多言語大規模言語モデルにおける 英語指示文と対象言語指示文の公平な比較

榎本大晟 [1]　金輝燦 [1]　陳宙斯 [2]　小町守 [2]
[1] 東京都立大学　[2] 一橋大学
enomoto-taisei@ed.tmu.ac.jp

概要

多言語大規模言語モデルでタスクを解く際，非英語のデータを扱う場合であっても，英語指示文が対象言語指示文より効果的である傾向が報告されている．しかし，それらの研究では英語から翻訳されたデータセットや指示文が用いられていることが多く，翻訳特有のバイアス（Translationese）が指示文の言語間の公平な比較を妨げている可能性がある．この問題に対して，本研究では Translationese の影響を排除し，指示文の公平な比較を実現する．結果として，先行研究と異なり，どちらの指示文がより効果的であるかはタスクや分類ラベルによって異なることを示す．また，各指示文を用いる際の生成テキストの特徴や活性化ニューロンの違いを分析する．

1　はじめに

近年，大規模言語モデル（LLM）はさまざまな自然言語処理タスクにおいて優れた性能を示している．その能力を最大限に引き出すためには，LLM に適切な指示を与えることが必要不可欠である [1, 2]．特に，**多言語大規模言語モデル（MLLM）**を用いて英語以外の言語（対象言語）のタスクを解く際，そのモデルへの指示文を英語で与えるべきか，それとも対象言語で与えるべきかについてはいくつかの研究で議論されてきた [3, 4, 5]．この背景には，MLLM の学習データは多くの場合，英語を中心に構築されているという事実がある．このことから，たとえタスクが英語以外の言語であっても，英語で指示を与える方が MLLM の能力をより効果的に引き出せる可能性が指摘されている．実際に多くの先行研究が，対象言語よりも英語で指示文を与える方が高性能になる傾向を報告している [4, 5]．

しかしながら，これらの先行研究では，対象言語のテストデータセットや指示文として英語から翻訳されたものが使用されている，という問題がある．翻訳によって作成された文は，情報の欠落や不自然さ，母語話者が書いた文とは著しく異なる文体や構造を持つ可能性（**Translationese**）がある [6, 7, 8]．これにより，英語から翻訳された対象言語のデータセットでは，表現が英語の文体に近づいたり，内容が英語圏の文化や背景に影響を受けている場合がある．また，対象言語の指示文において翻訳の前後で含んでいる情報が異なる可能性がある．これらの要因から，先行研究では英語指示文が潜在的に有利な設定になっており，英語指示文と対象言語指示文の公平な比較ができていないと考えられる．

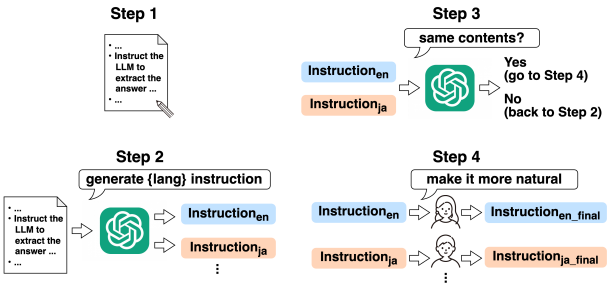

図 1　公平な指示文作成の手順．

この問題を解決するために，本研究では，Translationese の影響を排除して，MLLM において英語指示文と対象言語指示文の公平な比較を行う．具体的には，**翻訳に基づかない対象言語のデータセットや，言語的に自然で同じ内容を伝える公平な指示文（図 1）を使用し，MLLM の性能の違いをさまざまなタスクで調査する**．特に分類タスクでは，複数のラベルセットを使用し，ラベルセットの違いによる結果の変化についても調べる．実験結果から，先行研究とは異なり，英語指示文と対象言語指示文のどちらがより優れているかは，タスクや分類ラベルによって異なる傾向があることを明らかにする．さらに，それぞれの指示文を用いた場合の MLLM が生成するテキストの特徴や活性化するニューロンの違いについての詳細な分析も行う．本

