What Language Do Japanese-specialized Large Language Models Think in?

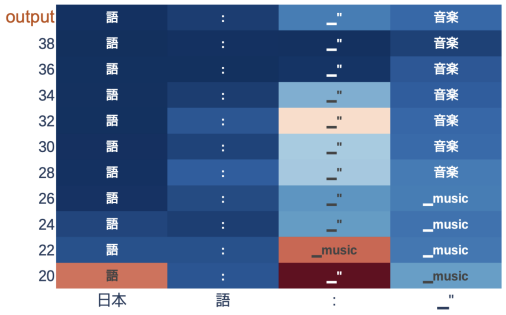
Chengzhi Zhong[1] Fei Cheng[1] Qianying Liu[2] Junfeng Jiang[3]
Zhen Wan[1] Chenhui Chu[1] Yugo Murawaki[1] Sadao Kurohashi[1,2]
[1] Kyoto University, Japan
[2] National Institute of Informatics, Japan
[3] The University of Tokyo, Japan
{zhong, feicheng, wan, chu, murawaki,kuro}@nlp.ist.i.kyoto-u.ac.jp
ying@nii.ac.jp
jiangjf@is.s.u-tokyo.ac.jp

Abstract

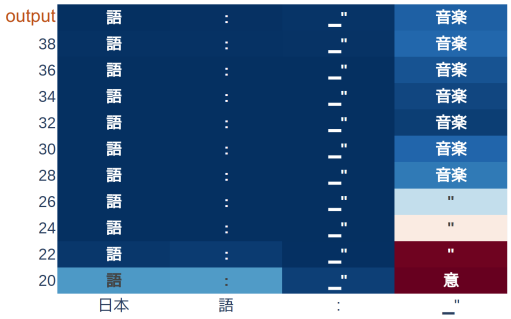
In this study, we investigate whether large language models (LLMs) trained with substantial Japanese data exhibit higher probabilities for Japanese in their intermediate layers when projected onto the vocabulary space (a.k.a latent languages). Focusing on Llama2 (English-centric), Swallow (continued in Japanese), and LLM-jp (balanced English-Japanese), we find Llama2 relies mainly on English, while Swallow and LLM-jp use both Japanese and English as latent languages. Moreover, input and target languages both influence the probability distribution between latent languages.

1 Introduction

Recent studies have shown that English-centric large language models (LLMs) display distinct patterns in their intermediate layers, where the language distribution is heavily skewed towards English when generating underrepresented languages [1]. This raises our interest in investigating whether LLMs utilize the dominant non-English languages from their training corpora in their intermediate layers during generation. We examine three typical categories of models that are used to process Japanese: Llama2 [2], an **English-centric model**; along with two Japanese-specialized models Swallow [3], an English-centric model with **continued pre-training (CPT) in Japanese**; and LLM-jp [4], a model pre-trained on **balanced corpora of English and Japanese**. More details of these models are shown in Table 1.



(a) Japanese CPT: Swallow



(b) Balanced English and Japanese: LLM-jp

Figure 1: Logit lens results of Japanese-specialized models, (a) Swallow, (b) LLM-jp. The input prompt is "Français: "musique" - 日本語: "", a French-to-Japanese translation task with the answer "音楽" (music). The figure shows the highest probability token from the intermediate layers, starting from layer 20.

To investigate how the LLMs' behaviour in the intermediate layers, we employ the logit lens method [5], which unembeds each layer's latent representation into the vocabulary space. We verify the latent languages of the three types of models when processing Japanese: While Llama2

Table 1: Categorization of multilingual models based on language proportion and training strategy.

Model Category	Model	Proportion in pre-training data			Token	From scratch
		En	Ja	Other		
English-centric	Llama 2	89.70%	0.10%	10.20%	2,000B	Yes
Japanese CPT	Swallow	10%	90%	0%	100B	Llama-2 based
Balanced English and Japanese	LLM-jp	50%	50%	0%	300B	Yes

uses English as its latent language [1], in contrast, the Japanese CPT model Swallow utilizes both English and Japanese within its intermediate layers, as shown in Figure 1 (a). Meanwhile, Figure 1 (b) shows LLM-jp primarily utilizes Japanese as the latent language in this case.

