What Language Do Non-English-Centric Large Language Models Think in?

Chengzhi Zhong[1] **Qianying Liu**[2] **Fei Cheng**[1] **Junfeng Jiang**[3]
Zhen Wan[1] **Chenhui Chu**[1] **Yugo Murawaki**[1] **Sadao Kurohashi**[1,2]

[1] Kyoto University, Japan
[2] National Institute of Informatics, Japan
[3] The University of Tokyo, Japan
{zhong, feicheng, wan, chu, murawaki,kuro}@nlp.ist.i.kyoto-u.ac.jp
ying@nii.ac.jp
jiangjf@is.s.u-tokyo.ac.jp

Abstract

In this study, we investigate whether non-English-centric large language models, 'think' in their specialized language. Specifically, we analyze how intermediate layer representations, when projected into the vocabulary space, favor certain languages during generation—termed as latent languages. We categorize non-English-centric models into two groups: CPMs, which are English-centric models with continued pre-training on their specialized language, and BLMs, which are pre-trained on a balanced mix of multiple languages from scratch. Our findings reveal that while English-centric models rely exclusively on English as their latent language, non-English-centric models activate multiple latent languages, dynamically selecting the most similar one based on both the source and target languages. This also influences responses to culture difference questions, reducing English-centric biases in non-English models. This study deepens our understanding of language representation in non-English-centric LLMs, shedding light on the intricate dynamics of multilingual processing at the representational level. Our code is publicly available.[1]

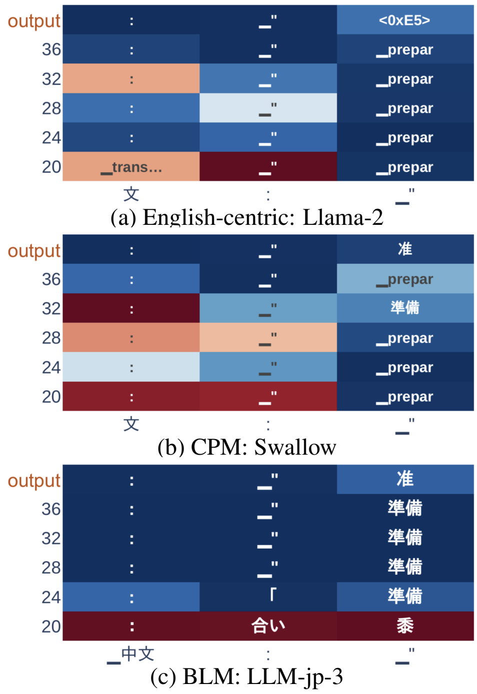


Figure 1: **Logit-lens for intermediate layers of three models when doing translation.** The input is "Français: "préparation" - 中文: "." The figure shows the highest probability token from the intermediate layers starting from layer 20. The Chinese answer "准备" is expected, where 0xE5 is the first UTF-8 byte of "准."

1 Introduction

Large Language Models (LLMs) need multilingual capability to effectively serve a global audience by facilitating communication and task execution across diverse languages. Nevertheless, state-of-the-art LLMs remain predominantly English-centric (Dubey et al., 2024) (Workshop et al., 2022). Despite their robust performance in English, these models often exhibit reduced proficiency in non-English languages, and their outputs may reflect an inherent bias toward English-centric perspectives. Recent studies on the Llama-2 family suggest that these English-centric models 'think' in English (Wendler et al., 2024). Specifically, as shown in Figure 1 (a), when employing logit-lens (Nostalgebraist, 2020) to examine the probability distributions of tokens in their intermediate layers, a pronounced internal preference for English token '*_prepar*' is observed—even when processing French to Chinese translation inputs. This phenomenon is defined as these English-centric models 'thinking' in English **latent language**. This reliance on an English latent language not only undermines performance in other languages but also may introduce unintended biases.

To mitigate these challenges, researchers have

[1] https://github.com/ku-nlp/latent_language_of_multilingual_model

Findings of the Association for Computational Linguistics: ACL 2025, pages 26333–26346
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

developed non-English-centric models designed to enhance performance in a specialized language and reduce biases. Two primary strategies have emerged. One approach adapts English-centric models by continuing pre-training with language-specific corpora (CPMs) (Fujii et al., 2024) (Cui et al., 2023), while the other constructs a model from scratch using a balanced corpus that contains English and the specialized language (BLMs) (Gouvert et al., 2025) (Aizawa et al., 2024).

While these models demonstrate improved performance, it remains unclear how their internal processing differs from English-centric models. To explore this, we investigate the open question: **When processing their specialized language, in what latent language do these models 'think'?** Specifically, we seek to determine whether these models employ the dominant language of their training data as latent language when processing monolingual cloze tasks. To address this question, we conduct experiments on four languages—Japanese, Chinese, French, and Arabic. An example finding of Japanese-specialized models indicate that while the English-centric model predominantly processes information in English, the BLM model primarily utilizes Japanese as a single latent language in its intermediate layer; the CPM model exhibits a mixed pattern of both English and specialized language utilization as latent languages.

While non-English-centric models 'think' in their specialized language when processing tasks in that language, an intuitive question arises: **What latent language do these models employ when handling cross-lingual tasks?** To address this, we systematically vary both the source and target languages across various non-English-centric models in a translation task. Our experiments reveal that the latent language in intermediate layers of these models follows a dynamic pattern: earlier layers tend to reflect latent language similar to the source language, while later layers increasingly utilize latent language similar to the target language—eventually yielding outputs in the target language. Notably, BLMs exhibit a noticeable tendency to adopt a single latent language (i.e., '準備' in Figure 1(c)), whereas CPMs tend to intermix activations across languages (i.e. '準備' and '$\_prepar$'). We refer to this phenomenon—where the model's probability distribution shifts stepwise from a language akin to the source to one more similar to the target, culminating in the final output—as the **'Probabilistic Cascade'**.

Given that non-English-centric models have been shown to reduce biases (Nie et al., 2024), it is crucial to understand how their internal latent language patterns contribute to shaping cultural biases. In particular, we investigate: **How do the latent language patterns influence semantic representations when processing culturally specific questions?** To address this, we analyze the models' internal responses when handling culture difference questions. When asked about the longest river in Japan, English-centric model initially produces latent representations biased toward English-centric cultural narratives (i.e., referencing the Mississippi River). Although later layers gradually adjust the output toward the target language context, the final answer remains culturally inappropriate. In contrast, non-English-centric models realign their latent language more effectively toward the target culture, resulting in more accurate and culturally relevant outputs. This investigation thus elucidates how latent language patterns in intermediate layers can shape cultural bias.

In summary, we demonstrate the aforementioned experiments, the model subjects, and covered non-English languages in Figure 2. Our contributions are threefold:

1. We investigate non-English-centric LLMs for Japanese, Chinese, French, and Arabic, confirming that these models employ their respective specialized languages—along with English—as latent languages in their intermediate layers when processing tasks in their designated languages.

2. We observe that when processing cross-lingual tasks, these models exhibit a dynamic latent language pattern between English and their specialized languages. The probability distribution of these latent languages reflects the similarity between the source/target language and the latent languages.

