Disentangling Language and Culture for Evaluating Multilingual Large Language Models

Jiahao Ying [1, 5]*, Wei Tang [2, 5], Yiran Zhao [3, 5], Yixin Cao [4], Yu Rong [5], Wenxuan Zhang [6]†

[1] Singapore Management University
[2] University of Science and Technology of China
[3] National University of Singapore
[4] Institute of Trustworthy Embodied AI, Fudan University
[5] DAMO Academy, Alibaba Group
[6] Singapore University of Technology and Design

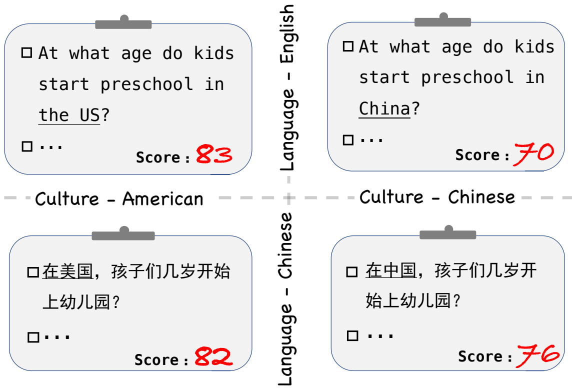
Abstract

This paper introduces a Dual Evaluation Framework to comprehensively assess the multilingual capabilities of LLMs. By decomposing the evaluation along the dimensions of linguistic medium and cultural context, this framework enables a nuanced analysis of LLMs' ability to process questions within both native and cross-cultural contexts cross-lingually. Extensive evaluations are conducted on a wide range of models, revealing a notable "Cultural-Linguistic Synergy" phenomenon, where models exhibit better performance when questions are culturally aligned with the language. This phenomenon is further explored through interpretability probing, which shows that a higher proportion of specific neurons are activated in a language's cultural context. This activation proportion could serve as a potential indicator for evaluating multilingual performance during model training. Our findings challenge the prevailing notion that LLMs, primarily trained on English data, perform uniformly across languages and highlight the necessity of culturally and linguistically model evaluations. Our code can be found at `https://yingjiahao14.github.io/Dual-Evaluation/`.

1 Introduction

With the rapid development of large language models (LLMs), increasing efforts have been made to make these models beneficial for people worldwide. To achieve this, non-English corpora are also incorporated into the training data, enabling LLMs to understand and generate text in various languages (i.e., multilingual capabilities) (Xue et al., 2021; Grattafiori et al., 2024; OpenAI et al., 2024; Nguyen et al., 2024; Zhang et al., 2024).

To evaluate the LLMs' multilingual capabilities, researchers primarily rely on translating English-centric benchmarks into target languages, such as translating MMLU (Hendrycks et al., 2021) into MMMLU (OpenAI, 2024). While this approach allows for efficient cross-lingual comparisons, it limits the evaluation to scenarios rooted in English-speaking cultural contexts, as the original data was predominantly collected from perspectives prevalent in English-speaking countries. In contrast, recent work has developed culture-specific benchmarks such as M3Exam (Zhang et al., 2023) and BLEnD (Myung et al., 2024), where evaluation data are sourced from authentic, real-world scenarios in native-speaking regions. While these better capture the majority of local usage, they also overlook that multilingual users frequently ask questions across cultural boundaries. For example, a Spanish speaker might inquire about Chinese tea usage in Spanish, while a user from China could seek details about Diwali celebrations in Chinese. These existing evaluations on multilingual capabilities, however, treat language and cultural context as inseparable dimensions, restricting analyses to single-language scenarios.



□ At what age do kids start preschool in the US?
□ ...

□ At what age do kids start preschool in China?
□ ...

□在美国，孩子们几岁开始上幼儿园？
□...

□ 在中国，孩子们几岁开始上幼儿园？
□ ...

Figure 1: Dual Evaluation Framework for evaluating multilingual capabilities of LLMs. The figure is divided into four quadrants, each showing the model's performance on questions framed in different **languages** (horizontal-axis) and **cultural contexts** (vertical-axis). The score refers to the aggregated performance of the model Claude-3.5-Sonnet on these four question sets.

[*] This work was performed when Jiahao Ying, Wei Tang and Yiran Zhao were interns at Alibaba DAMO Academy.

[†] Corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 22230–22251
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

To comprehensively evaluate multilingual capability, especially considering the real-world usage, we propose a Dual Evaluation framework in this paper, which decomposes the multilingual capability evaluation along two critical dimensions: (1) **linguistic medium** (the language used for questioning) and (2) **cultural context** (the regional and cultural knowledge being tested). As illustrated in Figure 1 through a preschool enrollment example, this framework generates four distinct evaluation scenarios from a single question template. This structured decomposition enables multiple essential multilingual capability assessments, including native cultural-linguistic alignment (same language and culture), cross-lingual understanding (different language, same culture), and cross-cultural ability (same language, different culture).

With such a dual evaluation framework design, we construct a dataset by adopting and extending the BLEnD dataset (Myung et al., 2024), which contains every-day questions across different cultural contexts. We then evaluate a wide range of open-source and close-source models with this newly constructed benchmark. Our findings indicate that: 1) Models generally perform better on scenarios rooted in English-speaking culture, a pattern that persists cross-lingually (Section 3.2), and 2) LLMs perform better when questions are posed in the language that corresponds to the cultural context of the question, rather than in English (Section 3.3). The second finding, in particular, draws our attention because most existing models are primarily trained on English data and have demonstrated strong performance in other multilingual evaluations like MMMLU. However, when faced with real-world culturally relevant questions in the corresponding language, these models perform better in that language than in English. We refer to this phenomenon as "**Cultural-Linguistic Synergy**" (as shown in Figure 1, Claude-3.5-Sonnet has better performance on the Chinese test than the English test when asking about Chinese culture questions, vice versa).

To understand the underlying causes of this phenomenon, we conduct interpretability probing by analyzing the activation status of neurons when answering questions in different languages and cultural contexts, we find that: 1) The proportion of specific neurons tends to be higher when the question is in the corresponding language and cultural context, which could explain the observed "Cultural-Linguistic Synergy" (Section 4.3); 2) Additionally, this proportion of specific neurons could serve as a potential indicator for comparing multilingual capabilities during model training (Section 4.3.1); 3) The number of neurons activated in the model is strongly correlated with the model's performance in the corresponding language. Specifically, when the question is in the English-speaking cultural context, the model tends to activate more neurons, leading to better performance (Section 4.3.2).

Our main contributions can be summarized as:

- We propose a Dual Evaluation Framework, which decomposes the multilingual capability evaluation along two critical dimensions, linguistic medium and cultural context.
- Through extensive experiments, we find the Cultural-Linguistic Synergy phenomenon: the selected models perform better on native cultural scenario questions when asked in the corresponding language, compared to English.
- We demonstrate that the proportion of specific neurons activated for a given language can explain the observed Cultural-Linguistic Synergy, and that this proportion can serve as a potential indicator for comparing multilingual capabilities.

2 Dual Evaluation Framework

To comprehensively assess the multilingual capabilities of LLMs, we propose a Dual Evaluation framework that evaluates along two critical dimensions: (1) **linguistic medium** (the language used to pose questions) and (2) **cultural context** (the regional and cultural knowledge being tested). This dual-axis approach reflects three fundamental requirements for real-world applications: first, the ability to handle native language queries within their cultural context (e.g., answering "*What is a common children's snack in Spain?*" in Spanish), second, the capability for cross-lingual understanding, (e.g., answering questions about Spanish culture in Spanish and English); and third the capability to address cross-cultural inquiries through a single linguistic medium (e.g., answering "*What is a traditional festival in Japan?*" in English). By evaluating LLMs in both dimensions, we can measure how well models adapt to language-specific usage scenarios while maintaining cross-lingual and cross-cultural competence.

Formally, we represent the evaluation question as $Q_{i,j}$, where i denotes the cultural context and j specifies the linguistic medium of the question. To construct a question set, we conduct a systematic adaptation of a culture-specific benchmark, BLEnD (Myung et al., 2024), testing everyday knowledge across diverse cultures and languages. Specifically, for native cultural-linguistic pairs (i.e., $Q_{i,i}$), we used the localized questions in BLEnD, which are constructed based on a template question with three key modifications: replacing country or region references, adapting phrasing to match linguistic norms, and curating culture-specific answer sets. The localized evaluation sets $Q_{i,i}$ for language i are denoted by:

$$Q_{i,i} = \{(q_i, a_i) | (q_i, a_i) = Adapt_i(q), q \in \text{Template}\}, \quad (1)$$

where $Adapt_i$ represents the localized modifications for language i, and q represents the template question from the Template set in BLEnD. For example, by adapting the original question *"What is the most popular sports team in your country?"* into *"What is the most popular sports team in the US?"*, where $i = en$, we can test the model's ability to handle English in the US context. Using these adapted questions for different languages, we can assess the model's ability to handle native-language queries within various cultural environments.

