ACL2024読み会@名大（2024.9.30）

https://aclanthology.org/2024.acl-long.820.pdf

ACL 2024

Do Llamas Work in English? On the Latent Language of Multilingual Transformers

Chris Wendler*, Veniamin Veselovsky*, Giovanni Monea*, Robert West*
EPFL
{chris.wendler, veniamin.veselovsky, giovanni.monea, robert.west}@epfl.ch

Code: https://github.com/epfl-dlab/llm-latent-language

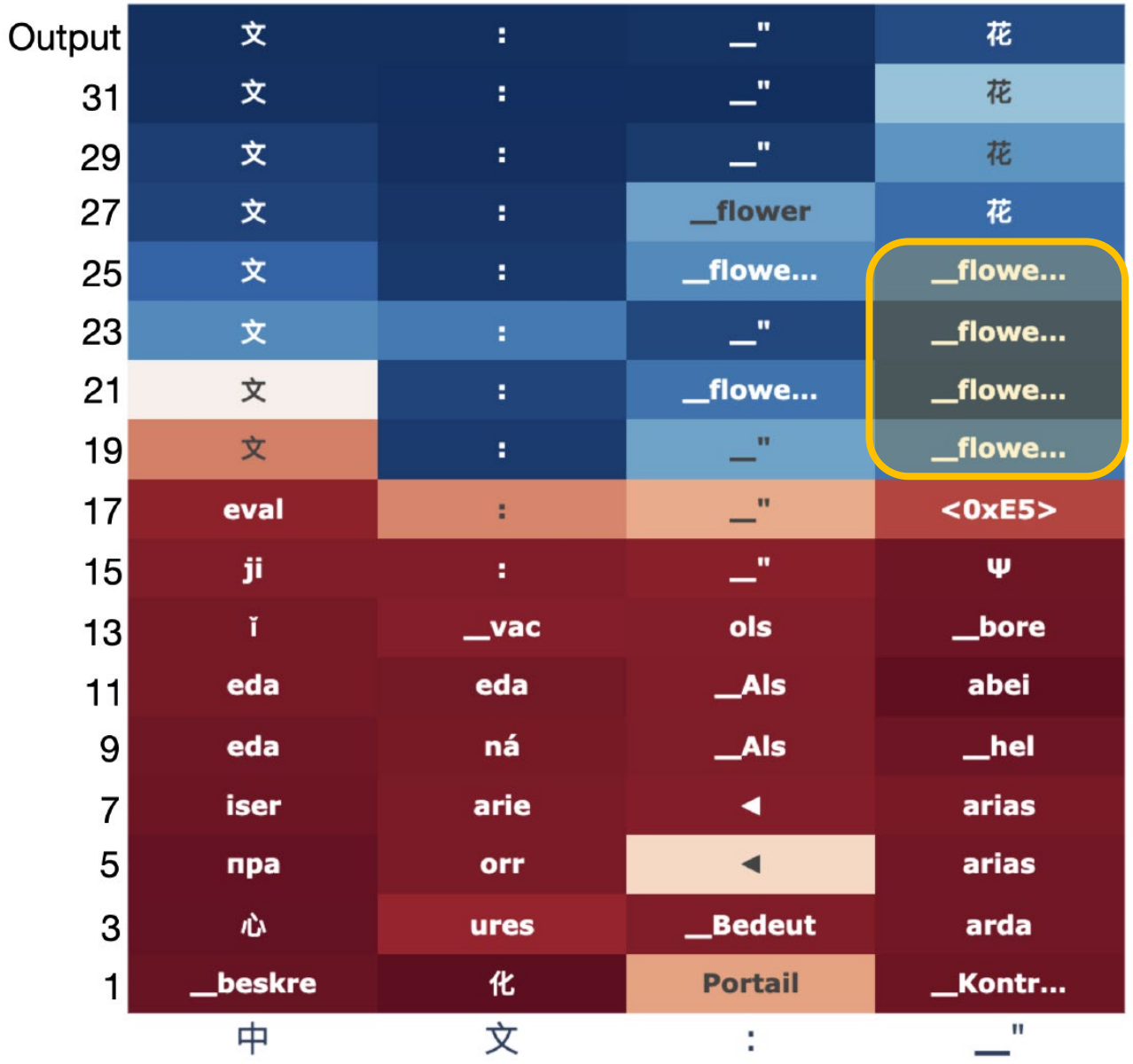
紹介者: 笹野遼平（名大）

論文の概要

- Llama等のLLMは学習データの大半は英語であるにも関わらず英語以外でも利用可能
- 英語以外を扱う場合も内部では英語がpivot言語として使われているという説を検証

結論

