) and preference tuning (Lai et al., 2023a; She et al., 2024). Especially, extra translation training is commonly used (Zhang et al., 2023; Yang et al., 2023b; Li et al., 2024; Ranaldi et al., 2024; Zhu et al., 2024; Lu et al., 2024). In this paper, we examine the effect of some of these techniques by comparing various SOTA models.

6.2 Cross-lingual machine reading comprehension

xMRC is a relatively new task of natural language understanding. Cui et al. (2019) proposed the task, in order to improve non-English MRC performance by introducing English resources. There are some representative datasets in this area, such as XQA (Liu et al., 2019), BiPaR (Jing et al., 2019), MLQA (Lewis et al., 2020), and XQuAD (Artetxe et al., 2020). Ushio et al. (2023) also proposes a pipeline for multilingual QA generation.

Previous work on the xMRC task mainly focuses on enhancing the performance of task-specific models using techniques such as data augmentation (Bornea et al., 2021; Xiang et al., 2024), knowledge injection (Duan et al., 2021), constractive learning (Chen et al., 2022) and knowledge transfer (Cao et al., 2023; Xu et al., 2023). However, the xMRC performance of LLMs has not been studied.

7 Conclusion

We investigate the cross-lingual context retrieval ability of LLMs with the xMRC task. We first evaluates the xMRC performance of existing open- and closed-sourced LLMs, and find that post-trained 7-9B models show high performance and little gap across English and non-English languages. Then, we conduct analysis on the LLaMA models and identifies the two main phases (question encoding and answer retrieval) of the xMRC process, and find a possible explanation to the language gap for larger post-trained models. We hope our research will inspire future study to foster the cross-lingual alignment of LLMs in a broader scope.

8 Limitations

While this study provides valuable insights into the cross-lingual context retrieval abilities of LLMs and identifies a two-phased mechanism underlying this process, it is important to acknowledge certain limitations.

First, the scope of our empirical evaluation is constrained by available resources and time. This necessarily limits the breadth of our testing, preventing us from exhaustively covering the rapidly expanding landscape of LLMs. Furthermore, while we test across 12 diverse languages, a more comprehensive analysis would ideally include an even wider range of languages in order to ensure the generalizability of our findings across linguistic diversity.

Second, although we successfully identify a two-phased feature of cross-lingual machine reading comprehension and confirm its correlation with pre-training and post-training stages, the precise factors within these training processes that drive this outcome remain unclear. Future work could delve deeper into the specifics of pre-training objectives, data composition, and post-training techniques to pinpoint the exact elements that contribute to the emergence and effectiveness of these two phases.

Finally, within the two major phases we discover – question encoding and answer retrieval – we observe hints of more fine-grained changes in model behavior, particularly in the hidden state similarity curves. These preliminary observations suggest the potential for a more nuanced understanding of the xMRC process. Future studies could further investigate these finer-grained dynamics within each phase to gain a more detailed and complete picture of how LLMs achieve cross-lingual context retrieval.

Beyond these limitations, it is also important to consider potential risks associated with this work. While our research is foundational and not directly tied to specific applications, advancements in cross-lingual context retrieval, like any technology, could be misused. For example, improved cross-lingual capabilities might inadvertently contribute to the spread of misinformation if models are used to retrieve and amplify biased or inaccurate information across languages. Furthermore, if deployed without careful consideration, these technologies could exacerbate existing inequalities by favoring languages and knowledge systems already dominant in LLM training data, potentially marginalizing less-represented languages and perspectives. Future work should consider these dual-use aspects and explore mitigation strategies to ensure responsible development and deployment of cross-lingual NLP technologies, paying special attention to fairness and inclusivity across diverse linguistic communities.

9 Ethics Statements

This research adheres to ethical principles in its use of language models and data. All language models evaluated and finetuned in this study are accessed and utilized in compliance with their respective licenses and terms of service. Furthermore, the XQuAD dataset employed for evaluation, and the TULU-v3 dataset used for finetuning LLaMA-3.1-8B, are both publicly available datasets intended for research purposes. Based on our review and the documented nature of these datasets, we have determined that they are not designed to collect or contain personally identifiable information or offensive content. To the best of our knowledge, and as indicated in their public documentation, neither dataset includes data that names or uniquely identifies individual people, nor do they present offensive content.

Our use of these existing artifacts, including both language models and datasets, is aligned with their intended use within research contexts. Specifically, derivatives of data accessed for research purposes, such as model outputs and analysis results, are used solely within the bounds of academic inquiry and are not disseminated or utilized outside of these research contexts, in accordance with responsible data handling practices. We have strived to conduct this research responsibly and ethically, focusing on understanding and improving the cross-lingual capabilities of language models for the benefit of the NLP community, while respecting the intended use and access conditions of all resources utilized.

References

Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. 2024. AttnLRP: Attention-Aware Layer-Wise Relevance Propagation for Transformers. In *Proceedings of the 41st International Conference on Machine Learning*, pages 135–168. PMLR.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024. MEGAVERSE: Benchmarking large language models across languages, modalities, models and tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2598–2637, Mexico City, Mexico. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Mihaela Bornea, Lin Pan, Sara Rosenthal, Hans Florian, and Avi Sil. 2021. Multilingual Transfer Learning for QA using Translation as Data Augmentation. In *AAAI Conference on Artificial Intelligence*.

Samuel Cahyawijaya, Holy Lovenia, Tiezheng Yu, Willy Chung, and Pascale Fung. 2023. InstructAlign: High-and-low resource language alignment via continual crosslingual instruction tuning. In *Proceedings of the First Workshop in South East Asian Language Processing*, pages 55–78, Nusa Dua, Bali, Indonesia. Association for Computational Linguistics.

Tingfeng Cao, Chengyu Wang, Chuanqi Tan, Jun Huang, and Jinhui Zhu. 2023. Sharing, Teaching and Aligning: Knowledgeable Transfer Learning for Cross-Lingual Machine Reading Comprehension. *Preprint*, arXiv:2311.06758.

Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xiannian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, and Zhoujun Li. 2024. xCoT: Cross-lingual Instruction Tuning for Cross-lingual Chain-of-Thought Reasoning. *Preprint*, arXiv:2401.07037.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Nuo Chen, Linjun Shou, Ming Gong, and Jian Pei. 2022. From Good to Best: Two-Stage Training for Cross-Lingual Machine Reading Comprehension. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10501–10508.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2019. Cross-lingual machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1586–1595, Hong Kong, China. Association for Computational Linguistics.

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. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Zhichao Duan, Xiuxing Li, Zhengyan Zhang, Zhenyu Li, Ning Liu, and Jianyong Wang. 2021. Bridging the Language Gap: Knowledge Injected Multilingual Question Answering. In *2021 IEEE International Conference on Big Knowledge (ICBK)*, pages 339–346.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in English? In *Proceedings of the 2024 Conference of the North American Chap-*

ter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564, Mexico City, Mexico. Association for Computational Linguistics.

Javier Ferrando, Gerard I. Gállego, and Marta R. Costa-jussà. 2022. Measuring the mixing of contextual information in the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8698–8714, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Changjiang Gao, Hongda Hu, Peng Hu, Jiajun Chen, Jixing Li, and Shujian Huang. 2024. Multilingual pretraining and instruction tuning improve cross-lingual knowledge alignment, but only shallowly. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6101–6117, Mexico City, Mexico. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho

Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

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. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2021. Self-Attention Attribution: Interpreting Information Interactions Inside Transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12963–12971.

Xiaoze Jiang, Yaobo Liang, Weizhu Chen, and Nan Duan. 2022. XLM-K: Improving Cross-Lingual Language Model Pre-training with Multilingual Knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10840–10848.

Yimin Jing, Deyi Xiong, and Yan Zhen. 2019. Bipar: A bilingual parallel dataset for multilingual and cross-lingual reading comprehension on novels. *Preprint*, arXiv:1910.05040.

Tannon Kew, Florian Schottmann, and Rico Sennrich. 2024. Turning English-centric LLMs into polyglots: How much multilinguality is needed? In *Findings* *of the Association for Computational Linguistics: EMNLP 2024*, pages 13097–13124, Miami, Florida, USA. Association for Computational Linguistics.

Hele-Andra Kuulmets, Taido Purason, Agnes Luhtaru, and Mark Fishel. 2024. Teaching llama a new language through cross-lingual knowledge transfer. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3309–3325, Mexico City, Mexico. Association for Computational Linguistics.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023a. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023b. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. *Preprint*, arXiv:2304.05613.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. 2025. Tulu 3: Pushing frontiers in open language model post-training. *Preprint*, arXiv:2411.15124.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual Language Model Pretraining. *Preprint*, arXiv:1901.07291.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023a. Bactrian-X: Multilingual Replicable Instruction-Following Models with Low-Rank Adaptation. *Preprint*, arXiv:2305.15011.

Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2024. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. *Transactions of the Association for Computational Linguistics*, 12:576–592.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. 2023b. M^{3}IT: A Large-Scale Dataset towards Multi-Modal Multilingual Instruction Tuning. *Preprint*, arXiv:2306.04387.

Jiahua Liu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2019. XQA: A cross-lingual open-domain question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2358–2368, Florence, Italy. Association for Computational Linguistics.

Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. 2024. LLaMAX: Scaling linguistic horizons of LLM by enhancing translation capabilities beyond 100 languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10748–10772, Miami, Florida, USA. Association for Computational Linguistics.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-Lingual Consistency of Factual Knowledge in Multilingual Language Models. *Preprint*, arXiv:2310.10378.

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. *Preprint*, arXiv:2412.15115.

Leonardo Ranaldi, Giulia Pucci, and Andre Freitas. 2024. Empowering cross-lingual abilities of instruction-tuned large language models by translation-following demonstrations. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7961–7973, Bangkok, Thailand. Association for Computational Linguistics.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2304–2317, Bangkok, Thailand. Association for Computational Linguistics.

Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. 2024. MAPO: Advancing multilingual reasoning through multilingual-alignment-as-preference optimization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10015–10027, Bangkok, Thailand. Association for Computational Linguistics.

Asahi Ushio, Fernando Alva-Manchego, and Jose Camacho-Collados. 2023. A Practical Toolkit for Multilingual Question and Answer Generation. *Preprint*, arXiv:2305.17416.

Elena Voita, Javier Ferrando, and Christoforos Nalmpantis. 2024. Neurons in large language models: Dead, n-gram, positional. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1288–1301, Bangkok, Thailand. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2021. Analyzing the source and target contributions to predictions in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1126–1140, Online. Association for Computational Linguistics.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, Ai Ti Aw, and Nancy F. Chen. 2023. SeaEval for Multilingual Foundation Models: From Cross-Lingual Alignment to Cultural Reasoning. *Preprint*, arXiv:2309.04766.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy Chen. 2024. SeaEval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 370–390, Mexico City, Mexico. Association for Computational Linguistics.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, Tianxiang Hu, Shangjie Li, Binyuan Hui, Bowen Yu, Dayiheng Liu, Baosong Yang, Fei Huang, and Jun Xie. 2023. PolyLM: An Open Source Polyglot Large Language Model. *Preprint*, arXiv:2307.06018.

Junyi Xiang, Maofu Liu, Qiyuan Li, Chen Qiu, and Huijun Hu. 2024. A cross-guidance cross-lingual model on generated parallel corpus for classical Chinese machine reading comprehension. *Information Processing & Management*, 61(2):103607.

Weiwen Xu, Xin Li, Wai Lam, and Lidong Bing. 2023. mPMR: A Multilingual Pre-trained Machine Reader at Scale. *Preprint*, arXiv:2305.13645.

Sen Yang, Shujian Huang, Wei Zou, Jianbing Zhang, Xinyu Dai, and Jiajun Chen. 2023a. Local interpretation of transformer based on linear decomposition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10270–10287, Toronto, Canada. Association for Computational Linguistics.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023b. BigTranslate: Augmenting Large Language Models with Multilingual Translation Capability over 100 Languages. *Preprint*, arXiv:2305.18098.

Kexun Zhang, Jane Dwivedi-Yu, Zhaojiang Lin, Yuning Mao, William Yang Wang, Lei Li, and Yi-Chia Wang. 2025. Extrapolating to unknown opinions using LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7819–7830, Abu Dhabi, UAE. Association for Computational Linguistics.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, and Yang Feng. 2023. BayLing: Bridging Cross-lingual Alignment and Instruction Following through Interactive Translation for Large Language Models. *Preprint*, arXiv:2306.10968.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024. Question translation training for better multilingual reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8411–8423, Bangkok, Thailand. Association for Computational Linguistics.

A Evaluation

A.1 Full list of models evaluated

A comprehensive list of all models evaluated during this study, including older versions and alternative sizes, is available in Table 4 to supplement the main list.

Name	Modes	Sizes
LLaMA 2	Base / Chat	7B / 13B / 70B
LLaMA 3	Base / Instruct	8B / 70B
LLaMA 3.1	Base / Instruct	8B / 70B
LLaMAX-2-Alpaca	-	7B
LLaMAX-3-Alpaca	-	8B
Mistral V0.1	Base / Instruct	7B
Mistral V0.3	Base / Instruct	7B
Qwen 1.5	Base / Chat	7B / 14B / 72B
Qwen 2	Base / Instruct	7B / 72B
Qwen 2.5	Base / Instruct	7B / 72B
DeepSeek V2	Base / Chat	Lite (16B)
Gemma 2	Base / IT	9B
GPT-3.5-Turbo-0125	-	-
GPT-4o	-	-

Table 4: Full list of models evaluated. This table presents a complete list of all models tested in this study, encompassing older versions and alternative sizes.

A.2 Prompt used in error type detection

The prompt for error type detection is:

```
<|im_start|>system
You are Qwen, created by Alibaba Cloud. You are a helpful assistant.<|im_end|>
<|im_start|>user
You are tasked with identifying the type of a given raw answer. You will be provided with a question and a raw answer. Your job is to determine whether the raw answer falls into one of the following categories based on the given question:

0. Reasonable Answer: The answer seems like some attempt to answer the question, regardless of whether it is correct or not.

1. Blank Answer: No response is provided.

2. Gibberish: Incoherent text with no clear meaning or cannot be seen as some kind of answer to the question, e.g. "{Your Answer}".

3. Denial of Answer: A statement indicating inability to answer, such as "I apologize, but I cannot answer this question because...".

You must provide your response as a SINGLE number representing the category (0, 1, 2, or 3) without extra output.
```

The v1 prompt format is:

```
{system prompt}

Below is a reading comprehension task. There will be paragraphs of context, each followed by a question related to its content. You should only present your answer to the last question by strictly copying the corresponding part of the context. Please provide a direct answer in English without extra output. Your answer should be in the form of "Answer: {Your Answer}"

Context: {demo context 1}

Question: {demo question 1}

Answer: {demo answer 1}

Context: {demo context 2}

Question: {demo question 2}

Answer: {demo answer 2}

Your task starts here:

Context: {text context}

Question: {text question}
```

The v2 prompt format is:

```
{system prompt}

Context: {demo context 1}

Question: {demo question 1}

Answer: {demo answer 1}

Context: {demo context 2}

Question: {demo question 2}

Answer: {demo answer 2}

Your task starts here:

Context: {text context}

Question: {text question}
You should only present your answer to the last question by strictly copying the corresponding part of the context. Please provide a direct answer in English without extra output. Your answer should be in the form of "Answer: {Your Answer}"
```

A.3 Detailed evaluation results

Table 5 presents the complete evaluation results from our 0-shot experiments, encompassing both the English-to-NonEnglish (en-x) and non-English monolingual (x-x) tasks in all models and languages tested. Table 6 further presents detailed F1 scores for en-x and x-x tasks in the 2-shot setting.

A.4 Detailed language error and generation failure error rates

Tables 7 and 8 show detailed language and generation failure error rates across all tested languages. Meanwhile, Table 9 provides a more granular view of the generation failure errors discussed in the main text. While Table 8 presents the aggregated rate of these errors, Table 9 is further subdivided into three separate tables: Table 9a, Table 9b, and Table 9c. These tables individually display the error rates for gibberish errors, refusal errors, and blank errors, respectively, across all tested models and languages in the 2-shot en-x xMRC task setting.

B Two-phased xMRC Analysis

B.1 Further analysis on MRD

B.1.1 Example of Attribution

Figure 8 shows an example of the attribution outcome for LLaMA-3.1-Instruct-8B.

B.1.2 MRD for other LLaMA models

Figures 9 and 10 provide further illustrative examples of the mean MRD for context and question

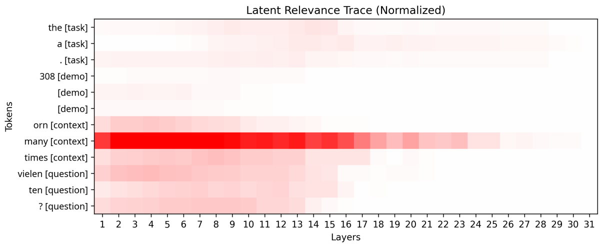
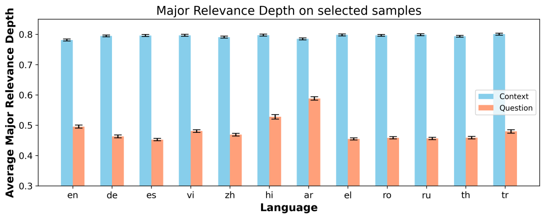
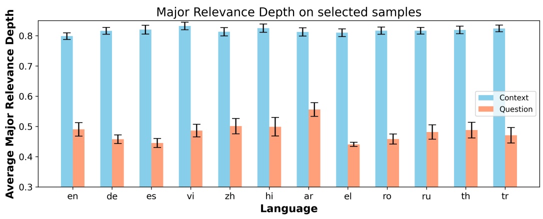


Figure 8: An example output of layer-wise attribution with LLaMA-3.1-Instruct-8B, where only the top 3 tokens from each input part are shown.

components, specifically for LLaMA-3.1-Instruct-70B and LLaMA-2-Chat-7B. These figures complement the MRD analysis presented in the main body of this paper.



