デコーダベースの事前学習済み言語モデルの多言語能力に関する分析：言語固有ニューロンの検出と制御

小島 武[1] 沖村 樹[1] 岩澤 有祐[1] 谷中 瞳[1] 松尾 豊[1]
[1] 東京大学
t.kojima@weblab.t.u-tokyo.ac.jp

概要

近年のデコーダベースの事前学習済み言語モデル（PLM）は，多言語能力の発現に成功している．しかし，モデル内部でそれぞれの言語がどのように扱われているかは明らかではない．我々は，多言語に対応したデコーダベースの PLM 内における，言語ごとに独自に発火する言語固有のニューロンの内部挙動を解析した．具体的には，英語，ドイツ語，フランス語，スペイン語，中国語，日本語の 6 言語を分析し，言語固有のニューロンは言語間でわずかな重複（＜ 5 %）があるものの言語固有であり，その多くはモデルの最初と最後の数層に分布していることを示した．この傾向は，言語やモデルを問わず一貫していることを確認した．また，推論時に言語固有ニューロンの発火値を改ざんすることで，テキスト生成において指定した対象の言語が生起する確率を顕著に変更できることを示した．

1 はじめに

近年の研究では，Transformer 構造を持つ事前学習済み言語モデル（Pre-trained language model: PLM）の優れた多言語能力が頻繁に報告されている．一部の PLM は，明示的に多言語コーパスを混合して事前学習を行っている [1, 2] が，英語主体のテキストコーパスを使用して学習したモデルでも，多言語のテキストが低い割合で含まれていたために，意図せずに多言語能力を獲得する場合もある．例えば，Llama 2 [3] がその一例である．これらのモデルはどのように多言語能力を表現するのか．この問いに答えるため，先行研究は複数言語にわたって活性化する言語普遍的なニューロンの検出に焦点を当てており，主にエンコーダベースの PLM に注目している [4, 5, 6, 7, 8]．入力の抽象化を検証するにはエンコーダベースのモデルで十分かもしれないが，一方でデコーダベースの PLM は生成の後半部分で言語固有の情報を回復して言語化する必要があるため，これらのモデル内での言語固有の処理は，エンコーダベースのものよりも複雑で本質的な機能であるはずである．しかし，デコーダベースの PLM における言語固有のニューロンの存在と発火に焦点を当てた研究は限られている．

本研究は，デコーダベースの PLM における言語固有のニューロンの挙動を調査する．具体的には，XGLM，BLOOM，および Llama 2 を含む複数のデコーダベースの PLM を，6 つの言語（英語，ドイツ語，フランス語，スペイン語，中国語，日本語）について分析する．言語固有のニューロンを調査するために，[9] によって提案されたアプローチを採用する．このアプローチは，あるグループの文（ポジティブ文）に対して活性化するが，他のグループ（ネガティブ文）に対しては活性化しないニューロンを特定する．対象とする言語のテキストをポジティブ，それ以外の言語のテキストをネガティブとして扱い，ポジティブな文に統計的に活性化する言語固有のニューロンを特定する．実験では，特定された言語固有のニューロンが主にモデルの最初の数層と最後の数層に分布し，この傾向は複数の言語とモデルの種類にわたって一貫していることを示す．また，検出したニューロンの効果を検証するために，推論時にモデル内の言語固有ニューロンに介入することで，生成テキストの対象言語の生起確率を制御することができることを示す．

2 提案手法

我々は，[9] のアプローチに基づき，各言語に固有のニューロンを検出する．このアプローチは元々，同音異義語や性別バイアスなど，特定の単語レベルの概念に反応するニューロンを見つけ出し制御するために開発された．しかし，我々はより広範な文レ