3. We analyze how latent language usage correlates with cultural bias. Specifically, when addressing culture difference questions, while English-centric models tend to generate latent representation biased towards English culture, non-English-centric models realign their latent language more effectively toward the specialized language's culture, resulting in outputs that better reflect the culturally appropriate context.

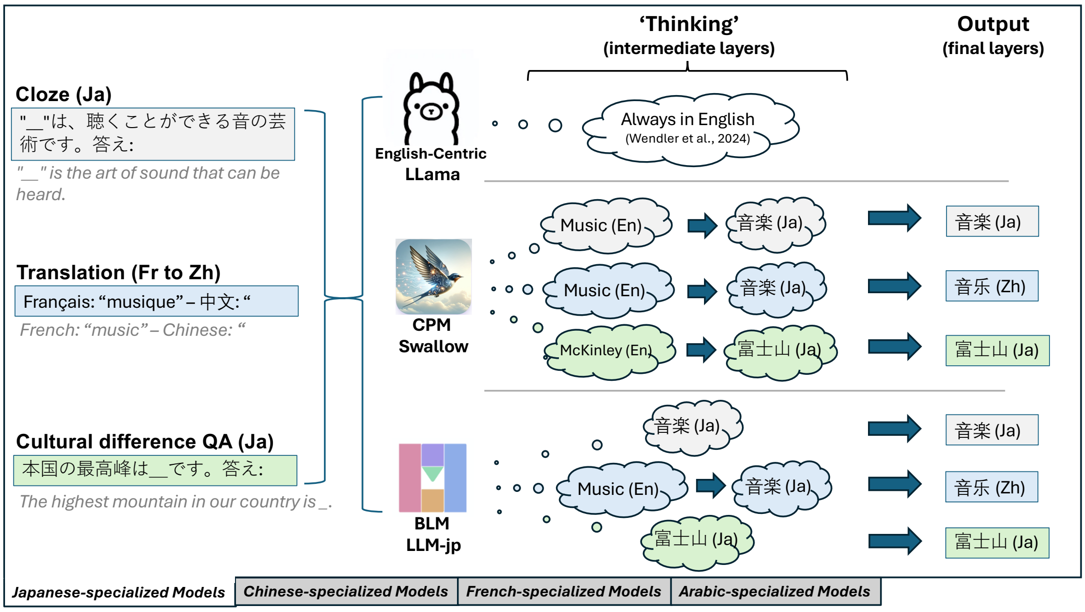


Figure 2: **An overview of detecting latent languages of two categories of non-English-centric models in three experiments across four languages: Japanese, Chinese, French, and Arabic**

2 Related work

2.1 Non-English-centric LLMs

Current frontier large language models, such as GPT-4 (Achiam et al., 2023), Gemini (Team et al., 2023), and Llama-2 (Touvron et al., 2023), are primarily trained with English-centric corpora, with other languages constituting only a small portion of the training data. Significant research efforts have been contributed to enhance these models' multilingual capabilities through various methods. One approach involves continued pre-training on specialized language data (Sun et al., 2020; Brown et al., 2020; Hunter et al., 2023), as demonstrated by models like Japanese Swallow (Fujii et al., 2024), ChineseLlama (Cui et al., 2023), Claire (Hunter et al., 2023) and SambaLingo (Csaki et al., 2024), all of which are based on Llama-2. Another approach is training from scratch with bilingual data (Sengupta et al., 2023; Yang et al., 2024; Gouvert et al., 2025), exemplified by models such as LLM-jp (Aizawa et al., 2024), Baichuan (Yang et al., 2023), Lucie (Faysse et al., 2024)and Jais (Sengupta et al., 2023). While these two approaches have proven effective, the community still knows little about the underlying mechanism. Our research substantially fills this gap.

2.2 Cultural Bias in LLMs

Existing research has demonstrated that LLMs exhibit biases related to culture, race, gender, and social values, among other factors (Nie et al., 2024) (Fang et al., 2024). Studies assessing word embeddings and generated text indicate that LLMs' biases correspond to the cultural and regional contexts of their training data (AlKhamissi et al., 2024) (Naous et al., 2023). Given that many LLMs are predominantly trained on English-language corpora, they tend to reflect cultural norms and values prevalent in English-speaking regions. Various approaches, such as data curation, model fine-tuning (Gallegos et al., 2024), and prompt engineering (Tao et al., 2023), have been employed to mitigate these biases. While previous studies have explored cultural bias in LLM-generated outputs, relatively little attention was paid to the underlying cause of such biases. In this work, we analyze how cultural biases manifest in the intermediate layers of English-centric and non-English-centric models, providing insights into the cause of bias.

2.3 Interpretability Techniques

Mechanistic interpretability is the study of understanding how models work by analyzing their internal components and processes to elucidate the mechanisms that give rise to their behavior and predictions, encompassing research lines like superposition (Elhage et al., 2022), sparse autoencoders (Huben et al., 2023), circuit analysis (Wang et al., 2022) and so on. Studies on multilingual models have identified language-specific neurons by analyzing their activation patterns across different languages (Tang et al., 2024). Similar to probing methods, this approach reveals structured multilingual representations by examining intermediate activations. Likewise, Logit Lens (Nostalgebraist, 2020) and Tuned Lens (Belrose et al., 2023) focus on decoding the probability distribution over the vocabulary from hidden vectors of the model. These methods help analyze the model's 'thinking' process. In this work, we follow the study (Wendler et al., 2024) to employ Logit Lens to analyze the internal behavior of non-English-centric models when processing multiple languages, examining the rich combination patterns of multiple latent languages.

3 Methodology

In this section, we first introduce Logit lens, which is used to detect the latent language of certain LLMs. We define two categories of non-English-centric LLMs and collect models across four non-English lanuguages—Chinese, Japanese, French, and Arabic—as our research subjects. We design three tasks including monolingual cloze, cross-lingual translation, and culture difference QA tasks to examine the three research questions described in the introduction.

3.1 Logit Lens

Logit Lens (Nostalgebraist, 2020) is a tool designed to reveal token information of the intermediate layers. LLMs use softmax to project the hidden vectors onto the dimensions of the vocabulary in the output layer, which is called unembedding. As the hidden vectors passed between the intermediate layers of the model have the same dimensions as the output vectors. By applying the same unembedding operation to those intermediate hidden vectors, we can obtain the vocabulary probability of certain intermediate layers. In this work, we use Logit Lens to obtain the predicted token probability distribution from the intermediate layers.

3.2 Measuring Multi-token Probability

The existing work (Wendler et al., 2024) limited its data construction to single-token words and calculated the single-token probability only. However, more words contain multiple sub-tokens and the single-token probability does not meet the practical usage. In this work, we measure the generation probability of multiple tokens in the intermediate layers.

After a word is tokenized into sub-token IDs $[x_1, x_2, \ldots, x_n]$, the probability p_1 of token x_1 is first obtained using logit lens on the hidden vector of a certain layer. Subsequently, the ground truth token x_1 is fed into the model as input to calculate the probability p_2 of token x_2. This process is conducted iteratively. The final probability of generating the token sequence $[x_1, x_2, \ldots, x_n]$ at layer i is then determined as the product of individual probabilities, $p_1 \times p_2 \times \cdots \times p_n$.