研究は，MLLM における指示文言語の公平な比較を行うことで， MLLM の能力を効果的に引き出すための新たな知見を提供する．

2 公平な比較の実現

この節では，英語指示文と対象言語指示文の公平な比較を実現するために Translationese の影響を排除する方法と，実験設定について説明する．

2.1 公平な指示文の作成

英語指示文と対象言語指示文の公平な比較をするためには，両方の指示文が十分に流暢であることと，同じ内容を伝えることが不可欠である．我々はそのような指示文を以下の手順で作成する（図 1）：

1. 各タスクの指示文に含まれるべき内容を人手で定義する．
2. Step 1 の定義に基づき，GPT-4（`gpt-4o-2024-05-13`）を用いて各言語の指示文を生成する．
3. 英語指示文と対象言語指示文が同じ内容を伝えているかを GPT-4 を用いて検証する．内容に違いがあると判断された場合は，Step 2 に戻る．
4. 各言語の指示文が自然な表現や言い回しになるように母語話者が修正を行う．

Step 1 の定義文から各言語の母語話者が指示文を作成する方法も検討したが，この方法では指示文間で内容や形式の差異が確認された．一方で我々の作成手順では，言語間で指示文が同じ内容を伝え，言語的に自然であることを保証する[1)]．

2.2 タスクとデータセット

本研究では 3 つのタスクで英語指示文と対象言語指示文の比較を行う．以降，それぞれのタスクの概要と，実験に使用する英語からの翻訳に基づかないテストデータセットについて説明する．

語彙平易化タスク 語彙平易化タスク（LS）は，ある文中の対象となる単語を，より簡単で理解しやすい同義語に置き換えることで，文の意味を保ちながら平易にするタスクである．本研究では，対象単語ごとにより平易な同義語を 1 つ生成し，それがゴールドスタンダードの解答に含まれるかに基づいて Accuracy を測定する．LS においての対象言語は，`de`，`es`，`fr`，`ja`，`zh` の 5 言語である．テストデータセットとしては，MultiLS [9]（`de`，`es`，`fr`，`ja`），Chinese-LS [10]（`zh`）を用いる．

機械読解タスク 機械読解タスク（MRC）は，質問と参照テキストが与えられ，その質問に対する答えを参照テキストから抽出するタスクである．本研究では，質問に対する答えを参照テキストから抽出するように指示し，生成されたテキストがゴールドスタンダードの答えと完全に一致するかに基づいて Accuracy を測定する．MRC においての対象言語は，`de`，`es`，`fr`，`id`，`ja`，`ko`，`zh` の 7 言語である．テストデータセットとしては，GermanQuAD [11]（`de`），SQAC [12]（`es`），FQuAD [13]（`fr`），TyDiQA-Gold [8]（`id`，`ja`，`ko`），DRCD [14]（`zh`） を用いる．

レビュー分類タスク 本研究におけるレビュー分類タスク（RC）は，レビュー文が肯定的な評価をしているか否定的な評価をしているかを分類するタスクである．分類ラベルセットの違いによる結果の変化を分析するために，英語のラベルセットを使用する設定と対象言語のラベルセットを使用する設定のそれぞれで英語指示文と対象言語指示文の macro-F1 を比較する．RC においての対象言語は，`de`，`es`，`fr`，`id`，`ja`，`ko`，`zh` の 7 言語である．テストデータセットとしては，MARC [15]（`de`，`es`，`fr`，`ja`，`zh`），NSMC [16]（`ko`），PRDECT-ID [17]（`id`）を用いる．

2.3 MLLM

本研究は，指示文の言語による MLLM の性能の変化を分析することを目的としており，指示チューニング済みモデルに焦点を当てている．用いるモデルは suzume-multilingual 8B [18]，Qwen2-Instruct 7B [19]，Mistral-NeMo-Instruct 12B [20] である．これらはそれぞれ Llama 3，Qwen2，Mistral-NeMo の多言語指示チューニング済みモデルである．以降，それぞれを ‘llama3-i’，‘qwen2-i’，‘mistraln-i’ と表記する．