(a). balanced samples



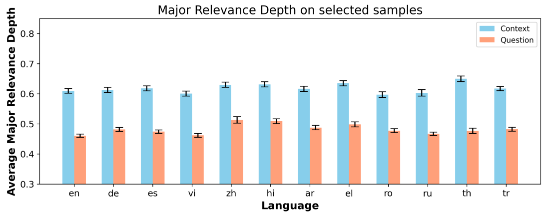
(b). en-superior samples

Figure 9: Mean MRD of the context and question parts for LLaMA-3.1-Instruct-70B.

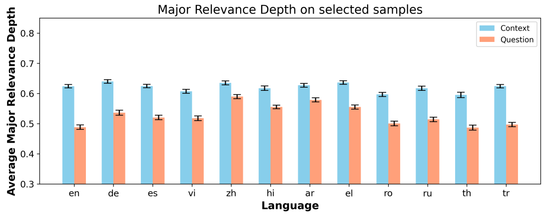
B.1.3 Analysis of task descriptions and demonstrations

Analyzing the MRD of task descriptions and demonstrations in our 2-shot setting (Figures 11-13) reveals a general trend where demonstrations tend to exhibit a comparable or slightly higher MRD than task descriptions across the LLaMA model family, suggesting demonstrations are at least as important as, if not slightly more impactful than, task descriptions in guiding the models. This could indicate that providing concrete examples is a particularly effective way to communicate the desired behavior for cross-lingual context retrieval to these models.

However, the precise relationship is not uniform and varies across models. For example, while LLaMA-3.1-Instruct-8B shows a relatively balanced MRD between task descriptions and demonstrations, LLaMA-2-Chat-7B consistently demonstrates a higher MRD for demonstrations, which implies that older or smaller models might lean more heavily on the provided in-context examples. In contrast, LLaMA-3.1-Instruct-70B exhibits the most pronounced difference, with a significantly elevated MRD for task descriptions across all languages and sample types, suggesting that larger models can become highly attuned to and reliant on user-specified task commands.

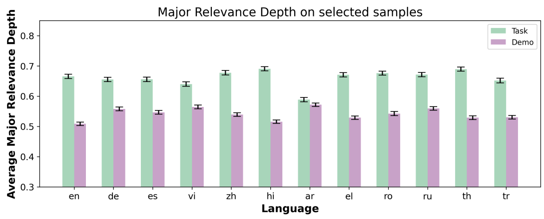


(a). balanced samples

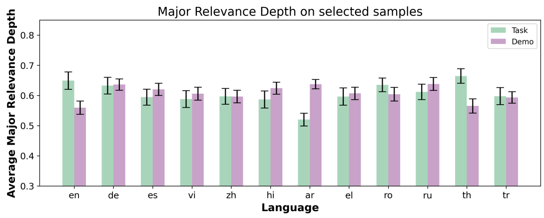


(b). en-superior samples

Figure 10: Mean MRD of the context and question parts for LLaMA-2-Chat-7B.



(a). balanced samples

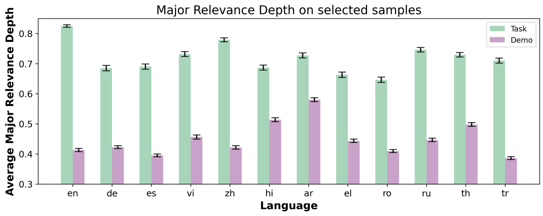


(b). en-superior samples

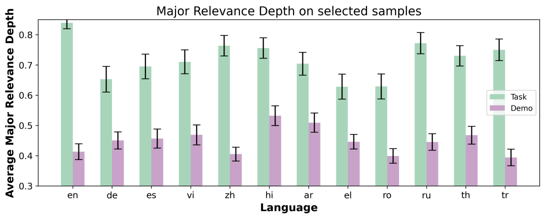
Figure 11: Mean MRD of the task descriptions and demonstrations parts for LLaMA-3.1-Instruct-8B.

B.1.4 Influence of prompt format on MRD pattern

We test the influence of different prompt formats (v1, v2) on LLaMA-3.1-Instruct-8B, and by com-

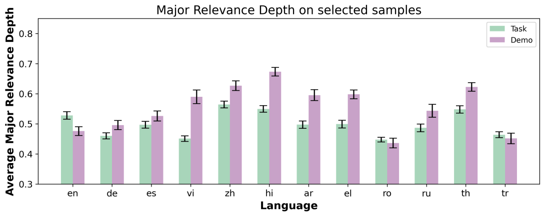


(a). balanced samples

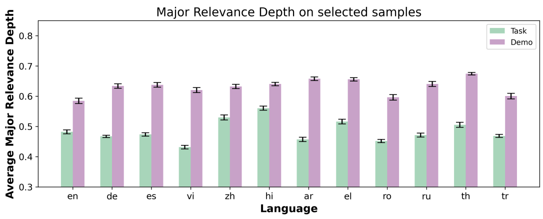


(b). en-superior samples

Figure 12: Mean MRD of the task descriptions and demonstrations parts for LLaMA-3.1-Instruct-70B.



(a). balanced samples



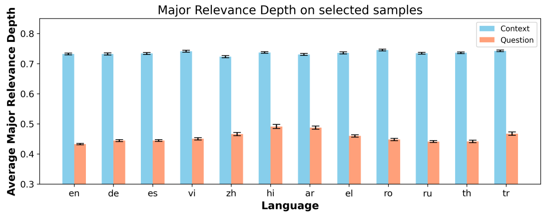
(b). en-superior samples

Figure 13: Mean MRD of the task descriptions and demonstrations parts for LLaMA-2-Chat-7B.

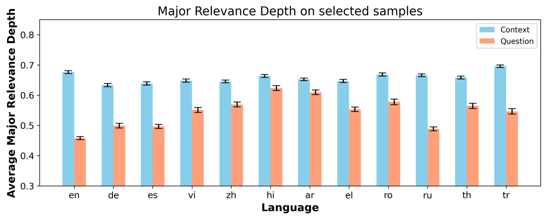
paring the results in Figure 14 and Figure 3, which present results obtained using the prompt format v1 and v2, respectively, it is clear that the fundamental pattern observed in the mean MRD is consistent across both formats. Therefore, the trend of the mean question MRD being consistently and substantially lower than the mean context MRD is maintained regardless of the prompt format employed.

B.2 Hidden state similarity results for other LLaMA models

Figures 15–18 present the hidden state similarity results for additional LLaMA models, complementing the analysis of the LLaMA-3.1-Instruct-8B model discussed in the main body of the paper.



(a). balanced samples



(b). en-superior samples

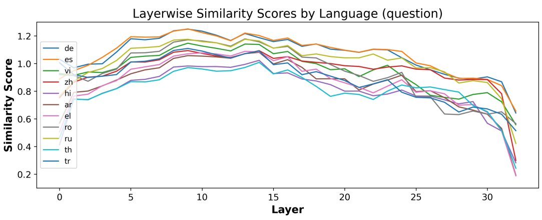
Figure 14: Mean MRD for LLaMA-3.1-Instruct-8B on both "balanced" and "en-superior" samples in v1 prompting format. Only the results of context and question parts of the prompt are displayed.

Meanwhile, Figure 19 provides an overview of the last-input-token hidden state similarity for base LLaMA models, as discussed in §5.3.

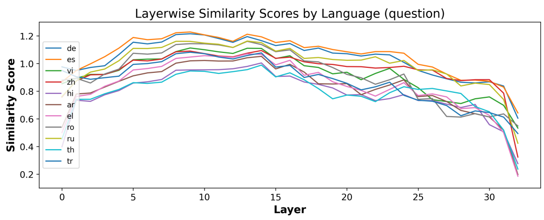
B.3 Training details and evaluation results of our finetuned LLaMA-3.1-8B

We tune the LLaMA-3.1-8B base model on TULU-V3 for 1 epoch with 8 * H800 GPUs for 15 hours using the LLaMA-Factory repository. The data cut-off length is 2048, batch size per device is 8, learning rate is 1.0e-5, and the warm-up ratio is 0.1 with cosine learning rate scheduling.

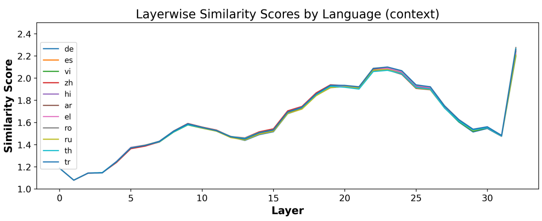
Regarding evaluation, Table 10 summarizes the performance of our finetuned model on both en-x cross-lingual and x-x monolingual MRC tasks. Furthermore, Figure 20 illustrates the hidden state similarity between English and other tested languages across layers, focusing on question, context, and last-input-token representations derived from balanced samples.



