Kanji lens: Which Script Do Japanese LLMs Traverse Internally?

Anonymous ACL submission

Abstract

The Japanese language uses three scripts with different typical functions: kanji, hiragana, and katakana. When a Japanese LLM generates output in hiragana, which script type is reflected in its intermediate representations? We analyze layerwise representations using logit lens and identify the latent kanji-fication phenomenon in Japanese-centric models; kanji representations become dominant in later intermediate layers even when the final output is in hiragana. We show that kanji-fication is pronounced in Japanese-centric models, but subtle in English-centric and continual pre-training models. We also find that kanji-fication weakens for loanwords, where katakana becomes dominant. These results suggest that internal script representations reflect preferred script forms in the vocabulary, as well as the language distribution of pre-training data.

1 Introduction

The Japanese language uses three scripts: kanji, hiragana, and katakana. These scripts have different typical functions. Kanji often expresses lexical stems, hiragana is widely used for inflectional endings and grammatical elements, and katakana is common for loanwords and foreign names. For example, the Japanese word corresponding to "learn" can be written as "学習" (kanji), "がくしゅう" (hiragana), or "ガクシュウ" (katakana), and many words and expressions permit multiple script variants. The internal processing of large language models (LLMs) for such a language is still not well understood. This paper poses the question of which script type is reflected in the internal processing of Japanese LLMs. When a model outputs hiragana, as in Figure 1, does it internally maintain hiragana, or does it transiently rely on another script type such as kanji? We refer to the tendency for kanji representations to emerge in intermediate layers as latent kanji-fication.

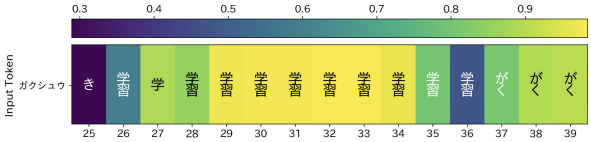


Figure 1: Results of analysis of the katakana-to-hiragana conversion task in a Japanese-centric model (Top-1 token heatmap). The most likely token at each layer is visualized. Kanji representations appear in intermediate layers, converging to hiragana in the final layers.

To examine which tokens are represented inside LLMs, we use logit lens (Nostalgebraist, 2020). Logit lens maps the hidden state at each layer to the vocabulary space and performs next-token prediction from that state, making it possible to inspect which tokens are generated by each layer. We use this method to trace the tokens produced across layers and to compute the probability of kanji at each layer, thereby investigating latent kanji-fication. In particular, we examine the layers in which latent kanji-fication appears, the model types that exhibit it, and whether it weakens for loanwords.

2 Related Work

Prior work has investigated how English-centric LLMs process non-English languages and what latent language they use internally (Wendler et al., 2024; Zhong et al., 2024). Two main approaches have been proposed to improve multilingualism: continual pre-training of English-centric models (Fujii et al., 2024) and training on balanced multilingual data from the outset (Aizawa et al., 2024).

Research on mechanistic interpretability (Elhage et al., 2021, 2022) has advanced the analysis of model internal representations. Methods such as logit lens (Nostalgebraist, 2020) and tuned lens (Belrose et al., 2023) have been proposed to visualize representations at each layer. Saji and Nakashole (2026) proposed RomanLens, which combines logit lens with activation patching (Conmy et al.,

 # Katakana → Hiragana (script conversion task)
Original: “ドクショ” *(dokusho / reading)* Hiragana: “とくしょ”
Original: “ベンキョウ” *(benkyou / studying)* Hiragana: “べんきょう”
Original: “ケンガク” *(kengaku / field trip)* Hiragana: “けんがく”
Original: “ジッケン” *(jikken / experiment)* Hiragana: “じっけん”
Original: “ガクシュウ” *(gakushuu / learning)* Hiragana: “

Figure 2: Example 4-shot prompt for the script conversion task. Romanizations and English glosses are added for readability; they do not appear in the actual prompt. The model generates the hiragana continuation after the opening quote of the 5th example.

2023), and reported “latent romanization,” where roman script representations temporarily appear in intermediate layers when English-centric models process non-roman-script languages. Building on this line of work, our study investigates how LLMs process Japanese across multiple script types and analyzes how internal representations are converted and integrated during that process.

3 Analysis of Kanji-fication

We use the script conversion task to analyze the transition of script types, and define the kanji-fication degree to quantify the reliance on kanji.

3.1 Script Conversion Task

To investigate the presence or absence of latent kanji-fication, we use a “katakana-to-hiragana” conversion task in which the input is fixed to katakana and the output to hiragana. The analysis is performed by comparing next-token predictions at the end of the prompt using logit lens. The prompt is given in a 4-shot format as shown in Figure 2, with the hiragana side of the 5th example stopped at the opening double quote to generate the continuation. The analysis targets the next-token prediction at the end of the prompt (immediately after the opening quote on the hiragana side), comparing the distribution across layers. We use the following two types of datasets, created using Claude 4[1]:

Kanji-writable words (182 pairs): Katakana-to-hiragana pairs for ordinary Japanese vocabulary that normally has a conventional kanji form, although it is written in katakana here for the conversion task (e.g., ガクシュウ→がくしゅう).
Loanwords (190 pairs): Katakana-to-hiragana pairs for imported words, mainly from Western languages, that are conventionally written in katakana and generally do not have a natural kanji form (e.g., コンピューター→こんぴゅーたー).

[1] https://www.anthropic.com/news/claude-4

The first category consists of ordinary Japanese vocabulary that usually has a conventional kanji form, even though it is written in katakana here for the conversion task. The second category instead consists of imported words, mainly from languages such as English, that are conventionally written in katakana and generally lack a natural kanji form.