3 実験結果

表 1 に zero-shot 設定での各タスクにおける全ての対象言語の平均の性能を示す．

語彙平易化タスク 実験結果から，対象言語指示文が英語指示文の性能を上回る傾向があることが確認された．また，日本語では，英語から翻訳された指示文の性能が大幅に低下した．これは，英語指示文に含まれる数値情報が翻訳の過程で失われたためである（付録 C）．この結果は，先行研究のような英語指示文と英語から翻訳された対象言語指示文の比較が，必ずしも公平ではない可能性があることを示

1) 最終的な指示文を用いたプロンプトの例は付録 B に示す．

表 1 en（英語指示文），tgt（対象言語指示文），tgt-mt（Bing Translator を用いて英語から翻訳された対象言語の指示文）の性能の比較．全対象言語間の平均スコアを示す．タスクごとに各モデルの最高性能を太字で強調する．

タスク	指示文	性能		
		llama3-i	qwen2-i	mistraln-i
LS	en	26.95	44.38	48.68
	tgt	**28.31**	**46.52**	**52.78**
	tgt-mt	23.33	40.64	46.12
MRC	en	**25.47**	**32.33**	**39.48**
	tgt	20.07	22.19	31.47
	tgt-mt	18.01	18.47	32.91
RC (en label)	en	**87.66**	**90.58**	**89.15**
	tgt	77.57	90.56	80.47
	tgt-mt	83.96	88.82	79.06
RC (tgt label)	en	66.72	86.49	65.34
	tgt	**70.14**	**89.46**	**65.47**
	tgt-mt	69.22	81.58	61.17

している．そのような偏った条件では英語指示文が効果的であると不当に評価されることになる．

機械読解タスク　実験結果から，英語指示文が対象言語指示文の性能を上回る傾向があることが確認された．この傾向は，LS の傾向とは対照的であり，英語指示文と対象言語指示文のどちらがより効果的なのかはタスクにより変化することを示している．

レビュー分類タスク　実験結果から，英語の分類ラベルを使用する設定では，英語指示文が対象言語指示文の性能を上回る傾向があることが確認された．一方で，対象言語の分類ラベルを使用する設定では，対象言語指示文が英語指示文の性能を上回る傾向がある．これらの結果は，分類タスクにおいて最適な指示文の言語は分類ラベルの言語に依存し，ラベルの言語と同じ言語の指示文がより高い性能になる傾向があることを示している．

4　指示文による違い

4.1　生成テキストの特徴

MRC において，英語指示文を用いる設定と対象言語指示文を用いる設定間で MLLM が生成するテキストが同じであるインスタンスの割合は llama3-i が約 30 %，qwen2-i が約 37 %，mistraln-i が約 48 % である．これらの結果から，同じ内容を伝えるが異なる言語で書かれている 2 つの指示文に対して，MLLM は異なるテキストを生成することが多くあることがわかる．以下では，各指示文を用いたときに MLLM が生成するテキストの特徴を分析する．

表 2　MLLM が対象言語以外の言語のテキストを生成するインスタンスの割合．全対象言語の平均の割合を示す．

タスク	指示文	llama3-i	qwen2-i	mistraln-i
LS	en	9.94	8.23	7.08
	tgt	7.13	6.43	6.22
MRC	en	4.33	4.36	2.98
	tgt	2.16	1.47	1.76

表 3　スペイン語と日本語の MRC において，MLLM が未検出テキストを生成するインスタンスの数．

対象言語	指示文	llama3-i	qwen2-i	mistraln-i
es	en	0	1	0
	tgt	8	18	2
ja	en	3	5	0
	tgt	28	15	3

英語指示文は非対象言語の生成が増加する　ここでは MLLM によって生成されたテキストの言語を判別する．言語の判別には FastText [21] を用いる．先行研究 [22, 23] を参考に，FastText の言語判別の確信度が 50 %以上の結果のみを用いる．表 2 に MLLM が非対象言語のテキストを生成したインスタンスの割合を示す．これらの結果は，英語指示文は非対象言語で生成することを増加させる傾向を示している．この観測は Marchisio ら [24] の報告に類似している．特に，英語指示文を用いると，MLLM は英語のテキストを生成することの増加が確認された．また，qwen2-i では英語指示文を用いると，中国語のテキストを生成することも増加する．