ベルおよび言語固有の概念を把握するニューロンを見つけることを目指しているため，元のアプローチを我々の目的に合わせて修正する．

まず，$|L|$ 言語の集合を考え，各言語のテキストを準備する．各言語 $l \in L$ について，N_l 個のテキストを準備することで，全言語で合計 $N = N_1 + ... + N_l + ... + N_{|L|}$ 個のテキストが得られる．全テキストの集合を $x = \{x_i\}_{i=1}^{N}$ とする．我々の目標は，対象とする言語 $l_t \in L$ のテキストに対して活性化するが，他の言語 $L \setminus l_t$ のテキストには活性化しないニューロンを検出することである．各テキスト $x_i \in x$ に対して，そのテキストが対象言語である場合（すなわち，$l = l_t$）はラベル $b_i = 1$ を割り当て，そうでない場合は $b_i = 0$ を割り当てる．$N_{l_t}^{+}$ を対象言語 l_t のテキストから成るポジティブ文（すなわち，$b_i = 1$），$N_{l_t}^{-}$ を他の言語のテキストから成るネガティブ文（すなわち，$b_i = 0$）とすると，合計で $N = N_{l_t}^{+} + N_{l_t}^{-}$ となる．例えば，対象言語 l_t がフランス語である場合，フランス語のテキストにはラベル 1 が割り当てられ，英語や中国語など他の言語のテキストにはラベル 0 が割り当てられる．

次に，入力テキストが与えられた際のモデル内の各ニューロンの活性化値を観察する．各ニューロンには一意のインデックス $m \in M$ を割り当てる．$|M|$ はモデル内のニューロンの総数である．テキスト $x_i \in x$ がモデルに入力されたときのニューロン m の出力値を $z_{m,i} \in z_m$ とする．この値の計算方法について詳細を説明する．具体的には，テキスト x_i は T 個のトークンのシーケンス $x_i = \{w_{i,1}, ..., w_{i,t}, ..., w_{i,T}\}$ で構成される．したがって，入力テキストが与えられた場合，デコーダベースの Transformer モデル内には T 個のニューロン出力値 $\{z_{m,i,1}, ..., z_{m,i,j}, ...z_{m,i,T}\}$ が存在する．我々は文中の各トークンにおけるニューロン出力値の平均を取る: $z_{m,i} = f(z_{m,i,1}, ..., z_{m,i,t}, ...z_{m,i,T})$ ここで，f は平均演算子としての集約関数である．元のアプローチ [9] では f を最大プーリング演算子として定義しているが，我々のアプローチでは言語識別の目的でトークン間で一貫して活性化するニューロンを特定するために，f を平均演算子として定義した．[PAD] トークン位置の出力値はノイズとみなされるため，例外として集約から除外した．

最後に，言語固有のニューロンを検出する．データセット $\{x_i, b_i, z_{m,i}\}_{i=1}^{N}$ を予測タスクのサンプルとみなす．具体的には，テキスト $\{x_i\}_{i=1}^{N}$ をモデルの入力，ラベル $\{b_i\}_{i=1}^{N}$ を出力の正解，ニューロンの出力値 $\{z_{m,i}\}_{i=1}^{N}$ を正解の予測スコアとみなす．異なる予測閾値における PR 曲線下の面積である平均適合率（$AP_m = AP(z_m, b) \in [0, 1]$）を用いて，タスクに対するニューロン m の性能を測定する．全てのニューロンに対して AP_m を測定し，降順に並べる．元のアプローチでは，降順で上位 k 個のニューロンのみが対象のニューロンとして定義されている．しかし，これはラベルとの強い正の相関（すなわち，平均適合率が最も高い上位 k 個のニューロン）のみを考慮しており，ラベルとの強い負の相関（すなわち，平均適合率が最も低い下位 k 個のニューロン）は考慮していない．我々は，上位 k 個のニューロンだけでなく，下位 k 個のニューロンも特定の言語と強く関連していると仮定し，元のアプローチを拡張して，上位 k 個と下位 k 個のニューロンの両方を言語固有のニューロンと定義する．実験全体でデフォルト値として $k = 1000$ を設定する．モデルの入力層（単語埋め込み）と出力層（投影層）のニューロンは，これらの層が言語固有のモジュール（言語固有の文字またはサブワード）で構成されていることが明らかであるため，検出対象から除外する．

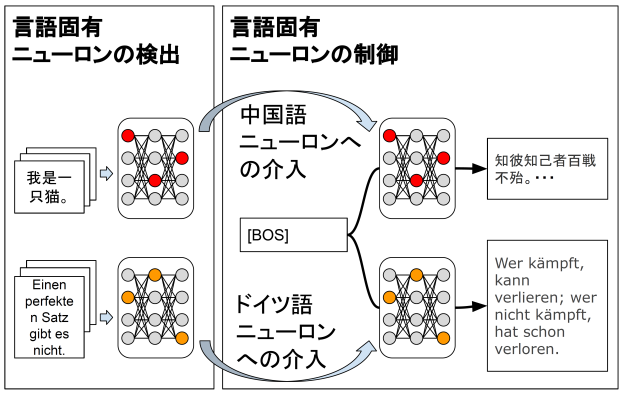

図 1: 提案手法の概要. 特定の言語に対して活性化する傾向がある言語固有ニューロンを検出する．推論時に，検出したニューロンを強制的に活性化する介入を行い，対象の言語文の生起確率を操作する．