These datasets allow us to verify how the internal processing of models differs between kanji-writable vocabulary and loanwords.

3.2 Procedure and Kanji-fication Degree

Logit lens (Nostalgebraist, 2020) is a method that maps hidden states at each layer to the vocabulary space to obtain next-token distributions. In this study, we aggregate the probabilities from the distribution at each layer obtained by logit lens by script type (kanji, hiragana, katakana, and alphabetic characters), computing the script-type probability distribution for each layer.

For the next-token prediction at the end of the prompt, let $p_l(\cdot)$ be the logit lens distribution at layer l, and let $\mathcal{K}$ be the set of tokens containing kanji. We define the kanji probability at layer l as $\sum_{t \in \mathcal{K}} p_l(t)$ and average it over all L layers to obtain the kanji-fication degree for that instance:

$$D_{\text{kanji}} = \frac{1}{L} \sum_{l=1}^{L} P_{\text{kanji}}(l) \quad (1)$$

We report the dataset average of this instance-level value as the model’s kanji-fication degree.

3.3 Analysis with Hiragana Sentences

The script conversion task described above uses a setting in which the candidate output tokens are restricted. To investigate latent kanji-fication in more general cases, we instead input hiragana sentences to the LLM with an instruction to output only hiragana and analyze the internal states. Specifically, for a short sentence s, the prompt is given in the format “Output only in hiragana: s.” We then visualize the internal states at each token position using logit lens.

4 Experiments and Results

Our primary analysis uses models in the 7–13B parameter range to compare the effect of the distribution of languages in the training data across different architectures. We use three models that differ in their primary training language: an English-centric model, Llama-3.1-8B-Instruct (Grattafiori et al.,

Model	Kanji-writable	Loanwords
Llama-3.1-8B-Instruct	0.022	0.021
Llama-3.1-Swallow-8B-Instruct-v0.1	0.031	0.024
LLM-jp-3-13B-Instruct	0.405	0.296
Sarashina2-7B	0.273	0.265
Yi-6B-Chat	0.069	0.095
Qwen3-8B-Base	0.265	0.249
Qwen3-Swallow-8B-SFT-v0.2	0.161	0.142

Table 1: Kanji-fication degree (average over all layers) for each model. Computed on the kanji-writable words dataset and the loanwords dataset.

2024); a Japanese continual pre-training model, Llama-3.1-Swallow-8B-Instruct-v0.5 (Fujii et al., 2024); and a Japanese-centric model, LLM-jp-3-13B-Instruct (Aizawa et al., 2024). To verify generalizability across different architectures, we also analyze Sarashina2-7B (SB Intuitions Corp., 2024) (a Japanese-centric model), Qwen3-8B-Base (Yang et al., 2025) and Qwen3-Swallow-8B-SFT-v0.2 (Yang et al., 2025; Fujii et al., 2024) (a multilingual base model and its Japanese-adapted variant), and Yi-6B-Chat (Young et al., 2024) (a Chinese-centric model). The effect of model scale is investigated separately in Section 4.3.

4.1 Results of Script Conversion Task

For each layer ℓ, we compute the script-type probability by aggregating the next-token distribution from logit lens over script classes. Figure 3 visualizes these layerwise script-type probability distributions for the three primary models on kanji-writable words and loanwords. In the English-centric model (Llama-3.1-8B-Instruct) and the continual pre-training model (Llama-3.1-Swallow-8B-Instruct-v0.5), the probability of alphabetic tokens is dominant throughout all layers, with hiragana probability rising sharply only in the final layer. In contrast, in the Japanese-centric model (LLM-jp-3-13B-Instruct), kanji probability increases in intermediate layers (around Layers 25–35), and converges to hiragana in the final layers, exhibiting "latent kanji-fication."

Table 1 shows the kanji-fication degree for all models. The Japanese-centric LLM-jp-3-13B-Instruct shows the highest degree (0.405 / 0.296), confirming that latent kanji-fication is strongest in models trained primarily on Japanese. Sarashina2-7B (0.273 / 0.265) and Qwen3-8B-Base (0.265 / 0.249) also show substantially elevated degrees, indicating that the phenomenon is not unique to LLM-jp but extends to other Japanese-centric and kanji-rich multilingual models. Qwen3-Swallow-8B-SFT-v0.2 shows a moderate degree (0.161 / 0.142), while the English-centric and continual pre-training Llama models show very low values (0.02–0.03). Yi-6B-Chat shows a low but non-negligible degree (0.069 / 0.095), consistent with its Chinese training data containing kanji characters. Script-type probability distributions for Sarashina2-7B, Qwen3-8B-Base, Qwen3-Swallow-8B-SFT-v0.2, and Yi-6B-Chat are shown in Appendix B.

