言語ニューロン介入は非目標言語の出力を抑制できるか？

謝 素春 [1]　金 輝燦 [1]　佐々木 翔大 [1]　山田 康輔 [3]　鈴木 潤 [1,2,4]
[1] 東北大学　[2] 理化学研究所
[3] 株式会社サイバーエージェント　[4] 国立情報学研究所 LLMC
xie.suchun.p7@dc.tohoku.ac.jp　is-failab-research@grp.tohoku.ac.jp

概要

大規模言語モデルにおいて，多言語対話時に意図しない言語で回答してしまう「非目標言語出力」は重要な課題である．本研究では，目標言語を明示的に指定するという実用的なシナリオで言語ニューロン介入手法の有効性を評価する．実験の結果，言語ニューロン介入による非目標言語出力の削減幅は限定的であり，モデルやタスク間で効果が不安定であることが明らかとなった．また，言語ニューロン介入は（1）言語属性のみならず生成内容の意味にも干渉し，情報の歪みを引き起こすこと，（2）トークンの生成確率の上昇幅は，最高確率トークンを逆転するには不十分であることを示した．

1　はじめに

昨今，大規模言語モデル (LLM) は多様な生成タスクにおいて顕著な性能向上を遂げている．しかし，Llama 3 [1] や GPT-4o [2] といった高性能なモデルにおいても，中国語や日本語などの英語以外の言語において，意図した目標言語で生成できないことがあるという課題が報告されている [3, 4, 5, 6, 7, 8]．この「非目標言語出力問題」は，LLM の多言語応用における大きな障壁となっている．

近年，言語モデルの内部に言語特有のニューロンが存在し，これらを操作することで出力言語を誘導できる可能性が示唆されている [9, 10]．これに基づき，追加学習を必要とせずに対象言語のニューロンを活性化させる言語ニューロン介入手法が提案された．既存研究では [9, 10]，事前学習済みモデルや，出力言語が明示されない設定において一定の制御効果が確認されている．しかし，実社会における LLM の利用形態は，指示調整済みモデルに対しプロンプトで直接言語を指定する運用が一般的である．このような実用的な条件での言語ニューロン介入の有効性に関する調査は未だ限定的である．

そこで本研究では，実用的な多言語生成における課題に焦点を当て，目標言語が明示的に指定された条件でも言語ニューロン介入が非目標言語出力を効果的に削減できるかを調査する．多様なモデルとタスクを用いた包括的な評価の結果，言語ニューロン介入による改善幅は限定的であり，その効果も不安定であることを示す．さらに本研究では，言語ニューロン介入が失敗するメカニズムを解明するために，（1）言語制御と生成内容の分析，および（2）トークン確率の推移を追跡する LogitFlow 分析を行う．分析の結果，言語ニューロン介入が内容の意味的な歪みを引き起こしていること，およびトークン生成確率の上昇が既存の最高確率トークンを逆転するには不十分であることを明らかにする．

この知見は，言語ニューロン介入手法の限界を浮き彫りにし，より効果的な言語制御手法の構築に向けた洞察を提供するものである．

2　言語ニューロン介入手法

言語ニューロン特定の手法としては主に 2 つの代表的なアプローチが存在するが [9, 10]，本研究では予備実験（付録 B）において優れた効果を示した Kojima ら [9] による平均適合率 (Average Precision; AP) に基づく手法を採用する．本手法は以下の 4 つのステップから構成される．

ステップ 1：コーパスのラベル付け　多言語コーパスに対し，目標言語のテキストを正例（1），それ以外の言語のテキストを負例（0）としてラベル付けを行う．

ステップ 2：活性化値の抽出　各テキストについて，中間層内の全ニューロンの活性化値を収集する．これらの活性化値を非パディングトークンにわたって平均化し，各テキストにおける各ニューロンの活性化値（スカラー値）を得る．

ステップ 3：スコアリングと選択　各ニューロンを目標言語の二値分類器と見なし，各テキストで

の平均活性化値を用いて，言語ラベルに対する AP スコアを算出する．算出した AP スコアに基づきニューロンを順位付けし，上位 k 個および下位 k 個のニューロンを選択する．

ステップ 4：言語ニューロン介入　推論時，選択されたニューロンの活性化値を，目標言語のテキストから算出された中央値へと置き換える．

3　関連研究

非目標言語出力問題　LLM が意図した目標言語で一貫してテキストを生成できない課題は，「オフターゲット言語生成（off-target language generation）」あるいは「言語混同（language confusion）」として広く報告されている [3, 4, 5, 6, 7, 8]．特に Marchisio ら [7] は，LLM における言語混同を体系的に評価し，英語中心の指示調整がモデルの英語への偏好を増幅させ，他言語のプロンプトに対しても英語で回答を生成する要因となることを示した．