To further find out the models' latent language when generating languages other than the dominant Japanese and English. We introduce a new setting in which non-Japanese and non-English languages are used as input and target languages to explore the behaviors of the intermediate layers. Our experiments show that in intermediate layers of the models, the latent language of Japanese-specialized models is a distribution over English and Japanese, with the probabilities of these distributions depending on their similarity to both input and target language. In the final layers , the internal predictions transform into the corresponding target language.

In summary, we confirm that Japanese-specialized models Swallow and LLM-jp exhibit two latent languages, English and Japanese. The utilization of these latent languages depends on their similarity to the input and target languages, reflecting a dynamic adjustment in internal language processing.

2 Related work

2.1 Multilingual Large Language Models

Current frontier large language models, such as GPT-4 [6], Gemini [7], and Llama-2 [2], are primarily trained with English-centric corpora, with other languages constituting only a small portion of the training data. Researchers have sought to enhance these models' multilingual capabilities through various methods. One approach involves continued pre-training with second-language data [8, 9, 10, 11, 12], as demonstrated by models like Swallow [3] based on Llama-2. While these approaches have proven effective, ongoing research aims to discover more efficient techniques to further improve the multilingual capabilities of large language models.

2.2 Mechanistic Interpretability

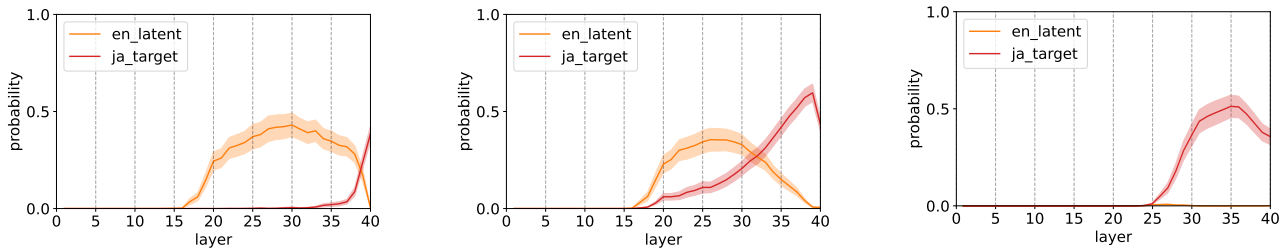
Mechanistic interpretability is the study of understanding how machine learning models work by analyzing their internal components and processes to elucidate the mechanisms that give rise to their behavior and predictions. It encompasses research lines like superposition [13], sparse autoencoders [14], circuit analysis [15] and so on. Within these studies, logits lens [5] and tuned lens [16] focus on decoding the probability distribution over the vocabulary from intermediate vectors of the model, aiding in the comprehension of how the model generates text in the target language. Previous study [1] showed that Llama-2 models have an abstract "concept space" that lies closer to English than to other languages. When Llama-2 models perform tasks such as translation between non-English languages, the probabilities in the intermediate layers initially focus on the English version of the answer and gradually shift to the target language.

In this work, we expand previous work and utilize these tools to study the distribution of latent languages in different categories of Japanese-specialized LLMs and examined how the probability of internal latent languages is associated with the target language.

3 Method

3.1 Logit Lens

In the last layer, LLMs use an unembedding matrix to project the hidden vectors onto the vocabulary dimensions. Then, a softmax function is applied to determine the output token. This process is called unembedding. By applying the same unembedding operation to the hidden vectors passed between the intermediate layers, we can obtain tokens generated by intermediate layers. *Logit lens* is a tool designed to achieve this purpose. Therefore, we leverage logit lens to calculate the probability for the model's inter-



Llama2-13b Swallow-13b LLM-jp-v2.0

Figure 2: **Comparison of English-centric and Japanese-specialized models when processing Japanese Cloze.** X-axes denote layer's index of the model, and y-axes denote probability of answer in each language. Translucent area show 95% Gaussian confidence intervals.

mediate layers to generate a specific token sequence.