3.3 Categorization of non-English-centric Large Language Models

Based on their training data, we classify non-English-centric LLMs into two types and include the original *English centric* one:

English-Centric Models: These models, such as Llama2, the majority of their training data is in English, making them highly proficient in generating and understanding English text.

CPMs: These models are built upon an English-centric model and undergo continued pre-training on a specialized language to enhance multilingual ability.

BLMs: These models are trained on a roughly equal amount of tokens from two or more languages, aiming to achieve balanced proficiency across these languages.

We selected non-English-centric models for Chinese, Japanese, French, and Arabic. Chinese and Japanese share a part of common Kanji characters. French is closely similar to English. Arabic is relatively distinct from all the other languages. This setting allows us to analyze the experimental results from the perspective of language similarity.

3.4 Dataset Construction

After selecting the models, we constructed three tasks across four languages (Japanese, Chinese, French, and Arabic), each task corresponding to one research question. Because Chinese and Japanese share common characters (Chu et al., 2012), we first prepared a set of non-overlapping Chinese-Japanese word pairs that have the same meaning but different characters. This is based on *Database of Japanese Kanji Vocabulary in Contrast to Chinese* (JKVC) (達彦 et al., 2020). Then, we use GPT-4 to translate from Japanese and obtain the corresponding English and French words or phrases, and correcting any mistakes. Finally, we obtained 166 parallel word pairs.

Prompt design: We then define three tasks: the monolingual cloze task, the cross-lingual translation task, and the culture difference QA task, using the following prompt format:

Cloze task: We use the prompt format following the previous work (Wendler et al., 2024). For the Cloze task, we use GPT-4 to generate a description for each word in each language. Each described word is placed at the beginning of the description. We then mask the word in the description and make the models generate the target word. We present a Japanese example (English meaning in Figure 2):

"__"は、聴くことができる音の芸術です。答え*:* "音楽"。

Translation task: When constructing translation prompts, we use a hyphen to connect the input language word and the target language word to form a one-shot example. We demonstrate an example of translating a French word into Chinese:

Français: "musique" - 中文*:*"音乐"

Culture difference QA task: For this task, we manually constructed 49 questions, each formulated in the five languages while explicitly including the name of a specific country. In English, the questions refer to the United States; in Japanese, to Japan; in Chinese, to China; in French, to France; and in Arabic, to Saudi Arabia. The following is an example. When the question is asked in different languages, referring to their respective countries, the answers vary. Furthermore, the process does not require manual answer collection, as elaborated in Section 4.3. Below, we present a Japanese example (English meaning in Figure 2):

本国の最高峰は_です。答え*:* "

4 Experimental Settings

To derive general conclusions considering linguistic diversity, we selected one CPM and one BLM of comparable size for Japanese, Chinese, French, and Arabic, respectively, and conducted our experiments alongside the English-centric Llama 2 family to investigate how training data influences latent language probabilities. Details of the selected models are presented in Appendix 1.

To ensure that the model can complete the task successfully, we use few-shot prompting (in the same language setting) to teach the LLM the task format in all experiments, with each shot structured as described in Section 3.4. We then monitored the probabilities of different language versions of the answers being generated at each layer and visualized the results in graphical form.

4.1 Design of Cloze Task

The first experiment aimed to determine our first research question: whether non-English-centric models could effectively utilize their specialized languages within its intermediate layers. To this end, we conduct monolingual cloze tasks in the corresponding languages on models specialized for Japanese, Chinese, French, and Arabic. We use two-shot prompting in this task, followed the previous work (Wendler et al., 2024).

4.2 Design of Translation Task

To investigate our second research question: which latent language is used when processing cross-lingual tasks, we conduct the translation tasks on these models and observe changes in the latent language probability by varying the source and target languages. Our dataset includes four languages: English, French, Japanese, and Chinese. Among these, En-Fr and Zh-Ja form two pairs of linguistically similar languages, allowing us to investigate how input source and output target language similarities to latent language influence the latent language usage on Japanese- and Chinese-specialized models. We use four-shot prompt in this task, followed by the previous work (Wendler et al., 2024).

4.3 Design of Culture Difference QA Task

When interacting with LLMs, users typically communicate in their native language without explicitly specifying their identity, nationality, or cultural background. Ideally, LLMs should generate responses that align with the cultural context associated with the language being used.

Because the cloze task demonstrated that non-English-centric models predominantly rely on their specialized language when processing tasks in that language, this experiment compares the biases in the intermediate layers of English-centric models and non-English-centric models when answering culture difference questions.

As described in Section 3.4, we design questions in five languages, each referring to its respective country. The experiment follows the two steps below.

1. **Querying the model with country-specific questions.** We separately query a non-English-centric model in English and its specialized language (e.g., Japanese) about the

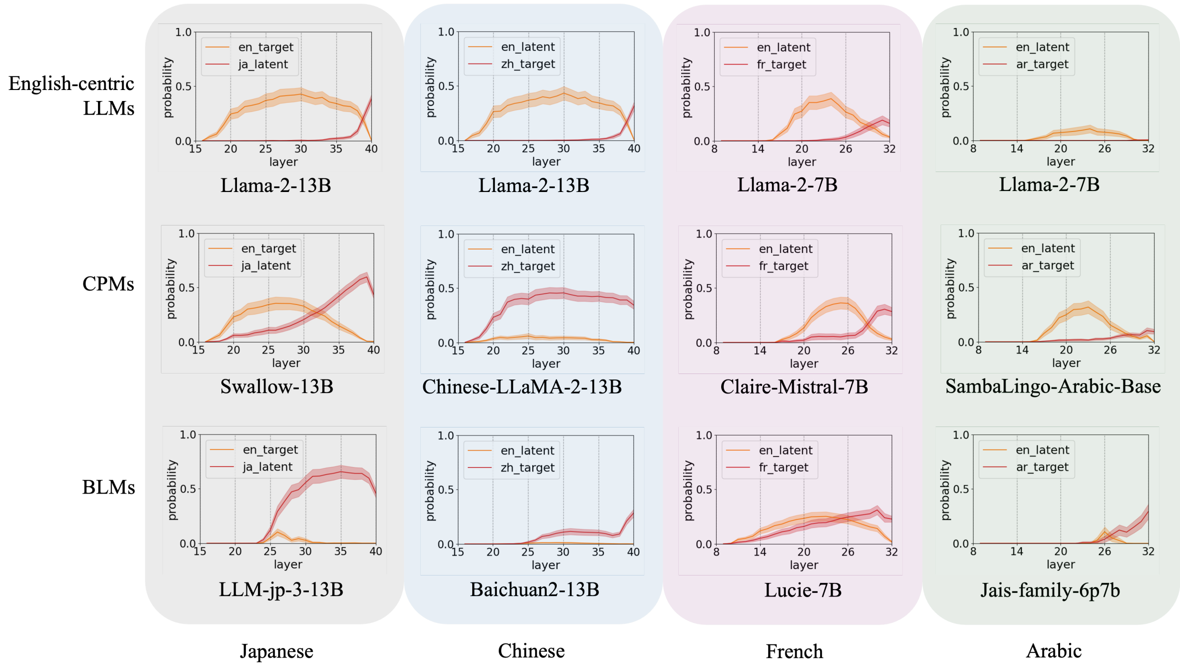


Figure 3: **Cloze task results of three kinds of LLMs in its specialized language.** Each row represents a model of the same category, while each column corresponds to the language used in the cloze task evaluation. The orange line represents the probability of English answers, the red line represents the probability of the models' specialized language answers. The x-axes denote the model's layer index, while the y-axes represent the probability of the answer in each language. The translucent areas indicate 95% Gaussian confidence intervals.