4.2 Results of Hiragana Sentence Analysis

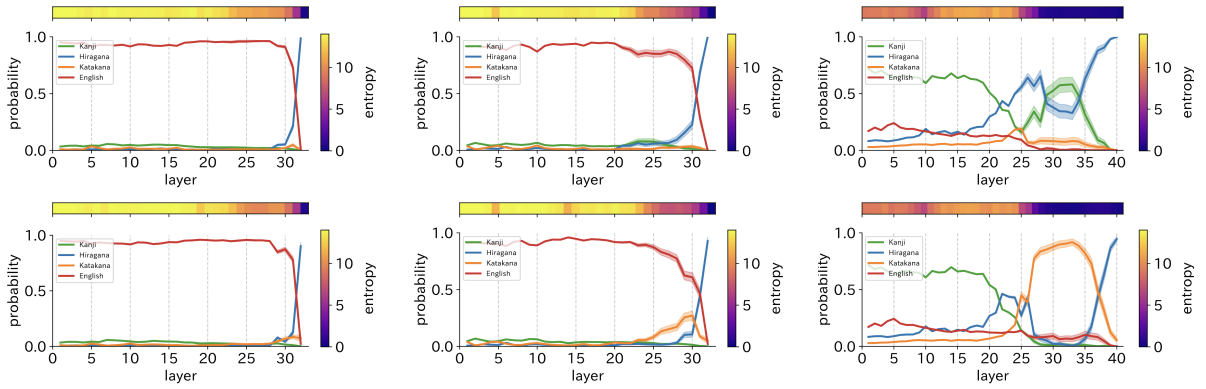
We present results for the case where the input sentence s is "あさごはんをたべました" (I ate breakfast). For each token position p and each layer ℓ, we compute the script-type probability $P_{\ell,p}(c)$ for script type c based on the next-token distribution obtained from logit lens and average it over positions to obtain $\bar{P}_\ell(c) = \frac{1}{N}\sum_p P_{\ell,p}(c)$, where N is the number of target tokens.

Figure 4 shows the results for the three primary models. In the English-centric model (Llama-3.1-8B-Instruct) and the continual pre-training model (Llama-3.1-Swallow-8B-Instruct-v0.5), alphabetic tokens dominate throughout most layers with hiragana rising only at the end. In contrast, in LLM-jp-3-13B-Instruct, kanji probability is elevated in intermediate layers, and the heatmap (Figure 5 in Appendix) confirms that kanji tokens such as "起き" (wake up) and "食べ" (eat) are predicted in intermediate layers before converging to hiragana in the final layers. The kanji-fication degrees in Table 1 indicate that this tendency holds broadly: models with higher kanji-fication degree (Sarashina2-7B, Qwen3-8B-Base) also exhibit kanji-dominant intermediate representations, while models with lower degrees (Llama, Qwen3-Swallow-8B-SFT-v0.2) show weaker or no such signal. Full results are provided in Appendix C.

4.3 Effect of Model Scale

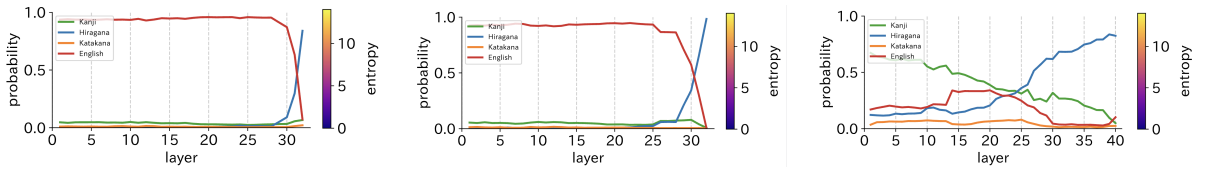
To investigate whether latent kanji-fication is also observed in larger models, we analyze larger variants from the same families: Sarashina2-70B, Qwen3-32B-Base, Qwen3-Swallow-32B-SFT-v0.2, Yi-34B-Chat, Llama-3.1-70B-Instruct, and Llama-3.1-Swallow-70B-Instruct-v0.3.

Across model families, larger scale mostly preserves and sometimes sharpens the qualitative behavior already seen at smaller sizes. Model families that already show latent kanji-fication at 7–13B, such as Sarashina2 and Qwen3-Base, continue to do so at 32B–70B, whereas the English-



(a) Llama-3.1-8B (b) Swallow-8B (c) LLM-jp-13B

Figure 3: Script-type probability distributions for the three primary models. Top: kanji-writable words. Bottom: loanwords.



(a) Llama-3.1-8B (b) Swallow-8B (c) LLM-jp-13B

Figure 4: Hiragana sentence analysis (script-type probability distributions) for the three primary models. See Figure 5 for corresponding heatmaps.

centric and continual pre-training Llama models remain largely alphabetic until the final layers. This pattern suggests that scale strengthens pre-existing language- and vocabulary-dependent structure rather than creating latent kanji-fication on its own. Detailed model-by-model results are provided in Appendix D.

4.4 Discussion

The experimental results suggest that models trained primarily on Japanese temporarily traverse kanji representations internally when producing hiragana. This is a phenomenon analogous to "latent romanization" reported by Saji and Nakashole (2026) for English-centric models, suggesting a connection between the language distribution of training data and the internal representations. On the other hand, although the continual pre-training model is additionally trained on Japanese data, similar kanji-dominant intermediate representations are not observed in our experiments. This result suggests that latent kanji-fication does not automatically arise merely from an increase in Japanese data volume, but rather that multiple factors may be involved, including the representation space and vocabulary distribution formed during the pre-training stage, as well as the design of the tokenizer.

5 Conclusion

We demonstrated that in Japanese LLMs, even when the output is in hiragana, latent kanji-fication occurs in which the corresponding kanji representations (e.g., "学習" for "がくしゅう") temporarily become dominant in intermediate layers. This result suggests that even when kanji is not needed in the output, corresponding concepts are represented in kanji internally.

This finding provides practical implications for the development of Japanese LLMs. Namely, if the model has sufficient hiragana-to-kanji conversion knowledge, it can be expected that the model will automatically perform processing through kanji representations internally even when input and output are in hiragana. Therefore, there is no need for users to consciously input in kanji or force kanji through prompts; instead, it may be more important to appropriately acquire script conversion knowledge during training. However, the extent to which such internal representation behavior contributes to inference efficiency or output quality is beyond the scope of this study and remains a topic for future investigation. As future work, we plan to analyze the impact of latent kanji-fication on downstream task performance.