3.2 Task Design

Dataset Construction. We first collect parallel words in four languages—English, French, Japanese, and Chinese. To obtain word pairs with different spellings but identical meanings, we construct this dataset based on part of the *Database of Japanese Kanji Vocabulary in Contrast to Chinese* (JKVC) [17]. Then, we use GPT-4 to do translation and obtain the corresponding English and French words or phrases, and then manually review and correct errors. The total size is 166. Based on the parallel words and following previous studies, we demonstrate the following prompts for two tasks, and the corresponding answers for examples will be the same Japanese word "原則" (principle). Models are asked to predict the answer, and we calculate the probability of the answer in the language we want to monitor. We use 4-shot for translation task and 2-shot for cloze task.

Translation task:

Français: "principe" - 日本語:"

Cloze task:

"__"は、基本的なルールや信念です。答え: "

4 Results

4.1 Analysis on Processing Dominant Language – Japanese

To investigate which latent language is used when processing Japanese, we conduct experiments to compare the latent language behaviors of three models when processing cloze task with Japaneses set as the target language.

The average result of cloze task is shown in Figure 2. Llama2, which is an English-dominant model, exhibits using English as latent language in its intermediate layers. In contrast, Swallow, which underwent CPT in Japanese, demonstrates a noticeable probability of Japanese in its intermediate layers. For LLM-jp, English probabilities are nearly absent in the intermediate layers. This indicates that these Japanese-specialized models lean to utilize Japanese more as the latent language when processing Japanese.

4.2 Analysis on non-Dominant Languages

We further investigate which latent language the models use when generating non-dominant languages, such as French and Chinese, compared to dominant languages. For this part, we test the models on translation tasks between different languages.

The average result is shown in Figure 3, the source language is always English. When the target language is also English, it becomes a repetition task. Following a left-to-right order, we gradually change the target language. It is observed that for both Swallow and LLM-jp, as the target language gets closer to Japanese, the probability of Japanese in the intermediate layers increases while that of English decreases. Additionally, for Swallow, English and Japanese are consistently intermixed in the intermediate layers, whereas for LLM-jp, the usage of English and Japanese in the intermediate layers is more isolated.

We also investigate how the source language affects the probability distribution of latent languages. We show those results in Appendix Figure 5. In this case, the target language is Japanese. When the source language is also Japanese, it becomes a repetition task. Following a left-to-right order, we gradually change the source language to increase its similarity to Japanese. The results are similar that the probability of Japanese in the intermediate layers