On the other hand, to assess the model's ability to handle questions across multiple cultural contexts when asked in a single language, we extend the $Q_{i,i}$ sets into localized transformations $Q_{i,j}$ for each language pair (i, j), where $i \neq j$. The original BLEnD includes, for each language-specific evaluation set $Q_{i,j}$ (except for English), an English translation evaluation dataset $Q_{i,en}$. Specifically:

$$Q_{i,en} = \{\text{Trans}_{en}(q_i, a_i) \mid (q_i, a_i) \in Q_{i,i}, i \neq \text{en}\}. \quad (2)$$

For other language pairs (i, j), we use the GPT-4o model, known for its strong multilingual capabilities, to construct these cross-lingual datasets.

$$Q_{i,j} = \{\text{Trans}_j(q_i, a_i) \mid (q_i, a_i) \in Q_\text{i}, j \neq \text{en}\}. \quad (3)$$

This setup enables assessing how well the model can adapt to answering questions posed in one language about the cultural context of another.

Combining the two evaluation scenarios, the complete evaluation set Q is thus represented as:

$$Q = \textstyle\bigcup_{i \neq j} Q_{i,i} \cup Q_{j,j} \cup Q_{i,j} \cup Q_{j,i}. \quad (4)$$

Linguistic Medium	Cultural Context	# Data Sample
English (en)	**United States (US)**, CN, ES, ID, KR, IR, JB	3,500
Chinese (zh)	**China (CN)**, US	1,000
Spanish (es)	**Spain (ES)**, US	1,000
Indonesian (id)	**Indonesia (ID)**, US	1,000
Korean (ko)	**South Korea (KR)**, US	1,000
Persian (fa)	**Iran (IR)**, US	1,000
Sundanese (su)	**West Java (JB)**, US	1,000
Total		9,500

Table 1: Overview of the evaluation dataset, detailing the language, cultural context of certain countries/regions, and the number of data samples. **Bolded** countries/regions indicate where the corresponding language is spoken natively, while the others are transformed for cross-cultural evaluation. Each language has 500 data samples per country/region. The parts we added to the original BLEnD are marked in green.

This Dual Evaluation framework, where questions are tailored to the linguistic medium and the corresponding cultural contexts of usage, not only assesses LLMs' multilingual abilities from both native usage scenarios ($Q_{i,i}$) and cross-cultural contexts ($Q_{j,i}$) but also employs a completely dual-format question approach. Specifically, $Q_{i,i}$ and $Q_{j,j}$ are constructed using the same template question, and tailored to different linguistic and cultural contexts. This approach allows us to quantitatively compare the multilingual capabilities cross-culturally within the same language (by comparing $Q_{i,i}$ and $Q_{j,i}$), and cross-lingually (by comparing $Q_{i,i}$ with $Q_{j,j}$, or $Q_{i,i}$ with $Q_{i,j}$). An example of this dual evaluation sample is shown in Figure 1, and the details of the completion for $Q_{i,j}$ and human evaluation for the quality are presented in Appendix A.4.

3 Multilingual Capabilities Evaluation

3.1 Experiment setting

We select a wide range of LLMs of different sizes to evaluate their multilingual capabilities, including GPT-4o (OpenAI et al., 2024), Claude-3.5-Sonnet (Anthropic, 2024), CommandR (Cohere, 2024), Llama-3-8B-Instruction, Llama-3-70B-Instruction(Grattafiori et al., 2024), Gemma-2-9B (Team et al., 2024), Qwen2.5-7B-Instruct (Qwen et al., 2025), and Bloomz-7B (Muennighoff et al., 2022). The experiment is conducted across seven languages, with cultural content sourced from one of the typical countries where each language is widely spoken. Considering the current performance of the model (primarily strong in English), and taking cost and time constraints into account, we only construct evaluation data $Q_{i,j}$ for the language pairs (i, j) where either language i or j is English. The language and culture information for the evaluation dataset are provided in Table 1. The questions used in evaluation are short-answer questions (SAQs) aligned with the BLEnD (Myung et al., 2024) benchmark. We apply non-weighted scores for the evaluation metrics. During inference, all models are tested in a zero-shot setting. The question prompts are derived from the original BLEnD instruction set. More details are shown in Appendix B.

3.2 Finding 1: LLMs' Performance Declines as the Cultural Context Shifts from English to Cross-Cultural Scenarios.

Using our Dual Evaluation Framework, we evaluate the performance of the selected LLMs. As mentioned in Section 2, one of the key advantages of this framework is its ability to compare models' adaptability in cross-cultural contexts (comparing $Q_{i,i}$ & $Q_{j,i}$), given that the questions are presented in a dual format. Since most of the selected models primarily use English as their training corpus, we first compare the performance on $Q_{en,en}$ and $Q_{i,en}$ (where $i \neq$ en). The results in Figure 2 (full result in Append A.2), for each bar, represent the performance in specific culture-contexts. By comparing the bars' height across different context, we observe that models perform best for English-speaking culture contexts, when asking questions in English, and performance declines in other language-speaking cultural context, with the drop becoming more pronounced as the language's resource availability decreases.

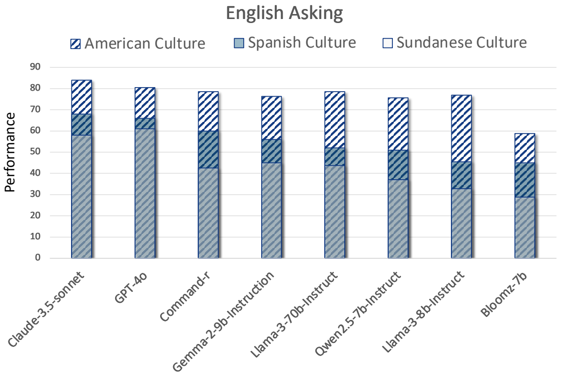


Figure 2: The performance of the selected models on American, Spanish, and Sundanese culture questions when asked in English. We find that models perform best on American culture.

Question Content	Spanish Culture	American Culture
Claude-3.5-Sonnet	81.0	**82.0**
GPT-4o	76.5	**77.6**
Command-r	69.9	**73.4**
Gemma-2-9b-Instruction	70.9	**72.7**
Llama-3-70b-Instruct	72.0	**79.6**
Qwen2.5-7b-Instruct	62.0	**70.5**
Llama-3-8b-Instruct	58.9	**74.5**
Bloomz-7b	**53.6**	52.8

Table 2: The performance of selected models on everyday questions about Spain and the US when asked in **Spanish**. Generally, models perform best when asking questions about US culture in Spanish.

While this trend echoes previous findings with translated datasets from English culture (Myung et al., 2024; OpenAI, 2024), it raises a further question: Does this phenomenon also hold in other languages? To explore more, we expanded the comparison to include $Q_{i,i}$ and $Q_{i,en}$, especially when $i \neq$ en. The results for $i =$ es (Spanish) are shown in Table 2, considering the high availability of resources of it. Additional results, demonstrating the same phenomena, are available in Appendix A.2. The results indicate that, in general, the selected models perform better on English-speaking culture questions compared to other languages when asked in the respective language. Since the training data for these models is not fully open-source, we hypothesize that, for each language in the training corpus, the models are trained on a larger volume of English-language usage scenarios. As a result, the models exhibit better performance on English-speaking culture questions across all languages. In the following interpretability Section 4.3.2, we delve deeper into the model's internal workings to explore the reasons behind this observed behavior.

3.3 Finding 2: LLMs perform better when asked in the corresponding language.

In addition to enabling comparisons of behavior in cross-cultural contexts within the same language (as discussed in Section 3.2), we can evaluate how models perform cross-lingually using Dual Evaluation Framework. Specifically, by comparing $Q_{i,i}$ and $Q_{i,j}$ $(i \neq j)$, we get the result shown in Figure 3. We surprisingly find that asking culture-related questions in the corresponding language outperforms asking in English, as indicated by the bars with patterns being higher than those without. Specifically, across the eight selected models, the average performance for questions related to Chinese culture when asked in Chinese exceeded that

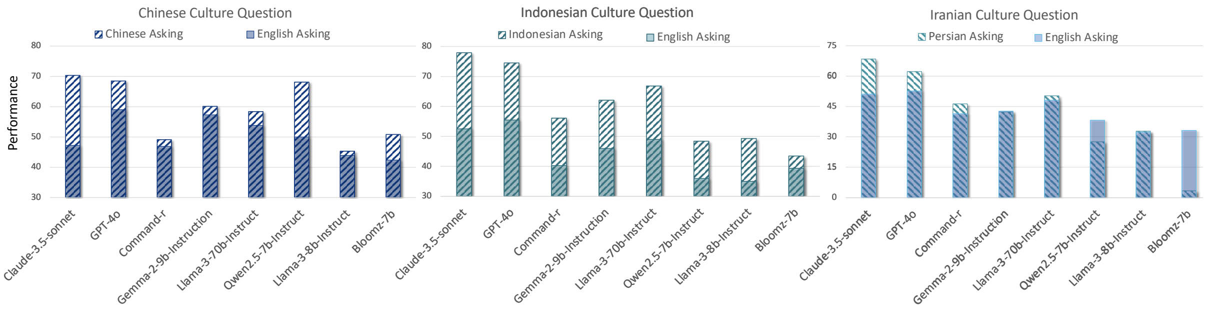


Figure 3: The performance of the selected models on Chinese, Indonesian, and Iranian culture questions when asked in the corresponding language versus English. We generally find that models perform better when the questions are posed in the corresponding major spoken language compared to English.

of asking in English by 8.8 points. Similarly, questions related to Indonesian culture posed in Indonesian outperformed those asked in English by 15.7 points. While this advantage diminishes when dealing with lower-resource languages. For instance, in Persian, the performance gap is -0.95 points. This can be attributed to models like Bloomz-7B, which have limited or no training data for Persian, resulting in better performance when asking questions in English instead. On the other hand, this corresponding advantage also appears in American culture questions, as shown in Figure 17.