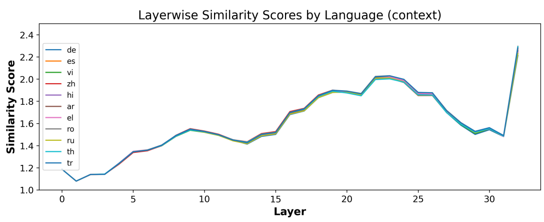
(a). Question hidden state similarity for balanced samples.



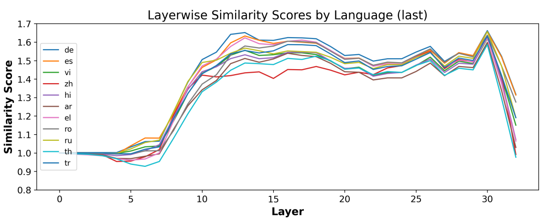
(b). Question hidden state similarity for en-superior samples.



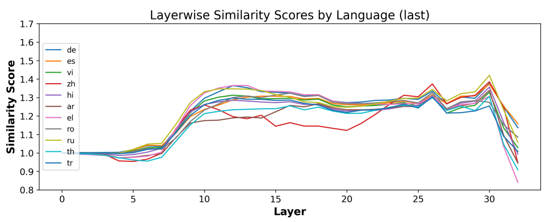
(c). Context hidden state similarity for balanced samples.



(d). Context hidden state similarity for en-superior samples.



(e). Last-input-token hidden state similarity for balanced samples.



(f). Last-input-token hidden state similarity for en-superior samples.

Figure 15: Hidden state similarity between English and other languages on different parts of the selected samples in each layer of the LLaMA-3.1-8B model.

	en-en	en-de	en-es	en-vi	en-zh	en-hi	en-ar	en-el	en-ro	en-ru	en-th	en-tr
LLaMA-3.1-8B	28.49	26.39	23.70	16.97	21.76	21.68	19.66	24.96	22.27	22.69	27.73	15.71
LLaMA-3.1-70B	49.92	38.32	36.22	41.34	45.13	24.62	27.40	37.73	38.49	34.87	47.90	33.95
Mistral-V0.3-7B	20.50	21.43	19.50	10.84	8.66	11.60	14.87	14.20	12.44	18.32	16.13	10.67
Qwen-2.5-7B	38.66	36.05	37.31	48.32	42.02	38.91	42.44	34.87	35.55	38.57	45.63	41.60
Qwen-2.5-72B	63.95	56.55	56.72	53.19	52.86	55.97	56.30	52.27	54.03	55.55	55.29	51.68
DeepSeek-V2-Lite-16B	12.35	11.26	4.12	4.20	9.58	9.58	4.79	7.73	7.06	10.84	5.80	4.96
Gemma-2-9B	15.21	5.97	14.87	14.87	26.47	7.40	11.43	4.62	13.70	20.42	32.27	17.31
LLaMA-2-Chat-7B	34.96	26.81	24.54	19.50	22.52	19.58	20.00	20.50	22.10	25.04	13.95	16.55
LLaMA-3.1-Instruct-8B	40.92	25.55	28.82	28.24	33.95	27.65	24.87	20.42	19.24	29.66	29.75	30.42
LLaMA-3.1-Instruct-70B	57.82	47.23	47.23	45.88	49.16	39.33	42.86	46.72	43.19	41.51	51.93	44.20
Mistral-V0.3-Instruct-7B	3.19	2.61	2.69	5.13	2.86	2.02	4.12	3.11	2.35	3.87	4.54	3.36
Qwen-2.5-Instruct-7B	53.28	40.17	35.97	36.13	40.67	36.22	35.46	33.70	38.99	35.71	36.22	38.24
Qwen-2.5-Instruct-72B	36.89	27.98	26.81	23.53	23.28	22.94	23.45	23.19	31.01	24.37	23.03	23.95
DeepSeek-V2-Chat-Lite-16B	16.30	18.24	13.87	8.24	15.13	11.01	11.09	13.95	13.87	12.61	9.75	12.61
Gemma-2-IT-9B	57.06	46.30	42.77	42.86	42.35	44.37	40.25	41.85	39.24	42.69	44.96	43.87
GPT-3.5-Turbo-0125	32.18	24.12	22.61	21.34	21.93	17.48	22.44	23.70	23.87	21.43	20.34	20.42
GPT-4o	51.01	39.58	36.97	40.50	41.18	37.48	36.05	37.82	36.97	39.33	39.16	39.66

(a). 0-shot Exact Match (EM) scores (%) on en-x tasks.

	en-en	en-de	en-es	en-vi	en-zh	en-hi	en-ar	en-el	en-ro	en-ru	en-th	en-tr
LLaMA-3.1-8B	37.60	34.11	33.87	21.86	35.12	36.97	30.11	36.66	30.82	30.37	39.43	21.79
LLaMA-3.1-70B	68.03	58.29	56.22	60.33	55.20	50.63	41.23	56.41	57.13	55.87	62.87	54.32
Mistral-V0.3-7B	40.07	27.17	31.26	19.26	13.19	17.94	23.98	19.16	18.23	28.26	21.86	18.63
Qwen-2.5-7B	59.02	55.58	55.61	64.37	60.26	56.32	60.61	53.14	51.21	56.72	63.23	55.52
Qwen-2.5-72B	80.50	73.73	74.48	72.48	71.37	73.60	73.23	70.59	71.44	73.36	72.78	68.18
DeepSeek-V2-Lite-16B	24.88	25.14	11.07	13.25	18.55	14.01	13.69	15.32	14.72	23.10	9.28	13.73
Gemma-2-9B	24.08	11.12	24.96	20.17	36.04	10.03	16.04	8.68	21.19	32.67	43.05	24.34
LLaMA-2-Chat-7B	56.83	45.97	44.45	37.78	36.50	33.65	32.13	34.03	39.14	45.51	25.18	30.64
LLaMA-3.1-Instruct-8B	64.47	50.69	53.41	52.33	57.10	50.09	48.61	45.36	43.46	54.14	53.88	52.72
LLaMA-3.1-Instruct-70B	78.28	70.13	68.39	64.43	68.91	56.20	60.19	67.79	65.78	64.72	67.69	62.13
Mistral-V0.3-Instruct-7B	35.19	32.41	32.68	30.23	32.15	28.49	30.21	29.42	28.93	32.51	31.48	29.23
Qwen-2.5-Instruct-7B	73.03	59.69	56.82	56.99	60.64	56.99	55.54	53.93	58.27	56.62	57.08	59.22
Qwen-2.5-Instruct-72B	59.84	46.79	46.86	43.46	36.20	43.82	40.01	44.46	50.91	43.98	44.68	44.28
DeepSeek-V2-Chat-Lite-16B	43.24	35.99	32.46	26.34	38.20	27.90	29.16	35.33	29.12	32.53	30.33	29.19
Gemma-2-IT-9B	76.80	67.91	64.91	64.74	64.83	65.31	62.35	63.89	61.29	64.94	65.84	64.10
GPT-3.5-Turbo-0125	60.08	51.06	50.45	47.23	49.06	41.58	48.90	50.57	50.90	49.32	46.66	46.22
GPT-4o	74.24	64.38	58.94	65.71	65.40	62.48	58.81	64.00	62.15	64.82	64.60	64.95

(b). 0-shot F1 Scores on en-x tasks.

	de-de	es-es	vi-vi	zh-zh	hi-hi	ar-ar	el-el	ro-ro	ru-ru	th-th	tr-tr
LLaMA-3.1-8B	24.87	19.16	18.66	33.95	16.05	17.31	10.00	18.15	17.90	33.87	14.37
LLaMA-3.1-70B	42.94	37.23	42.86	50.08	30.00	40.17	36.30	40.76	35.13	57.06	37.40
Mistral-V0.3-7B	27.31	25.29	20.17	34.62	7.98	21.09	17.73	22.69	11.93	23.45	12.44
Qwen-2.5-7B	29.50	35.04	38.57	57.23	23.11	42.94	21.09	31.85	32.18	53.95	31.51
Qwen-2.5-72B	46.89	46.39	51.93	78.32	41.18	50.67	36.64	50.67	43.03	63.78	41.34
DeepSeek-V2-Lite-16B	4.87	2.44	1.01	7.56	2.02	1.43	4.03	3.19	2.02	5.13	2.94
Gemma-2-9B	2.02	14.03	8.91	12.10	8.66	1.43	1.43	4.62	10.92	9.66	13.28
LLaMA-2-Chat-7B	22.86	18.82	15.55	4.54	1.68	4.79	6.05	22.86	11.93	3.78	10.59
LLaMA-3.1-Instruct-8B	25.71	26.97	36.64	48.40	27.06	33.03	13.70	28.32	26.22	35.63	29.83
LLaMA-3.1-Instruct-70B	40.25	35.29	47.31	57.65	35.71	44.96	30.92	40.76	38.66	59.50	40.84
Mistral-V0.3-Instruct-7B	2.69	2.44	4.29	0.92	0.84	3.70	1.51	3.11	1.43	4.29	2.61
Qwen-2.5-Instruct-7B	32.86	30.92	30.42	47.40	24.71	27.90	21.68	36.05	27.82	42.44	30.67
Qwen-2.5-Instruct-72B	29.41	24.45	26.97	45.71	20.42	32.18	15.88	32.02	25.29	36.30	21.51
DeepSeek-V2-Chat-Lite-16B	12.18	7.48	10.08	7.82	7.65	9.24	8.40	7.73	9.41	11.93	7.31
Gemma-2-IT-9B	42.94	40.92	46.97	51.09	41.93	43.03	37.98	45.97	42.52	56.13	36.30
GPT-3.5-Turbo-0125	23.53	23.19	30.50	38.32	25.71	28.74	24.03	27.31	25.80	39.41	26.47
GPT-4o	40.34	34.87	44.37	54.29	32.10	42.10	26.22	38.94	37.82	52.79	30.59