3 実験

3.1 モデルとデータセット

分析対象のモデルは XGLM（564M，1.7B，2.9B）[1]，BLOOM（560M，1.7B，3B）[2]，および Llama 2（7B，13B）[3] の 3 つとする．XGLM と BLOOM は，明示的に宣言された多言語モデルである．Llama 2

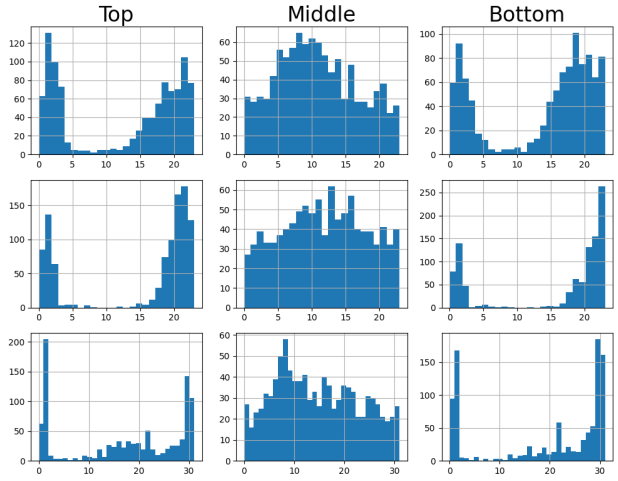

図 2: 平均適合率で高い順に並べたときの上位 1000 個，中位 1000 個，下位 1000 個のニューロンの各層における分布．1 行目：XGLM-564M, de．2 行目：BLOOM-1b7, fr．3 行目：Llama2-13b, zh.

は大半が英語のテキストコーパスで訓練されたモデルあり，他の言語の含有は最小限である（付録 A）．

分析対象の言語は英語（en），ドイツ語（de），フランス語（fr），スペイン語（es），中国語（zh），日本語（ja），の 6 言語とする．言語固有のテキストコーパスは，PAWS-X [10] と FLORES-200 [11] を混合して作成する．PAWS-X は，上記の言語を含む 7 言語における 2 つのテキスト間のパラフレーズ同定用のデータセットである．FLORES-200 は，200 以上の言語に対する機械翻訳タスクのためのデータセットである．これらのデータセットにおけるサンプル文は品質が高く，テキストの種類も多様であり，実験に必要な 6 言語をカバーしているため，これらを混合して使用する．本実験では，各言語のテキストコーパスを作成するために，2 つのデータセットからランダムに 1:1 の比率でテキストをサンプリングする．各言語について 500 件のテキストを準備し，6 つの言語全体で 3000 件のテキストを構成する．

3.2 結果と考察

3.2.1 言語固有ニューロンの検出

図 2 は，2 節の手法を用いて検出されたニューロンの各層における分布を表したヒストグラムである．平均適合率順の上位 1000 個，下位 1000 個のニューロンの大部分がモデルの最初と最後の数層に分布している．対照的に，平均適合率順で中間 1000 個のニューロン（中央値の周辺）は主に中間層に分布している．これらの分布特性は言語，モデルのサイズ，種類を問わず同じであることが示された．

表 1: 言語間で重複する言語固有ニューロンの数（XGLM-564M）.

	de	en	es	fr	ja	zh
de	2000	41	74	39	44	34
en	41	2000	34	41	49	40
es	74	34	2000	57	77	22
fr	39	41	57	2000	21	93
ja	44	49	77	21	2000	27
zh	34	40	22	93	27	2000

さらに，ある言語で発火する言語固有のニューロンは，他の全ての言語との重複が少ないことが確認された．表 1 は，各言語間で重複するニューロンの数を集計した結果だが，言語間のニューロンの重複はどの言語間でも 5 %未満であることがわかる．