From these observations, we can generally summarize that asking culture-specific questions in their corresponding language tends to outperform answering them in English. We refer to this counterintuitive phenomenon as **"Cultural-Linguistic Synergy"**. That is, aligning the cultural context with the appropriate linguistic medium, we can achieve superior performance — even for models primarily trained on English data, which perform better on English-specific tasks than on other language benchmarks like MMMLU and translated GSM8K (Shi et al., 2022). An intuitive explanation for this Cultural-Linguistic Synergy could lie in the training data. However, due to the lack of access to the training data and the massive scale of the training corpus, further exploration in this direction can be challenging. Thus, in the following sections, we proceed with interpretability analysis to understand the mechanisms of this Cultural-Linguistic Synergy, beginning with preliminary insights.

4 Interpreting Cultural-Linguistic Synergy

4.1 Preliminary

Neurons in FFN Module: Recent interpretability studies suggest that factual knowledge is stored in the FFN memories and represented by neurons in the network (Geva et al., 2021). Given the input token x, the FFN module of layer l in a decoder-only Transformer can be represented as (outer activation functions and bias terms are omitted for clarity):

$$\text{FFN}^l(h^l) = \left(W^l_{down} \cdot Activation(W^l_{up} \cdot h^l)\right) \quad (5)$$

where h^l is the input to the FFN, W^l_{up} and W^l_{down} are the weight matrices, and $Activation$ is the activation function. Following previous works, the i-th element of $Activation(W^l_{up} \cdot h^l) \in \mathbb{R}^{dm}$ is considered the i-th neuron in layer l (a simple illustration of neuron in Figure 4). The value of this neuron for the input token x can be represented by its corresponding activation value $v^x_{(i,l)}$.

Key Neuron Set: Following previous work, neurons with higher activation values when answering a question are considered more important (Tang et al., 2024a; Zhao et al., 2024; Hong et al., 2024; Cao et al., 2025b). Therefore, given a question q , we can identify the "Key Neurons" N_q by selecting neurons that are highly activated in the model's response $r = \{r_1, r_2, \ldots, r_n\}$, where r_i denotes the i-th token in the response, based on a threshold function ($threshold$) as:

$$N_q = \left\{(i,l) \mid \text{v}^{r_i}_{(i,l)} > threshold, r_i \in r\right\} \quad (6)$$

By aggregating these key neurons for each question q in the dataset $Q = \{q\}$, we obtain the Key Neuron set for the entire dataset Q as (ref Figure 4 for illustration for getting key neurons):

$$N_Q = \{(i,l) | (i,l) \in N_q, q \in Q\} \quad (7)$$

4.2 Experiment Setup

Considering that the Cultural-Linguistic Synergy arises from variations in cultural context, we investigate how the model's internal behavior differs

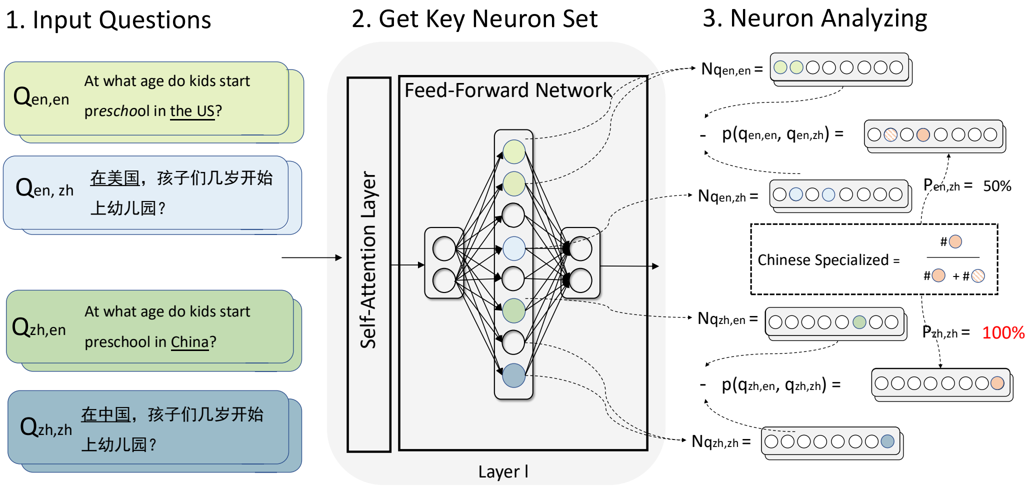


Figure 4: Workflow for interpreting Cultural-Linguistic Synergy. For every dual-format question pair, we trace neuron activations and measure the share of language specialized neurons (Chinese in the above figure) that fire when each question is posed.

when asking questions in language i. Specifically, we focus on two types of question content: one related to the American cultural context and the other to the cultural context of language i. By comparing these two contexts, we aim to uncover why, in the latter case, asking questions in the corresponding language leads to better performance than asking them in English. To explore this, we focus on calculating the "specialized neurons" activated in each context. These specialized neurons refer to the Key Neurons that activate when answering in language i , as opposed to English. For American cultural context, we obtain the Key Neuron sets $N_{q_{en,en}}$ and $N_{q_{en,i}}$ for the dual-ed questions $q_{en,en}$ and $q_{en,i}$ from $Q_{en,en}$ and $Q_{en,i}$, respectively. By gathering the neuron only activate when asking $q_{en,i}$, we can determine the proportion of specialized language i neurons for the question pair $(q_{en,en}, q_{en,i})$ as:

$$p(q_{en,en}, q_{en,i}) = \frac{|N_{q_{en,i}} - N_{q_{en,en}}|}{|N_{q_{en,i}}|} \tag{8}$$

For example, as shown in Figure 4, we calculate the key neurons for the paired question *"At what age do kids start preschool in the US?"* in both English and Chinese, to identify the specialized Chinese neurons (depicted in red). We then repeat for every question pair $(q_{en,en}, q_{en,i})$, and compute the average proportion of specialized neurons $p(q_{en,en}, q_{en,i})$ across all dual-ed question pairs. This gives us the proportion of specialized neurons for language i in the American cultural context, denoted as $P_{en,i}$. Similarly, we calculate the proportion of specialized neurons for i in the cultural context of language i, denoted as $P_{i,i}$. By comparing the proportions of specialized neurons between these two contexts, we aim to find the underlying factors contributing to Cultural-Linguistic Synergy.

For time and cost efficiency considerations, we deploy Qwen2.5-7B-Instruction and Llama-3-8B-Instruction models as the target model. To obtain the Key Neuron Set, we use the instruction 2 (Appdenx B) to get the response r and apply the top-k ($k = 5$) threshold for each layer, as defined in Equation 6 (more details about the threshold function is shown in Appendix A.5. The details of the selection for this hyperparameter can be found in Section 5.

4.3 Analyzing

We compare $P_{en,i}$ between $P_{i,i}$ ($i \neq$ en) across the six languages. As shown in Figure 5, generally, $P_{i,i}$ (bars with patterns) is higher than $P_{en,i}$ (bars without patterns) in the scenarios where model demonstrates the Cultural-Linguistic Synergy (e.g., Llama-3-8B in Chinese, Indonesian, Persian, and Korean, Qwen2.5-7B in Chinese, Korean). Conversely, when no Cultural-Linguistic Synergy is observed, $P_{i,i}$ is lower than $P_{en,i}$ (e.g., Llama-3-8B in Sundanese, Qwen2.5-7B in Persian, Sundanese). This suggests that **models tend to activate a higher proportion of neurons specialized for the target language when the cultural context aligns with the corresponding linguistic medium, compared to when this alignment is absent.** The activation of these specialized neurons allows the model to better utilize knowledge specific to the culture and the target language. This knowledge, which may not be fully accessed when asking in English, contributes to the model's better performance in the target language. However, Spanish stands as an exception, which we attribute to the high similarity between Spanish and English in terms of language structure, and thus may have greater overlap of the knowledge-storing neurons.

