# 論文解説: Language-Specific Neurons: The Key to Multilingual Capabilities in Large Language Models

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, Ji-Rong Wen

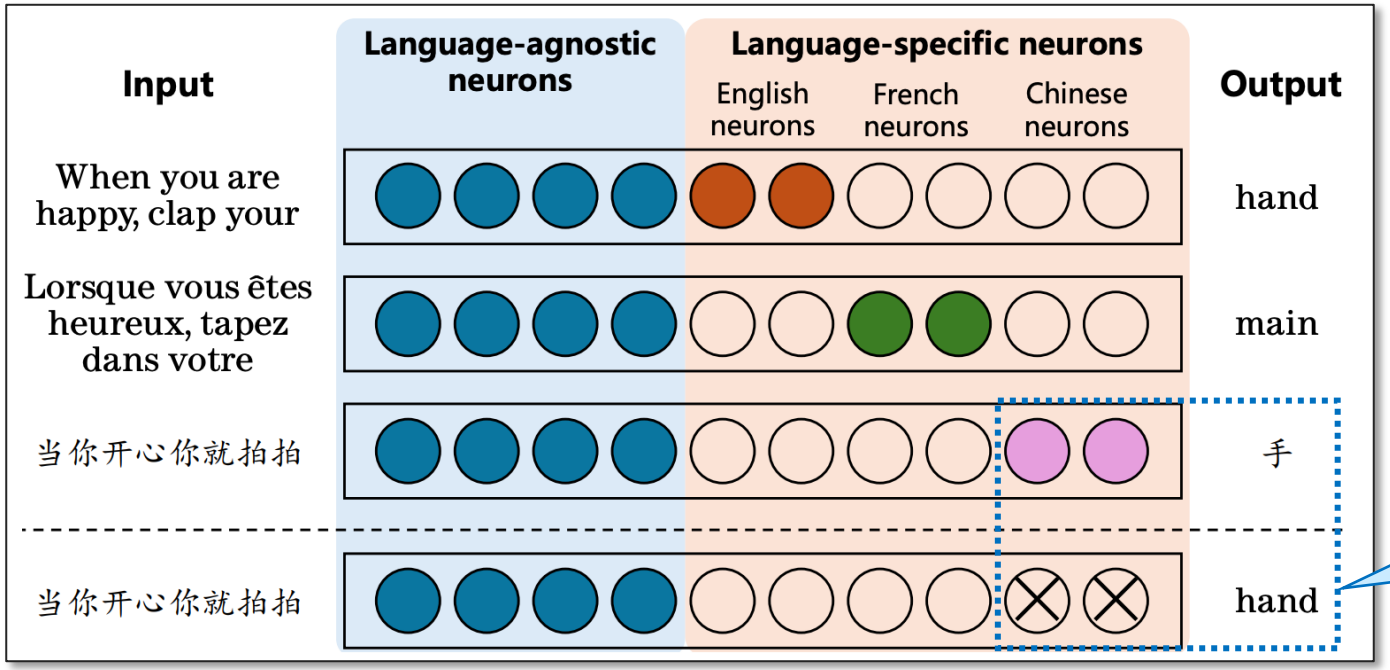
ACL2024 Long, arXiv link

**発表者: 稲葉 達郎**
M2, Audio and Speech Processing Lab, Kyoto University
2024/8/25 第16回最先端 NLP 勉強会

※特に言及がない場合は本スライド中の図表は元論文からの引用です

# 論文概要

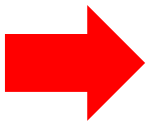
- 大規模言語モデル内の言語特有のニューロン (Language-specific neurons) を特定
  - ある言語特有のニューロンを人為的に非活性化するとその言語の処理能力が悪化
  - 逆に，ある言語特有のニューロンを活性化させるとその言語で出力を行うようになる




中国語特有のニューロンを非活性化させると出力が手からhandに変化

# モチベーション

- 大規模言語モデルは高い多言語能力を持つが，内部的にどのように多言語を処理しているかはほとんど分かっていない
- 人間の脳ではブローカ野やウェルニッケ野等の多数の領域が言語処理に関わっており，各領域が特定の言語機能に関与する，即ち役割分担が行われていると言われている
- 言語モデルにおいても内部で役割分担が行われているのではないか
  - Language-agnostic 領域: 普遍的な世界知識や誤用論に関連する領域
  - Language-specific 領域: 言語特有の語彙や文法，構文に関連する領域



大規模言語モデルの多言語能力に直結する Language-specific な領域は存在するのか，存在するならどんな役割を果たすのか

# ニューロンとは (1/2)

- 本論文では言語モデル内の領域として FFN 内の中間層をニューロンとみなし分析

# ニューロンとは (1/2)

- 本論文では言語モデル内の領域として FFN 内の中間層をニューロンとみなし分析

**一般的な言語モデルのFFN**

$i$ 層目の FFN の入力を $\tilde{h}^i \in \mathbb{R}^d$ とするとその出力 $h^i \in \mathbb{R}^d$ は,

$$h^i = \text{act\_fn}\left(\tilde{h}^i W_1^i\right) \cdot W_2^i$$

と定義される．ただし，$W_1^i, \in \mathbb{R}^{d \times 4d}, W_2^i \in \mathbb{R}^{4d \times d}$ は学習可能パラメータであり，

act_fn は RELU や GELU 等の活性化関数.

