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

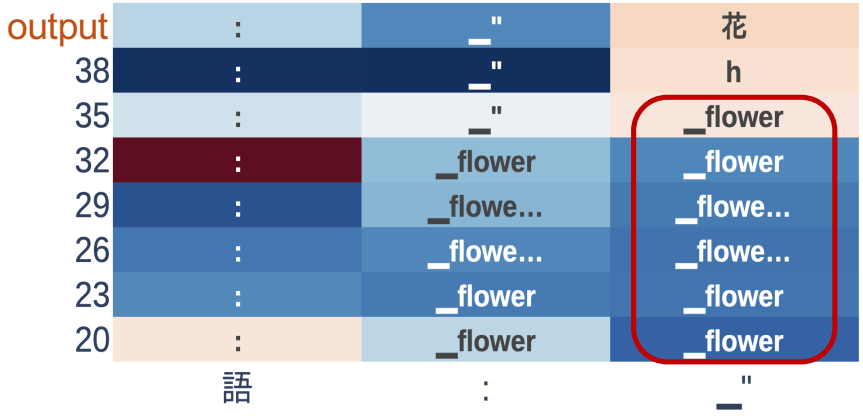
Chengzhi Zhong[1] Fei Cheng[1] Qianying Liu[2] Junfeng Jiang[3]
Zhen Wan[1] Chenhui Chu[1] Yugo Murawaki[1] Sadao Kurohashi[1,2]

[1]Kyoto University
[2]National Institute of Informatics
[3]The University of Tokyo

# Background

Mainstream LLMs (e.g., Llama-2) are trained mainly on English.

Even when processing other languages, the intermediate layers language (a.k.a latent language) is English (Wendler et al., 2024).




Model: Llama2 - 13b
Prompt: Français: "fleur" – 日本語: "
This picture shows the argmax token at each layer.

**What's the latent language of Japanese-specialized LLMs?**

# Background

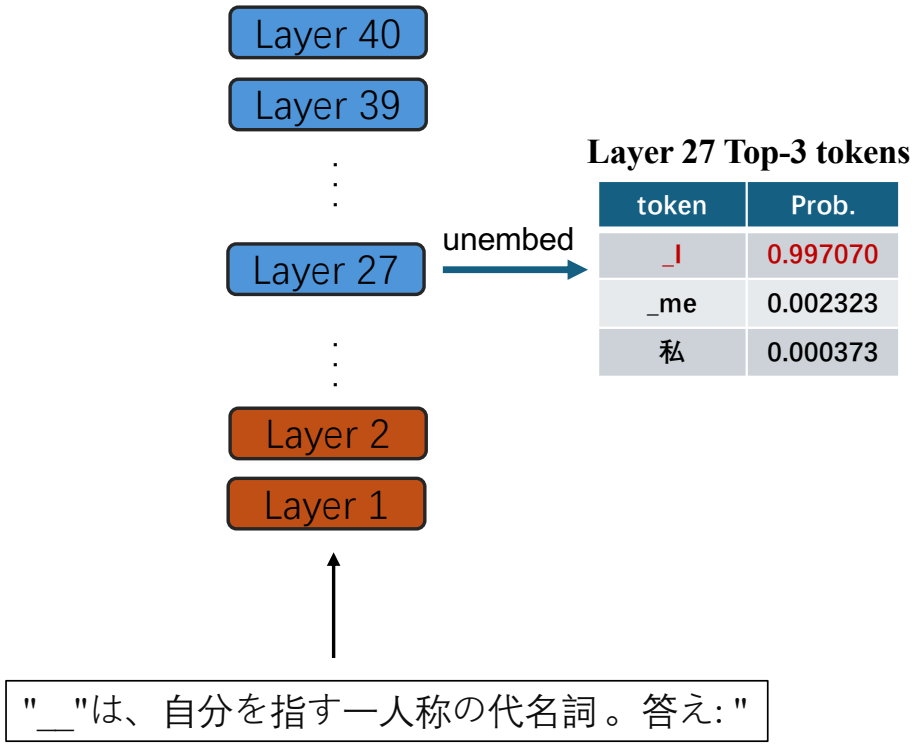
There are two types of Japanese-specialized LLM:

1. LLMs based on English-centric model with continued pretraining in Japanese (e.g., Swallow).
2. LLMs trained from scratch on balanced English-Japanese data (e.g., LLM-jp).