4.3.1 Hypothesis 1 and Validation

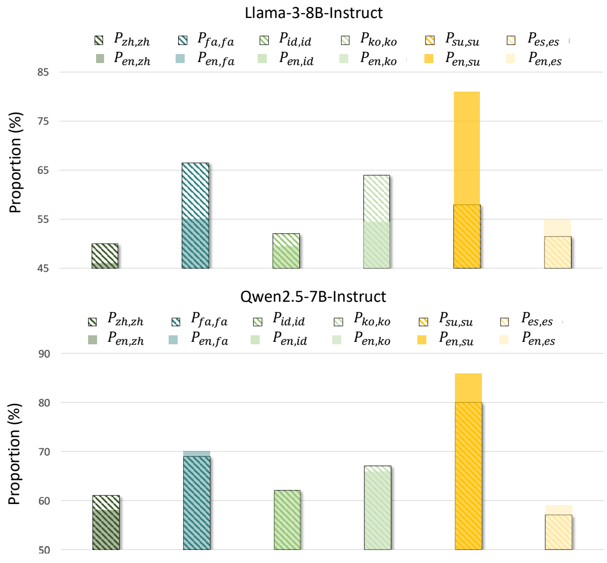


Figure 5: Comparison of the proportion of specialized neurons for language Chinese(zh), Spanish(es), Indonesian(id), Korean(ko), Persian(fa), and Sundanese(su) between different cultural context. It indicates that when Cultural-Linguistic Synergy happens, models generally activate a higher proportion of specific neurons.

Previous analysis (Section 4.3) suggests that when Cultural-Linguistic Synergy occurs, the model activates a higher proportion of neurons specialized for the language and culture. This ability to better utilize knowledge aligned with the corresponding cultural context helps guide the model to perform better than when asking in English. Building on this, we further consider whether more powerful multilingual models have a better ability to utilize culture and language-specific knowledge. This could, in turn, serve as a valuable metric for evaluating model performance during training.

> **Hypothesis 1:** Models have a better ability to utilize cultural knowledge will activate a **higher proportion** of specialized neurons when the cultural context aligns with the linguistic medium.

Figure 5 indicates that Qwen2.5 utilizes more specialized neurons (66 %) than Llama-3 (57%) across the six languages, which may provides evidence for this hypothesis. However, note that differences in training data and model architectures between different series may limit the direct comparability.

On the other hand, validating this hypothesis by improving one model with additional training data may present challenges. This is due to the limited availability of language resources and the potential risk of benchmark leakage, which could affect the analysis. Thus we use well-recognized multilingual same series models with distinct language capabilities, such as the open-source multilingual extension of the Llama-3 model, Llama-3.1-8B-Instruction, for comparative analysis with Llama-3-8B-Instruction in the validation experiment.

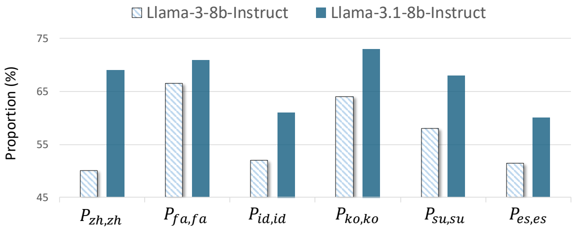


Figure 6: Comparison of proportion of specialized neurons for Llama-3-8B-instruct and Llama-3.1-8B-instruct. The result shows that Llama-3.1-8B-instruct, the multilingual extension for Llama-3-8B-instruction, has a higher proportion of specialized neurons.

The results shown in Figure 6 indicate that Llama-3.1-8B-Instruction activates a higher proportion of specialized neurons (67%) compared to Llama-3-8B-Instruction (57%), supporting our hypothesis that models with stronger capabilities in the corresponding language are better at leveraging language-specific neurons. Furthermore, this proportion of the specific neurons could be utilized as a potential **indicator for evaluating a model's ability to effectively leverage multilingual knowledge** during the training phase.

4.3.2 Hypothesis 2 and Validation

Through previous analysis (Section 4.3), we find that the proportion of specific neurons may be indicative of the Cultural-Linguistic Synergy. It left us thinking: If a higher proportion of neurons corresponds to greater knowledge neuron utilization by the model, then assuming a consistent increase in the proportion of language-specific neurons for one specific model, we expect that an increase in the number of neurons should lead to better performance for the corresponding language.

Hypothesis 2: The greater the **number of neurons** activated for questions in a given language, the better the performance.

Since there is no consensus on how to definitively measure the importance of individual neurons, we take a different perspective. Instead of focusing on neuron quantity directly, we explore whether the total number of neurons activated across the dataset is correlated with the model's performance.

From Section 4.3, we notice that knowledge representation may vary across languages. Therefore, in this validation study, we focus on comparisons within the same language. Specifically, we investigate the relationship between the set of Key Neurons set, $|N_{Q_{i,i}}|$ and $|N_{Q_{j,i}}|$, and the model's performance on the corresponding evaluation data. The results, shown in Figure 7, indicate that the total number of activated neurons is highly correlated with the model's performance, with a Pearson correlation coefficient of 0.95 for English questions (more results are in Appendix A.2). This suggests that when more neurons are activated, the model is likely utilizing more relevant knowledge, leading to better performance. This finding aligns with the observation in Section 3.2, where, in American cultural context, the model activates the most neurons.

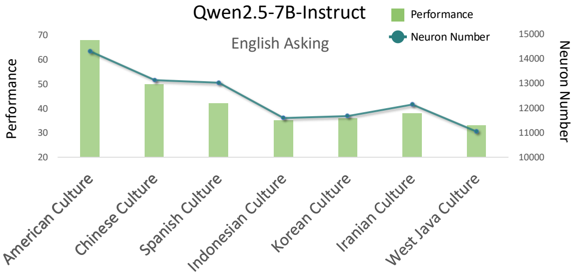


Figure 7: The performance and the number of Key Neurons for the Llama-3-8B on cross-cultural contexts.

5 Ablation Study

In our experimental setup (Section 4.2), we select $k = 5$ as the threshold. The threshold is set to ensure that the selected key neurons accurately represent the model's knowledge on the given question. To determine the optimal threshold, we measure the performance drop when masking the corresponding neurons on the selected task. We choose the threshold where the performance drops significantly for the masked neurons, while the performance on out-of-distribution (OOD) knowledge (here we use the ARC (Clark et al., 2018) dataset), is largely unaffected. As shown in Figure 8, we draw in blue and red, respectively. We also use a random mask with the same number of neurons as a baseline (dash line). The selected threshold is depicted in green.

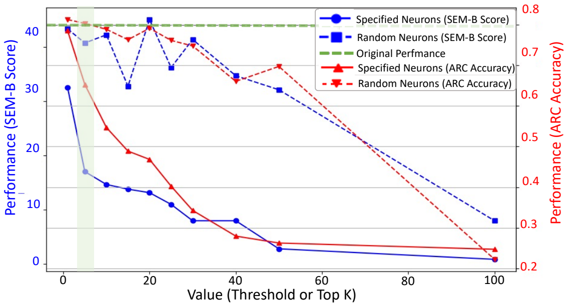


Figure 8: Performance for Llama-3-8B-Instruction on ARC and $Q_{en,en}$ when masking Key Neurons with different threshold.

6 Related Work

Multilingual Capabilities Evaluation To evaluate the LLMs' multilingual capabilities, researchers use translating English-centric benchmarks such as MMMLU (OpenAI, 2024), MGSM (Shi et al., 2022) and Multilingual MT-Bench (Zheng et al., 2023). Recent work also developed culture-specific benchmarks (Cao et al., 2025a; Zhang et al., 2023; Myung et al., 2024; Leong et al., 2023; Liu et al., 2025). For example, M3Exam (Zhang et al., 2023) sourced from real and official human exam questions, BLEnD (Myung et al., 2024) where evaluation data are crafted from real-world scenarios and CulturalBench (Chiu et al., 2024) with human-written questions covering 45 global regions. These existing evaluations, however, treat language and cultural context as inseparable dimensions, restricting analyses to single-language scenarios.

Multilingual Capabilities Interpretation Recently, some work (Wang et al., 2024a; Kojima et al., 2024; Wang et al., 2024b) use the Mechanistic interpretability to analyze the model's multilingual capabilities. Tang et al. (2024b) shows that proficiency in processing a particular language is predominantly due to a small subset of neurons. Wendler et al. (2024) projects the hidden state into vocabulary to investigate the Latent Language. Zhao et al. (2024) further proposed the multilingual workflow to understand how LLMs Handle Multilingualism. However, these studies do not investigate the model's behavior across different cultural contexts and languages.