United States and its respective country (e.g., Japan) with country names explicitly attached. The model generates responses using a greedy decoding algorithm, and the generated two answers are recorded as two references, representing the cultural knowledge associated with the two countries.

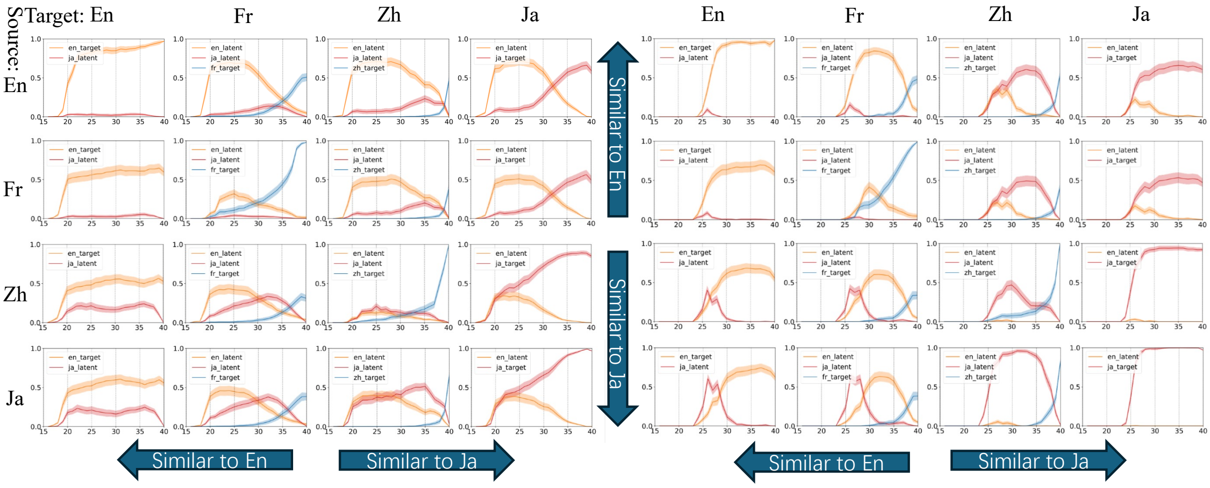
2. **Querying the model with country-free questions in its specialized language.** We modify the original question by replacing the explicit country name with "our country" and query the model in its specialized language again (e.g., Japanese). By monitoring the probability of two reference answers in the intermediate layers, we can recognize how cultural bias is internally encoded within the model's reasoning process.

5 Results

5.1 Cloze Task: Analysis of Input in Specialized Languages

To address our central question—whether non-English-centric large language models (LLMs) use English as a latent language or rely on their specialized language—we conducted cloze tasks in four languages (Japanese, Chinese, French, and Arabic). Figure 3 presents the intermediate layer latent language probabilities of English-centric LLMs and eight non-English-centric models spanning two categories (CPMs and BLMs). Consistent with findings from a previous study(Wendler et al., 2024), our results confirm that English-centric LLMs rely on English in their intermediate representations, even when processing tasks in other languages.

In contrast, CPMs exhibit a bilingual latent language pattern: the specialized language appears in the early layers, but most of these models predominantly rely on English. BLMs, meanwhile, predominantly rely on its specialized language from their early layers, using English only minimally. One outlier is the BLM Lucie-7B, which occasionally assigns a higher probability to English terms; this likely stems from lexical overlap between English and French, where certain English words used in the cloze tasks also appear in French, thereby influencing the model's intermediate representations. In summary, these findings suggest that **language-specific models (CPMs and BLMs) incorporate their specialized language—either partially or entirely—in their latent representations.**



(a) *Japanese* CPM: Swallow-13B

(b) *Japanese* BLM: LLM-jp-3-13B

Figure 4: **Comparison of Translation Task Patterns Between CPMs and BLMs.** (a) results for Japanese CPM Swallow-13B, (b) results for BLM LLM-jp-3-13B. Each row represents a source language in the translation task, while each column corresponds to a target language. The **orange line** represents the probability of **English** answers, the **red line** represents the probability of **Japanese** answers, and the **blue line** represents the probability of answers in **other languages**. The x-axes denote the model's layer index, starting from the 15th layer, while the y-axes represent the probability of the answer in each language. The translucent areas indicate 95% Gaussian confidence intervals.

5.2 Translation Task: Analysis of Input in non-specialized Languages

To investigate which latent language these non-English-centric models employ when handling the cross-lingual translation task, we vary both the input source and output target languages.

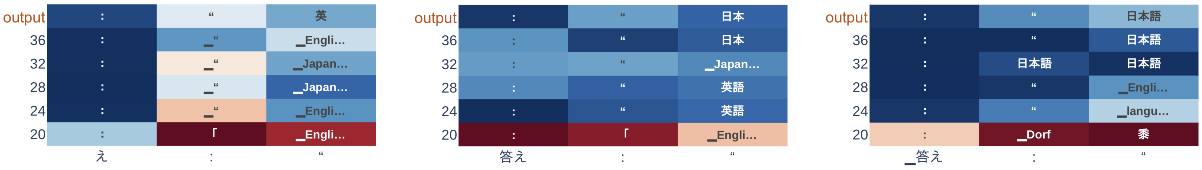
We specifically focus on Japanese-specialized models here. We investigate the latent language on translation task on two Japanese-specialized models: the CPM-based architecture (left) and the BLM-based one (right), as illustrative examples in Figure 4. Additional results for other languages are provided in the Appendix B.1, where we observe similar behaviors for Chinese-specialized models.

Within each subfigure (a) or (b), the diagonal cells represent scenarios in which the source and target languages coincide (i.e., repetition rather than translation). Examining each row (fixed source language) from left to right shows an increasing similarity of the target language to Japanese, and accordingly, both models exhibit a rising probability of Japanese in later intermediate layers. Likewise, scanning each column (fixed target language) from top to bottom reveals that a gradually more Japanese-like source boosts the activation of Japanese in earlier intermediate layers. These observations indicate that models with multiple latent languages choose which latent language to activate based on its similarity to the source or target. The two categories of models also have distinct patterns: **non-English-centric CPMs consistently utilize both Japanese and English as latent language, while BLMs exhibit a stronger propensity toward a single latent language.**

Furthermore, we observe a distinct phenomenon—here referred to as the "Probabilistic Cascade" for BLMs: during multilingual processing, the probability of a latent language closer to the source first surges, then transitions to another latent language more akin to the target, and finally culminates in the target language output. Overall, this study shows that language-specific models—here, specialized for Japanese—leverage both English and their specialized language as latent languages across intermediate layers when handling multilingual content, suggesting they could be adaptable to typologically similar languages.