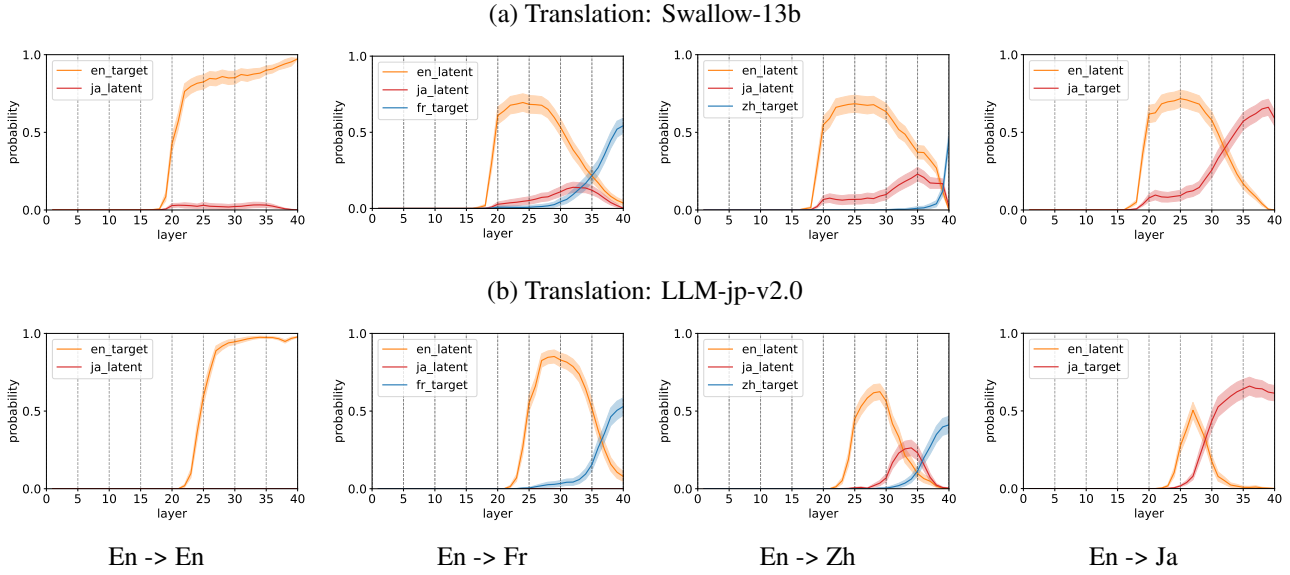


Figure 3: **Translation task results of two models with a fixed target language and varying source languages.** (a) results for Swallow-13b, (b) results for llm-jp-v2.0. X-axes denote layer's index of the model, and y-axes denote probability of answer in each language. Translucent area show 95% Gaussian confidence intervals.

increases while that of English decreases. In the selection of latent languages in the intermediate layers, the source language has a similar influence to the target language.

The results indicate that the activation of latent languages in LLMs depends on their similarity to the input and target languages.

4.3 How Is Culture Conflict QA Solved?

Because the models 'think' in latent languages, whether this affects the model's reasoning in QA tasks is a question worth discussing. Because some questions can have different answers in different cultural contexts across languages. Thus, We conduct a case study on this topic and use the logit lens to observe the intermediate layers of the models.

As shown in Figure 4, we ask the models about the start date of the school year in Japan with Japanese prompt. In Japan, the new school term begins in April. Llama-2's English-dominant intermediate layers prefer the answer "September/nine," which is the typical start date for American schools. The correct answer for Japan only appears in the latter layers where the probability is concentrated on the target language. In Swallow, the wrong answer "九" (nine) only appear once in layer 36. In contrast, the bilingual-centric LLM-jp does not exhibit this issue. You can see in the early layers that other numbers like "八" and 1 appear. But it is likely just due to the chaotic state in the early layers before the answer is determined. This indicates that, for such questions, the knowledge in the primary language context significantly influences the model's predictions. This provides an internal perspective on why operations like knowledge editing should focus on the model's primary language.

5 Conclusion

In this study, we demonstrate that the latent language of LLMs is majorly determined by the language of its training corpora. We confirm that Japanese-specialized Swallow and LLM-jp both utilize Japanese as their latent language when processing Japanese input.

Given that Swallow and LLM-jp exhibit the use of two internal latent languages, the degree to which each latent language is utilized depends on its similarity to the input and target languages. When the input language is more similar to Japanese, the proportion of Japanese in the intermediate layers increases, and the same applies to the target language. Additionally, For Swallow, the internal latent language distribution consistently includes both English and Japanese, with English being more dominant. In contrast, LLM-jp tends to favor a single language.