# ニューロンとは (1/2)

- 本論文では言語モデル内の領域として FFN 内の中間層をニューロンとみなし分析

**一般的な言語モデルのFFN**

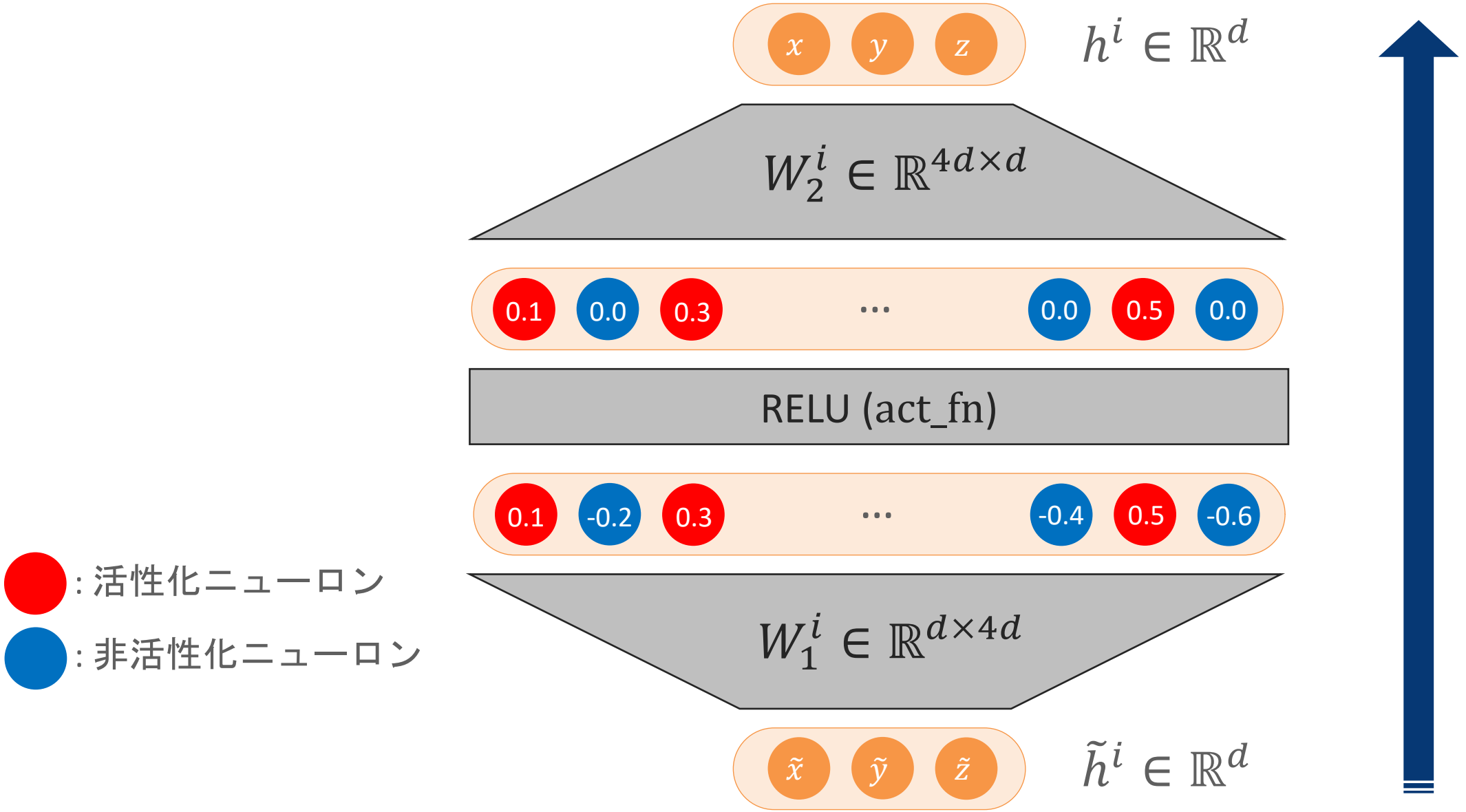
$i$ 層目の FFN の入力を $\tilde{h}^i \in \mathbb{R}^d$ とするとその出力 $h^i \in \mathbb{R}^d$ は,

$$h^i = \text{act\_fn}\left(\tilde{h}^i W_1^i\right) \cdot W_2^i$$

と定義される．ただし，$W_1^i, \in \mathbb{R}^{d \times 4d}, W_2^i \in \mathbb{R}^{4d \times d}$ は学習可能パラメータであり，act_fn は RELU や GELU 等の活性化関数.