図 2 の所見と多言語モデルに関する先行研究 [12] に基づいて，デコーダベースの PLM の内部動作について以下の解釈が可能である．PLM の最初の数層は，各言語の語彙的または構文的表現を言語非依存の意味論的表現に変換する．モデルの中間層は主に言語非依存の意味理解と表現の変換処理である．モデルの最後の数層は，主に意味論的表現を目標言語の構文と語彙情報に逆変換する．

3.2.2 言語固有ニューロンの制御

検出された言語固有のニューロンの有効性を示すために，言語固有のニューロンに介入することでモデルがテキスト生成において言語を制御できるかどうかを調査する．我々は，推論中に上位 1000 個および下位 1000 個のニューロンの出力値を固定値で上書きすることにより，テキスト生成を制御する．具体的には，各ニューロン m に対して，事前に以下のように固定値を計算する: $\bar{z}_m = Median(\{z_m|b = 1\})$. これは，対象言語のテキストに対するニューロン出力の中央値を意味する．推論中には，順伝播において上位 1000 個および下位 1000 個のニューロンの出力をこの固定値で置き換えることにより介入し，モデルの対象言語のテキスト生成確率を観察する．

本実験では，入力に [BOS] トークンのみをモデルに与え，テキストを生成させる．各モデルは，1 から 100 までランダムシードを変更しながら，ランダムサンプリングデコーディングによりテキスト生成を 100 回繰り返す．図 3 は，XGLM-564M モデルを

xglm-564M : en
Some of the issues that we are gonna have here are: the NSA is investigating whether the program is leaking in to the public and the government is trying to stop it as of late as it is possible. In the meantime the NSA is going to run the Panama Papers to find out what the UAE(UAE is the official of the U.E.E.)

xglm-564M : de
Vorträge unter der Überschrift 'War für Trojä und ihr jahrhundert' zu nutzen und abzuschließen.
Urlaub für Menschen mit Schmerzen

xglm-564M : fr
«Il serait dommage de réécrire l'histoire au lieu de donner à entendre qu'une personne est une personne vivant dans l'état dans lequel elle est présente», ajoute le Kentou. «La plupart des médias dans le monde ne donnent pas suffisamment de voix, et qu'un jour il n'y sérieux en ligne seulement sur la.

xglm-564M : es
Chile, Colombia, Paraguay, Uruguay, Bolivia, Chile, Ecuador, Perú, Uruguay, Colombia, Paraguay, Paraguay, Colombia
Utilizamos cookies para asegurar que damos la mejor experiencia al usuario en nuestro sitio web. Si continúa utilizando este sitio asumiremos que está de acuerdo.Estoy de acuerdo

xglm-564M : zh
三是(一)有权与允诺的机关有权予以采纳。
地点:深圳市高新区瑞华路科创新大厦1006室

xglm-564M : ja
ただいま(25日の遅れのため)この商品は、注文確認日の翌営業日に発送致します。
いつものモテテレムはいつもと同じのにしちゃうの。

図 3: XGLM-564M モデルで推論時に各言語固有ニューロンに介入して生成したテキストサンプル.

用いて推論時に各言語固有ニューロンに介入して生成したテキスト文のサンプルである. 各言語固有のニューロンに介入することにより，出力テキストの言語が変更できることが定性的に示されている.

対象言語の生起確率を定量的に測定するために，生成されたテキストの対象言語を FastText の言語識別分類器 [13, 14] によって推定する. 分類スコアが閾値 0.5 を超える場合，各テキストを対象言語に分類することとする [15, 3]. 表 2 は，各言語固有ニューロンへの介入によるテキスト生成における対象言語の生起確率の変化を定量的に測定した結果である. 介入によって，テキスト生成における対象言語の生起確率が増加する傾向があることが示された. また我々は，上位 1000 個のニューロンのみ，

表 2: 生成されたテキストにおける対象言語の生起確率. ”-”行の値は 6 つの言語の平均値.