(c). 0-shot Exact Match (EM) scores (%) on x-x tasks

	de-de	es-es	vi-vi	zh-zh	hi-hi	ar-ar	el-el	ro-ro	ru-ru	th-th	tr-tr
LLaMA-3.1-8B	37.62	34.52	36.04	43.66	41.18	34.54	32.58	34.61	35.68	47.38	28.02
LLaMA-3.1-70B	62.66	59.02	61.54	55.26	56.81	63.43	56.52	59.81	55.99	70.20	57.39
Mistral-V0.3-7B	41.31	47.37	39.40	41.73	24.04	42.28	35.43	37.91	29.51	37.70	28.32
Qwen-2.5-7B	49.22	58.04	62.82	68.32	46.19	63.63	44.78	52.10	52.96	68.59	54.30
Qwen-2.5-72B	70.77	72.81	74.95	84.33	68.09	72.94	64.72	73.17	68.56	78.50	67.79
DeepSeek-V2-Lite-16B	12.56	8.24	7.64	9.73	8.35	6.98	9.59	9.39	7.79	9.03	8.89
Gemma-2-9B	4.16	20.57	13.66	14.74	15.58	2.53	3.07	6.96	18.84	14.14	23.02
LLaMA-2-Chat-7B	40.12	40.37	32.22	10.68	13.62	17.59	18.83	38.52	26.51	10.64	23.93
LLaMA-3.1-Instruct-8B	51.95	58.67	62.55	54.15	52.86	56.45	45.21	55.02	52.99	56.89	55.40
LLaMA-3.1-Instruct-70B	68.60	68.79	73.54	66.32	62.66	70.97	66.97	68.90	65.53	73.91	66.35
Mistral-V0.3-Instruct-7B	24.81	26.71	26.83	11.22	15.57	20.21	22.81	25.11	20.37	30.99	21.98
Qwen-2.5-Instruct-7B	57.65	57.89	56.71	57.29	50.26	53.38	50.18	59.65	53.25	63.81	55.85
Qwen-2.5-Instruct-72B	54.08	51.89	52.80	57.27	45.64	57.66	45.27	56.80	50.74	62.11	48.48
DeepSeek-V2-Chat-Lite-16B	35.78	34.43	34.37	15.75	25.36	29.23	33.21	32.04	34.46	28.91	28.62
Gemma-2-IT-9B	68.38	68.84	71.81	65.03	67.91	66.82	67.78	69.48	66.19	73.71	64.21
GPT-3.5-Turbo-0125	53.58	56.26	58.50	52.01	50.36	57.77	58.48	56.60	55.95	57.45	54.36
GPT-4o	67.90	67.79	71.45	65.72	61.33	69.67	62.12	66.61	66.59	74.43	62.68

(d). 0-shot F1 Scores on x-x tasks.

Table 5: 0-shot evaluation results on en-x and x-x tasks.

	en-en	en-de	en-es	en-vi	en-zh	en-hi	en-ar	en-el	en-ro	en-ru	en-th	en-tr
LLaMA-3.1-8B	75.97	45.97	43.90	50.71	47.70	48.41	50.98	50.40	47.22	49.12	59.77	44.94
LLaMA-3.1-70B	82.39	60.46	60.00	59.57	62.68	57.38	51.50	59.43	54.28	56.15	66.54	57.48
Mistral-V0.3-7B	79.57	66.94	67.57	59.82	53.66	48.36	52.56	47.83	67.30	70.62	48.56	62.94
Qwen-2.5-7B	62.42	56.68	56.45	59.15	58.84	55.85	62.94	50.62	54.43	58.37	61.61	57.72
Qwen-2.5-72B	86.03	77.24	79.22	80.16	80.14	79.09	78.70	76.72	80.41	80.77	78.48	77.16
DeepSeek-V2-Lite-16B	73.81	46.34	47.65	52.75	41.77	34.45	44.61	46.18	50.39	51.91	34.58	40.57
Gemma-2-9B	80.42	60.13	61.74	64.76	71.33	64.55	68.84	75.81	65.49	72.09	70.59	59.72
LLaMA-3.1-Instruct-8B	77.89	74.81	73.50	73.29	72.78	68.59	69.66	70.41	72.24	73.40	73.60	71.20
LLaMA-3.1-Instruct-70B	83.29	73.58	72.98	73.97	73.79	70.87	75.96	72.91	71.69	72.73	75.30	70.00
Mistral-V0.3-Instruct-7B	62.01	59.06	60.98	54.81	58.84	47.16	60.51	57.49	59.80	60.81	55.63	47.86
Qwen-2.5-Instruct-7B	81.83	77.37	77.02	76.06	78.82	73.70	76.11	74.70	76.80	77.44	77.05	75.69
Qwen-2.5-Instruct-72B	77.12	67.65	68.89	64.34	52.67	70.35	59.59	69.52	69.81	70.19	67.01	66.40
DeepSeek-V2-Chat-Lite-16B	70.30	55.96	58.96	51.21	62.05	48.39	52.04	54.80	52.86	57.04	50.01	51.01
Gemma-2-IT-9B	83.69	78.72	78.13	79.38	79.17	77.86	76.53	79.82	79.96	79.80	79.28	77.24
GPT-3.5-Turbo-0125	81.74	71.98	72.81	71.53	68.63	63.20	65.77	63.05	70.86	70.70	65.21	72.54
GPT-4o	83.29	78.31	74.51	80.29	79.40	77.64	78.29	80.03	78.10	80.23	79.56	80.00

(a). 2-shot F1 scores on en-x tasks.

	de-de	es-es	vi-vi	zh-zh	hi-hi	ar-ar	el-el	ro-ro	ru-ru	th-th	tr-tr
LLaMA-3.1-8B	71.67	74.17	73.19	64.96	71.91	68.20	69.35	74.65	67.17	70.89	66.89
LLaMA-3.1-70B	76.24	78.68	76.90	71.67	75.85	76.43	72.41	78.06	68.43	76.70	70.67
Mistral-V0.3-7B	71.02	72.91	69.91	66.65	56.73	58.25	62.81	71.73	63.57	62.08	58.44
Qwen-2.5-7B	58.74	57.36	72.29	73.38	63.74	72.74	67.24	58.82	64.12	77.89	60.84
Qwen-2.5-72B	81.29	81.15	82.82	89.08	79.12	79.53	77.83	82.18	77.49	85.00	77.29
DeepSeek-V2-Lite-16B	65.26	67.33	64.81	63.68	48.64	46.98	51.04	65.73	56.19	49.43	55.20
Gemma-2-9B	75.62	76.34	73.60	66.92	73.90	72.51	71.87	78.14	67.38	74.20	71.43
LLaMA-3.1-Instruct-8B	66.22	69.74	69.38	61.99	66.00	66.19	58.81	68.18	61.08	66.18	61.43
LLaMA-3.1-Instruct-70B	75.26	76.36	78.83	71.09	74.05	72.34	72.10	77.23	70.01	76.23	71.98
Mistral-V0.3-Instruct-7B	55.84	53.27	57.29	39.27	34.57	45.94	44.52	59.13	52.33	55.19	45.97
Qwen-2.5-Instruct-7B	73.75	75.24	78.26	70.21	67.08	70.82	67.65	75.61	67.99	73.89	67.23
Qwen-2.5-Instruct-72B	73.09	71.26	73.36	71.12	64.14	69.76	64.70	75.14	69.13	73.92	67.60
DeepSeek-V2-Chat-Lite-16B	56.24	59.33	56.18	50.06	41.19	42.46	44.03	54.63	56.10	40.71	48.52
Gemma-2-IT-9B	76.22	77.12	79.86	72.25	75.47	74.44	74.89	77.32	72.64	78.57	72.04
GPT-3.5-Turbo-0125	75.68	77.58	73.09	70.49	67.64	70.09	71.03	77.17	71.55	67.00	71.12
GPT-4o	76.94	78.78	77.25	71.37	72.02	76.17	73.40	77.66	77.34	80.00	71.58