対象言語指示文は未検出テキストの生成が増加する　MRC では，参照テキスト中に質問に対する解答が必ず含まれる．しかしながら，“与えられた参照文には質問に対する情報がありません．” のような，情報が見つからなかったことを示すテキスト（未検出テキスト）を MLLM が生成する現象を確認した．我々は es と ja において，MLLM がそのような未検出テキストを生成するインスタンスを人手で数えた．表 3 に MLLM が未検出テキストを生成するインスタンスの数を示す．これらの結果は，対象言語指示文を用いることは未検出テキストの生成を増加させることを示している．特に，対象言語指示文の場合は未検出テキストを生成する一方で，英語指示文の場合は正しい回答を生成するようなインス

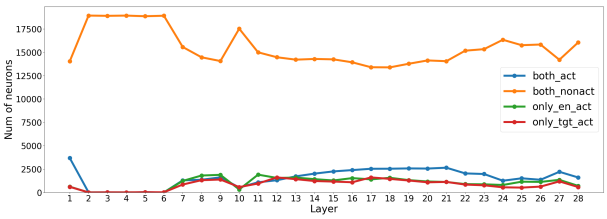

図 2　英語指示文を用いるプロンプトおよび対象言語指示文を用いるプロンプトの最終トークン処理時のニューロンの状態の比較.

タンスがいくつか確認された．このことは，英語指示文を用いることは MLLM の読解能力を引き出すのにより効果的であることを示唆している．

4.2　活性化ニューロン

本節では，指示文の言語が MLLM 内部に与える影響を調査するために，活性化するニューロンについて分析する [2]．以降ではタスクは RC（en label），モデルは qwen2-i に注目する．

指示文と活性化ニューロン　同一のインスタンスに対して英語指示文を用いるプロンプトおよび対象言語指示文を用いるプロンプトの最終トークン [3] を処理する際のニューロンを比較する．ニューロンの活性化のパターンは以下の 4 つに分類できる：

`both_act`：両方の指示文で活性化する
`both_nonact`：両方の指示文で活性化しない
`only_en_act`：英語指示文でのみ活性化する
`only_tgt_act`：対象言語指示文でのみ活性化する

図 2 に qwen2-i の各層においてそれぞれの活性化パターンに該当するニューロンの数を示す．結果として，英語指示文または対象言語指示文のみで活性化するニューロンが一定数存在することが確認された．このことは，指示文の言語に依存して活性化するニューロンが存在することを示唆している．

この結果を踏まえると，ある言語の指示文を用いる際に活性化するニューロンが，言語固有ニューロンと関連している可能性が考えられる．以降では，言語固有ニューロンを説明し，それらが指示文の言語の影響をどのように受けるのか分析する．

言語固有ニューロン　MLLM 内部で特定の言語に強く関連づけられて機能するニューロンの存在が示されており，それらは ‘言語固有ニューロン’ と呼ばれる [23, 25]．言語固有ニューロンは，MLLM が特定の言語を処理する際に主に活性化し，他の言語ではほとんど活性化しない特徴を持つ．

2)　本研究のニューロンと活性化の定義は付録 D.1 で述べる．
3)　各指示文のプロンプト間で最後のトークンは同一である．

表 4　qwen2-i における各言語の $P(l)$．‘tgt’ は対象言語を示し，インスタンスの言語と一致する．$P(en)$ に水色を，$P(tgt)$ に橙色をつける．

指示文	tgt	$P(l)\times100$					
		en	fr	es	de	zh	ja
en	fr	48.66	19.05	16.83	12.59	12.35	13.21
	es	48.68	15.94	22.00	12.60	12.06	13.12
	de	49.12	14.87	15.48	17.20	12.13	13.31
	zh	46.70	10.55	11.34	10.09	21.26	17.23
	ja	46.28	12.05	12.39	11.41	17.26	20.73
tgt	fr	31.30	72.14	34.94	21.81	11.90	15.27
	es	31.42	34.99	68.30	21.10	12.38	15.74
	de	33.28	26.48	24.69	66.10	13.29	18.06
	zh	24.79	8.05	7.55	8.13	50.72	23.98
	ja	23.31	13.17	13.83	18.32	32.44	59.97