- 各 FFN ごとに $4d$ 個のニューロンがあり，$j$ 番目のニューロン $\text{act\_fn}\left(\tilde{h}^i W_1^i\right)_j$ が 0 より大きければ活性しているという

# ニューロンとは (2/2)


$x$ $y$ $z$
$h^i \in \mathbb{R}^d$
$W_2^i \in \mathbb{R}^{4d \times d}$
0.1 0.0 0.3 … 0.0 0.5 0.0
RELU (act_fn)
0.1 -0.2 0.3 … -0.4 0.5 -0.6
$W_1^i \in \mathbb{R}^{d \times 4d}$
$\tilde{x}$ $\tilde{y}$ $\tilde{z}$
$\tilde{h}^i \in \mathbb{R}^d$
: 活性化ニューロン
: 非活性化ニューロン

# Language-Specific ニューロンをどう探すのか (1/2)

**Language Activation Probability Entropy (LAPE): 言語別の活性化確率のエントロピー**

- $i$ 層目 $j$ 番目のニューロンが言語 $k$ が入力された時に活性化する確率を $p_{i,j}^{k}$ とする
- この確率を $l$ 個の言語全てで計算

$$\boldsymbol{p}_{i,j} = \left(p_{i,j}^{1}, \ldots, p_{i,j}^{k}, \ldots, p_{i,j}^{l}\right)$$

- L1 normalizationをした後に

$$\boldsymbol{p}'_{i,j} = \frac{1}{\sum_{k=1}^{l} p_{i,j}^{k}} \boldsymbol{p}_{i,j}$$

- エントロピーを計算する

$$\mathrm{LAPE}_{i,j} = -\sum_{k=1}^{l} p'^{k}_{i,j} \log\left(p'^{k}_{i,j}\right)$$

# Language-Specific ニューロンをどう探すのか (2/2)

**Language Activation Probability Entropy (LAPE) 続き**

- LAPE が小さく，言語 $k$ に対する活性化確率が閾値を超えている場合そのニューロンを言語 $k$ の Specific ニューロンとする
  - エントロピーが小さい

    → 1つか2つ程度の言語の活性化確率が比較的大きい

    → その言語が入力された時にのみ活性化する

# Language-Specific ニューロンをどう探すのか (2/2)

## Language Activation Probability Entropy (LAPE) 続き

- LAPE が小さく，言語 $k$ に対する活性化確率が閾値を超えている場合そのニューロンを言語 $k$ の Specific ニューロンとする
  - エントロピーが小さい