Limitations

This study has several limitations. First, our analysis covers a limited slice of the model space, even though we observe latent kanji-fication in multiple models and at multiple scales. Second, the datasets used in our experiments were generated by an LLM (Claude 4), which may introduce biases in the selection of words. Third, while we use logit lens as our primary analysis tool, it provides only an approximation of the internal processing, and more precise methods such as tuned lens or activation patching could provide complementary insights. Finally, the practical impact of latent kanji-fication on downstream task performance has not been evaluated in this work.

References

Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Goto, Chio Hanada, and 1 others. 2024. LLM-jp: Large Japanese language models through balanced multilingual training. *arXiv preprint arXiv:2407.03963*.

Nora Belrose, Zach Furman, and Logan Smith. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

Arthur Conmy, Augustine N. Mavor-Parker, and Aengus Lynch. 2023. Towards automated circuit discovery for mechanistic interpretability. *arXiv preprint arXiv:2304.14997*.

Nelson Elhage, Tristan Hume, and Catherine Olsson. 2022. Toy models of superposition. *Transformer Circuits Thread*.

Nelson Elhage, Neel Nanda, and Catherine Olsson. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

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*.

Aaron Grattafiori and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Nostalgebraist. 2020. Logit lens: Understanding language model representations.

Alan Saji and Ndapandula Nakashole. 2026. RomanLens: The role of latent romanization in multilinguality in LLMs. *arXiv preprint arXiv:2502.07424*.

SB Intuitions Corp. 2024. Sb intuitions, building unique Japanese LLM: Released Japanese LLM with 7 billion, 13 billion, and 65 billion parameters to contribute to R&D in academia and industry.

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*.

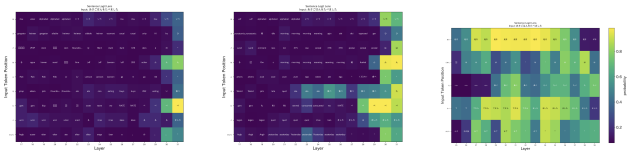
An Yang and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Alex Young and 1 others. 2024. Yi: Open foundation models by 01.AI. *arXiv preprint arXiv:2403.04652*.

Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. 2024. Beyond English-centric LLMs: What language do multilingual language models think in? *arXiv preprint arXiv:2408.10811*.

A Top-1 Token Heatmaps for Hiragana Sentence Analysis

Figure 5 shows the Top-1 token heatmaps for the hiragana sentence analysis of the three primary models. In LLM-jp-3-13B-Instruct, kanji tokens such as “起き” (wake up) and “食べ” (eat) appear in intermediate layers, confirming the latent kanji-fication phenomenon.



(a) Llama-3.1-8B (b) Swallow-8B (c) LLM-jp-13B

Figure 5: Top-1 token heatmaps for the hiragana sentence analysis: (a) Llama-3.1-8B-Instruct, (b) Llama-3.1-Swallow-8B-Instruct-v0.5, (c) LLM-jp-3-13B-Instruct.

B Script Conversion Results for Additional Models

Figures 6–7 show the script-type probability distributions for Sarashina2-7B, Qwen3-8B-Base, Qwen3-Swallow-8B-SFT-v0.2, and Yi-6B-Chat on kanji-writable words and loanwords. The main point of comparison in these figures is not only whether hiragana becomes dominant in the final layers, but also which script occupies the intermediate layers before that final convergence. In particular, we compare whether kanji remains competitive for kanji-writable words, and whether the same models shift toward katakana when the target item is a loanword.

Qwen3-8B-Base shows a clear intermediate-layer kanji region for kanji-writable words, followed by a late transition to hiragana. This behavior is qualitatively similar to that of LLM-jp-3-13B-Instruct in the main text and is consistent with its relatively high kanji-fication degree in Table 1. Qwen3-Swallow-8B-SFT-v0.2 preserves the same broad pattern, but the intermediate kanji region is weaker and shorter, while the final transition to hiragana is sharper. On loanwords, both Qwen models move probability mass away from kanji and toward katakana, supporting the interpretation that latent kanji-fication depends on whether a natural kanji-written lexical alternative exists for the target word.

Sarashina2-7B is another strong positive case: for kanji-writable words, kanji remains substantial through much of the computation and yields to hiragana only near the end. This makes Sarashina2-7B the closest additional model to the Japanese-centric behavior observed in the main text. Yi-6B-Chat shows a different mixed pattern in which alphabetic tokens remain dominant for a large portion of the network while kanji retains non-negligible mass. This is compatible with the values in Table 1: the Chinese-centric model exhibits some sensitivity to kanji characters, but does not display the same strength or stability of latent kanji-fication as the Japanese-centric models.

C Hiragana Sentence Analysis for Additional Models

Figures 8–9 show the hiragana sentence analysis results for Sarashina2-7B, Qwen3-8B-Base, Qwen3-Swallow-8B-SFT-v0.2, and Yi-6B-Chat. These figures should be read jointly: the line plots average script-type probabilities over token positions, whereas the heatmaps reveal whether the same tendency is shared across the sentence or driven by only a small subset of positions. Agreement between the two views provides stronger evidence that the intermediate representation is genuinely script-biased rather than an artifact of averaging.

Sarashina2-7B and Qwen3-8B-Base both reproduce the qualitative picture reported in Section 4.2. Sarashina2-7B keeps substantial kanji probability over many layers and then transitions to hiragana near the output layers; the corresponding heatmap indicates that this transition is not restricted to a single token position. Qwen3-8B-Base shows a noisier trajectory, but again the sentence-level plot and token-level heatmap agree that intermediate representations are more kanji-heavy than in the English-centric baselines, especially before the final shift to hiragana.