… the model’s internal lingua franca is not English but concepts— concepts that are biased toward English. Hence, English could still be seen as a pivot language, but in a semantic, rather than a purely lexical, sense.



“Latent language”分析のためのタスク

- 特定言語に明確に属する単語がnext tokenとなるよう設計
- ３種類のタスクで実験

a) Translation (翻訳) task (X->ZH)
  - 非英語言語から中国語への翻訳タスク
  - 4語の対訳ペアを与え5語目の約を予測

b) Repetition (繰り返し) task (ZH->ZH)
  - 入力をそのまま繰り返すタスク

c) Cloze (穴埋め) task (ZH)
  - マスク単語が答えとなる英文を生成
  - 対象言語に翻訳して使用
  - 2事例をdemonstrationとして使用

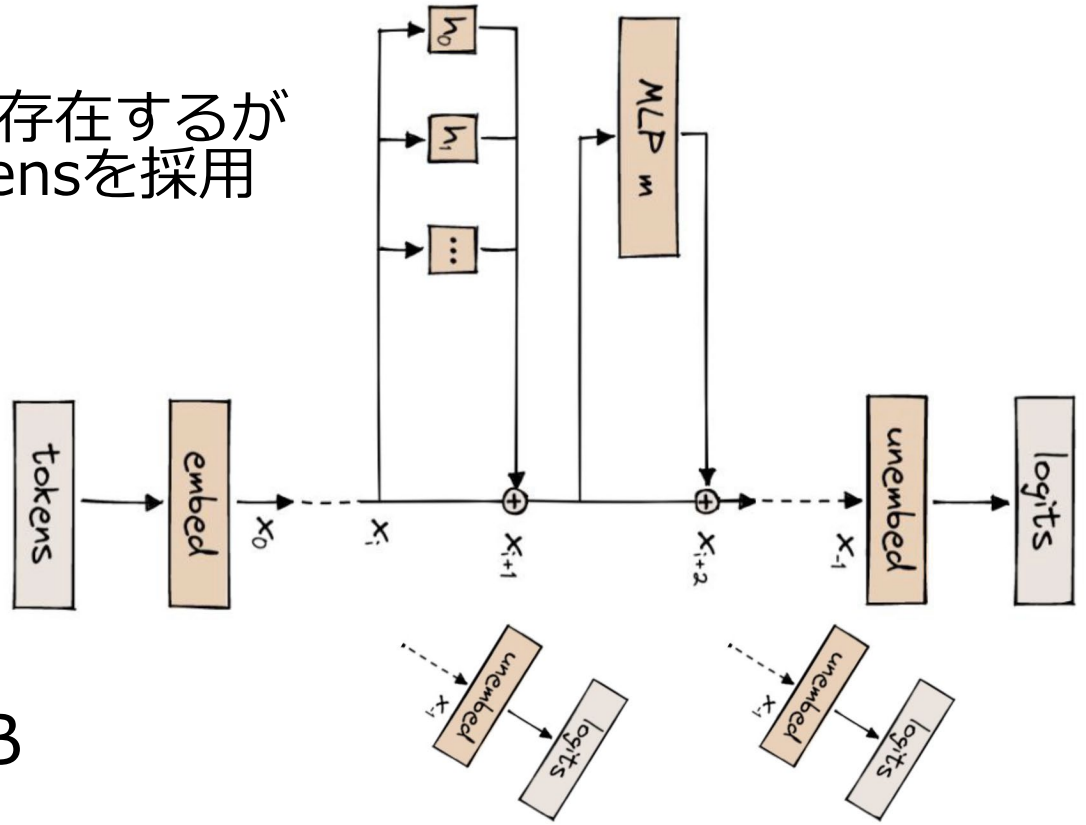
Français: "vertu" - 中文: "德"
Français: "siège" - 中文: "座"
Français: "neige" - 中文: "雪"
Français: "montagne" - 中文: "山"
Français: "fleur" - 中文: "