In future research, we aim to extend our investigation to models with other specific dominant languages, such as Chinese, French, and Arabic, to further explore the behavior and mechanisms of non-English-centric LLMs.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

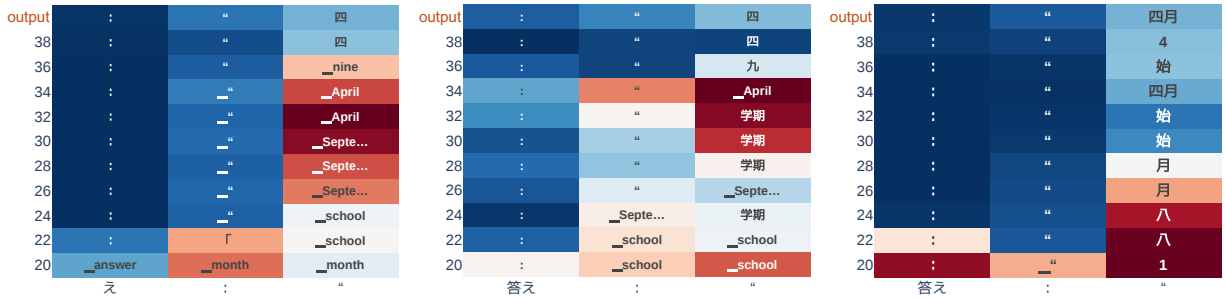
[14]Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In **The Twelfth International Conference on Learning Representations**, 2023.

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

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

[17]松下達彦, 陳夢夏, 王雪竹, 陳林柯. 日中対照漢字語データベースの開発と応用. 日本語教育, Vol. 177, pp. 62–76, 2020.

A Culture Conflict QA

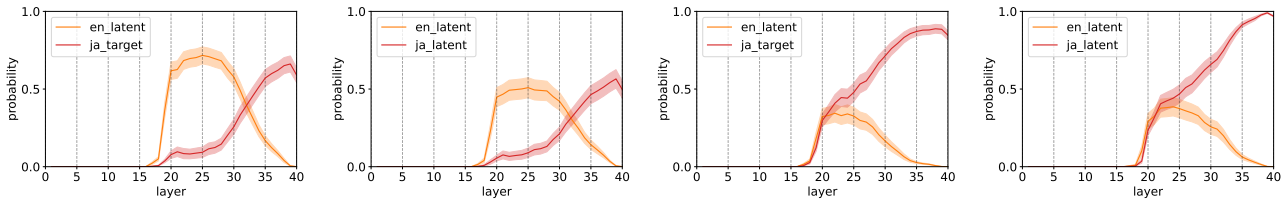


(a) English-centric: Llama-2-13b (b) Japanese CPT: Swallow-13b (c) Balanced English and Japanese: LLM-jp-v2.0

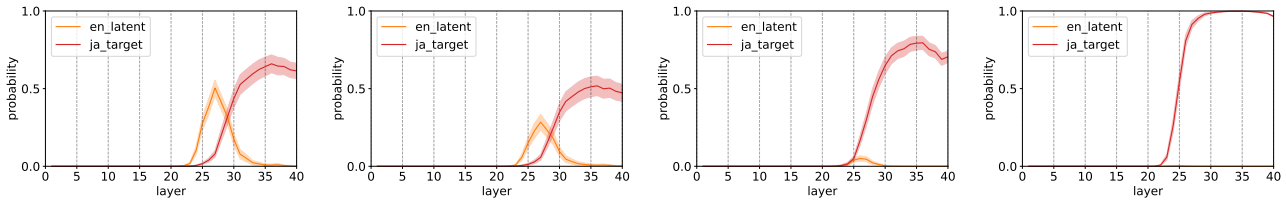
Figure 4: **Results of culture conflict question.** We use one-shot format prompts. The question is: 「日本の学校新学期が始まる月は：__月、答え：」(The month when the new school term starts in Japan is: _ month, answer:). The correct answer is 「四」(April). The colors in the figures represent entropy: blue indicates the probability is concentrated on the top tokens, while red means it is dispersed across the vocabulary.

B Extra Results

(a) Translation: Swallow-13b



(b) Translation: LLM-jp-v2.0



En -> Ja Fr -> Ja Zh -> Ja Ja -> Ja

Figure 5: **Translation task results of two models with a fixed source language and varying target languages.** (a) results for Swallow-13b, (b) results for llm-jp-v2.0. X-axes denote layer's index of the model, and y-axes denote probability of answer in each language. Translucent area show 95% Gaussian confidence intervals.