**Do these two types of models work similarly?**

# Logit Lens

By projecting the model‘s internal representations onto the vocabulary, logit lens allows us to observe how the model predicts the next token (nostalgebraist 2020).



# Experimental Settings

- **Models:**

| Model Category | Model | Proportion in pre-training data | | | Token | From scratch |
|---|---|---|---|---|---|---|
| | | En | Ja | Other | | |
| English-centric | Llama 2 | 89.7% | 0.1% | 10.2% | 2,000B | Yes |
| Japanese CPT | Swallow | 10.0% | 90.0% | 0.0% | 100B | Llama-2 based |
| Balanced En and Ja | LLM-jp-V2.0 | 45.0% | 51.6% | 3.4% | 256B | Yes |

- **Dataset:** A manually constructed word-level parallel dataset (En, Fr, Ja, Zh), 166 examples.
- **Task:**
  - Cloze (2-shot)
  - Translation (4-shot)

"__"は、新しいものを作ることです。答え: "開発"

…

"__"は、お金を預けたり借りたりする場所です。答え: "

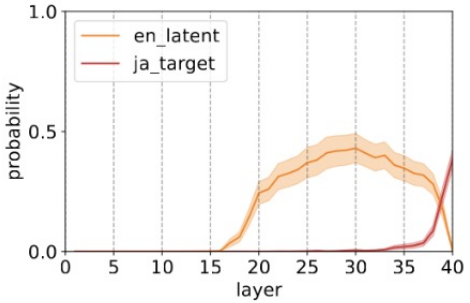
Français: "créativ" - 日本語: "開発"

…

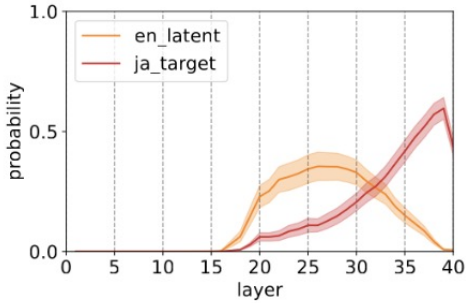
Français: "banque" - 日本語: "

# Results: Monolingual Cloze Task

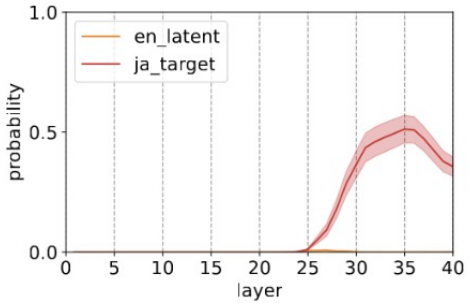
Average result of **Ja** cloze task on 166 examples