5.3 Culture Difference QA

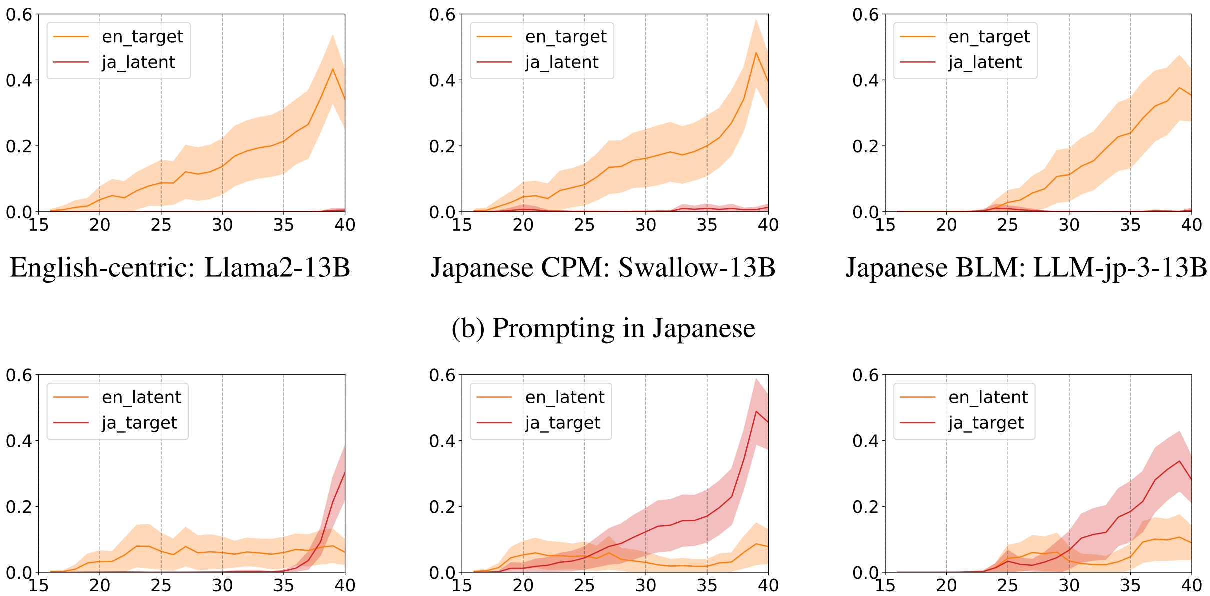
Given that non-English-centric models can help mitigate biases, it is important to examine how their internal latent language patterns shape cultural biases. In particular, we aim to understand how these latent patterns influence a model's semantic representations when it processes culturally specific questions. Figure 5 illustrates this phenomenon on Japanese via a logit lens analysis (panels (a), (b), and (c) show Llama-2, Swallow, and



(a) English-centric: Llama-2 (b) CPM: Swallow (c) BLM: LLM-jp-3

Figure 5: **Logit lens results of intermediate layers of three models,** (a) Llama-2, (b) Swallow, (c) LLM-jp. The input prompt is “本国の公用語は_です。 答え:",” which means “The official language of our country is _ . Answer:""” with the answer being “日本語” (Japanese). The figure shows the highest probability token from the intermediate layers starting from layer 20.

(a) Prompting in English



English-centric: Llama2-13B Japanese CPM: Swallow-13B Japanese BLM: LLM-jp-3-13B

(b) Prompting in Japanese

English-centric: Llama2-13B Japanese CPM: Swallow-13B Japanese BLM: LLM-jp-3-13B

Figure 6: **Comparison of English-centric and Japanese-specialized models when processing culture difference QAs.** (a) shows the intermediate layer probabilities of the reference answers generated from prompts explicitly mentioning “the United States” and “Japan,” when the model is prompted in **English** using the culturally neutral phrase “our country.” (b) shows the same setting using prompts written in **Japanese** with the corresponding term “本国.” The x-axes denote the model’s layer index, starting from the 15th layer, while the y-axes represent the probability of the cultural reference answers. The translucent areas indicate 95% Gaussian confidence intervals.

LLM-jp respectively). We prompt each model in Japanese for the official language of “our country” (“本国の公用語は_です。 答え:”) and examine the highest-probability tokens from layer 20 onward. Llama-2 initially generates an English-centric token sequence referencing the English; although the model later considered Japanese, it ultimately generated the incorrect answer “英語” (English). By contrast, the Japanese-specialized models (Swallow and LLM-jp) exhibit direct alignment with the Japanese context in the earlier layers, generating tokens for the correct answer “日本語” (Japanese) far earlier.

To further examine these patterns in terms of overall probability distributions, Figure 6 tracks the probability of each model generating a Japanese versus an English answer across intermediate layers. Figure 6 (a) shows that when prompting in English with culturally neutral expressions such as our country, all three models internally favor English-aligned representations. When prompting in Japanese using the culturally neutral term “本国,” from Figure 6 (b), we observe that Llama-2 maintains a high likelihood of producing English-centric responses in most mid-layers, only converging on the Japanese context near the end. In contrast, the Japanese-specialized models remain consistently aligned with the Japanese cultural context from earlier layers, highlighting their capacity to "think" in the specialized language more effectively. This result indicates that **non-English-centric models can reason directly in their specialized language from the outset, allowing them to generate more culturally appropriate responses.** Additional experimental results and comparisons of other languages are presented in Appendix B.2, where similar trends are observed.

6 Conclusion

In this study, we leverage Logit Lens to analyze the latent languages of non-English-centric LLMs. Our findings in the monolingual cloze task indicate that CPMs exhibit a mixture of latent languages, blending their specialized language with English, while BLMs activate the latent language most similar to the input dynamically. While conducting cross-lingual translation, both source and target languages influence latent language activation, with higher linguistic similarity leading to stronger activation. A typical pattern termed 'Probabilistic Cascade' is observed: the probability of latent languages peaks and then declines alternately, and ultimately shifts the peak to the target language. Finally, we observe that English-centric models introduce cultural biases, whereas non-English-centric models better capture their respective cultural contexts. These insights contribute to understanding multilingual bias and guiding future improvements.

7 Limitations

Despite our efforts to construct a high-quality dataset, certain limitations remain in our study. First, while we ensured that word pairs across languages do not overlap during dataset construction, the inherent lexical similarities between languages, such as English and French, pose a challenge. Specifically, although the English and French answers used in the cloze task were explicitly selected to avoid direct overlap, some chosen English words may also exist as valid French words with similar meanings. This unintended overlap may contribute to higher probabilities for English in the intermediate layers of the French model. A more rigorous dataset construction process could mitigate this issue, potentially leading to more reliable results in French model evaluations.

Second, the Arabic dataset was generated using translations from GPT-4, which limits our ability to manually verify the accuracy of the translations or determine whether the selected words are the most commonly used ones in Arabic-speaking regions. This limitation may explain the lower probability of Arabic responses when evaluating Arabic-specialized models.