7 Conclusion

This study introduced a Dual Evaluation Framework specifically designed to comprehensively assess LLMs across linguistic medium and cultural contexts. Our findings reveal “Cultural-Linguistic Synergy,” phenomenon where models perform optimally when questions are culturally aligned with the language, challenging the prevailing assumption that LLMs, primarily trained on English data, perform uniformly across different languages. Utilizing interpretative methods, we delved deeper into this phenomenon and found that it is related to the Key Neurons. As the field of interpretability in AI continues to evolve, we plan to further expand this framework to enable more comprehensive and nuanced evaluations of multilingual models.

8 Limitation

While the Dual Evaluation Framework is flexible enough to incorporate additional benchmarks, the prerequisite for conducting meaningful cross-cultural comparisons, especially to conduct neuron probing, lies in having dual-format question content. This content needs to capture both linguistic and cultural nuances. Without this dual-format structure, performing robust and quantitative cross-cultural comparisons remains limited.

In our current experimental design, we focus on a single cultural context for each language, based on typical countries or regions where the language is spoken. However, given the widespread usage of some languages, especially in regions with diverse cultural contexts, we plan to expand the framework in the future to incorporate more varied cultural contexts to make our conclusions more robust.

Due to time and computational cost constraints, we limited our probing validation to models like Qwen2.5-7B-Instruction and Llama-3-8B-Instruction. As LLMs interpretation techniques continue to evolve and improve, we plan to expand the range of models included in future studies, especially larger models with more parameters, to gain deeper insights into multilingual and cross-cultural model behavior.

9 Acknowledgement

This research is supported by the Ministry of Education, Singapore, under its Academic Research Fund (AcRF) Tier 1 grant, and funded through the SUTD Assistant Professorship Scheme (SAP 2025_001). We would also like to thank Yizhe Yang from Beijing Institute of Technology for helping us with human evaluation.

References

Anthropic. 2024. Claude 3.5 sonnet model card addendum.

Yixin Cao, Shibo Hong, Xinze Li, Jiahao Ying, Yubo Ma, Haiyuan Liang, Yantao Liu, Zijun Yao, Xiaozhi Wang, Dan Huang, Wenxuan Zhang, Lifu Huang, Muhao Chen, Lei Hou, Qianru Sun, Xingjun Ma, Zuxuan Wu, Min-Yen Kan, David Lo, Qi Zhang, Heng Ji, Jing Jiang, Juanzi Li, Aixin Sun, Xuanjing Huang, Tat-Seng Chua, and Yu-Gang Jiang. 2025a. Toward generalizable evaluation in the llm era: A survey beyond benchmarks.

Yixin Cao, Jiahao Ying, Yaoning Wang, Xipeng Qiu, Xuanjing Huang, and Yugang Jiang. 2025b. Model utility law: Evaluating llms beyond performance through mechanism interpretable metric.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2024. Culturalbench: a robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of llms.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.

Cohere. 2024. The command r model (details and application).

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

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.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Yihuai Hong, Lei Yu, Haiqin Yang, Shauli Ravfogel, and Mor Geva. 2024. Intrinsic evaluation of unlearning using parametric knowledge traces.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. 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 6912–6964.

Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William Chandra Tjhi. 2023. Bhasa: A holistic southeast asian linguistic and cultural evaluation suite for large language models.

Chaoqun Liu, Wenxuan Zhang, Jiahao Ying, Mahani Aljunied, Anh Tuan Luu, and Lidong Bing. 2025. Seaexam and seabench: Benchmarking llms with local multilingual questions in southeast asia.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Victor Gutierrez Basulto, Yazmin Ibanez-Garcia, Hwaran Lee, Shamsuddeen H Muhammad, Kiwoong Park, Anar Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, Jose Camacho-Collados, and Alice Oh. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. In *Advances in Neural Information Processing Systems*, volume 37, pages 78104–78146. Curran Associates, Inc.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2024. SeaLLMs - large language models for Southeast Asia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 294–304, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI. 2024. Multilingual massive multitask language understanding.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

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.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners.

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

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

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan

Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size.

Teng Wang, Zhenqi He, Wing-Yin Yu, Xiaojin Fu, and Xiongwei Han. 2024a. Large language models are good multi-lingual learners: When llms meet cross-lingual prompts. *CoRR*.

Weixuan Wang, Barry Haddow, Minghao Wu, Wei Peng, and Alexandra Birch. 2024b. Sharing matters: Analysing neurons across languages and tasks in llms.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Wenxuan Zhang, Hou Pong Chan, Yiran Zhao, Mahani Aljunied, Jianyu Wang, Chaoqun Liu, Yue Deng, Zhiqiang Hu, Weiwen Xu, Yew Ken Chia, Xin Li, and Lidong Bing. 2024. Seallms 3: Open foundation and chat multilingual large language models for southeast asian languages. *CoRR*, abs/2407.19672.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? *arXiv preprint arXiv:2402.18815*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

A More Experiment Detail

A.1 Response Generation Setting

Answer generation across the involved models is conducted in a zero-shot setting, with all models set to a temperature of 0.0 and a maximum token length of 1024.

A.2 Result Display Setting

The results presented in Section 3.2 and Section 3.3 are based on Instruction 1 (as shown in Appendix B). All other results are displayed in an averaged format in Figures 9, 10, 11, 12, 13, 14, 15, 16, and 17 for each model on every $Q_{i,j}$.

For the results in Section 4.3, Section 4.3.1, and Section 4.3.2, we conducted evaluations using Instruction 2 (as shown in Appendix B) to manage computational costs. This is particularly relevant for certain questions where the model might generate lengthy responses, making the interpretation of results impractical without these adjustments. For results in other languages, except English, which are not shown in Section 4.3.2, please refer to Figures 18 and 19.

A.3 The primarily experiment result

Table 3 shows the performance of the models Llama-3-8B-Instruct, Gemma-2-9B-Instruct, and Qwen2.5-7B-Instruct on multilingual benchmarks: GSM8K and MMMLU. The experiment is conducted in a zero-shot setting, and the results suggest that the models perform better when the questions are asked in English compared to other languages.

A.4 The completion for Dataset

The original BLEnD includes, for each language-specific evaluation set $Q_{i,j}$ (except for English), an English translation evaluation dataset $Q_{i,en}$. For the rest language (i, j) pair (when $i = en$), we deploy GPT-4o to conduct the translation. To ensure the translated question $q_{en,i}$ aligns $q_{i,i}$ with the dual-format structure, we prompt GPT-4 with a one-shot example using the question pair $q_{i,en}$, $q_{i,i}$ to obtain the translated version $q_{en,en}$ for language i , which we then use as $q_{en,i}$.

To further evaluate the quality of this complement, we conduct a human evaluation involving four senior computational linguistics researchers who have a research focus in multilingualism and are trained in advanced. From the constructed $Q_{en,i}$, we randomly sampled 100 cases, along with their dual cases from $Q_{en,en}$, $Q_{i,en}$, and $Q_{en,en}$. We presented these 100 paired cases ($q_{i,en}$, $q_{i,i}$, $q_{en,en}$, and GPT-o translated $q_{en,i}$) to the evaluators, asking them to score the translated content and format consistency: 1 point: The translation content is problematic or inaccurately expressed. 2 points: The translation content is accurate, but the format deviates significantly from the corresponding $q_{i,i}$. 3 points: The translation content is accurate, and the format aligns perfectly with the corresponding $q_{i,i}$. The results indicate that the average full mark rate (3 points) for translated content and format consistency is 97.8%, with scores above 2 points reaching 100%. The overall agreement rate is 95%. This prove the quality of the newly introduced dataset.

A.5 The threshold function for Interpretation

In our experiment, we deploy top-k (k = 5) threshold for each layer. Specifically, we compute the activation value for each corresponding response token r_i for each question q. We then aggregate the activation scores of each neuron for each response token across each layer l ($l \in \{1, 2, \ldots, L\}$), represented as: $V_l = \left[v_{(j,l)}^{r_i} \mid r_i \in r, j \in \{1, 2, \ldots, dm\} \right]$. To determine the key neurons for question q, we select the top-k neurons for each layer, forming the key neuron set N_q as:

$$N_q = \{ (j,l) \mid v_{(j,l)}^{r_i} \geq V_l^{\text{top-}k}, r_i \in r, \\ j \in \{1, 2, \ldots, dm\},\ l \in \{1, 2, \ldots, L\} \}. \quad (9)$$

When we conduct experiments, we also explore other threshold function. Including: 1) Layer-specific top-k (final adoption in the paper) 2) Global top-k, 3) Global top-k score, 4) Global top-k score. We determine the threshold by conducting the experiment shown in Section 5 and select the optimal ones.

B Instruction

We mainly use the instructions from the original benchmark BlEnD (Myung et al., 2024). However, some models' responses are longer due to the nature of the instruction, so to better match each question with candidate answers and help us conduct the interpretation experiment, we manually add additional instructions (instruction 2 for each language).