(b). 2-shot F1 scores on x-x tasks

Table 6: Detailed 2-shot F1 scores on en-x and x-x tasks in each language.

	en-de	en-es	en-vi	en-zh	en-hi	en-ar	en-el	en-ro	en-ru	en-th	en-tr
LLaMA-3.1-8B	0.68	0.00	1.28	0.00	0.10	0.10	0.00	0.00	0.19	0.10	1.10
LLaMA-3.1-70B	60.20	54.59	63.73	48.90	69.09	66.86	56.83	67.88	56.69	56.09	61.50
Mistral-V0.3-7B	2.11	1.78	16.39	23.80	61.82	54.17	6.96	2.65	1.02	46.35	16.63
Qwen-2.5-7B	2.42	1.01	0.43	0.43	0.00	0.00	0.00	1.15	0.00	0.00	5.17
Qwen-2.5-72B	7.39	6.35	11.22	0.98	5.74	2.82	19.78	11.36	3.59	27.01	23.64
DeepSeek-V2-Lite-16B	8.90	1.87	1.69	26.75	46.80	4.76	3.68	2.27	0.47	35.17	10.31
Gemma-2-9B	0.00	0.10	9.78	2.37	0.50	0.20	0.34	3.15	0.00	2.23	2.37
LLaMA-2-Chat-7B	3.07	0.45	0.72	2.25	0.27	0.00	0.32	1.55	0.00	0.12	1.62
LLaMA-3.1-Instruct-8B	1.15	1.87	0.37	0.36	0.59	0.00	0.00	1.75	0.36	0.37	3.01
LLaMA-3.1-Instruct-70B	0.79	0.76	0.00	0.00	0.68	0.00	0.00	0.00	0.00	0.00	0.34
Mistral-V0.3-Instruct-7B	4.30	1.73	12.79	0.00	0.35	0.00	0.44	3.12	0.00	0.67	7.03
Qwen-2.5-Instruct-7B	1.89	0.71	0.76	0.38	0.00	0.00	0.00	0.68	0.00	0.00	2.91
Qwen-2.5-Instruct-72B	6.31	1.13	5.68	9.73	2.46	7.51	2.61	4.41	1.96	0.35	8.18
DeepSeek-V2-Chat-Lite-16B	5.43	2.35	0.90	0.41	4.47	1.33	0.94	4.09	0.65	0.66	4.74
Gemma-2-IT-9B	0.96	0.00	0.31	0.00	0.00	0.00	0.00	0.28	0.00	0.00	0.28
GPT-3.5-Turbo-0125	1.12	0.36	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.32
GPT-4o	0.39	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.38	0.00	0.00

Table 7: 2-shot language error rates (%) on en-x xMRC tasks.

	en-en	en-de	en-es	en-vi	en-zh	en-hi	en-ar	en-el	en-ro	en-ru	en-th	en-tr
LLaMA-3.1-8B	5.60	6.77	8.55	10.97	5.00	9.59	8.37	10.56	9.64	7.94	9.74	10.52
LLaMA-3.1-70B	1.20	1.98	2.41	2.19	2.62	4.12	2.33	2.35	1.71	1.57	2.70	4.94
Mistral-V0.3-7B	0.49	4.85	4.93	17.66	10.00	32.00	13.97	18.50	4.86	5.25	26.43	18.33
Qwen-2.5-7B	1.51	2.03	1.30	5.72	1.96	2.26	3.11	6.09	2.10	3.99	2.85	4.77
Qwen-2.5-72B	0.00	1.19	1.35	0.97	1.82	3.69	2.62	3.16	1.91	2.94	4.93	3.29
DeepSeek-V2-Lite-16B	1.87	3.26	2.33	11.97	2.39	20.61	10.10	8.04	4.27	2.76	15.97	11.23
Gemma-2-9B	1.02	1.34	1.38	5.39	2.60	6.51	4.69	4.98	3.68	2.64	5.05	6.98
LLaMA-2-Chat-7B	6.18	9.59	16.79	22.45	17.69	60.33	55.46	46.73	10.01	8.93	62.89	21.38
LLaMA-3.1-Instruct-8B	0.85	2.51	1.73	1.36	1.99	1.66	2.58	2.37	6.48	2.67	1.03	3.43
LLaMA-3.1-Instruct-70B	1.85	1.06	0.68	1.78	2.94	1.53	1.54	1.75	2.91	2.07	2.26	2.10
Mistral-V0.3-Instruct-7B	1.77	2.03	2.18	7.79	2.25	1.81	4.00	1.80	2.71	2.78	5.61	3.37
Qwen-2.5-Instruct-7B	2.75	2.47	3.20	3.42	4.04	2.78	5.38	3.25	2.36	2.56	2.04	3.80
Qwen-2.5-Instruct-72B	0.38	1.58	1.09	1.19	0.54	1.71	1.68	1.35	2.77	1.65	1.30	3.00
DeepSeek-V2-Chat-Lite-16B	0.58	4.28	2.93	9.77	3.27	6.80	4.98	6.31	6.69	8.60	6.15	5.39
Gemma-2-IT-9B	1.95	2.26	1.76	1.92	3.30	3.03	3.73	1.47	3.43	2.84	0.94	2.52
GPT-3.5-Turbo-0125	0.00	0.65	2.35	1.61	3.49	2.15	2.60	1.44	2.89	3.70	5.17	4.73
GPT-4o	0.00	0.45	0.86	1.56	0.50	0.00	0.92	1.00	3.57	2.56	0.00	4.00

Table 8: 2-shot generation failure error rates (%) on en-x xMRC tasks.

	en-en	en-de	en-es	en-vi	en-zh	en-hi	en-ar	en-el	en-ro	en-ru	en-th	en-tr
LLaMA-3.1-8B	4.80	4.03	1.83	7.08	2.83	5.38	3.56	2.99	2.45	2.70	6.42	7.81
LLaMA-3.1-70B	0.60	1.32	2.19	1.31	2.14	3.50	2.15	1.93	1.33	1.18	2.43	4.12
Mistral-V0.3-7B	0.00	3.77	4.11	16.78	9.62	31.16	12.50	18.17	4.32	4.94	26.43	16.90
Qwen-2.5-7B	1.29	1.11	1.11	4.45	1.57	1.69	2.63	5.41	1.75	3.59	2.63	4.17
Qwen-2.5-72B	0.00	1.19	1.35	0.97	1.82	3.69	2.62	3.16	1.91	2.45	4.93	2.88
DeepSeek-V2-Lite-16B	1.87	2.77	2.00	10.31	2.24	19.72	8.24	6.27	3.20	2.02	15.84	10.79
Gemma-2-9B	1.02	1.12	0.92	5.16	1.95	5.51	3.52	3.83	2.89	2.31	3.79	6.54
LLaMA-2-Chat-7B	5.41	8.78	16.60	3.98	3.23	60.21	55.46	45.03	3.28	8.74	6.37	20.64
LLaMA-3.1-Instruct-8B	0.00	2.15	0.69	1.02	1.99	0.83	2.01	1.78	5.83	2.34	1.03	3.43
LLaMA-3.1-Instruct-70B	0.00	0.35	0.34	1.42	2.67	1.22	0.77	0.70	1.94	1.38	1.51	0.90
Mistral-V0.3-Instruct-7B	0.00	1.22	1.09	7.07	1.02	1.36	2.67	1.20	0.39	1.93	4.84	1.53
Qwen-2.5-Instruct-7B	1.65	1.65	2.80	3.04	3.14	1.74	3.85	3.25	1.97	2.13	2.04	3.80
Qwen-2.5-Instruct-72B	0.00	1.05	0.82	0.95	0.54	0.57	1.26	0.81	1.66	1.10	0.78	2.00
DeepSeek-V2-Chat-Lite-16B	0.58	3.70	2.30	8.25	2.80	5.67	4.81	5.75	6.19	7.89	5.65	4.72
Gemma-2-IT-9B	0.65	1.36	0.88	0.96	2.83	3.03	2.90	0.98	2.45	1.90	0.47	0.84
GPT-3.5-Turbo-0125	0.00	0.65	2.35	1.61	2.91	1.67	2.60	1.44	2.89	3.70	4.65	4.05
GPT-4o	0.00	0.45	0.86	1.56	0.50	0.00	0.92	1.00	3.57	2.56	0.00	4.00

(a). Gibberish error.