English-centric: Llama2




Japanese CPT: Swallow



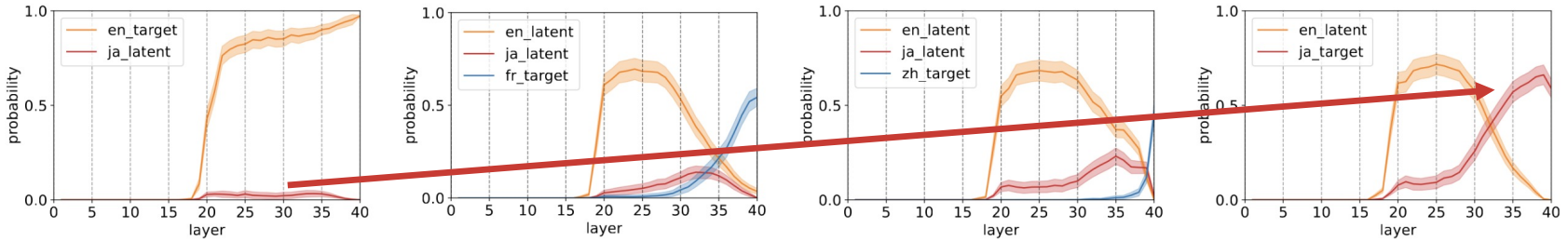

Balanced En and Ja: LLM-jp-v2.0

While Japanese-specialized models utilize Japanese as the latent language：

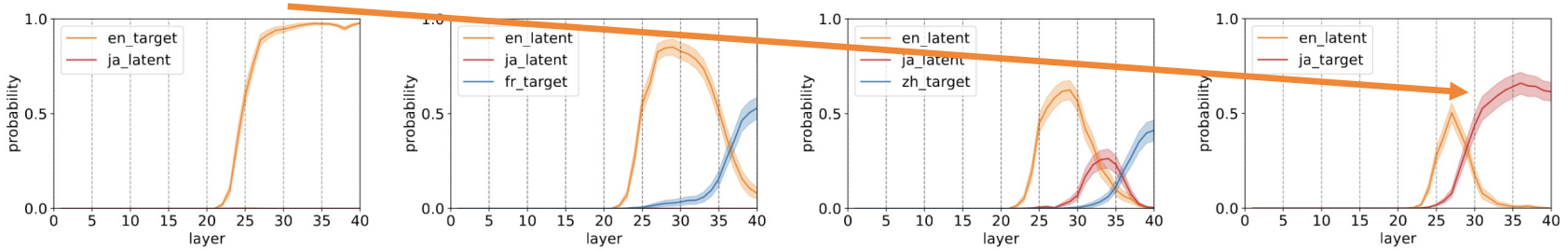
- Swallow uses a mix of English and Japanese.
- LLM-jp fully uses Japanese.

# Results: Translation From Fixed Source Language 1/3

Japanese CPT: Swallow




Balanced En and Ja: LLM-jp-v2.0



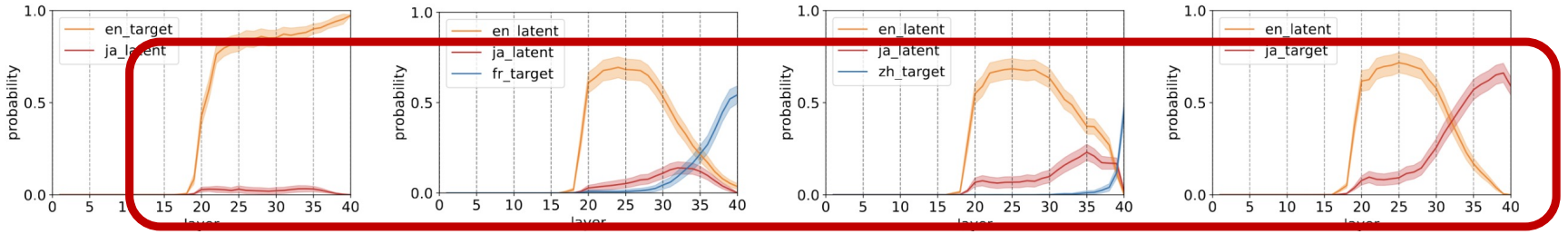

En -> En | En -> Fr | En -> Zh | En -> Ja

closer to En **target language** closer to Ja

From left to right, as the target language get closer to Japanese,
the probability of English decreases while the probability of Japanese increases.

# Results: Translation From Fixed Source Language 2/3

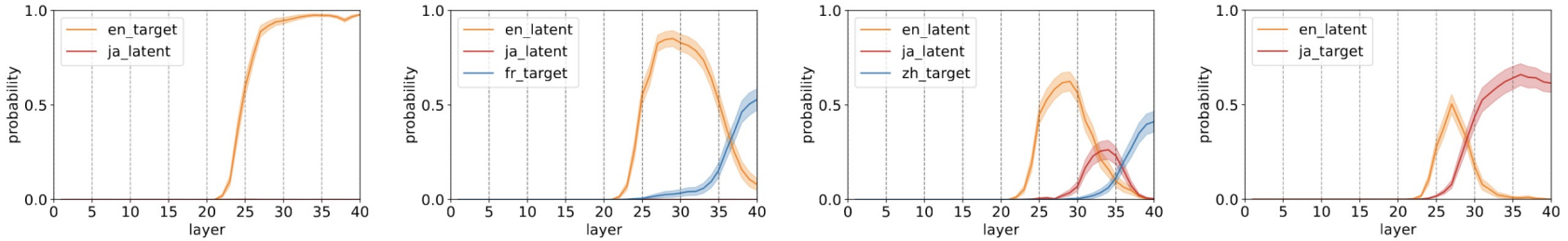
Japanese CPT: Swallow




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

Always a mixture of English and Japanese.