Model	GSM8K$_{en}$	GSM8K$_{cn}$	GSM8K$_{es}$	MMMLU$_{en}$	MMMLU$_{id}$	MMMLU$_{cn}$
Llama-3-8B-Instruct	**77.1**	60.2	66.7	**64.4**	52.4	54.5
Gemma-2-9B-Instruct	**81.2**	77.9	75.1	**73.4**	64.4	64.0
Qwen2.5-7B-Instruct	**84.3**	80.3	71.1	**71.3**	56.8	60.8

Table 3: The performance for mode Llama-3-8B-Instruct, Gemma-2-9B-Instruct and Qwen2.5-7B-Instruct on MMMLU (OpenAI, 2024), MGSM (Shi et al., 2022). The experiment is conducted in a zero-shot setting. The languages we select are English(en), Chinese(cn), Spanish(es), and Indonesian(id). We find that models have better performance when the question is asked in English.

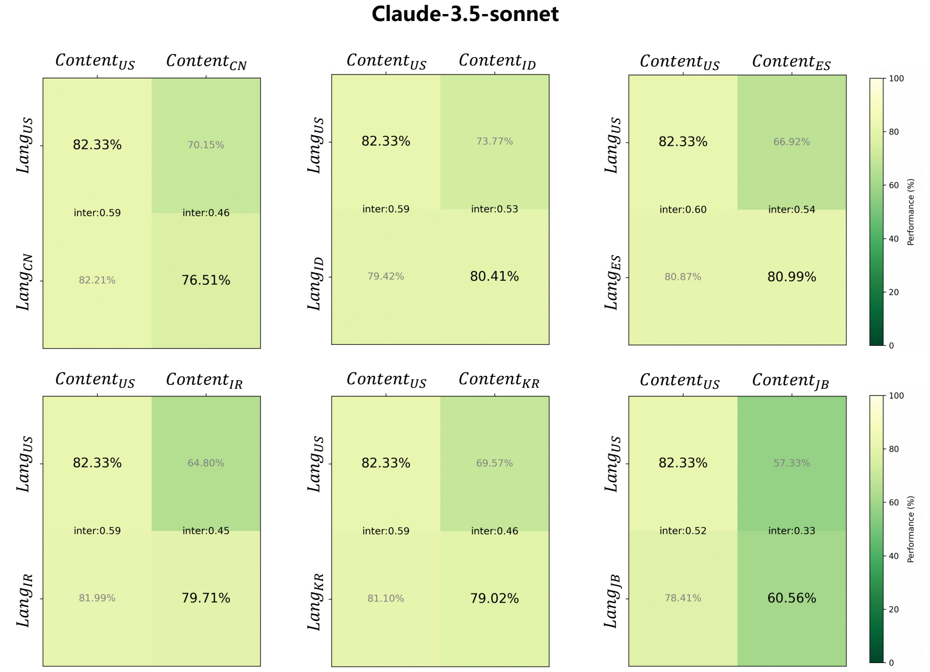


Figure 9: The average performance of Claude-3.5-sonnet on Instruction set B. The $Content_i$ represents langue$i$-speaking culture context, $Lang_i$ represents the linguistic medium for language i.

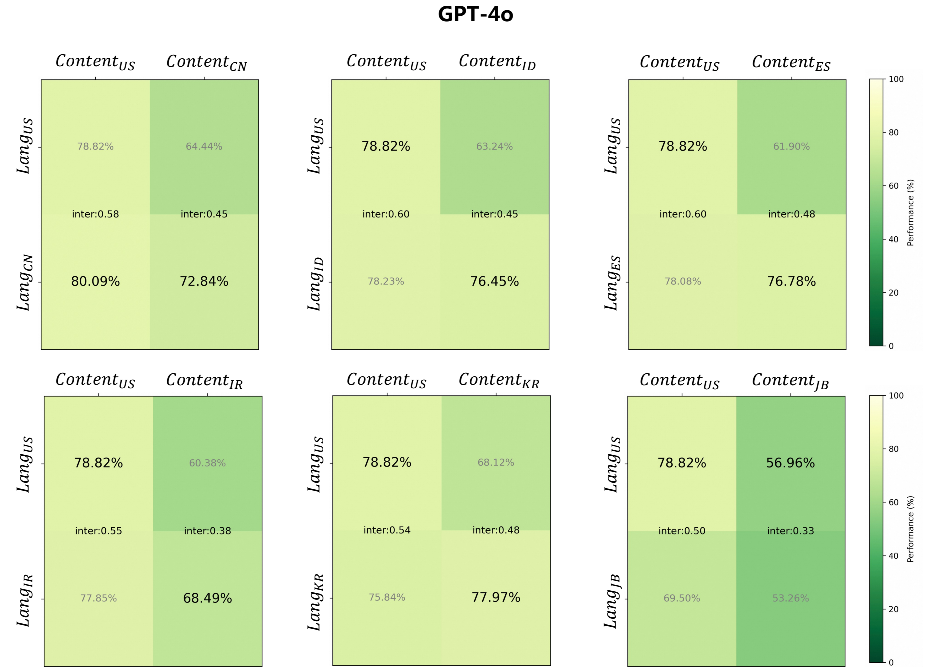


Figure 10: The average performance of GPT-4o on Instruction set B. The $Content_i$ represents the langue$i$-speaking culture context, $Lang_i$ represents the linguistic medium for language i.

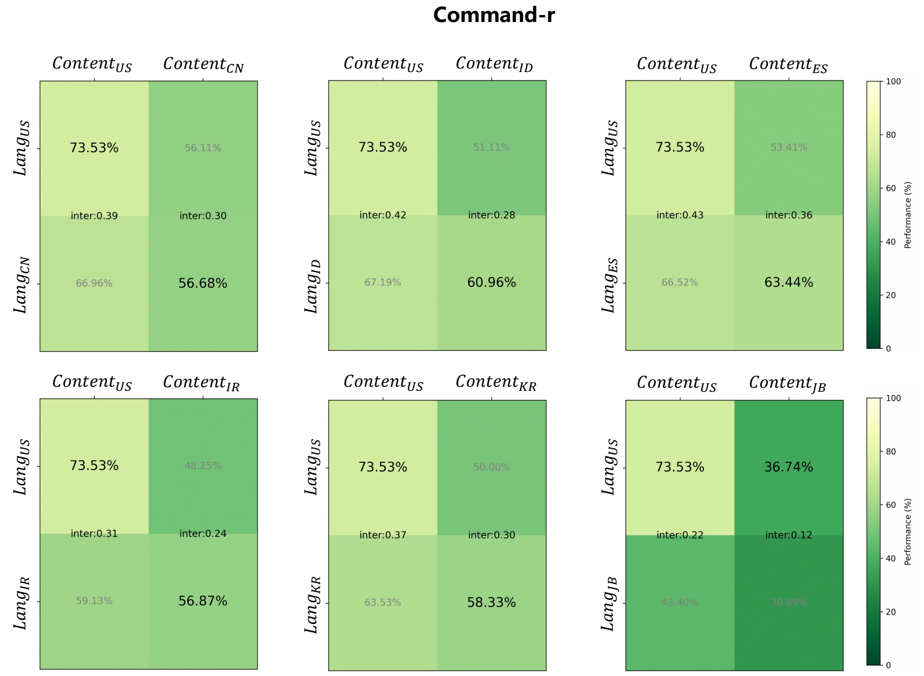


Figure 11: The average performance of Command-r on Instruction set B. The $Content_i$ represents the langue$i$-speaking culture context, $Lang_i$ represents the linguistic medium language.

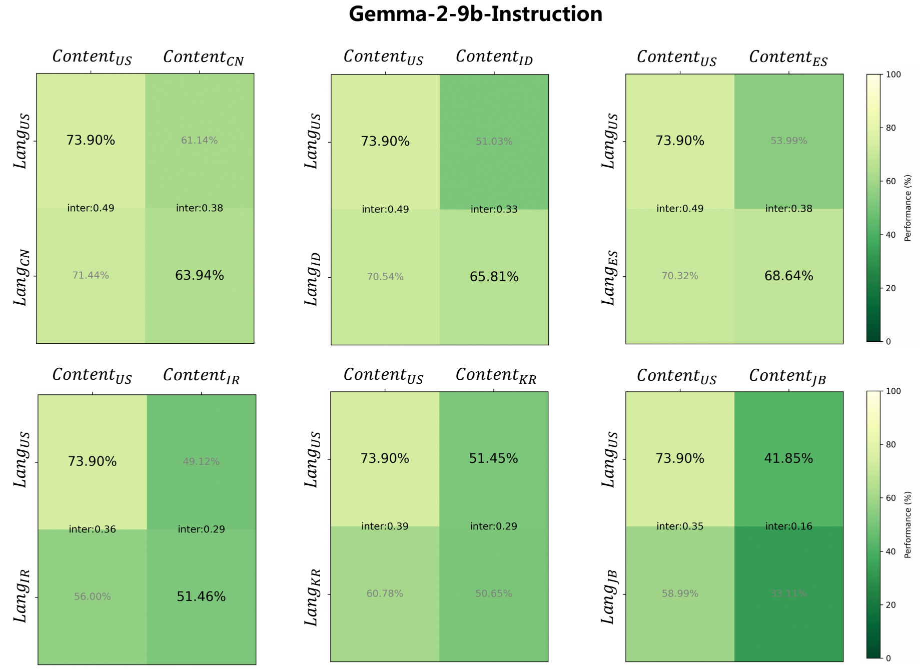


Figure 12: The average performance of Gemma-2-9b-Instruct on Instruction set B. The $Content_i$ represents the langue$i$-speaking culture context, $Lang_i$ represents the linguistic medium for language i.