中文: "德" - 中文: "德"
中文: "座" - 中文: "座"
中文: "雪" - 中文: "雪"
中文: "山" - 中文: "山"
中文: "花" - 中文: "

A "___" is used to play sports like soccer and basketball. Answer: "ball".
A "___" is a solid mineral material forming part of the surface of the earth. Answer: "rock".
A "___" is often given as a gift and can be found in gardens. Answer: "

分析方法・対象

- LLMの利用言語の特定ツール:
  - tuned lens [Belrose+'23]などの改良手法も存在するが真の出力予測が目的ではないためlogit lensを採用
- Logit lens [Nostalgebraist'20]:
  最終層に適用される隠れ層からトークン分布への変換操作を中間層に適用しトークンを予測
  - トークンから利用言語を特定
  - logitはsoftmaxに通す前の出力



- 分析対象モデル: Llama-2 7/13/70B
- 使用言語: 中国語、独語、仏語、露語

実験結果

- 翻訳タスク、穴埋めタスクでは　中間層付近で英語の確率がまず上昇
- 繰り返しタスクでは中国語の確率も英語と同時に上昇
- いずれも最終層付近で中国語確率が上昇
- 次単語予測のエントロピーは英語確率が高くなった時点で急激に低下(その前はほぼ一様分布, cf. $|V| \fallingdotseq 2^{15}$)