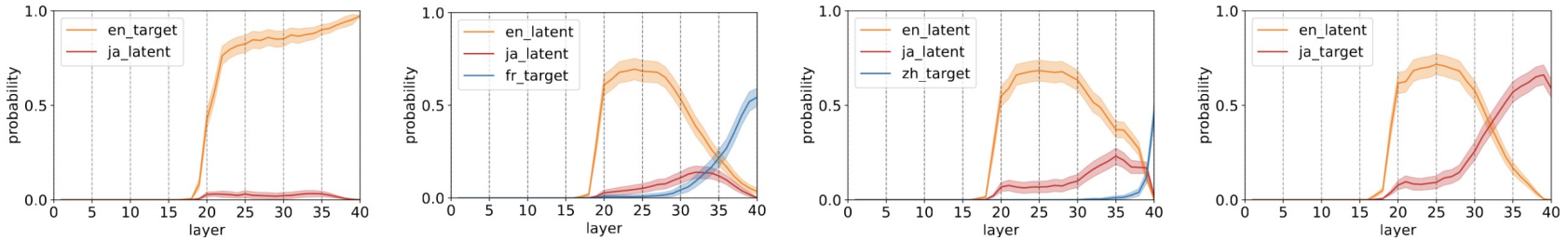
Balanced En and Ja: LLM-jp-v2.0




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

# Results: Translation From Fixed Source Language 3/3

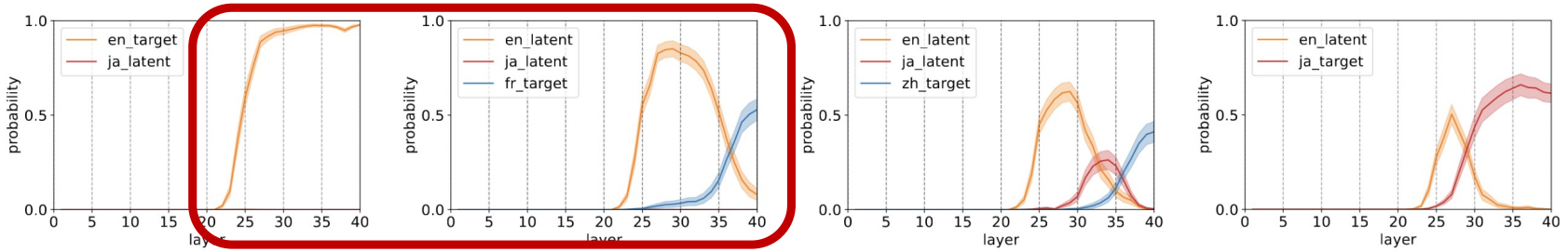
Japanese CPT: Swallow




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

LLM-jp can independently use one specific latent language.