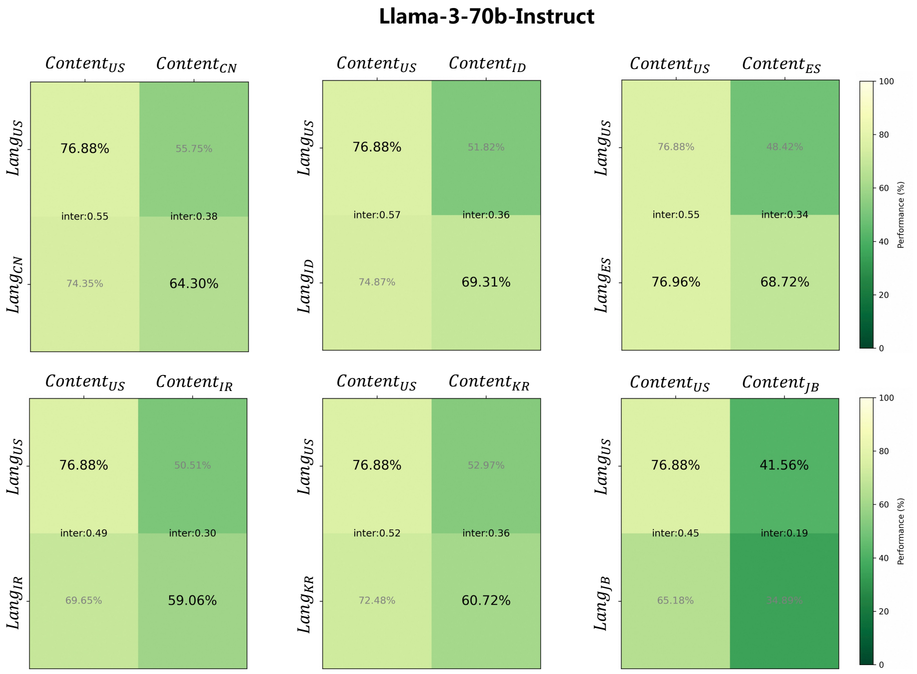


Figure 13: The average performance of Llama-3-70b-Instruct on Instruction set B. The $Content_i$ represents the langue$i$-speaking culture context, $Lang_i$ represents the linguistic medium for language i.

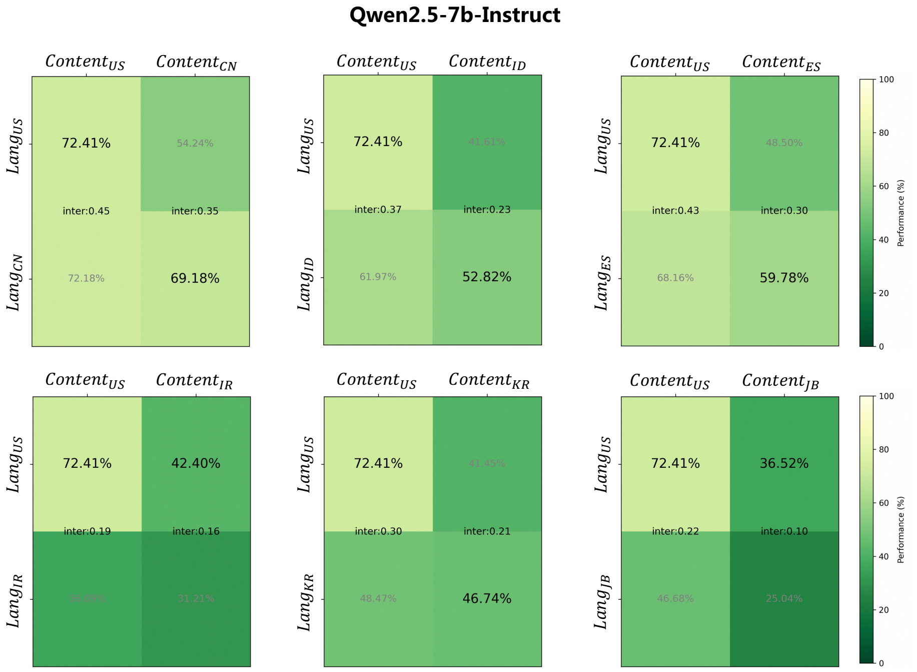


Figure 14: The average performance of Qwen-2.5-7b-Instruct on Instruction set B. The $Content_i$ represents the langue$i$-speaking culture context, $Lang_i$ represents the linguistic medium for language i.

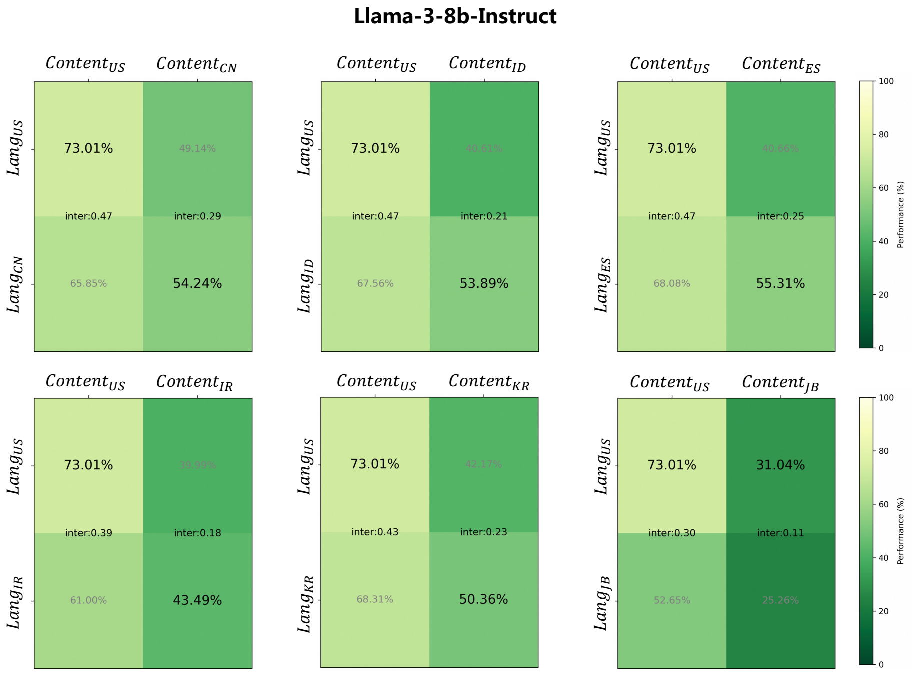


Figure 15: The average performance of Llama-3-8b-Instruct on Instruction set B. The $Content_i$ represents the langue$i$-speaking culture context, $Lang_i$ represents the linguistic medium for language i.

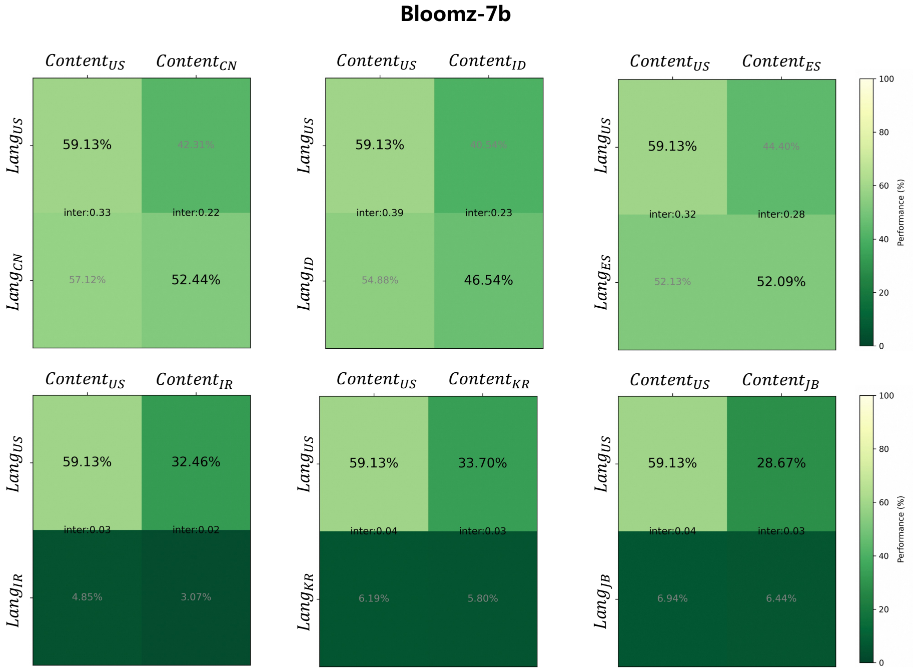


Figure 16: The average performance of Bloomz-7b on Instruction set B. The $Content_i$ represents the culture context, $Lang_i$ represents the linguistic medium language.