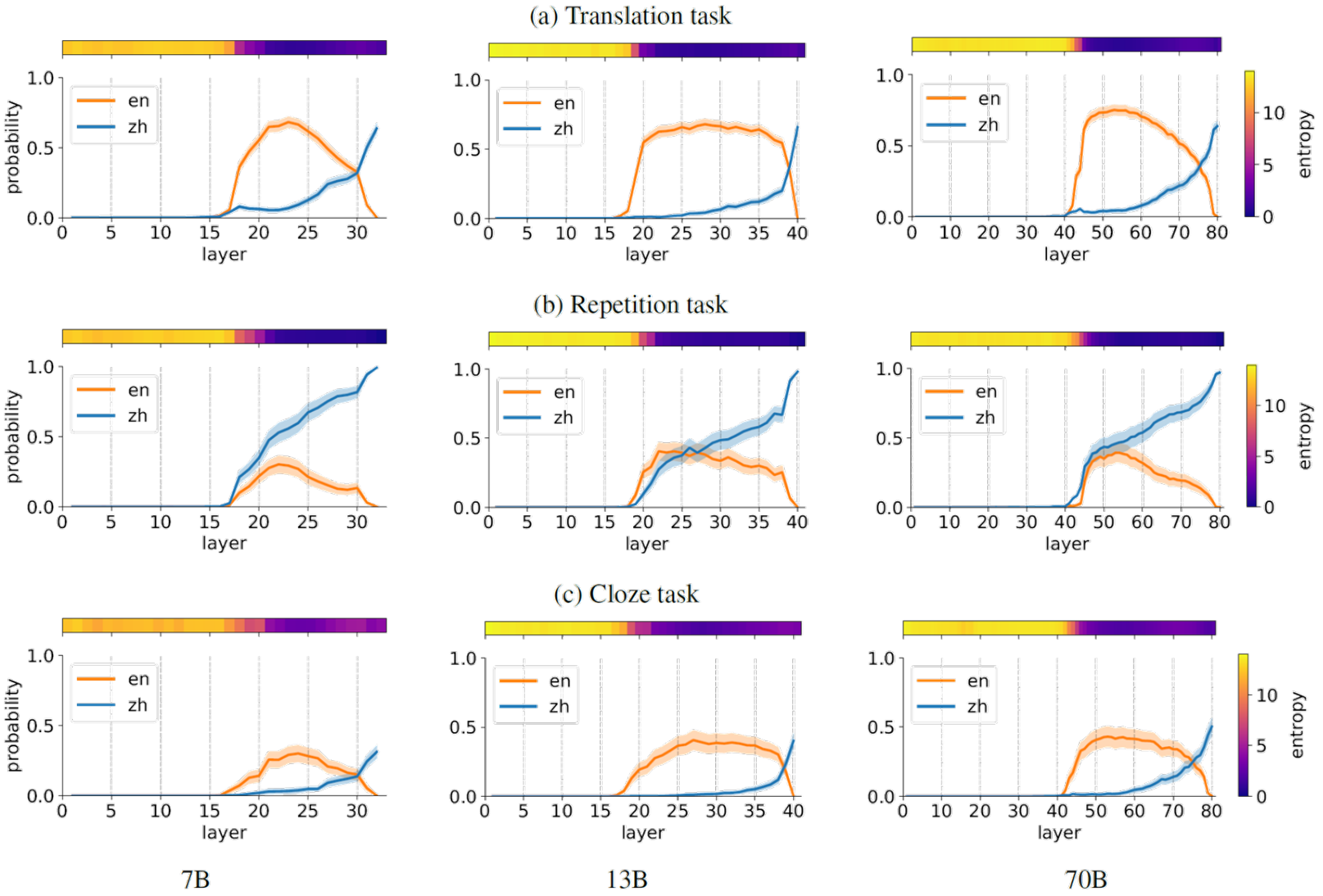


Figure 2: **Language probabilities for latents during Llama-2 forward pass,** for **(a)** translation task from union of German/French/Russian to Chinese, **(b)** Chinese repetition task, **(c)** Chinese cloze task. Each task evaluated for

幾何学的分析

- 隠れ状態hのうちtoken部分空間との直交成分はlogit lensで考慮されない
- どの程度hの持つenergy (≒情報量)がlogit scoreに反映されるか分析
- 隠れ状態hとトークンサブ空間のmean squared cosineを算出

1. **Phase 1** (layers 1–40): High entropy (14 bits, nearly uniform), low token energy, no language dominates.
2. **Phase 2** (layers 41–70): Low entropy (1–2 bits), low token energy, English dominates.
3. **Phase 3** (layers 71–80): Low entropy, high token energy (up from 20% to 30%), Chinese dominates.