    → 1つか2つ程度の言語の活性化確率が比較的大きい

    → その言語が入力された時にのみ活性化する

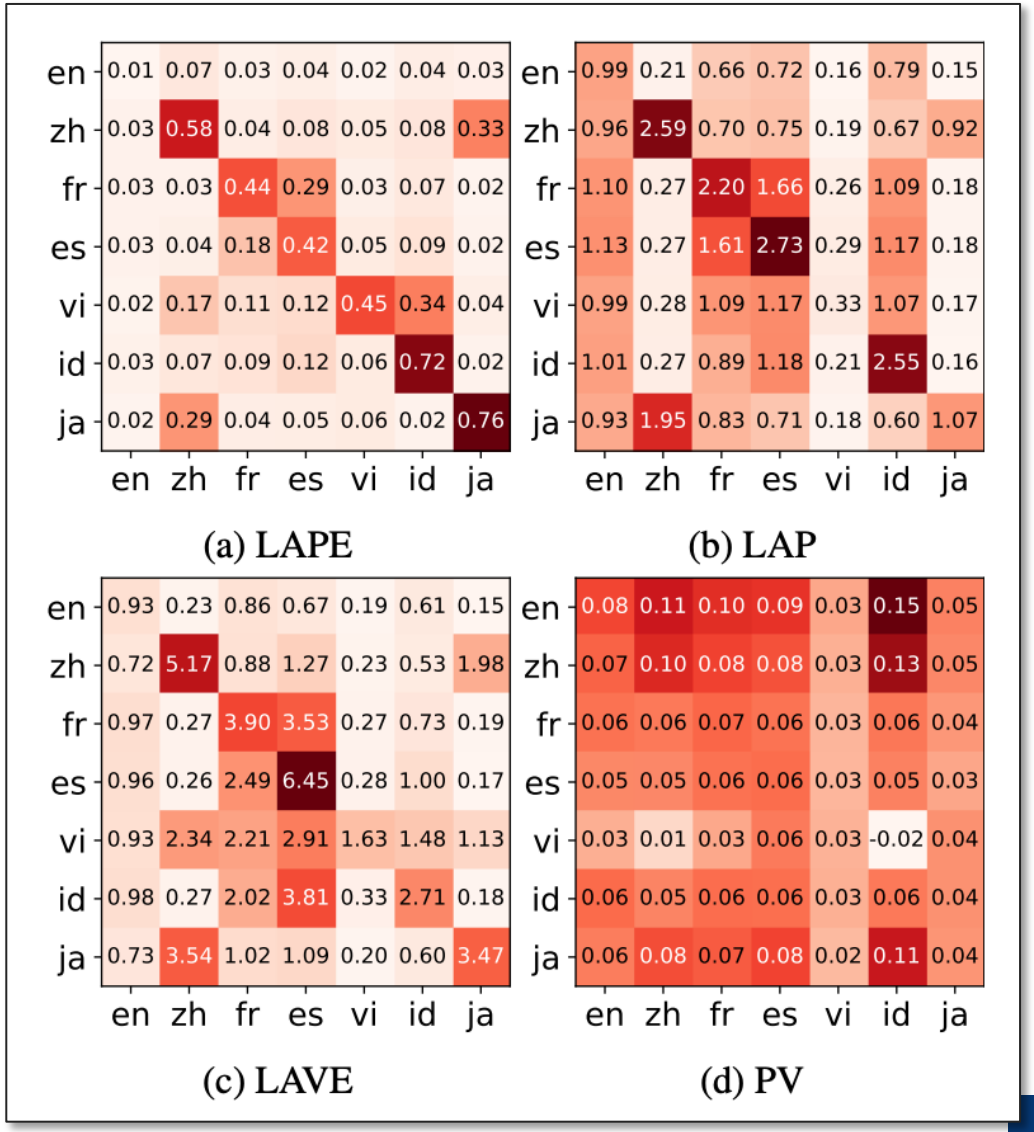
## LAPE 以外の特定手法

- Language Activation Probability (LAP): 言語 $k$ に対する活性化確率が 95 % を超えている
- Language Activation Value Entropy (LAVE): LAPE の活性化確率を活性化関数の出力値に置き換え
- Parameter Variation (PV): 言語 $k$ での Instruction tuning 前後では変化せず，他の言語での Instruction tuning 前後で大きく変化するパラメータを LAPE と同様に Entropy を用いて特定

# Language-specific ニューロンを非活性化すると…

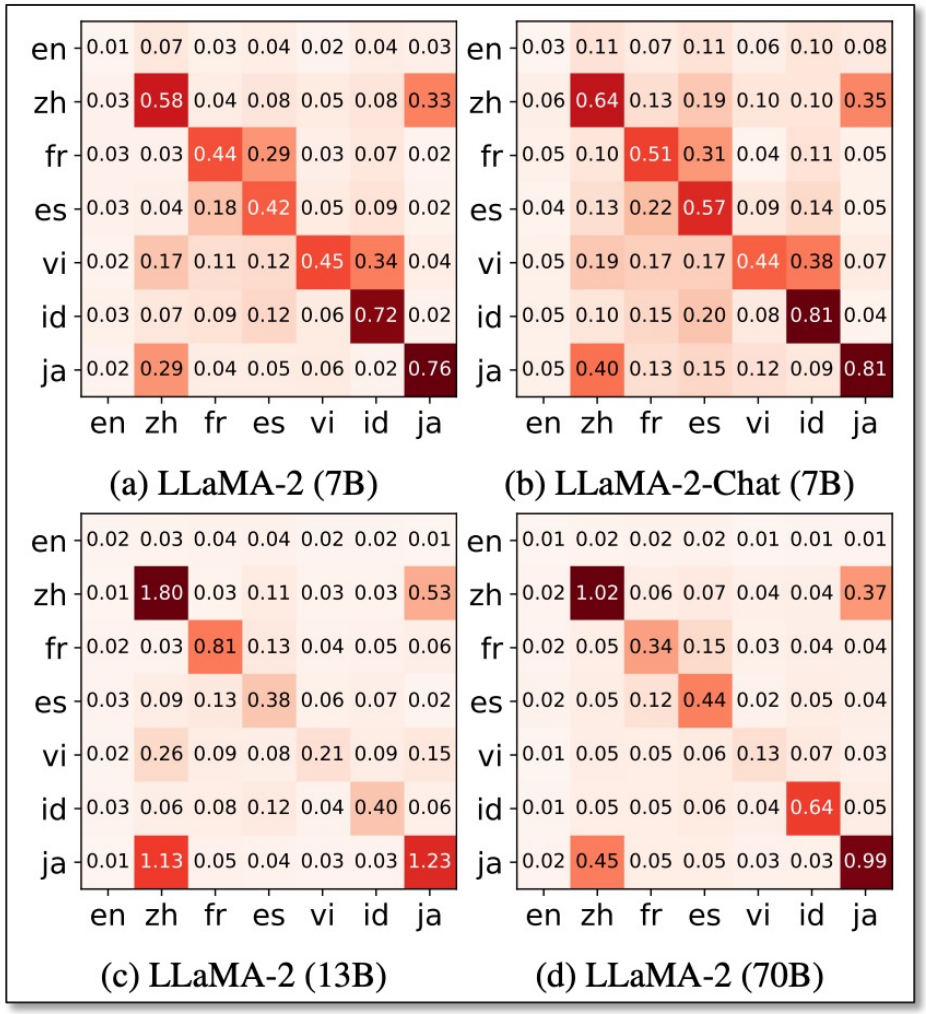
- Language-Specigic ニューロンを非活性化させた時にどれだけPPLが上昇するか
  - 横軸が非活性化する言語，縦軸が性能を評価する言語
- LAPE は対角成分において性能が悪化
  - “言語特有”のニューロンを特定できている
- LAVEでは対角成分以外でもPPLが上昇
  - 他の言語への影響が大きい，”言語特有”ではない