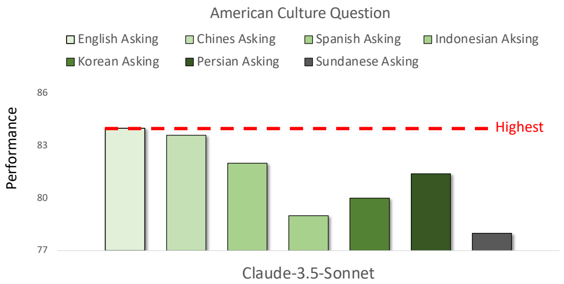


Figure 17: The performance of the selected models on the American culture question when asked in the other six languages versus English. Models perform the best when asking questions in English.

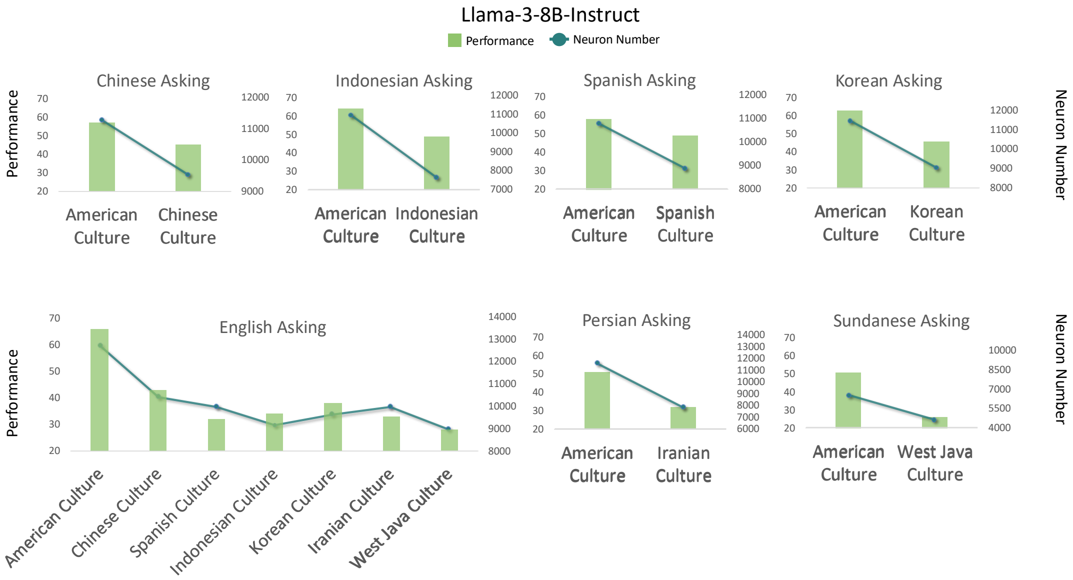


Figure 18: The performance and the number of Key Neurons for the Llama-3-8B-Instruction on cross-cultural contexts.

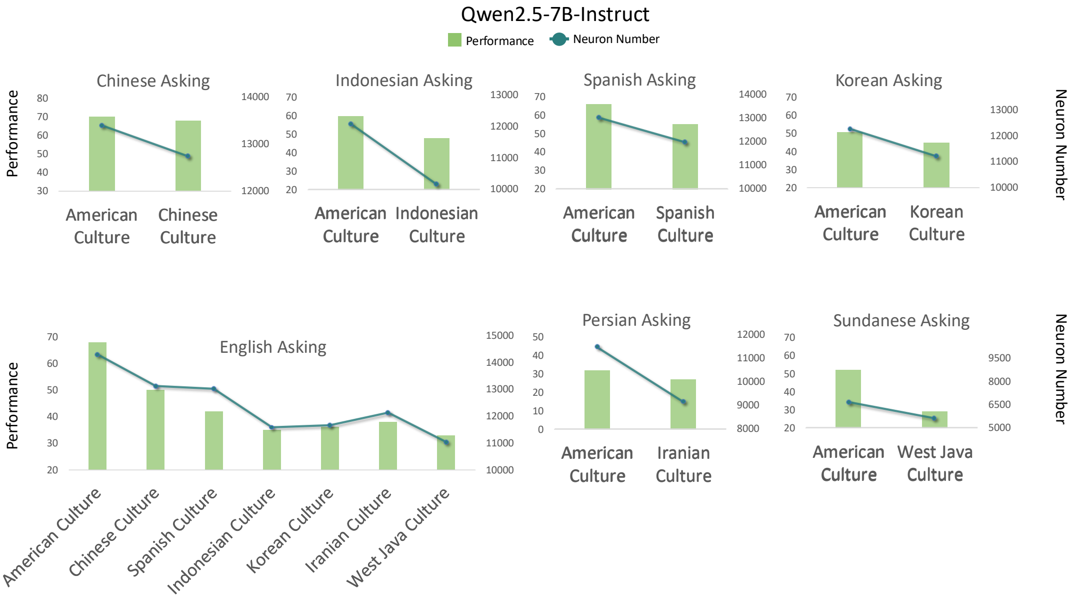


Figure 19: The performance and the number of Key Neurons for the Qwen-2.5-7B-Instruction on cross-cultural contexts.

English Instruction 1:
Read the following question and provide a single answer without any explanations.

English Instruction 2:
Read the following question and return only ONE answer in a single word or an unbroken phrase, avoiding splitting any proper nouns or named entities.

English Instruction 3:
Read the following question and provide the single answer you think is most appropriate without any explanations.

Chinese Instruction 1:
阅读以下问题并仅提供一个答案，无需任何解释。

Chinese Instruction 2:
阅读以下问题，并仅用一个单词或一个不间断的短语返回一个答案，避免拆分任何专有名词或命名实体。

Chinese Instruction 3:
阅读以下问题并仅提供一个你觉得最合适的答案，无需任何解释。

Spanish Instruction 1:
Lee la siguiente pregunta y contesta con una sola respuesta sin explicaciones.

Spanish Instruction 2:
Lee la siguiente pregunta y devuelve solo UNA respuesta en una sola palabra o una frase sin dividir ningún nombre propio o entidad nombrada.

。

Spanish Instruction 3:
Lee la siguiente pregunta y proporciona la respuesta única que consideres más apropiada sin ninguna explicación.

Indonesian Instruction 1:

Bacalah pertanyaan berikut dan berikan satu jawaban tanpa penjelasan apa pun.

Indonesian Instruction 2:

Bacalah pertanyaan berikut dan berikan hanya SATU jawaban dalam satu kata atau frasa utuh, hindari memisahkan nama diri atau entitas bernama.

Indonesian Instruction 3:

Read the following question and provide the single answer you think is most appropriate (one word) without any explanations.

Korean Instruction 1:

다음 질문을 읽고 설명 없이 단 하나의 답변만을 제공하시오.

Korean Instruction 2:

당신은 외국인에게 당신의 나라의 문화를 설명하려는 대한민국 사람입니다. 다음 질문을 읽고 설명 없이 가장 적절하다고 생각되는 단 하나의 답변을 제공하시오.

Korean Instruction 3:

다음 질문을 읽고 설명 없이 가장 적절하다고 생각되는 단 하나의 답변을 제공하시오.

Persian Instruction 1:

سوال زیر را بخوانید و یک پاسخ بدون هیچ توضیحی ارائه دهید.

Persian Instruction 2:

دستور زیر را بخوانید و فقط یک پاسخ را در قالب یک کلمه یا یک عبارت بدون شکستن اسامی خاص یا موجودیت‌های نام‌گذاری شده ارائه دهید.

Persian Instruction 3:

متن زیر را بخوانید و پاسخی که فکر می‌کنید مناسب‌ترین است بدون هیچ توضیحی ارائه دهید.

Sundanese Instruction 1:
Bacalah pertanyaan berikut dan berikan satu jawaban tanpa penjelasan apa pun.

Sundanese Instruction 2:
Bacakeun pananya di handap ieu jeung pasih hiji jawaban dina hiji kecap atawa frasa anu teu dipisah, ulah misahkeun ngaran sorangan atawa entitas anu dingaranan.

Sundanese Instruction 3:
Bacakeun pananya di handap ieu jeung pasih hiji jawaban anu anjeun anggap paling cocog tanpa penjelasan nanaon.