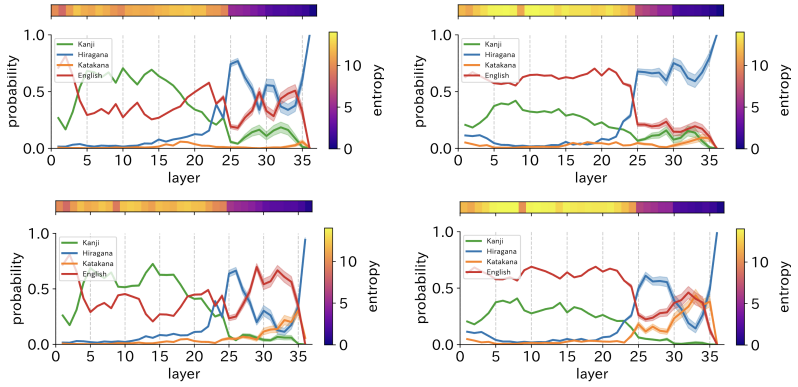
Qwen3-Swallow-8B-SFT-v0.2 and Yi-6B-Chat occupy a weaker intermediate regime. For Qwen3-Swallow-8B-SFT-v0.2, the kanji signal is still visible but shorter-lived than in Qwen3-8B-Base, suggesting that Japanese adaptation of a multilingual base model does not automatically produce the same degree of latent kanji-fication as models whose pre-training is more strongly Japanese-centric. Yi-6B-Chat again falls between the English-centric and Japanese-centric cases: kanji remains visible for some positions, but the pattern is less coherent and the final transition to hiragana is later and less stable than in LLM-jp-3-13B-Instruct or Sarashina2-7B.

D Results for Large-Scale Models

Figures 10–13 show the script conversion and hiragana sentence analysis results for large-scale models (32B–70B parameter range). These figures are intended to separate two questions: whether larger scale preserves a phenomenon that is already present in smaller models, and whether scale can create latent kanji-fication in model families where it is weak or absent. Because the compared models have different depths, the most relevant comparison is not the absolute layer index, but whether a distinct intermediate regime emerges before the final hiragana-dominant layers.

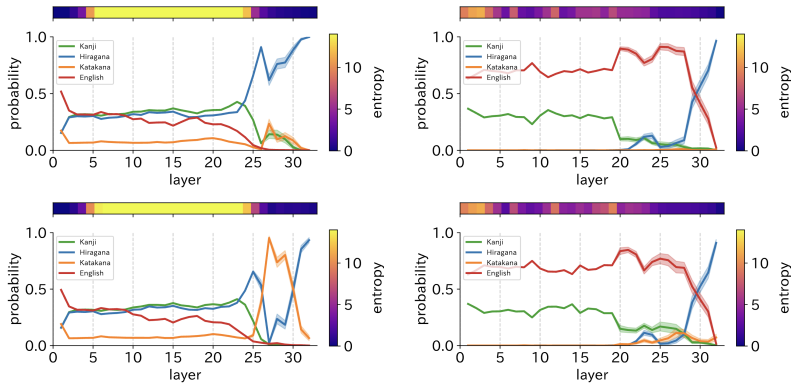
For the Japanese-centric and kanji-rich multilingual models, scale mainly strengthens or stabilizes the same qualitative behavior observed at smaller sizes. Sarashina2-70B keeps kanji and hiragana both active until very late layers, and Qwen3-32B-Base shows an even more pronounced kanji-dominant band in the script-conversion setting than its 8B counterpart. The sentence-level analyses in Figures 12 and 13 support the same interpretation: when the smaller model already exhibits latent kanji-fication, larger scale tends to sharpen or prolong that preference rather than remove it.

By contrast, the English-centric Llama-3.1-70B-Instruct and the continual pre-training model Llama-3.1-Swallow-70B-Instruct-v0.3 remain overwhelmingly alphabetic until the final few layers, just as in their 8B counterparts. Increasing scale therefore does not by itself induce latent kanji-fication. Qwen3-Swallow-32B-SFT-v0.2 and