本研究では，en，de，es，fr，zh，ja の言語固有ニューロンを特定するために LAPE [25]（付録 D.2）を用いる．言語固有のテキストコーパスには先行研究 [23] で用いられたデータを採用する．各言語固有ニューロンの分布は付録 D.3 に示す．

指示文と言語固有ニューロンの関係　各指示文を用いるプロンプトの最終トークン処理時に，各言語の言語固有ニューロンがどの程度活性化しているかを調べるために言語ごとに以下を計算する：

$$P(l) = \frac{\text{活性化した言語 } l \text{ の固有ニューロンの数}}{\text{言語 } l \text{ の固有ニューロンの数}} \quad (1)$$

表 4 に各対象言語においての $P(l)$ の結果を示す．結果として，インスタンスは対象言語であるにもかかわらず，英語指示文のプロンプトでは英語の言語固有ニューロンが強く活性化するのに対し，対象言語の言語固有ニューロンの活性化は弱い傾向が確認された．一方で，対象言語指示文のプロンプトでは対象言語の言語固有ニューロンが強く活性化する傾向が見られた．この結果は，指示文の言語が MLLM 内部のニューロン活性化パターンに強く影響を与え，モデルが内部で処理する際の言語的な重点が指示文の言語によって変化することを示唆している．

5　おわりに

本研究では，Translationese の影響を排除し，MLLM において英語指示文と対象言語指示文の公平な比較を行った．実験結果から，どちらの指示文がより効果的であるかはタスクや分類ラベルによって異なる傾向があることを明らかにした．また，それぞれの指示を用いた場合に生成されるテキストの特徴や活性化するニューロンに違いが生じることを示した．

謝辞

本研究の一部は JST さきがけ JPMJPR2366 の支援を受けたものである．

参考文献

[1] Xinyuan Wang, Chenxi Li, Zhen Wang, Fan Bai, Haotian Luo, Jiayou Zhang, Nebojsa Jojic, Eric P. Xing, and Zhiting Hu. PromptAgent: Strategic planning with language models enables expert-level prompt optimization. In **ICLR**, 2024.

[2] Ayana Niwa and Hayate Iso. AmbigNLG: Addressing task ambiguity in instruction for NLG. In **EMNLP**, pp. 10733–10752, 2024.

[3] Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. Few-shot learning with multilingual generative language models. In **EMNLP**, pp. 9019–9052, 2022.

[4] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning. In **ACL**, pp. 15991–16111, 2023.

[5] Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. MEGA: Multilingual evaluation of generative AI. In **EMNLP**, pp. 4232–4267, 2023.

[6] Sauleh Eetemadi and Kristina Toutanova. Asymmetric features of human generated translation. In **EMNLP**, pp. 159–164, 2014.

[7] Shuly Wintner. Translationese: Between human and machine translation. In **COLING**, pp. 18–19, 2016.

[8] Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. **TACL**, Vol. 8, pp. 454–470, 2020.

[9] Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Marcos Zampieri, and Horacio Saggion. An extensible massively multilingual lexical simplification pipeline dataset using the MultiLS framework. In **READI**, pp. 38–46, 2024.

[10] Jipeng Qiang, Kang Liu, Ying Li, Yun Li, Yi Zhu, Yun-Hao Yuan, Xiaocheng Hu, and Xiaoye Ouyang. Chinese lexical substitution: Dataset and method. In **EMNLP**, pp. 29–42, 2023.

[11] Timo Möller, Julian Risch, and Malte Pietsch. GermanQuAD and GermanDPR: Improving non-English question answering and passage retrieval. In **MRQA**, pp. 42–50, 2021.

[12] Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquin Silveira-Ocampo, Casimiro Pio Carrino, Carme Armentano-Oller, Carlos Rodriguez-Penagos, Aitor Gonzalez-Agirre, and Marta Villegas. MarIA: Spanish language models. **Procesamiento del Lenguaje Natural**, p. 39–60, 2022.