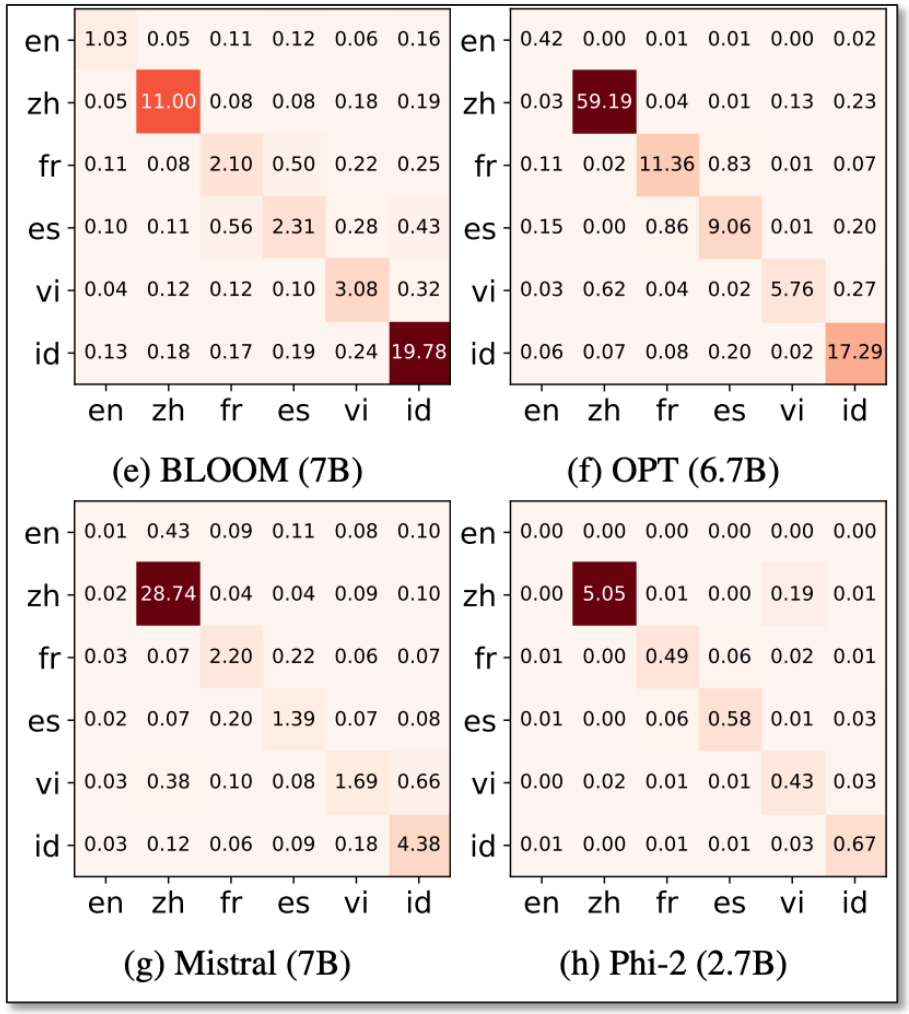
※この実験では全体のニューロンの1%のみを非活性化している



# 他のモデルでも同様の結果

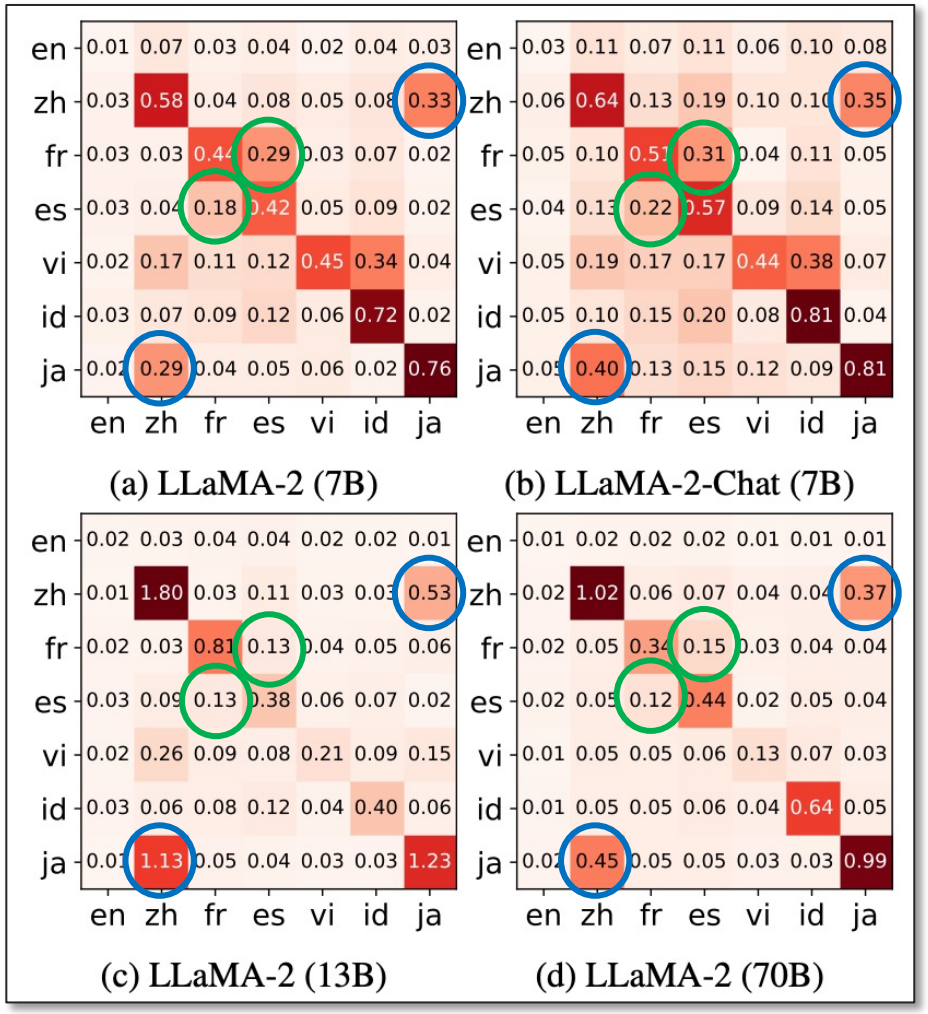
- LAPE により特定した Language-specific ニューロンを非活性化



# 他のモデルでも同様の結果

- LAPE により特定した Language-specific ニューロンを非活性化




- 中国語(zh)と日本語(ja)間や，フランス語(fr)とスペイン語(es)間での相互干渉が大きい
  - これらの似た言語間ではそれぞれのLanguage-Specific ニューロン集合に共通部分がある

# Open-ended 生成タスクでもLanguage-specific ニューロンの存在を確認 (2/2)

- Language-specific ニューロンを非活性化した時に Open-ended 生成タスクに対する性能がどう変わるか
- GPT-4により各言語へ翻訳したVicunaデータセットを使用，GPT-4による5段階評価

| | zh | fr | es | vi | id | ja |
|---|---|---|---|---|---|---|
| **Normal** | 4.30 | 4.19 | 3.51 | 3.70 | 4.16 | 2.86 |
| **Random** | 4.18 | 4.22 | 3.35 | 3.53 | 4.42 | 2.99 |
| **zh** | **2.46** | 3.56 | 2.96 | 3.64 | 3.56 | 2.31 |
| **fr** | 3.69 | **2.50** | 2.29 | 3.01 | 3.59 | 2.76 |
| **es** | 3.51 | 2.57 | **2.01** | 3.14 | 3.34 | 2.56 |
| **vi** | 3.93 | 3.19 | 2.49 | **2.74** | 3.59 | 2.74 |
| **id** | 3.67 | 3.10 | 2.67 | 3.21 | **2.84** | 2.80 |
| **ja** | 3.21 | 3.69 | 3.07 | 3.49 | 3.37 | **1.84** |