(a) Qwen3-8B-Base (b) Qwen3-Swallow-8B-SFT-v0.2

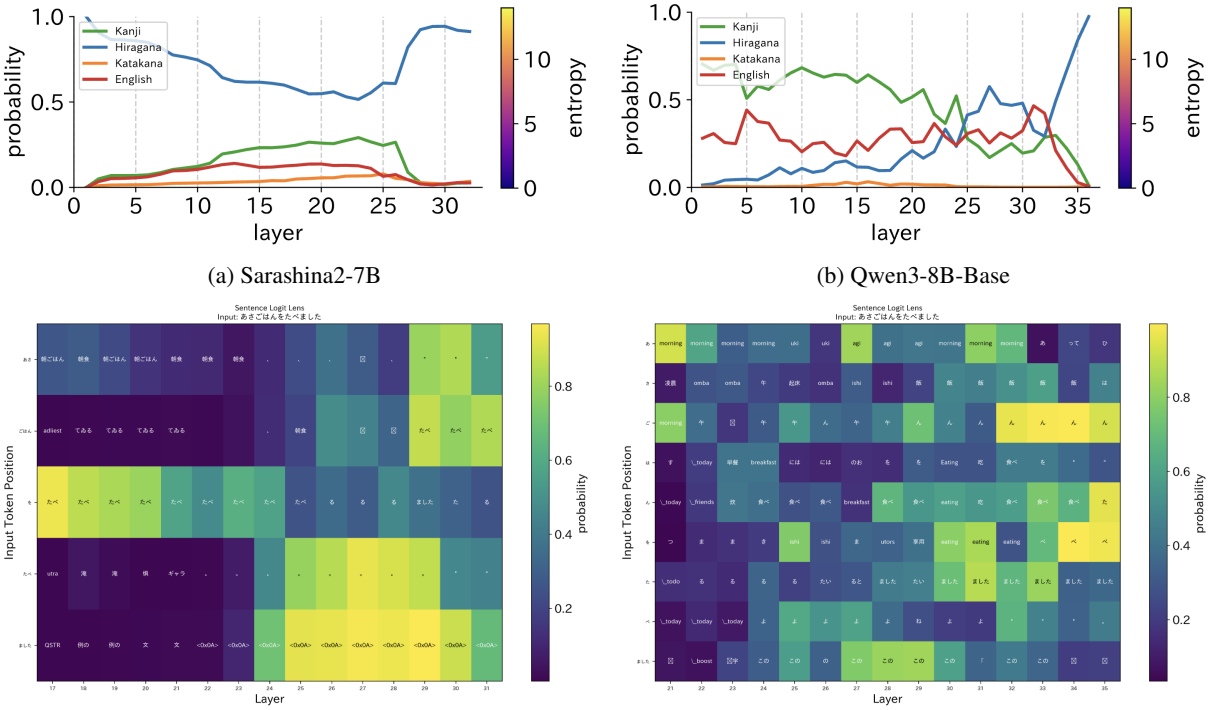
Figure 6: Script-type probability distributions for Qwen3 models. Top row: kanji-writable words. Bottom row: loanwords.



(c) Sarashina2-7B (d) Yi-6B-Chat

Figure 7: Script-type probability distributions for Sarashina2-7B and Yi-6B-Chat. Top row: kanji-writable words. Bottom row: loanwords.

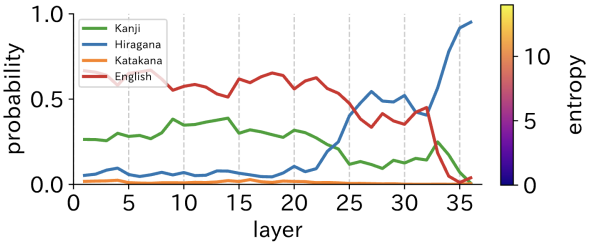
Yi-34B-Chat occupy an intermediate regime: both show non-trivial kanji mass in the middle layers, but their trajectories are less clean and less consistently kanji-dominant than those of Sarashina2-70B or Qwen3-32B-Base. Overall, the large-scale appendix supports the claim in Section 4.3 that scale amplifies pre-existing language- and vocabulary-dependent structure, rather than replacing it with a universal Japanese processing strategy.



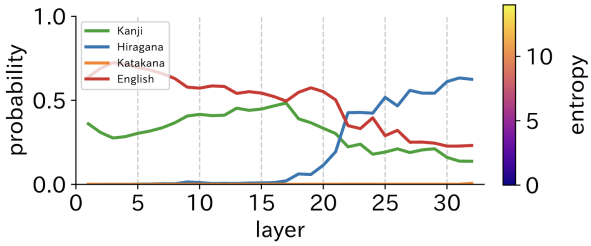
(a) Sarashina2-7B

(b) Qwen3-8B-Base

Figure 8: Hiragana sentence analysis: (a) Sarashina2-7B, (b) Qwen3-8B-Base. Top row: script-type probability distributions. Bottom row: Top-1 token heatmaps.



(c) Qwen3-Swallow-8B-SFT-v0.2



(d) Yi-6B-Chat

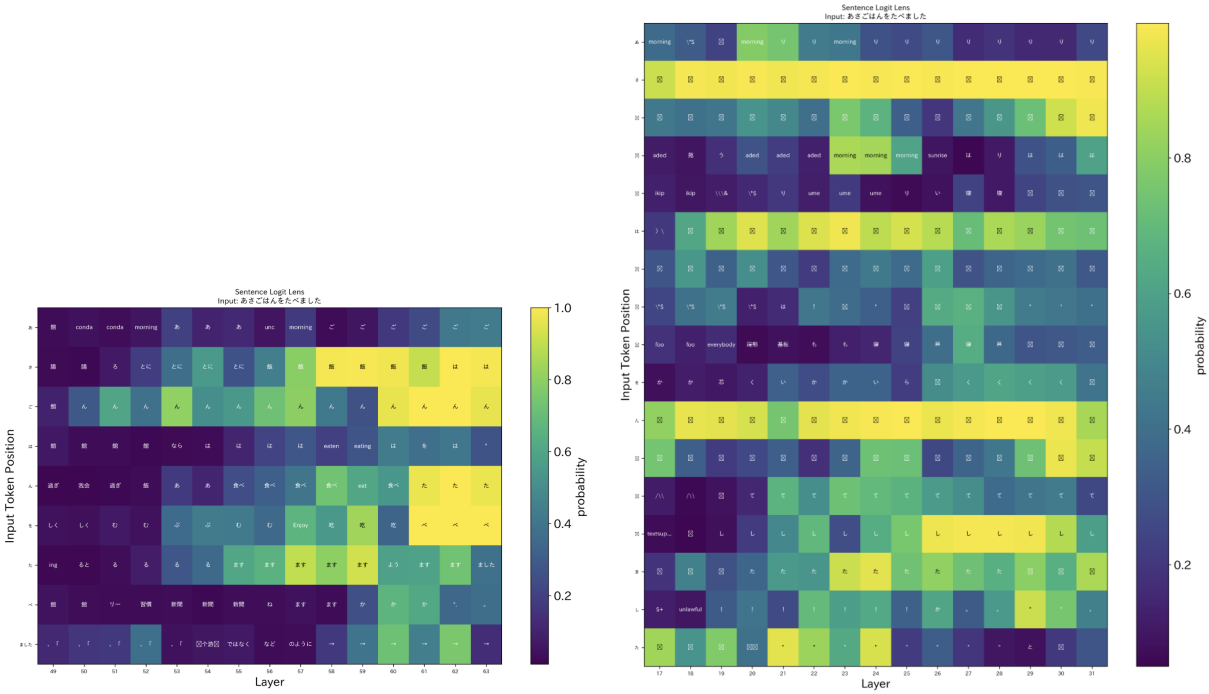
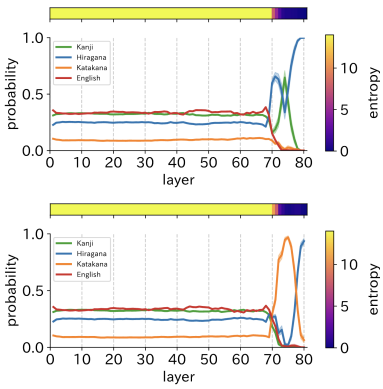
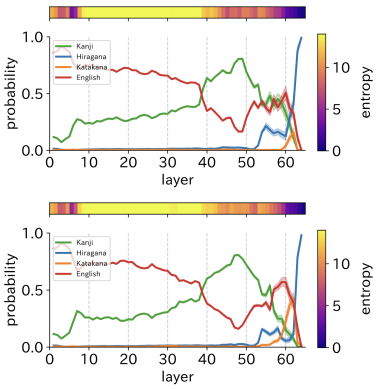