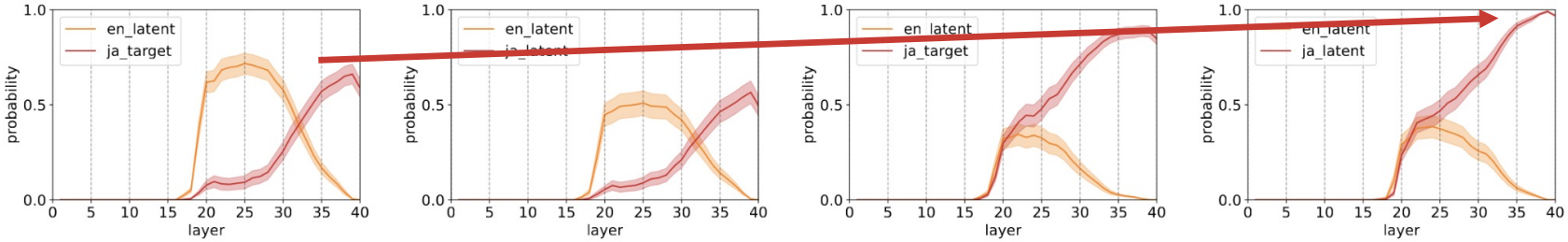
Balanced En and Ja: LLM-jp-v2.0



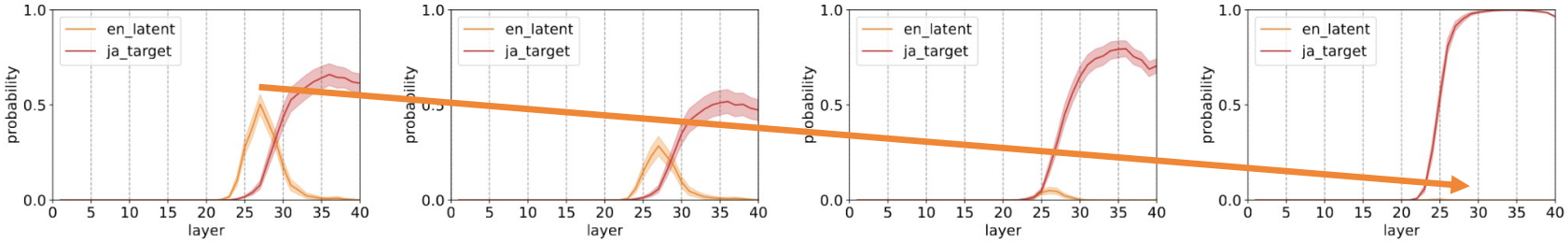

En -> En | En -> Fr | En -> Zh | En -> Ja

# Results: Translation To Fixed Target Language 1/3

Japanese CPT: Swallow




Balanced En and Ja: LLM-jp-v2.0



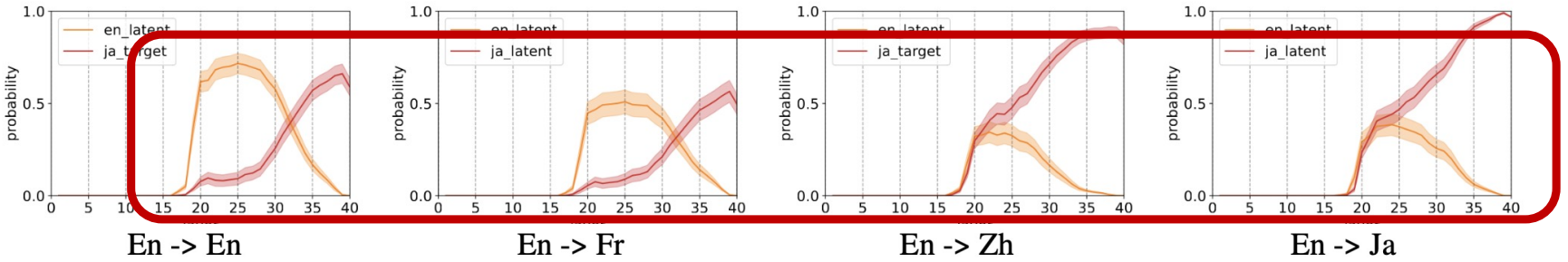

En -> Ja  Fr -> Ja  Zh -> Ja  Ja -> Ja

closer to En  **source language**  closer to Ja

From left to right, as the source language get closer to Japanese,
the probability of English decreases while the probability of Japanese increases.

# Results: Translation To Fixed Target Language 2/3

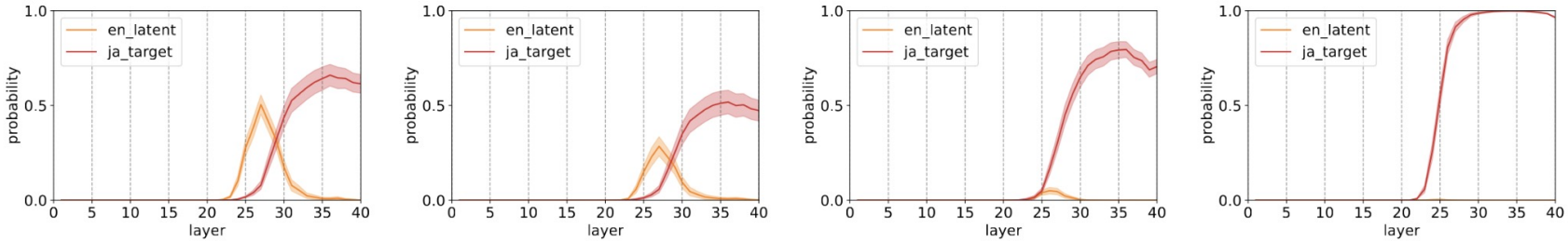
(a) Translation: Swallow-13b




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

Always a mixture of English and Japanese.