[13] Martin d'Hoffschmidt, Wacim Belblidia, Quentin Heinrich, Tom Brendlé, and Maxime Vidal. FQuAD: French question answering dataset. In **EMNLP Findings**, pp. 1193–1208, 2020.

[14] Chih Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. DRCD: a Chinese machine reading comprehension dataset, 2019.

[15] Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. The multilingual Amazon reviews corpus. In **EMNLP**, pp. 4563–4568, 2020.

[16] Lucy Park. Naver sentiment movie corpus v1.0, 2015.

[17] Rhio Sutoyo, Said Achmad, Andry Chowanda, Esther Widhi Andangsari, and Sani M. Isa. PRDECT-ID: Indonesian product reviews dataset for emotions classification tasks. **Data in Brief**, Vol. 44, p. 108554, 2022.

[18] Peter Devine. Tagengo: A multilingual chat dataset. In **MRL**, pp. 106–113, 2024.

[19] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024.

[20] MistralAI. Mistral NeMo, 2024.

[21] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. FastText.zip: Compressing text classification models, 2016.

[22] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting high quality monolingual datasets from web crawl data. In **LREC**, pp. 4003–4012, 2020.

[23] Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. In **NAACL**, pp. 6919–6971, 2024.

[24] Kelly Marchisio, Wei-Yin Ko, Alexandre Berard, Théo Dehaze, and Sebastian Ruder. Understanding and mitigating language confusion in LLMs. In **EMNLP**, pp. 6653–6677, 2024.

[25] Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. Language-specific neurons: The key to multilingual capabilities in large language models. In **ACL**, pp. 5701–5715, 2024.

[26] Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In **EMNLP Findings**, pp. 12365–12394, 2023.

[27] Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lacalle, and Mikel Artetxe. Do multilingual language models think better in English? In **EMNLP**, pp. 550–564, 2024.

[28] Yotam Intrator, Matan Halfon, Roman Goldenberg, Reut Tsarfaty, Matan Eyal, Ehud Rivlin, Yossi Matias, and Natalia Aizenberg. Breaking the language barrier: Can direct inference outperform pre-translation in multilingual LLM applications? In **NAACL**, pp. 829–844, 2024.

[29] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In **EMNLP**, pp. 2475–2485, 2018.

[30] Patrick Bareiß, Roman Klinger, and Jeremy Barnes. English prompts are better for NLI-based zero-shot emotion classification than target-language prompts. In **ACM Web Conference**, WWW '24, p. 1318–1326, 2024.

A　関連研究

指示チューニング済みモデルへのプロンプトはインスタンスと指示文の両方から構成されることが一般的である．MLLM へ入力するプロンプトが英語であるべきかそれとも対象言語であるべきかについての研究はインスタンスベースと指示文ベースに分類できる．

インスタンスベース　インスタンスベースのアプローチではインスタンスを英語に翻訳することに焦点を当てている．Huang ら [26] と Etxaniz ら [27] は，タスクを処理する LLM 自体を用いてインスタンスを英語に翻訳することの有効性を報告した．一方で，Intrator ら [28] は PaLM2 ではインスタンスを英語に翻訳することがタスクの性能を低下させる傾向があることを報告した．

指示文ベース　本研究が属する指示文ベースのアプローチは，指示文やプロンプトテンプレートの言語に焦点を当てており，インスタンスには変更を加えない．Lin ら [3] は MLLM に与えるプロンプトテンプレートの言語を比較し，英語のテンプレートがより高い性能になる傾向を報告した．Muennighoff ら [4] と Ahuja ら [5] は指示チューニング済みモデルにおいて，英語指示文と英語から翻訳された対象言語の指示文を比較し，英語指示文がより高い性能になる傾向を報告した．しかしながら，これらの研究は XNLI [29] のような英語から翻訳された多言語のデータセットをテストデータとして使用したり，英語から翻訳された対象言語の指示文を使用しており，Translationese の影響を一切考えていない．一方で，Barei ら [30] は英語に基づかない多言語データセットを用いているが，プロンプトテンプレートは機械翻訳に基づいている点と，自己回帰言語モデルでなくマスク言語モデルに焦点を当てている点において本研究と異なる．