Third, in the *culture difference QA* experiment, we constructed only 49 questions, which is a relatively small sample size. Expanding the dataset in future work would enhance the robustness of our findings. Additionally, in this experiment, we selected a single representative country for each language, yet in reality, these languages are spoken across multiple regions with potentially varying cultural contexts. Future work should consider a broader selection of representative regions to improve the generalizability of the results.

8 Ethical Considerations

This study analyzes the latent language dynamics of non-English-centric LLMs and how they influence cultural bias in the model's intermediate layers. While we examine bias in intermediate layers, we do not propose direct mitigation strategies, and biases in training data may still influence model behavior.

Our evaluation focuses on a limited set of languages, which may affect generalizability. Additionally, while non-English-centric models reduce English cultural bias, other biases may persist. Future work should explore broader linguistic contexts and bias mitigation techniques to promote fairness in LLMs.

9 Acknowledgment

This work was supported by the "R&D Hub Aimed at Ensuring Transparency and Reliability of Generative AI Models" project of the Ministry of Education, Culture, Sports, Science and Technology.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, et al. 2024. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms. *arXiv preprint arXiv:2407.03963*.

Badr AlKhamissi, Muhammad ElNokrashy, Mai AlKhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. *arXiv preprint arXiv:2402.13231*.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2012. Chinese characters mapping table of Japanese, traditional Chinese and simplified Chinese. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2149–2152, Istanbul, Turkey. European Language Resources Association (ELRA).

Zoltan Csaki, Bo Li, Jonathan Li, Qiantong Xu, Pian Pawakapan, Leon Zhang, Yun Du, Hengyu Zhao, Changran Hu, and Urmish Thakker. 2024. Sambalingo: Teaching large language models new languages. *Preprint*, arXiv:2404.05829.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.

Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. 2024. Bias of ai-generated content: an examination of news produced by large language models. *Scientific Reports*, 14(1):5224.

Manuel Faysse, Patrick Fernandes, Nuno Guerreiro, António Loison, Duarte Alves, Caio Corro, Nicolas Boizard, João Alves, Ricardo Rei, Pedro Martins, et al. 2024. Croissantllm: A truly bilingual french-english language model. *arXiv preprint arXiv:2402.00786*.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. *arXiv preprint arXiv:2404.17790*.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79.

Olivier Gouvert, Julie Hunter, Jérôme Louradour, Christophe Cérisara, Evan Dufraisse, Yaya Sy, Laura Rivière, and Jean-Pierre Lorré. 2025. The lucie-7b llm and the lucie training dataset: open resources for multilingual language generation.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*.

Julie Hunter, Jérôme Louradour, Virgile Rennard, Ismaïl Harrando, Guokan Shang, and Jean-Pierre Lorré. 2023. The claire french dialogue dataset. *arXiv preprint arXiv:2311.16840*.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2023. Having beer after prayer? measuring cultural bias in large language models. *arXiv preprint arXiv:2305.14456*.

Shangrui Nie, Michael Fromm, Charles Welch, Rebekka Görge, Akbar Karimi, Joan Plepi, Nazia Afsan Mowmita, Nicolas Flores-Herr, Mehdi Ali, and Lucie Flek. 2024. Do multilingual large language models mitigate stereotype bias? *arXiv preprint arXiv:2407.05740*.

Nostalgebraist. 2020. Interpreting gpt: The logit lens. `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`. Accessed: 2024-07-28.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Alham Fikri Aji, Zhengzhong Liu, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *Preprint*, arXiv:2308.16149.

Fan-Keng Sun, Cheng-Hao Ho, and Hung-Yi Lee. 2020. Lamol: Language modeling for lifelong language learning. In *International Conference on Learning Representations*.

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

Yan Tao, Olga Viberg, Ryan S Baker, and Rene F Kizilcec. 2023. Auditing and mitigating cultural bias in llms. *arXiv preprint arXiv:2311.14096*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. In *The Eleventh International Conference on Learning Representations*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. *arXiv preprint arXiv:2402.10588*.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

松下 達彦, 陳 夢夏, 王 雪竹, and 陳 林柯. 2020. 日中対照漢字語データベースの開発と応用. 日本語教育, 177:62–76.

A Model Details

Table 1 presents details of the models we tested. For CPMs, the language proportions refer to those used during the CPT process.

B Extra Results

B.1 Translation Task: Analysis of Input in non-specialized Languages

Figure 7 presents the results of the Chinese-specialized models in Experiment 2. These models exhibit the same pattern observed in the analysis of the Japanese model in the main text. For the BLM Baichuan2, we also observe the Probabilistic Cascade phenomenon.

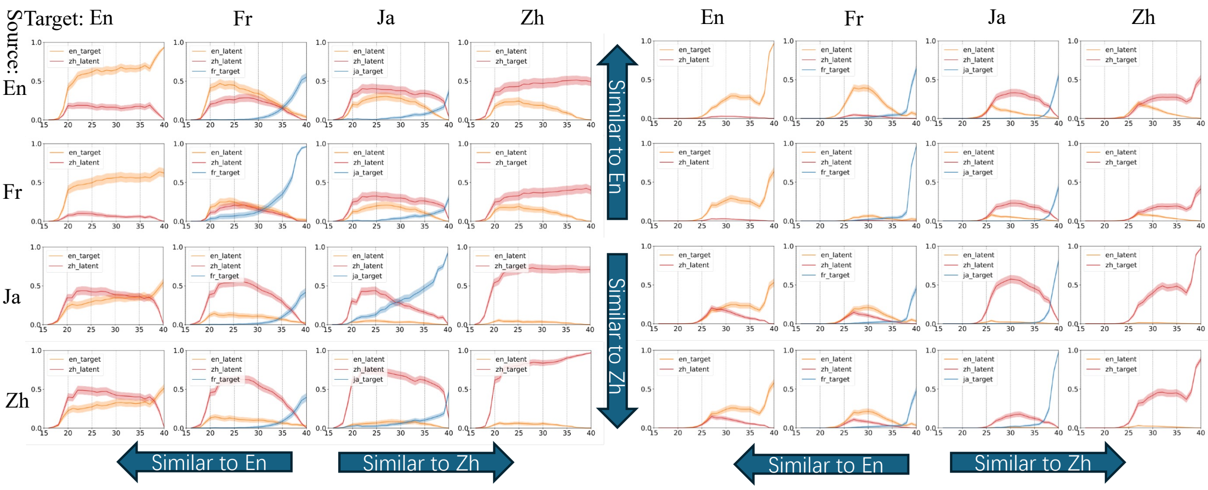
Figure 8 presents the results of the French-specialized model in Experiment 2. Across all settings, the intermediate layers of the French CPM show a low probability for French. In contrast, the French BLM exhibits a higher probability for French in its intermediate layers, achieving a more balanced representation. However, as noted in the Limitations 7, our dataset has shortcomings for evaluating French-specialized models.

B.2 Culture Difference QA: Analysis on Culture Difference Questions

As shown in Figure 9, 10, 11, for the English-centric Llama2, across all tested languages in the culture difference QA task, intermediate layers consistently first generate English answers aligned with the U.S. cultural context. In contrast, non-English-centric models do not exhibit this tendency when processing culture difference QAs in their specialized languages. This suggests that non-English-centric models demonstrate a reduced susceptibility to English cultural bias.