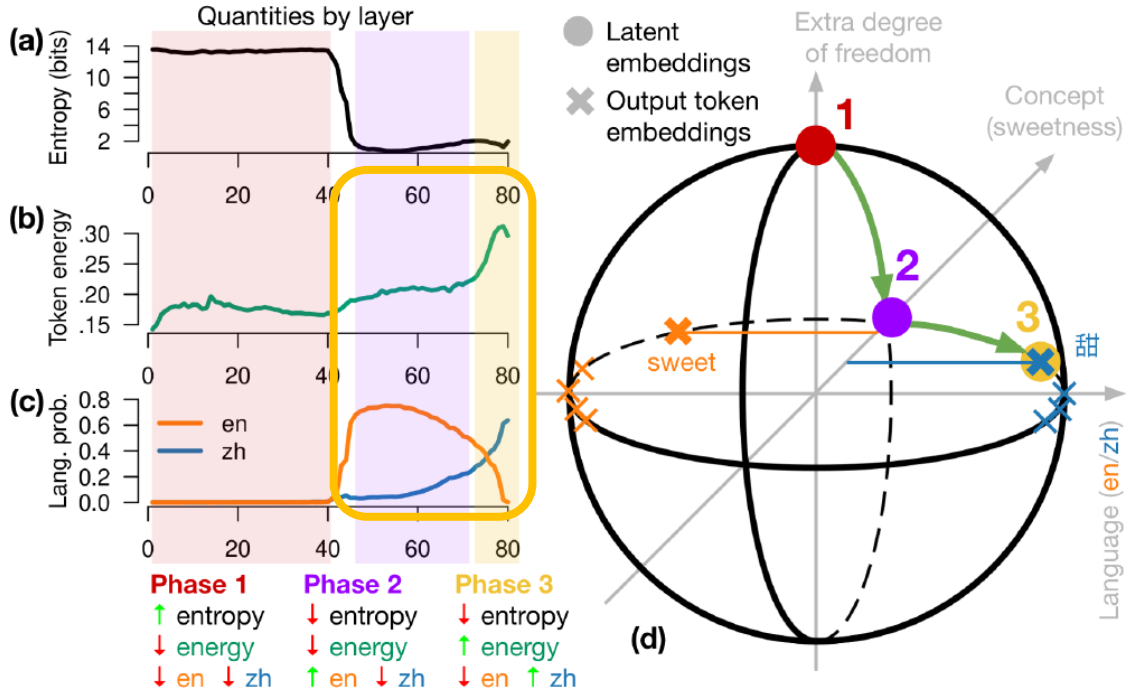
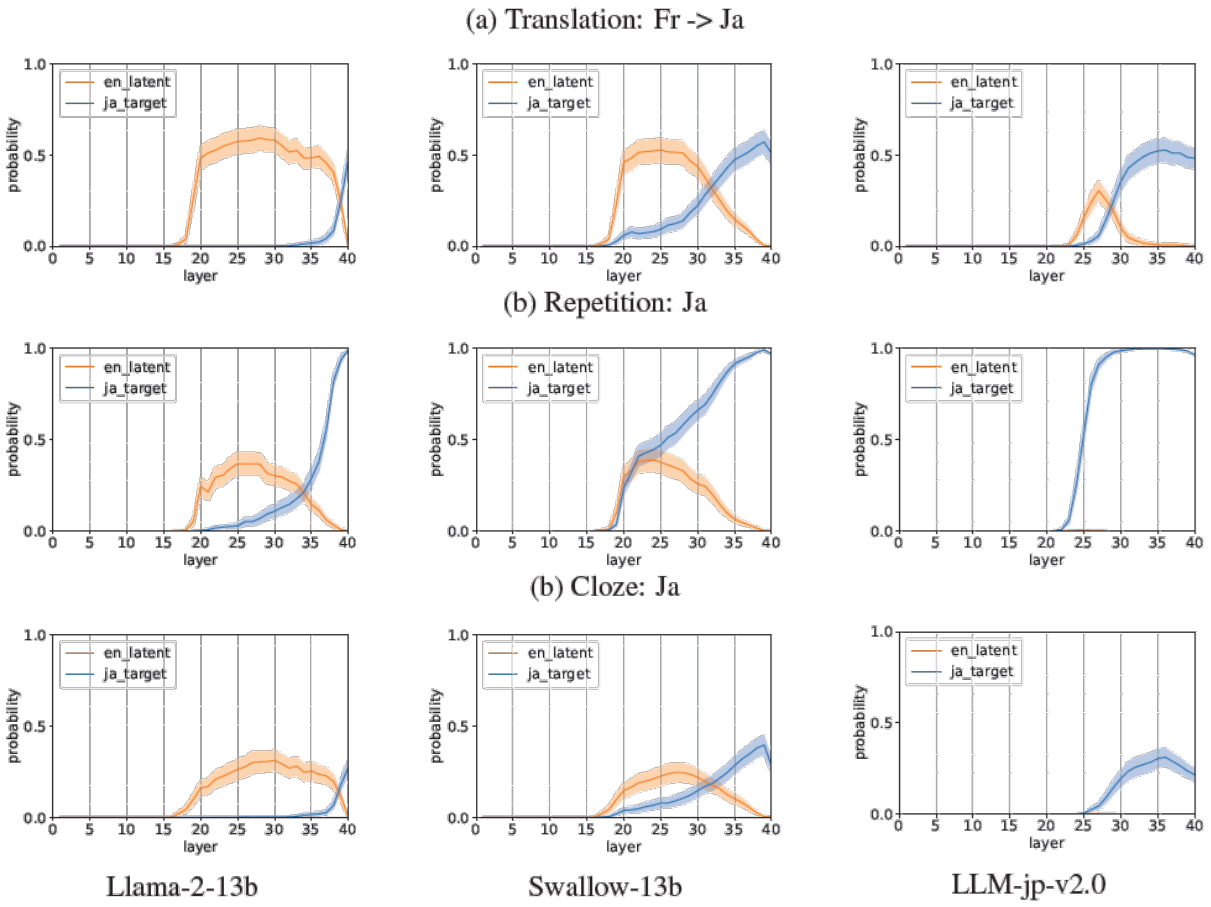