	en-en	en-de	en-es	en-vi	en-zh	en-hi	en-ar	en-el	en-ro	en-ru	en-th	en-tr
LLaMA-3.1-8B	0.00	0.16	0.00	0.00	0.00	0.34	0.18	0.18	0.00	0.00	0.22	0.00
LLaMA-3.1-70B	0.00	0.00	0.00	0.22	0.00	0.00	0.18	0.21	0.00	0.00	0.27	0.41
Mistral-V0.3-7B	0.00	0.00	0.00	0.00	0.19	0.17	1.10	0.00	0.00	0.31	0.00	0.00
Qwen-2.5-7B	0.22	0.92	0.19	0.21	0.39	0.38	0.48	0.34	0.35	0.40	0.22	0.20
Qwen-2.5-72B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DeepSeek-V2-Lite-16B	0.00	0.00	0.00	0.37	0.00	0.38	0.62	0.16	0.00	0.37	0.13	0.00
Gemma-2-9B	0.00	0.00	0.00	0.00	0.65	0.25	0.88	0.38	0.00	0.33	0.63	0.00
LLaMA-2-Chat-7B	0.77	0.61	0.19	18.31	14.31	0.12	0.00	1.70	6.56	0.19	56.42	0.59
LLaMA-3.1-Instruct-8B	0.85	0.36	1.04	0.00	0.00	0.55	0.57	0.59	0.65	0.00	0.00	0.00
LLaMA-3.1-Instruct-70B	1.23	0.00	0.00	0.00	0.27	0.00	0.00	0.35	0.32	0.00	0.00	0.60
Mistral-V0.3-Instruct-7B	1.77	0.61	0.65	0.54	0.82	0.45	1.33	0.60	2.13	0.64	0.00	1.84
Qwen-2.5-Instruct-7B	0.55	0.41	0.40	0.38	0.45	0.35	1.15	0.00	0.00	0.43	0.00	0.00
Qwen-2.5-Instruct-72B	0.38	0.53	0.27	0.24	0.00	1.14	0.21	0.54	0.83	0.55	0.52	1.00
DeepSeek-V2-Chat-Lite-16B	0.00	0.19	0.21	1.01	0.47	0.81	0.17	0.56	0.33	0.53	0.33	0.67
Gemma-2-IT-9B	0.00	0.45	0.00	0.00	0.00	0.00	0.00	0.00	0.49	0.47	0.00	0.84
GPT-3.5-Turbo-0125	0.00	0.00	0.00	0.00	0.29	0.24	0.00	0.00	0.00	0.00	0.26	0.00
GPT-4o	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

(b). Refusal error.

	en-en	en-de	en-es	en-vi	en-zh	en-hi	en-ar	en-el	en-ro	en-ru	en-th	en-tr
LLaMA-3.1-8B	0.80	2.58	6.72	3.89	2.17	3.87	4.63	7.39	7.19	5.24	3.10	2.71
LLaMA-3.1-70B	0.60	0.66	0.22	0.66	0.48	0.62	0.00	0.21	0.38	0.39	0.00	0.41
Mistral-V0.3-7B	0.49	1.08	0.82	0.88	0.19	0.67	0.37	0.33	0.54	0.00	0.00	1.43
Qwen-2.5-7B	0.00	0.00	0.00	1.06	0.00	0.19	0.00	0.34	0.00	0.00	0.00	0.40
Qwen-2.5-72B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.49	0.00	0.41
DeepSeek-V2-Lite-16B	0.00	0.49	0.33	1.29	0.15	0.51	1.24	1.61	1.07	0.37	0.00	0.44
Gemma-2-9B	0.00	0.22	0.46	0.23	0.00	0.75	0.29	0.77	0.79	0.00	0.63	0.44
LLaMA-2-Chat-7B	0.00	0.20	0.00	0.16	0.15	0.00	0.00	0.00	0.17	0.00	0.10	0.15
LLaMA-3.1-Instruct-8B	0.00	0.00	0.00	0.34	0.00	0.28	0.00	0.00	0.00	0.33	0.00	0.00
LLaMA-3.1-Instruct-70B	0.62	0.71	0.34	0.36	0.00	0.31	0.77	0.70	0.65	0.69	0.75	0.60
Mistral-V0.3-Instruct-7B	0.00	0.20	0.44	0.18	0.41	0.00	0.00	0.00	0.19	0.21	0.77	0.00
Qwen-2.5-Instruct-7B	0.55	0.41	0.00	0.00	0.45	0.69	0.38	0.00	0.39	0.00	0.00	0.00
Qwen-2.5-Instruct-72B	0.00	0.00	0.00	0.00	0.00	0.00	0.21	0.00	0.28	0.00	0.00	0.00
DeepSeek-V2-Chat-Lite-16B	0.00	0.39	0.42	0.51	0.00	0.32	0.00	0.00	0.17	0.18	0.17	0.00
Gemma-2-IT-9B	1.30	0.45	0.88	0.96	0.47	0.00	0.83	0.49	0.49	0.47	0.47	0.84
GPT-3.5-Turbo-0125	0.00	0.00	0.00	0.00	0.29	0.24	0.00	0.00	0.00	0.00	0.26	0.68
GPT-4o	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

(c). Blank error.

Table 9: Detailed percentages (%) of generation failure error types (gibberish error, refusal error, and blank error) on 2-shot en-x tasks.

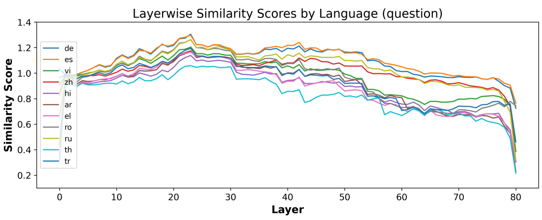
	en-en	en-de	en-es	en-vi	en-zh	en-hi	en-ar	en-el	en-ro	en-ru	en-th	en-tr
LLaMA-3.1-Tuned-8B	78.80	74.07	69.85	72.12	71.03	69.23	67.57	69.35	71.54	71.86	71.12	71.02

(a). en-x

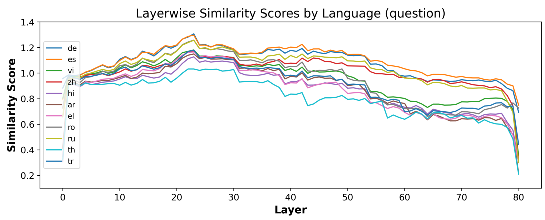
	de-de	es-es	vi-vi	zh-zh	hi-hi	ar-ar	el-el	ro-ro	ru-ru	th-th	tr-tr
LLaMA-3.1-Tuned-8B	69.28	71.16	73.73	63.71	67.48	64.35	61.84	73.03	65.66	62.34	62.63

(b). x-x

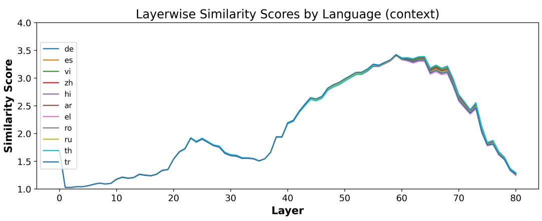
Table 10: 2-shot F1 scores on en-x and x-x tasks for our finetuned LLaMA-3.1-8B.



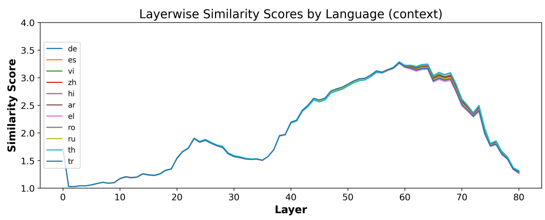
(a). Question hidden state similarity for balanced samples.



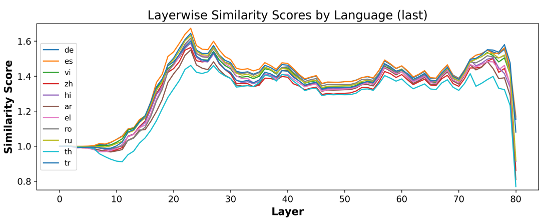
(b). Question hidden state similarity for en-superior samples.



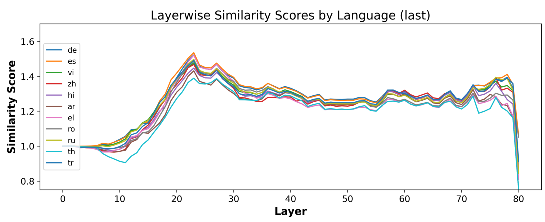
(c). Context hidden state similarity for balanced samples.



(d). Context hidden state similarity for en-superior samples.

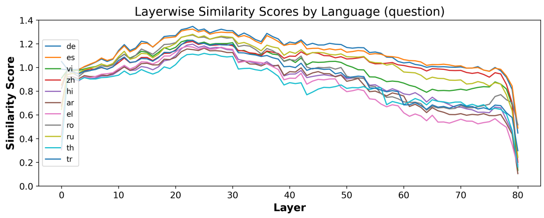


(e). Last-input-token hidden state similarity for balanced samples.

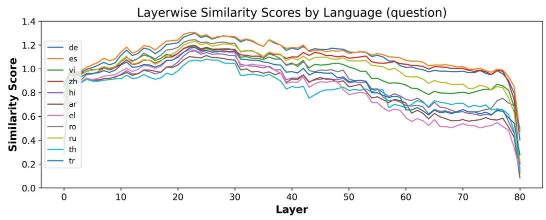


(f). Last-input-token hidden state similarity for en-superior samples.