Category	Model	Parameter	Proportion in pre-training data			From scratch
			En	Specialized language	Other	
English-centric	Llama 2	7/13B	89.7%	0.1%	10.2%	Yes
CPM (Ja)	Swallow	13B	10%	90%(Ja)	0%	Llama-2 based
BLM (En + Ja)	LLM-jp-3	13B	45.8%	48.6%(Ja)	7.4%	Yes
CPM (Zh)	ChineseLLaMA2	13B	0%	100%(Zh)	0%	Llama-2 based
BLM (En + Zh)	Baichuan2	13B	-%	-%(Zh)	-%	Yes
CPM (Fr)	Claire-Mistral	7B	0%	100%(Fr)	0%	Mistral based
BLM (En + Fr)	Lucie	7B	33.3%	32.1%(Fr)	34.6%	Yes
CPM (Ar)	SambaLingo-Arabic-Base	7B	25%	75%(Ar)	0%	Llama-2 based
BLM (En + Ar)	Jais-family	6.7B	59.0%	29.4%(Ar)	11.6%	Yes

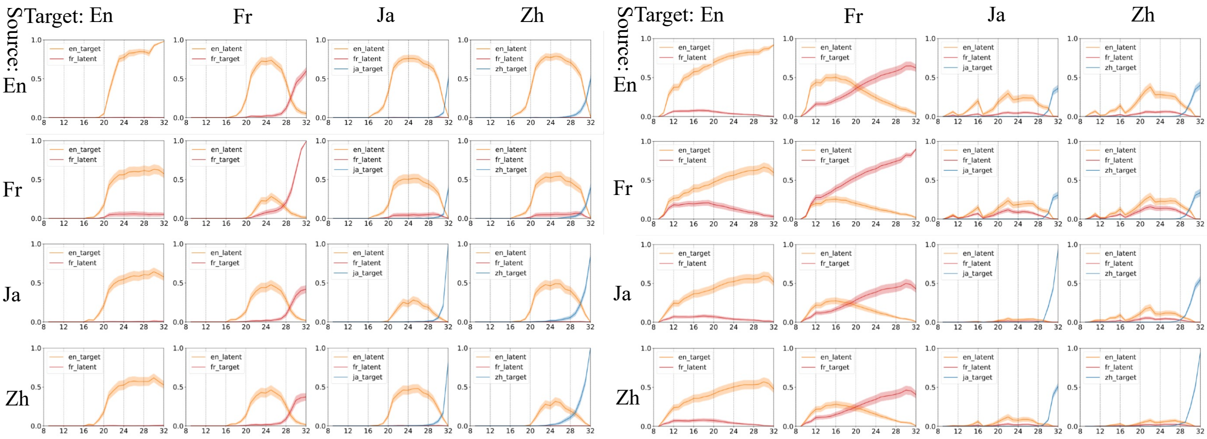
Table 1: Categorization of LLMs based on language proportion and training strategy. To be noted, Baichuan2 is primarily pre-trained on English and Chinese data, but the exact proportions have not been disclosed.



(a) *Chinese* CPM: ChineseLlaMA2

(b) *Chinese* BLM: Baichuan2

Figure 7: **Comparison of Translation Task Patterns Between CPMs and BLMs.** (a) results for Chinese CPM ChineseLlaMA2-13B, (b) results for Chinese BLM Baichaun2-13B. Each row represents a source language in the translation task, while each column corresponds to a target language. The **orange line** represents the probability of **English** answers, the **red line** represents the probability of **Chinese** answers, and the **blue line** represents the probability of answers in **other languages**. The x-axes denote the model's layer index, starting from the 15th layer, while the y-axes represent the probability of the answer in each language. The translucent areas indicate 95% Gaussian confidence intervals.



(a) *French* CPM: Claire-Mistral-7B (b) *French* BLM: Lucie-7B

Figure 8: **Comparison of Translation Task Patterns Between CPMs and BLMs.** (a) results for French CPM Claire-Mistral-7B, (b) results for French BLM Lucie-7B. Each row represents a source language in the translation task, while each column corresponds to a target language. The **orange line** represents the probability of **English** answers, the **red line** represents the probability of **French** answers, and the **blue line** represents the probability of answers in **other languages**. The x-axes denote the model's layer index, starting from the 15th layer, while the y-axes represent the probability of the answer in each language. The translucent areas indicate 95% Gaussian confidence intervals.

(a) Prompting in English

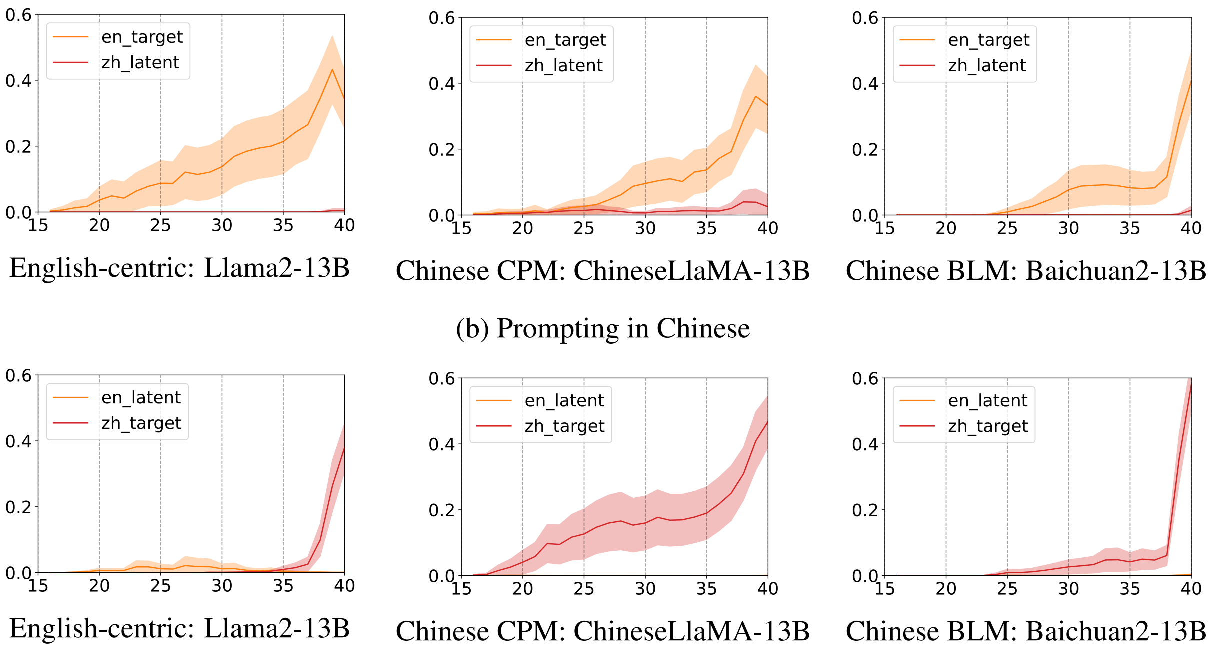


Figure 9: **Comparison of English-centric and Chinese-specialized models when processing culture difference QAs.** (a) shows the intermediate layer probabilities of the reference answers generated from prompts explicitly mentioning "the United States" and "China," when the model is prompted in **English** using the culturally neutral phrase "our country." (b) shows the same setting using prompts written in **Chinese** with the corresponding term "我国." The x-axes denote the model's layer index, starting from the 15th layer, while the y-axes represent the probability of the cultural reference answers. The translucent areas indicate 95% Gaussian confidence intervals.