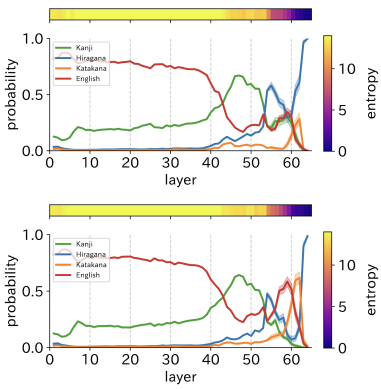
Figure 9: Hiragana sentence analysis: (c) Qwen3-Swallow-8B-SFT-v0.2, (d) Yi-6B-Chat. Top row: script-type probability distributions. Bottom row: Top-1 token heatmaps.



(a) Sarashina2-70B

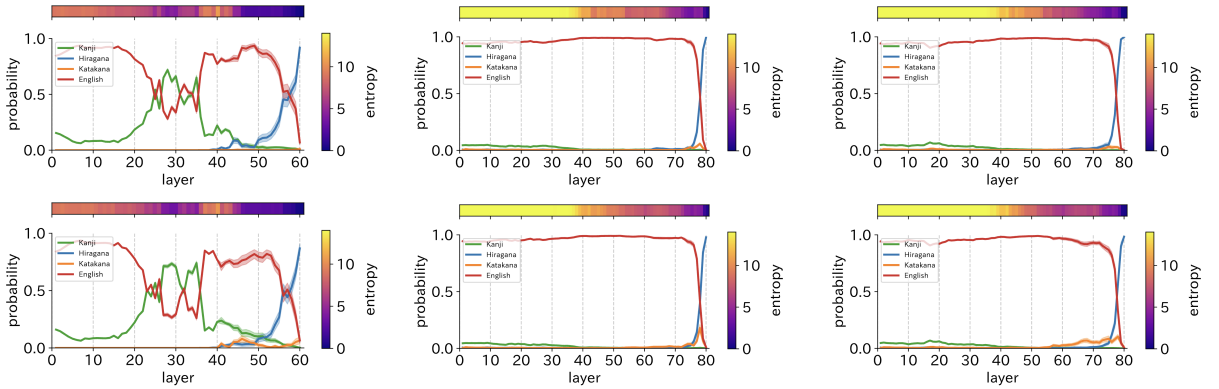


(b) Qwen3-32B-Base



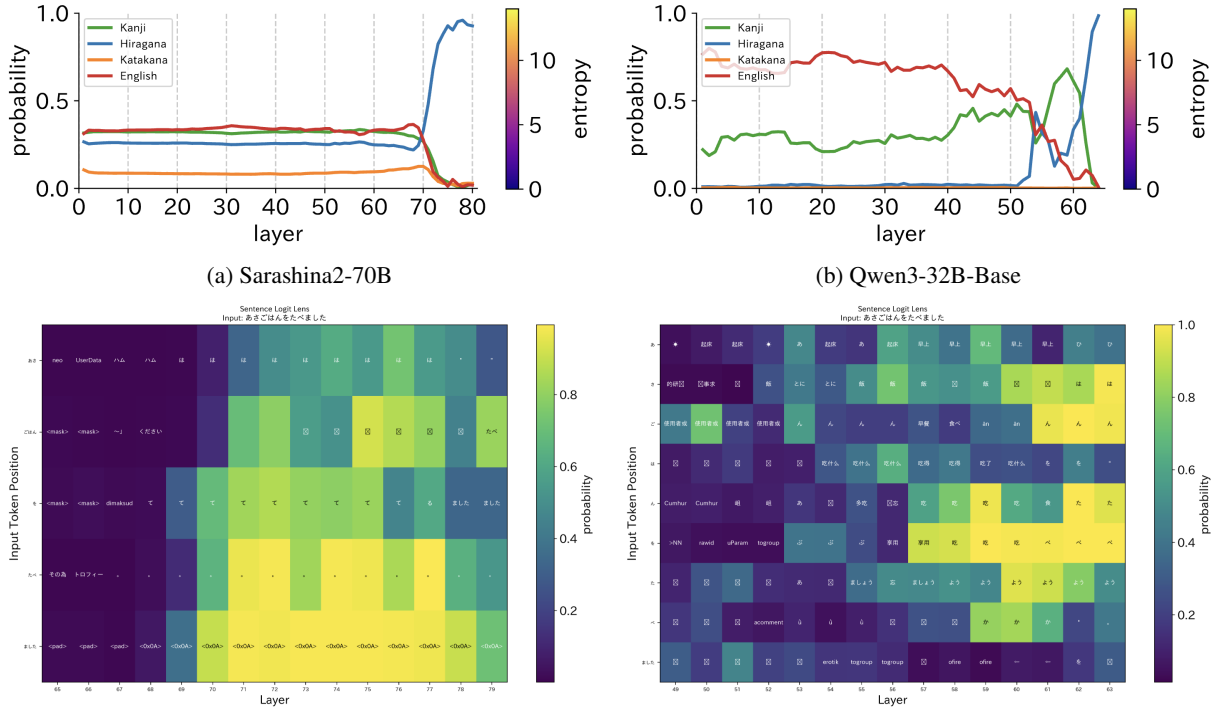
(c) Qwen3-Swallow-32B-SFT-v0.2

Figure 10: Script-type probability distributions for large-scale models (1/2). Top row: kanji-writable words. Bottom row: loanwords.



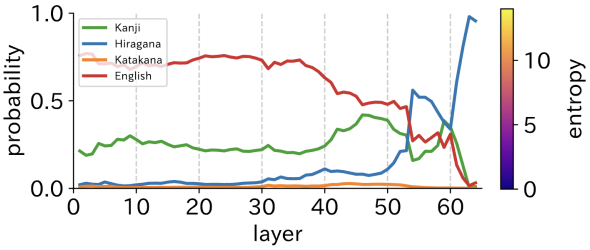
(d) Yi-34B-Chat (e) Llama-3.1-70B-Instruct (f) Llama-3.1-Swallow-70B-Instruct-v0.3