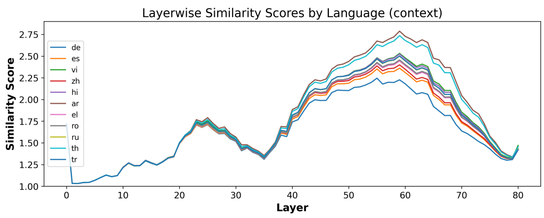
Figure 16: Hidden state similarity between English and other languages on different parts of the selected samples in each layer of the LLaMA-3.1-70B model.



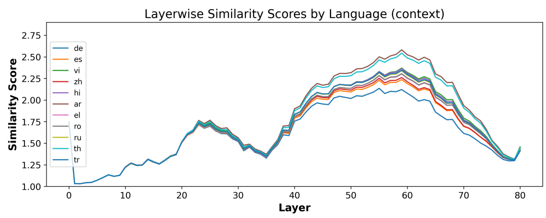
(a). Question hidden state similarity for balanced samples.



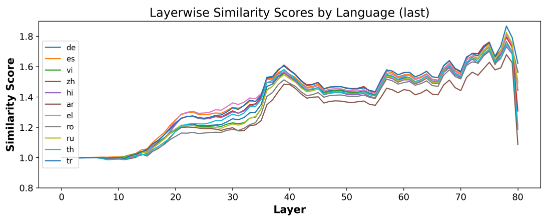
(b). Question hidden state similarity for en-superior samples.



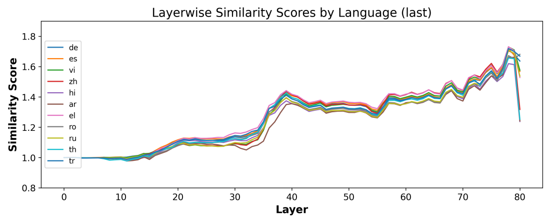
(c). Context hidden state similarity for balanced samples.



(d). Context hidden state similarity for en-superior samples.

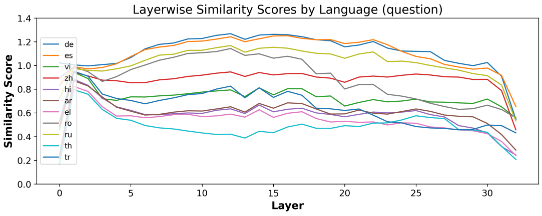


(e). Last-input-token hidden state similarity for balanced samples.

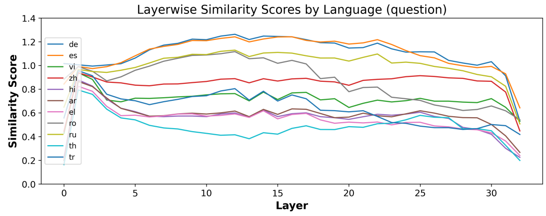


(f). Last-input-token hidden state similarity for en-superior samples.

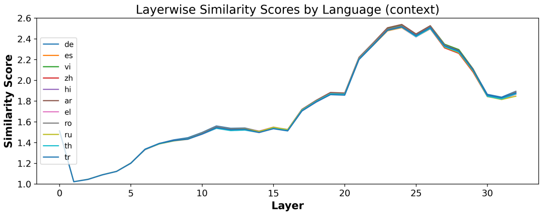
Figure 17: Hidden state similarity between English and other languages on different parts of the selected samples in each layer of the LLaMA-3.1-Instruct-70B model.



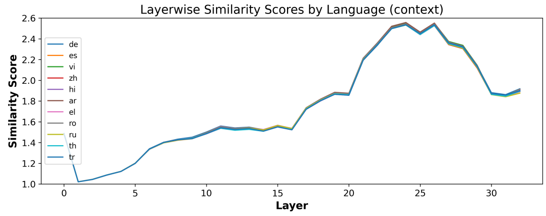
(a). Question hidden state similarity for balanced samples.



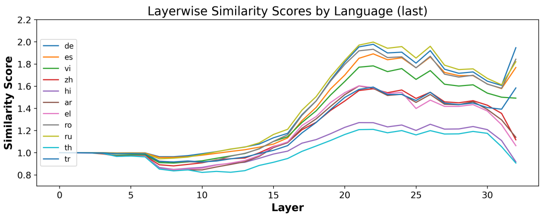
(b). Question hidden state similarity for en-superior samples.



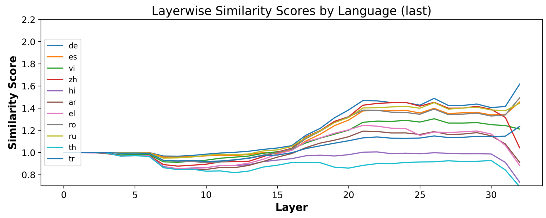
(c). Context hidden state similarity for balanced samples.



(d). Context hidden state similarity for en-superior samples.

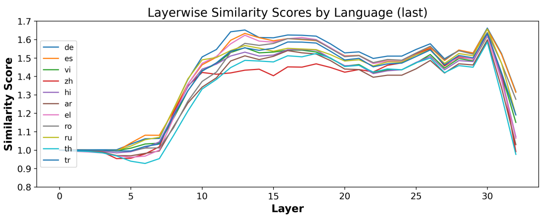


(e). Last-input-token hidden state similarity for balanced samples.

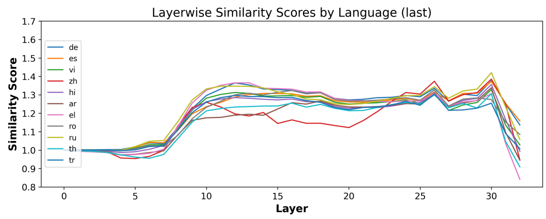


(f). Last-input-token hidden state similarity for en-superior samples.

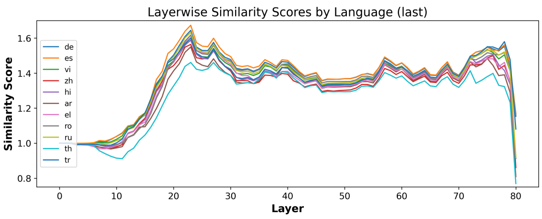
Figure 18: Hidden state similarity between English and other languages on different parts of the selected samples in each layer of the LLaMA-2-Chat-7B model.



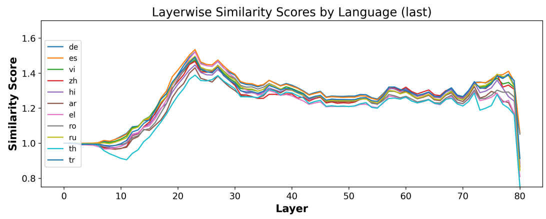
(a). "Balanced" samples for LLaMA-3.1-8B.



(b). "En-superior" samples for LLaMA-3.1-8B.

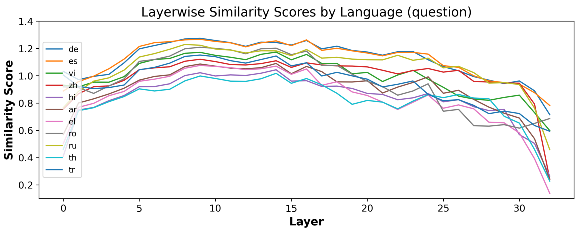


(c). "Balanced" samples for LLaMA-3.1-70B.

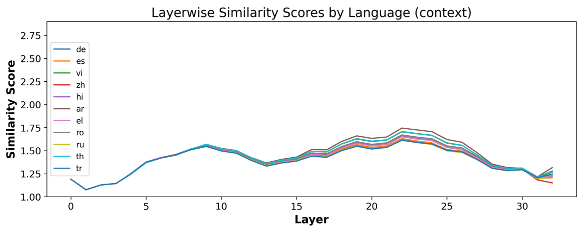


(d). "En-superior" samples for LLaMA-3.1-70B.

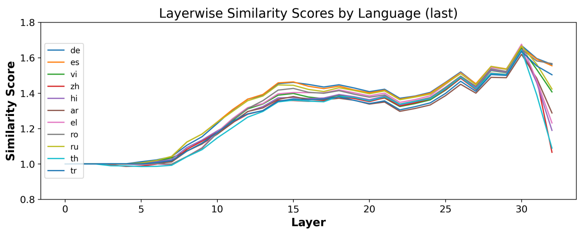
Figure 19: Last-input-token hidden state similarity between English and other languages in each layer of the base LLaMA models.



(a). Question hidden state similarity.



(b). Context hidden state similarity.



(c). Last-input-token hidden state similarity.

Figure 20: Hidden state similarity between English and other languages on different parts of the balanced samples in each layer for our finetuned LLaMA-3.1-8B model.