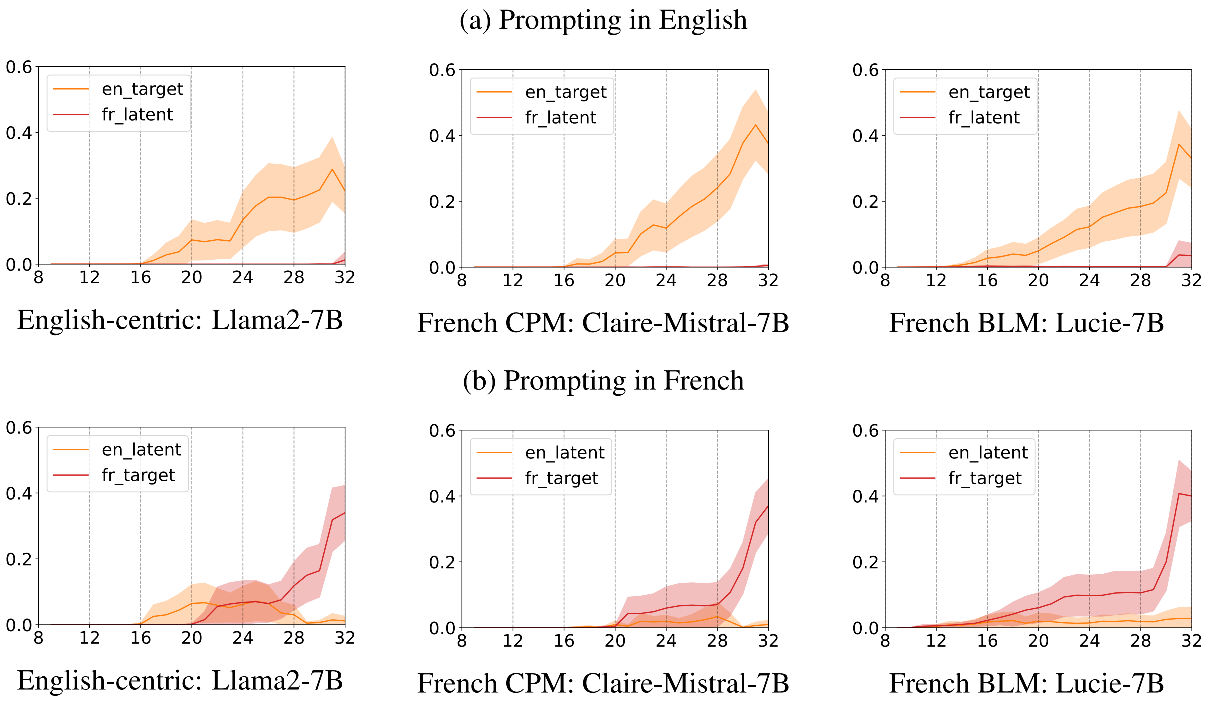


Figure 10: **Comparison of English-centric and French-specialized models when processing culture difference QAs.** (a) shows the intermediate layer probabilities of the reference answers generated from prompts explicitly mentioning “the United States” and “France,” when the model is prompted in **English** using the culturally neutral phrase “our country.” (b) shows the same setting using prompts written in **Chinese** with the corresponding term “de notre pays.” The x-axes denote the model’s layer index, starting from the 8th layer, while the y-axes represent the probability of the cultural reference answers. The translucent areas indicate 95% Gaussian confidence intervals.

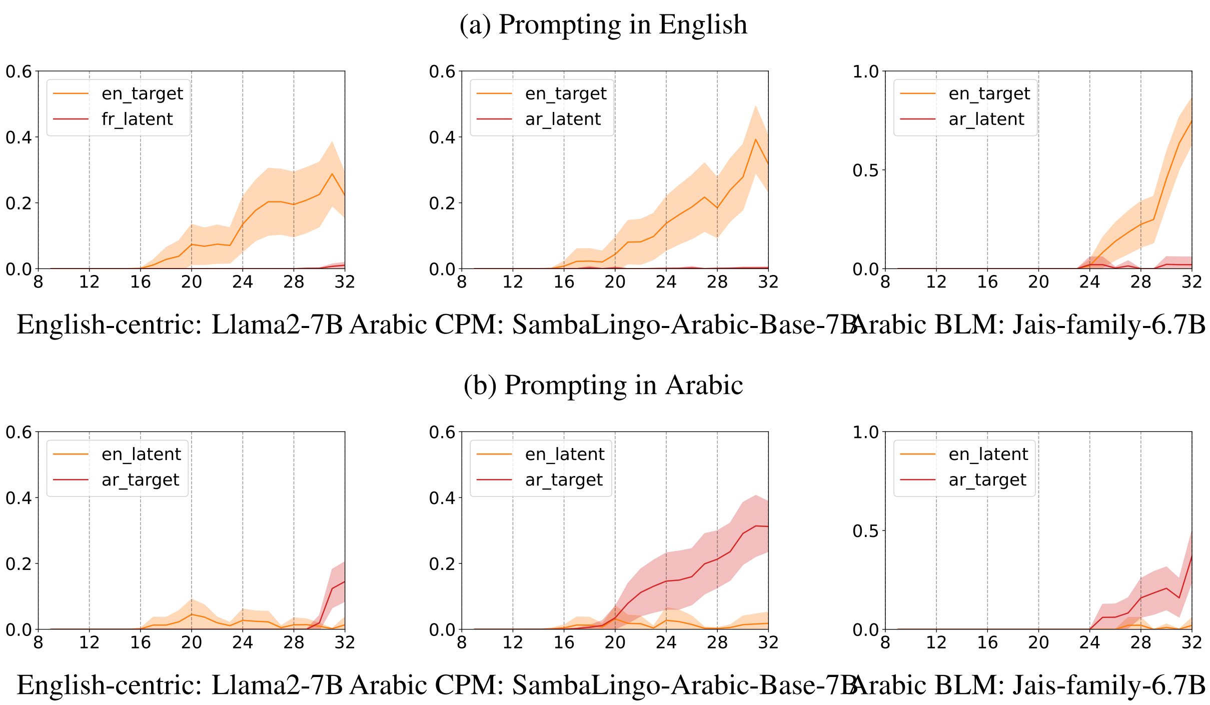


Figure 11: **Comparison of English-centric and Arabic-specialized models when processing culture difference QAs.** (a) shows the intermediate layer probabilities of the reference answers generated from prompts explicitly mentioning “the United States” and “Saudi Arbia,” when the model is prompted in **English** using the culturally neutral phrase “our country.” (b) shows the same setting using prompts written in **Chinese** with the corresponding term. The x-axes denote the model’s layer index, starting from the 8th layer, while the y-axes represent the probability of the cultural reference answers. The translucent areas indicate 95% Gaussian confidence intervals.