非活性化なし→
ランダムに1%非活性化→

# Open-ended 生成タスクでもLanguage-specific ニューロンの存在を確認 (2/2)

- 中国語の質問に対し，本来は中国語で流暢な回答を生成する
- 一方で，中国語ニューロンを非活性化させると英語と中国語混じりの意味の通らない回答を生成

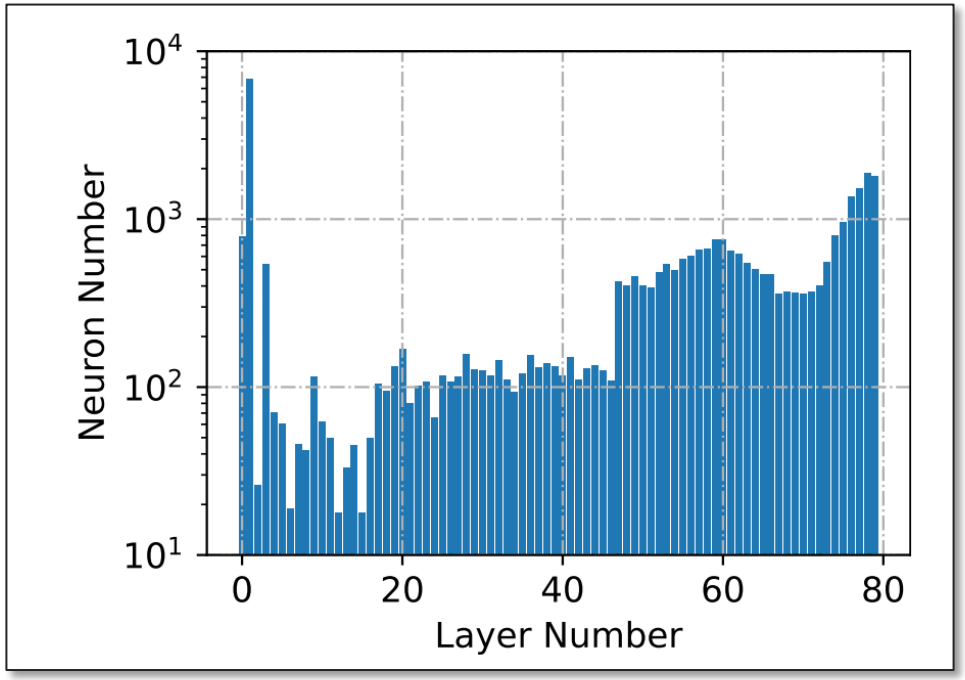
**Question**
你是一位登上珠穆朗玛峰顶峰的登山者。描述一下你的情绪和从高处看到的景色。
(*Translation*: You are a mountain climber reaching the summit of Mount Everest. Describe your emotions and the view.)

**Normal output**
我是一个登上珠穆朗玛峰顶峰的登山者。当我站在山顶时，我感到非常兴奋和自豪。...
(*Translation*: I am a climber who has reached the summit of Mount Everest. When I stood on the top of the mountain, I felt very excited and proud. ...)

**Deactivated output**
我是一個登上珠穆朗瑪峰頂峰的登山者。I am a mountaineer who has climbed to the top of Mount Everest. 當我站在珠my朗ma峰頂峰，我感到非常興奮和欣慰。...

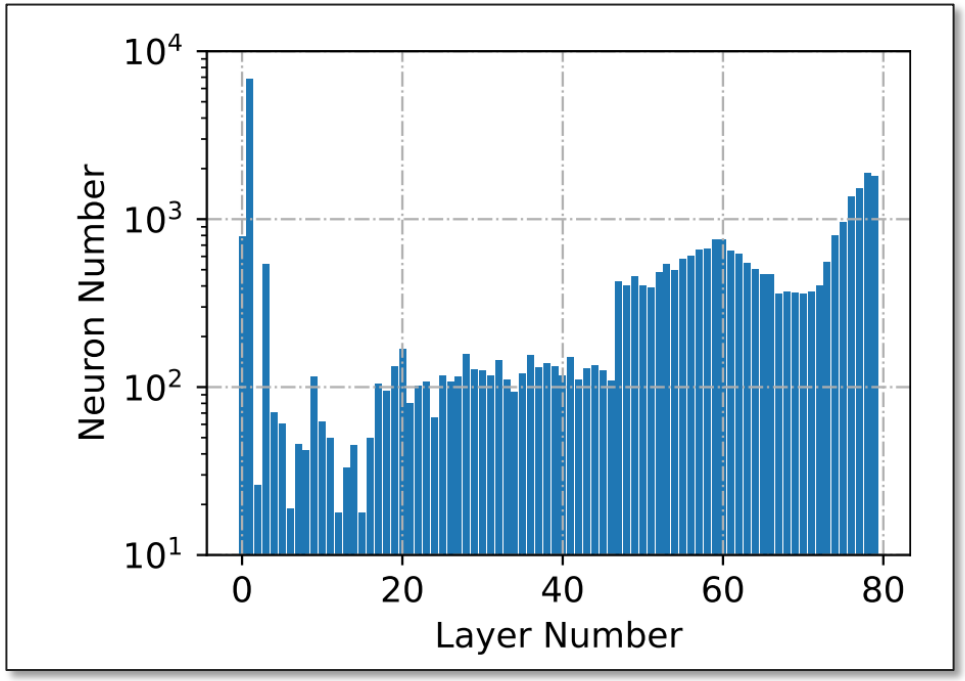
# どこに Language-specific ニューロンがあるのか

- 浅い層と深い層に多く存在 (U字を描く)