B　プロンプトの例

英語

I will provide a sentence and a word included in the sentence.
Please generate a simpler Japanese synonym for the word.
Generate nothing but the synonym.
Sentence: 芋づる式に窃盗団のメンバーが検挙された。
Word: 芋づる式に
Synonym:

日本語

これから文とその文に含まれる単語を与えます。
与えられた単語に対して、より簡単な日本語の同義語を一つ生成してください。
同義語以外は何も生成しないでください。
文: 芋づる式に窃盗団のメンバーが検挙された。
単語: 芋づる式に
同義語:

C　Translationese の事例

語彙平易化タスクの対象言語が日本語である設定において，指示文での Translationese がタスクの性能に悪影響を与える事例を観測した．以下に英語指示文（en）と英語から翻訳された対象言語指示文（tgt-mt）の一部を示す．

en　Please generate a simpler Japanese synonym for the word.
tgt-mt　より簡単な日本語の同義語を生成してください。

英語指示文では，生成するべき同義語の個数情報である ‘a’ が含まれていることがわかる．一方で，英語から翻訳された日本語指示文は，翻訳の過程で個数情報が失われてしまい，同義語をいくつ生成するべきなのか不透明になっている．その結果，英語から翻訳された日本語指示文を用いると，MLLM は ‘パトカー’ という単語に対して ‘交番車, 車両, 付近の警備車, 駆けつけ車, 警察車’ のような複数の同義語を生成することがあり，性能が大幅に低下した．このことは，先行研究のような，英語指示文と英語から翻訳された対象言語指示文の比較は必ずしも公平でないことを示している．このような公平でない設定では，英語指示文がより効果的であると不当に評価される可能性がある．

D　ニューロン

D.1　ニューロンの定義

本研究では，先行研究 [25] に基づき，各 Transformer 層の feed-forward network における活性化関数の出力をニューロンとする．また，ニューロンの値が正である場合に活性化しているとみなす．

D.2　言語固有ニューロンの求め方：LAPE

Tang ら [25] は言語固有ニューロンの検出指標として，言語活性化確率エントロピー（Language Activation Probability Entropy：LAPE）を提案した．LAPE は，各ニューロンが特定の言語に対してどの程度選択的に活性化するかを定量的に評価する指標である．具体的には，あるニューロンにおける各言語の活性化確率分布からエントロピーを計算する：

$$\mathrm{LAPE}(n) = -\sum_{l \in L} p(n,l) \log p(n,l) \tag{2}$$

ここで，n はニューロン，l は言語，L は MLLM が処理する全言語の集合，$p(n,l)$ はニューロン n が言語 l において活性化する確率を表す．$p(n,l)$ は MLLM が言語 l のテキストコーパスに含まれる各トークンを処理する際の，ニューロン n の活性化の頻度を平均したものである．LAPE が低いニューロンほど特定の言語に対する活性化が集中していることを示す．言語ごとに LAPE が下位 1 % に該当するニューロンのうち，$p(n,l)$ が事前に定義された閾値を超えるものが言語固有ニューロンとみなされる．

D.3　言語固有ニューロンの数

図 3 に qwen2-i においての各言語の言語固有ニューロンの数を示す．学習データの大部分を占める言語は言語固有ニューロンが少なくなることを Tang らは報告しており，本研究では qwen2-i は英語と中国語の言語固有ニューロンが比較的少ないことが確認された．

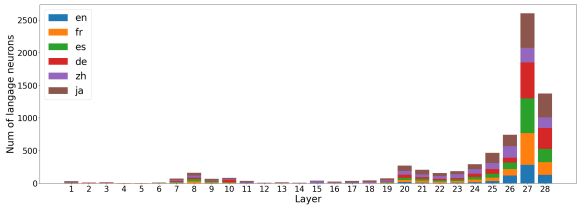

図 3　qwen2-i の各層における各言語の言語固有ニューロンの数．