		介入前	介入後		
			上位	下位	両方
xglm	en	40.0	62.0	77.0	**89.0**
(564M)	de	0.0	89.0	31.0	**95.0**
	fr	0.0	86.0	7.0	**90.0**
	es	2.0	71.0	5.0	**78.0**
	zh	7.0	**82.0**	50.0	79.0
	ja	7.0	92.0	61.0	**99.0**
	-	9.3	80.3	38.5	**88.3**
bloom	en	37.0	78.0	67.0	**88.0**
(1b7)	de	0.0	60.0	0.0	**86.0**
	fr	13.0	80.0	72.0	**98.0**
	es	18.0	44.0	94.0	**97.0**
	zh	6.0	1.0	89.0	**90.0**
	ja	0.0	67.0	35.0	**97.0**
	-	12.3	55.0	59.5	**92.7**
Llama-2	en	83.0	82.0	**89.0**	**89.0**
(7b)	de	0.0	2.0	6.0	**23.0**
	fr	2.0	1.0	**8.0**	7.0
	es	1.0	4.0	4.0	**35.0**
	zh	0.0	2.0	4.0	**50.0**
	ja	1.0	1.0	**12.0**	10.0
	-	14.5	15.3	20.5	**35.7**

下位 1000 個のニューロンのみ，そして上位下位両方のニューロンに介入するという 3 種類の実験を行い，上位下位両方のニューロンへの介入が，どちらか片方への介入よりも所望の言語の生起確率が高くなることを確認した. これは上位下位両方のニューロンが言語の制御に関係があることを示唆している. 上位のニューロンは対象言語に対して正値の発火，下位のニューロンは負値の発火をする傾向がある（付録 B.1）. また介入するニューロンの数を変化させて対象言語の生起確率と生成文の品質を測った結果，k=1000〜10000 が最適であった（付録 B.2）.

4 おわりに

本研究は，PLM 内で各言語固有に発火するニューロンの存在を検証した. 実験により，検出された言語固有ニューロンは言語やモデルの種類に関わらず，主に最初と最後の数層に存在することが示された. また，推論時に言語固有ニューロン出力値を固定の活性値に置き換える介入実験を行った結果，対象言語文の生起確率が増加することを確認した.

謝辞

本研究は，JST さきがけ JPMJPR21C8 の助成を受けたものである．

参考文献

[1]Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. Few-shot learning with multilingual language models. **arXiv preprint arXiv:2112.10668**, 2021.

[2]Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. **arXiv preprint arXiv:2211.05100**, 2022.

[3]Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. **arXiv preprint arXiv:2307.09288**, 2023.

[4]Omer Antverg and Yonatan Belinkov. On the pitfalls of analyzing individual neurons in language models. In **International Conference on Learning Representations**, 2022.

[5]Karolina Stańczak, Edoardo Ponti, Lucas Torroba Hennigen, Ryan Cotterell, and Isabelle Augenstein. Same neurons, different languages: Probing morphosyntax in multilingual pre-trained models. **arXiv preprint arXiv:2205.02023**, 2022.

[6]Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Journey to the center of the knowledge neurons: Discoveries of language-independent knowledge neurons and degenerate knowledge neurons. **arXiv preprint arXiv:2308.13198**, 2023.

[7]Karolina Stańczak, Lucas Torroba Hennigen, Adina Williams, Ryan Cotterell, and Isabelle Augenstein. A latent-variable model for intrinsic probing. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 37, pp. 13591–13599, 2023.

[8]Andrea Gregor de Varda and Marco Marelli. Data-driven cross-lingual syntax: An agreement study with massively multilingual models. **Computational Linguistics**, Vol. 49, No. 2, pp. 261–299, 2023.

[9]Xavier Suau Cuadros, Luca Zappella, and Nicholas Apostoloff. Self-conditioning pre-trained language models. In **International Conference on Machine Learning**, pp. 4455–4473. PMLR, 2022.

[10]Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In **Proceedings of the 2019 Conference on EMLNP and the 9th IJCNLP (EMNLP-IJCNLP)**, pp. 3687–3692, Hong Kong, China, November 2019.

[11]Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. **arXiv preprint arXiv:2207.04672**, 2022.

[12]Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. First align, then predict: Understanding the cross-lingual ability of multilingual BERT. In **Proceedings of the 16th Conference of the EACL: Main Volume**, pp. 2214–2231, Online, April 2021.

[13]Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In **Proceedings of the 15th Conference of the EACL: Volume 2, Short Papers**, pp. 427–431, Valencia, Spain, April 2017.

[14]Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Herve Jegou, and Tomas Mikolov. Fasttext.zip: Compressing text classification models, 2017.

[15]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 **Proceedings of the Twelfth Language Resources and Evaluation Conference**, pp. 4003–4012, Marseille, France, May 2020. European Language Resources Association.