Figure 4: **Anatomy of transformer forward pass** when translating to Chinese (cf. Sec. 3.3). Layer-by-layer evolution of **(a)** entropy of next-token distribution, **(b)** token energy, **(c)** language probabilities. As latents are transformed layer by layer, they go through three phases (Sec. 4.2), **(d)** traveling on a hypersphere, here in 3D instead of actual 8192D (Sec. 5). "甜" means "sweet".

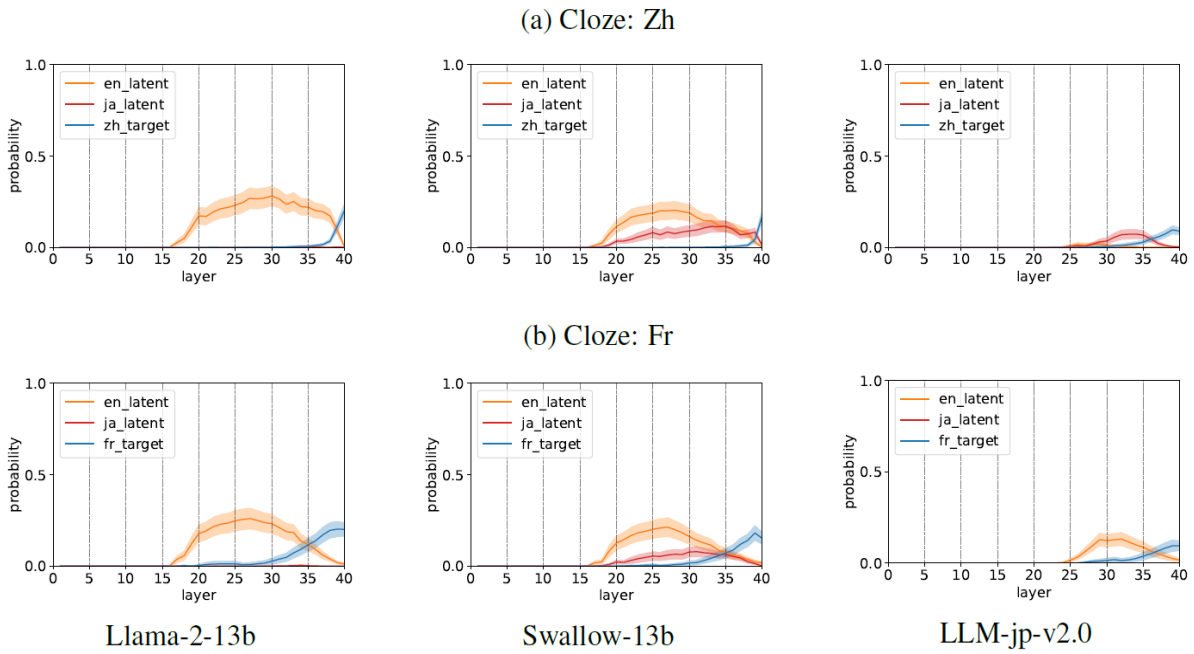
後続研究: Beyond English-Centric LLMs: What Language Do Multilingual Language Models Think in (1/2)