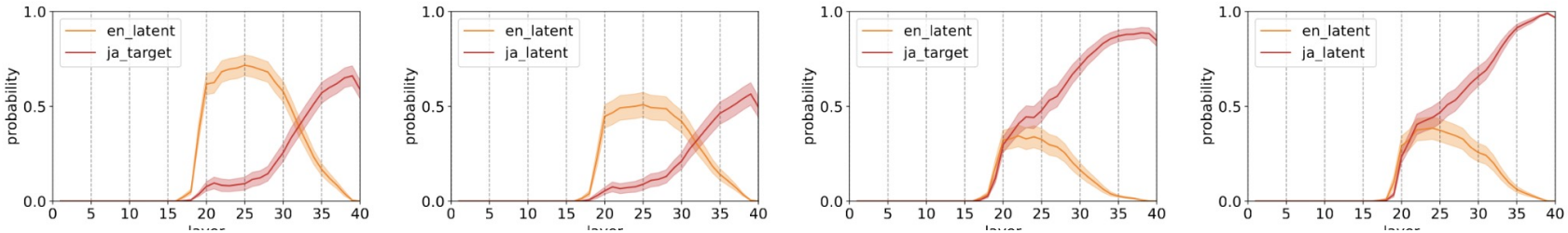
Balanced En and Ja: LLM-jp-v2.0




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

# Results: Translation To Fixed Target Language 3/3

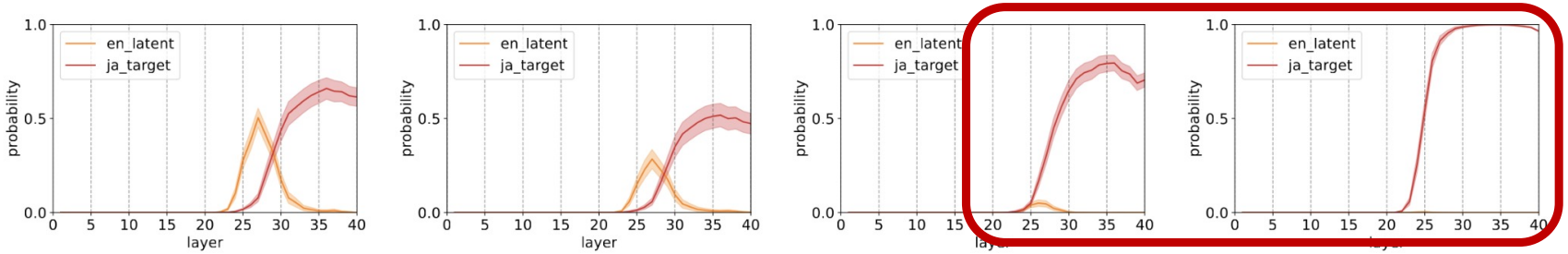
(a) Translation: Swallow-13b




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

LLM-jp can independently use one specific latent language.

Balanced En and Ja: LLM-jp-v2.0

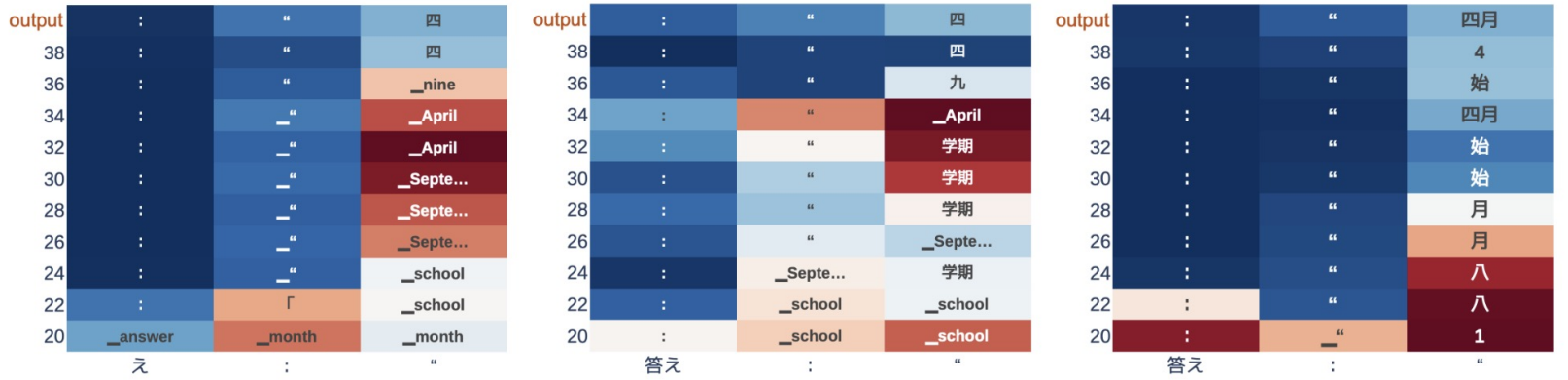



En -> Ja | Fr -> Ja | Zh -> Ja | Ja -> Ja

# How are Culture Conflict Questions Solved 1/2

Considering that each language represents a culture, does the latent language influence how the model answers cultural conflicts questions?