Figure 11: Script-type probability distributions for large-scale models (2/2). Top row: kanji-writable words. Bottom row: loanwords.

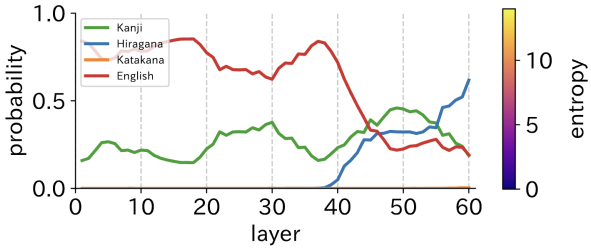


(a) Sarashina2-70B (b) Qwen3-32B-Base

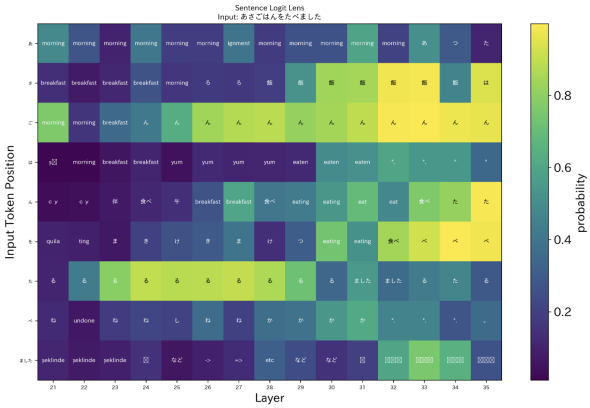
Figure 12: Hiragana sentence analysis for large-scale models (1/2): (a) Sarashina2-70B, (b) Qwen3-32B-Base. Top row: script-type probability distributions. Bottom row: Top-1 token heatmaps.



(c) Qwen3-Swallow-32B-SFT-v0.2



(d) Yi-34B-Chat



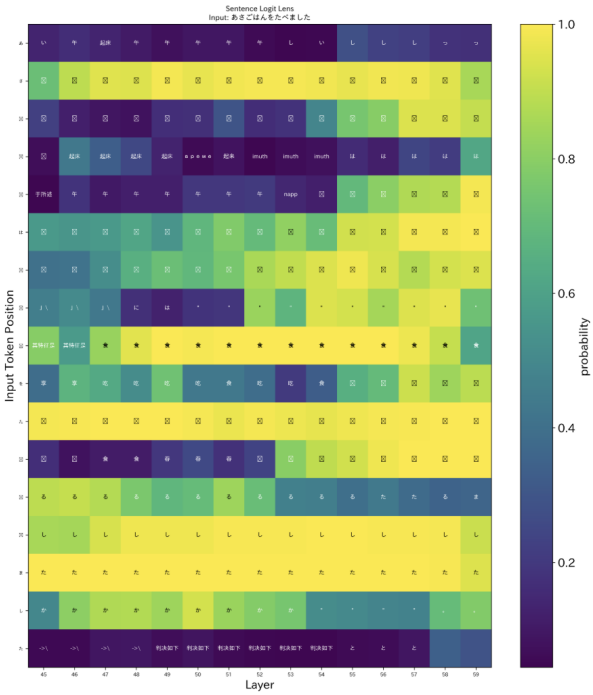


Figure 13: Hiragana sentence analysis for large-scale models (2/2): (c) Qwen3-Swallow-32B-SFT-v0.2, (d) Yi-34B-Chat. Top row: script-type probability distributions. Bottom row: Top-1 token heatmaps.