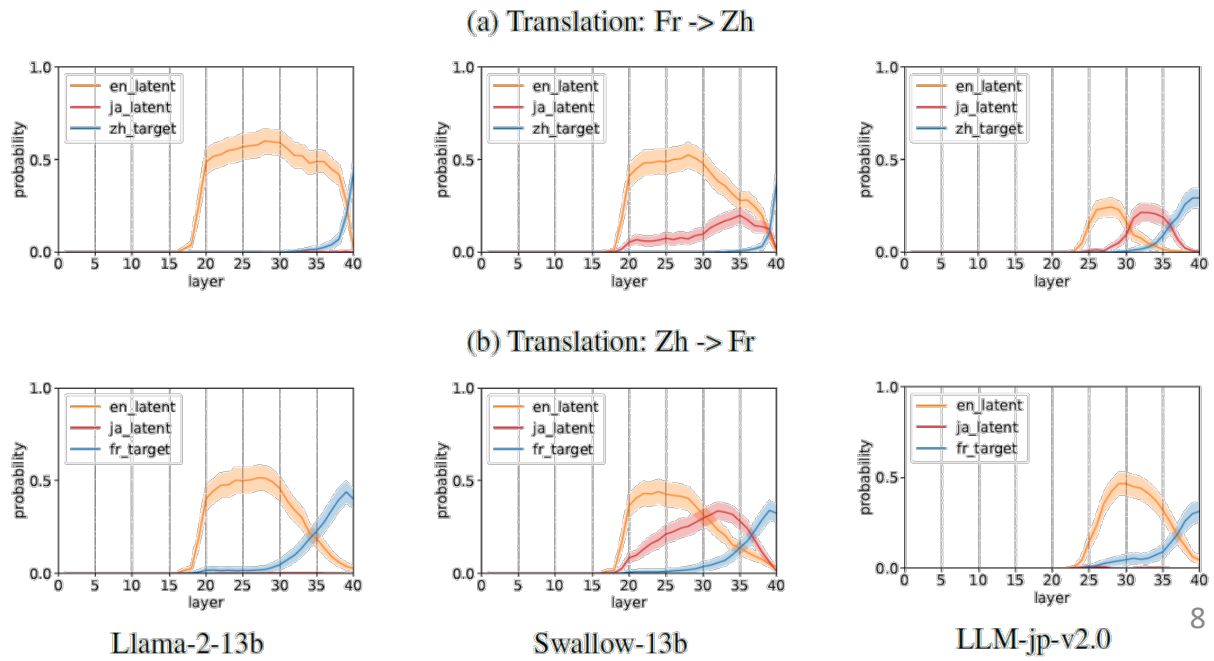
- LLM-jpチームによるwork in progressな報告（Zhong+'24, arXiv）
- ３タイプの日本語で利用可能な言語モデルで同種の実験を実施
  1. Llama2（英語を中心に学習）
  2. Swallow (日本語で継続事前学習)
  3. LLM-jp　(日本語を中心に学習)
- 日本語の解析時
  - Swallowにおける潜在言語は英語と日本語の両方（Llamaより明らかに日本語が多い）
  - LLM-jpにおける潜在言語は基本的に日本語のみ



後続研究: Beyond English-Centric LLMs: What Language Do Multilingual Language Models Think in (2/2)