# どこに Language-specific ニューロンがあるのか

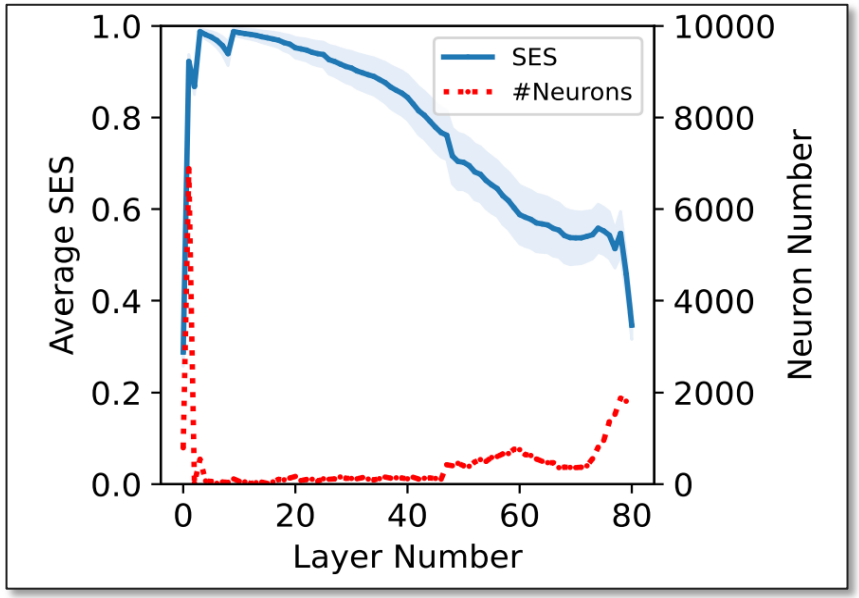
- 浅い層と深い層に多く存在 (U字を描く)




➡ なぜ中層は Language-specific ニューロンが少ないのか

# なぜ Language-specific ニューロンは U 字に分布するのか

- 各言語へ翻訳されたVicunaデータをモデルに入力し，層ごとに異なる言語間の潜在表現のsentence embedding similarity(SES) を計算すると，Uの上下反転の形
  - 異なる言語のテキストでも内容が同じならモデルの中層では同じような表現として扱っている




- 浅い層: 異なる言語の入力文を共通の潜在空間に mapping
- 深い層: 共通の潜在空間から各言語の語彙空間へ mapping

# Case Study: 入力と同じ言語で回答生成

- Off-target issue (モデルが入力と異なる言語で回答を生成してしまう問題) を入力言語の language-specific ニューロンを活性化させることで軽減
  - 生成される回答の質も向上

入力言語と同じ言語で回答を生成する確率

生成された回答の質（GPT-4により評価）

| Metrics | Settings | zh | fr | es | vi | id | ja |
|---|---|---|---|---|---|---|---|
| Language accuracy | normal | 0.87 | 0.73 | 0.81 | 0.60 | 0.40 | 0.79 |
| | steered | 0.99 | 0.90 | 0.93 | 0.97 | 0.99 | 1.00 |
| Content quality | normal | 4.30 | 4.19 | 3.51 | 3.70 | 4.16 | 2.86 |
| | steered | 4.57 | 4.35 | 4.02 | 3.57 | 4.28 | 2.91 |

# Case Study: 入力と異なる言語で回答生成

- 入力言語の Language-specific ニューロンを非活性化させ，別言語のLanguage-specific ニューロンを活性化

- 右図の例ではスペイン語の質問を与え，スペイン語ニューロンを非活性化&中国語ニューロンを活性化
- 流暢な中国語で回答を生成できている

**Question**

¿Cómo puedo mejorar mis habilidades de gestión del tiempo?

(*Translation*: How can I improve my time management skills?)

**Steered output in Chinese**

Time管理是一个非常重要的技能，因为它可以提高你的生产力并使你更有效地完成任务。下面是一些提示来改进你的时间管理技能：1、制定计划：每天开始时制定清单并将任务分配给特定时间段。这样可以确保你在每天都有足够的时间来完成所有任务。2、设置目标：设置明确而可实现的目标可以帮助你保持专注并提高效果。3、抵制分心：当你正在工作时，请不要被其他事情分心。4、利用技术：……

(*Translation*: Time management is a very important skill because it increases your productivity and allows you to complete tasks more efficiently. Here are some tips to improve your time management skills: 1. Make a plan: Make a list at the beginning of each day and assign tasks to specific time periods. This ensures that you have enough time each day to complete all your tasks. 2. Set goals: Setting clear and achievable goals can help you stay focused and improve your results. 3. Resist distractions: When you are working, please don't be distracted by other things. 4. Utilize technology: ...)

# まとめ

- Language-specific ニューロンは存在する
  - 全体のニューロン数に比べ少量な Language-specific ニューロンを非活性化させるとその言語の性能のみが著しく低下
- Language-Specific ニューロンはモデルの浅い層と深い層に多く存在
  - 浅い層: 異なる入力の言語を共通の潜在空間へ射影
  - 深い層: 共通の潜在空間から出力の言語の空間へ射影
- Language-specific ニューロンの活性化の有無を変化させることで出力言語を意図的に操作することが可能
  - Off-target issue の軽減に成功