Prompt: 一年の始まりの月は：__月、答え：“—”
日本の学校の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: ")




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

Llama’s thinking flow: September → 四

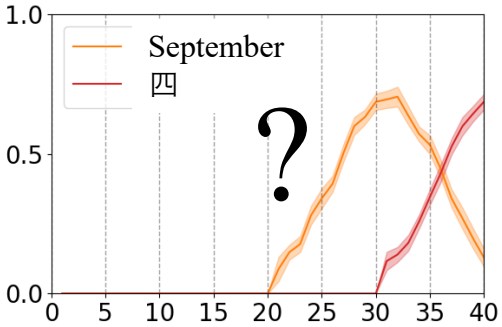
Answer for the U.S. Answer for JP

# How are Culture Conflict Questions Solved 2/2

Prompt: 一年の始まりの月は：__月、答え：“一”
日本の学校の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: ")

September → 四

Answer for the U.S　　Answer for JP




Will the model think in such procedure ?

# Experimental Settings

- **Models:** Same as before.
- **Dataset:** A manually constructed culture conflict QA dataset (En, Ja) with 49 examples.
- **Procedure:**

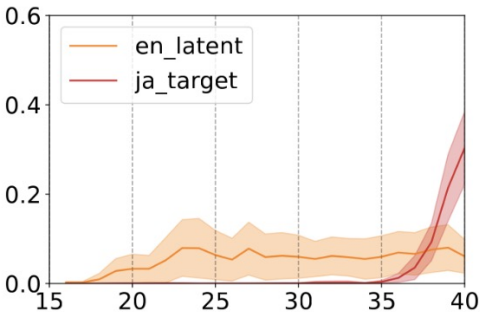
1. Ask the model about the U.S./Japan in their respective languages.
2. Record the responses as answers in the U.S. and Japanese contexts.
3. Ask again using “本国” in Japanese, and monitor the probability of answers in the U.S. and Japanese contexts.

The highest mountain in the United States is _.
日本の最高峰は__です。
…

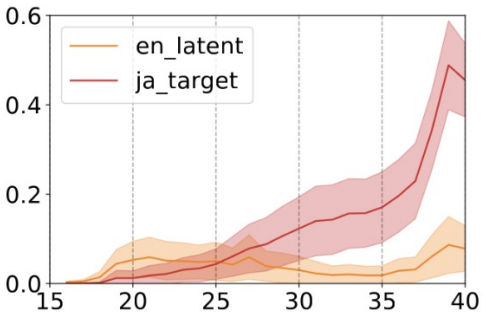
本国の最高峰は__です。
…

# Results: How are Culture Conflict Questions Solved

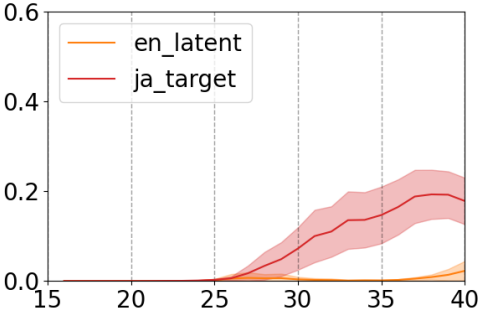
Average results over 49 examples




English-centric: Llama2




Japanese CPT: Swallow




Balanced En and Ja: LLM-jp-v2.0

Latent language affects the model's reasoning:

- **Llama** always arrives at the answer for the U.S. first.
- **Swallow** produces a mixture of answers for the U.S. and Japan.
- **LLM-jp** unaffected by U.S.-context answers.

# Conclusion

- Japanese-specialized models can use Japanese as their latent language
- The probability of a latent language is influenced by the source and target language; higher similarity increases the probability
- The latent language affects how the model answers culture conflict questions

Future works:

- We plan to evaluate whether Llama's intermediate layer English answers are more accurate than target-language answers in non-English benchmarks.

# Thanks