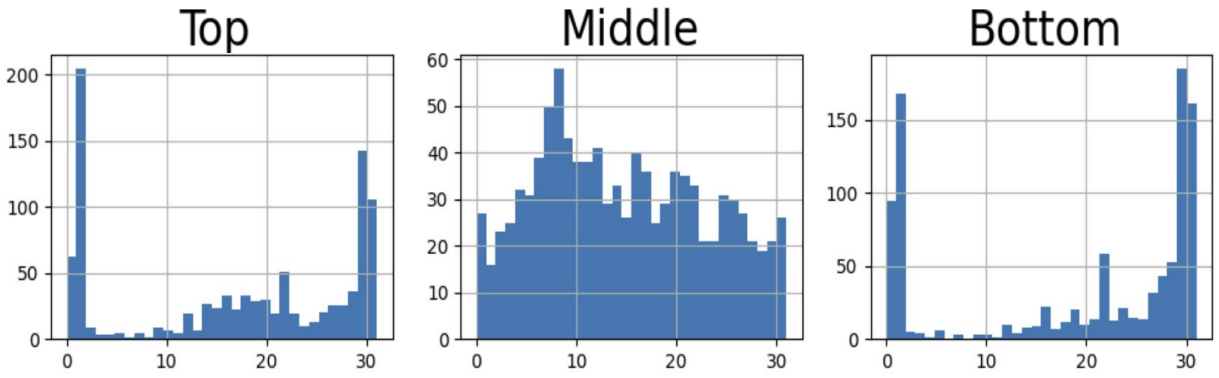
- 非データ支配言語 x 穴埋めタスク
  - Swallowの潜在言語は英語と日本語の両方
  - LLM-jpの潜在言語は中国語の解析時は主に日本語、仏語解析時は主に英語
- 非データ支配言語 x 翻訳タスク
  - Swallowの潜在言語は英語と日本語の両方
  - LLM-jpの潜在言語はZh→Frタスクでは英語、Fr→Zhタスクでは英語と日本語の両方



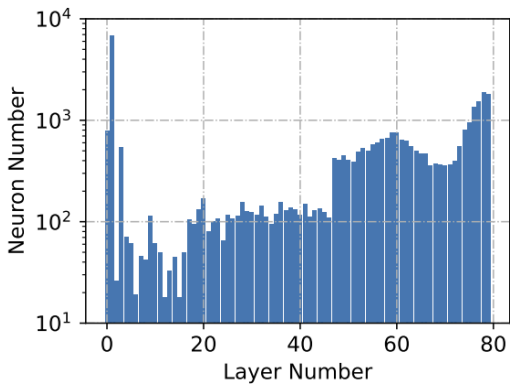


関連研究: Language-specific neurons

- 特定言語で重要なニューロン（=パラメータ）を特定・分析する研究
  - On the Multilingual Ability of Decoder-based Pre-trained Language Models: Finding and Controlling Language-Specific Neurons [Kojima+, NAACL2024]
  - Language-Specific Neurons: The Key to Multilingual Capabilities in Large Language Models [Tang+, ACL2024]
- いずれも言語固有のニューロンが入力層、出力層付近に多く存在すること、非活性化することで対象言語を扱う能力が大幅に低下することを報告
- 特定言語の処理は入出力付近の層で行われる点で本論文の主張と整合



Llama-2 13Bにおける言語特有ニューロン(=Top, Bottom 1000)の分布: 言語特有ニューロンは入力層、出力層付近に多く存在 [Kojima+'24]



Llama-2 70Bにおける言語特有ニューロンの数(1%)の分布 [Tang+'24]

まとめ

- モデルの中間層でトークン予測を行うと英語が出力される傾向
  ⇒ 英語はピボット言語と見なせる
- しかし埋め込みのうち次トークン予測に関係のある部分は限定的
- 英語の語彙を単純に利用していというより概念として英単語を利用している

… the model's internal lingua franca is not English but concepts— concepts that are biased toward English. Hence, English could still be seen as a pivot language, but in a semantic, rather than a purely lexical, sense.