[16]Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. Finding skill neurons in pre-trained transformer-based language models. In **Proceedings of the 2022 Conference on EMLNP**, pp. 11132–11152, Abu Dhabi, United Arab Emirates, December 2022.

A　モデル詳細

表 3 は実験に使用されたモデルの詳細である．すべてのモデルは Huggingface からダウンロードした．

表 3: 実験に用いたモデル一覧．

Model	# Params	# Layers	# Neurons
XGLM	564M	24	221,184
	1.7B	24	442,368
	2.9B	48	884,736
BLOOM	560M	24	221,184
	1.7B	24	442,368
	3B	30	691,200
Llama2	7B	32	1,359,872
	13B	40	2,129,920

表 4 は，各モデルの事前学習データセットにおける言語の分布である[1)]．

表 4: 各モデルの事前学習データ内の言語分布．

	en	de	fr	es	zh	ja
XGLM	49.0	5.4	4.7	5.3	8.1	4.0
BLOOM	30.0	-	12.9	10.8	16.2	-
Llama2	89.7	0.2	0.2	0.1	0.1	0.1

B　追加実験

B.1　言語固有ニューロンの活性値の傾向

原則として，平均適合率における上位 1000 個のニューロンは正の活性化値と相関している．対照的に，下位 1000 個のニューロンは負の活性化値と相関している．図 4 はこの主張を検証した結果である．活性値と正の相関があるニューロンだけでなく負の相関を持つニューロンも，言語固有ニューロンとして重要であることを示唆している [16]．

B.2　介入するニューロン数の変更

介入するニューロン数を変更するアブレーション実験を行い，対象言語の生起確率に対する効果を分析した．また，モデルによって生成されたテキストの品質を BLEU-4 スコアを用いて検証した．言語識別器によって対象言語であると識別されたテキストのみについて品質を評価した．具体的には，対象言語として識別された生成テキストそれぞれについて，そのテキストを仮説文とし，すべてのポジティブ文を参照文として，BLEU スコアを算出し，平均を取った．テキストを各モデルのトークナイザーでトークン化した後に BLEU スコアを測定した．BLEU スコアの測定には NLTK ライブラリを使用した．図 5 に示されるように，介入するニューロンの数 k を約 1000〜10000（この図では 10 の対数で 3〜4）まで増やすと，一般的に対象言語の生起確率が増加するが，それを超えて増加させるとテキストの品質が低下する．最終的には文が崩壊し，言語識別と品質が著しく低下した．

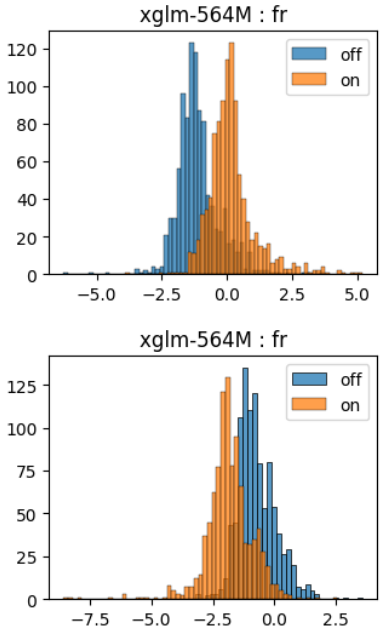

図 4: 【上】平均適合率上位 1000 ニューロンにおける対象言語（on）とその他言語（off）の活性値の分布．【下】下位 1000 ニューロンの活性化値の分布．

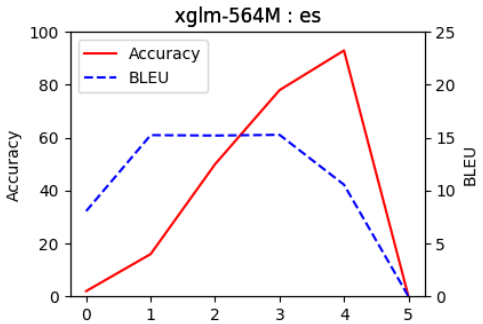

図 5: テキスト生成時に介入するニューロン数の変更実験．x 軸：$\log_{10}$(介入ニューロン数 k)．

1) XGLM の情報は https://huggingface.co/facebook/xglm-2.9B から引用されている．BLOOM の情報は https://huggingface.co/bigscience/bloom#languages から引用されている．