こうした言語混同を緩和するための手法として，few-shot の例示や多言語 SFT の適用が有効であることが示されている．また，推論時に直接言語ベクトルを操作する手法 [11] や，言語特有のニューロンを操作する手法 [10, 12, 9, 13] といった，モデルの出力を制御するアプローチも導入されている．

言語ニューロン介入手法　言語特有のニューロンに関する研究は，Mahowald ら [14] や Zhang ら [15] の知見を基礎としている．Mahowald ら [14] は，LLM が文法などの形式的な言語処理には優れる一方で，機能的な言語使用において不安定さを示すことを発見し，言語処理能力と認知能力の乖離を示唆した．また，Zhang ら [15] は，モデルパラメータのわずか 1%が多言語性能において決定的な役割を果たしており，このサブセットへの摂動が多言語性能の著しい低下を招くことを明らかにした．これらの研究に基づき，近年では言語特有ニューロンの特定とその介入効果を検証するアプローチが数多く提案されており [10, 12, 9, 13]，多言語処理におけるニューロンの重要性が実証されている．しかし，既存研究の多くは事前学習済みモデルを用いた暗示的な言語指定条件での評価 [9] や，小規模な事例研究 [10] に留まっている．実世界での利用シナリオにより近い「目標言語が明示的に指定された設定」における言語特有ニューロンの有効性については，未だ体系的な評価がなされていない．この実用上の空白を埋めることが，本研究の目的である．

4　非目標言語出力の評価実験

本節では，非目標言語出力の緩和における言語ニューロン介入の潜在能力を調査する．主な目的は，(1) 非目標言語出力の割合の変化，(2) タスク性能への影響 を測定することにある．

4.1　実験設定

言語ニューロン介入の実装　Kojima ら [9] の手法に従い，複数の言語（英語，フランス語，ドイツ語，スペイン語，中国語，日本語, ヒンディー語）から各 500 サンプルを抽出した均衡データセットを用いて言語ニューロンを特定する．データは PAWS-X [16] および FLORES-200 [17] から均等に収集した．ニューロン選択におけるハイパーパラメータ k については，$k = 1000$ に設定した．

モデル　言語ニューロン介入の効果を異なるモデルファミリーおよびスケールにわたって評価するため，英語中心の LLM と多言語 LLM の両方を用いて実験を行う．英語中心のモデルとして Llama2-Chat (7B, 13B) [18] および Llama3-8B Instruct [1] を，多言語モデルとして Bloomz-7b1-p3 [19] を使用する．これらすべてのモデルは，生成タスクにおいて指示調整済み（instruction-tuned）モデルである．

データセット　先行研究 [7] に基づき，非目標言語出力が顕著な日本語および中国語を中心に，フランス語，スペイン語およびヒンディー語の計 5 つの言語を対象とする．

本研究では生成タスクに焦点を当て，新聞要約データセットである **XL-Sum** [20]，および多様なタスクを含む **Dolly** [21]（主に QA サブセット）を用いる．データセットの詳細は付録 A に記す．

評価指標　以下の 2 つの指標を用いて評価する．

- **NT Ratio**: 非目標言語で生成された出力の割合．fastText [22, 23][1] を用いて，閾値 0.5 で検出する．
- **Task Performance**: テキスト生成の質．ROUGE-L [24] によって測定する．

4.2　実験結果

表 1 に実験結果を示す．要約すると，言語ニューロン介入は非目標言語出力を抑制する潜在能力を示すものの，全条件での平均削減率は 3.0%になり，性能向上も 1%以下であった．

1）https://github.com/facebookresearch/fastText

表 1　Dolly および XL-Sum データセットにおける実験結果．NT Ratio：非ターゲット言語出力率（%，低いほど良い）．RougeL：ROUGE-L F1 スコア（高いほど良い）．変化量 (Δ) は (After - Before) として算出している．NT Ratio において，負の Δ は改善（非ターゲット出力の減少，緑色）を示し，RougeL において，正の Δ は改善（スコアの上昇，緑色）を示す．色の濃淡は変化の大きさを表す（緑，赤）．

	Dolly						XL-Sum					
	NT Ratio (%)			RougeL			NT Ratio (%)			RougeL		
Model	Before	After	Δ	Before	After	Δ	Before	After	Δ	Before	After	Δ
日本語												
Llama2-Chat 7B	28	22	-6	0.09	0.10	+0.01	100	98	-2	0.01	0.01	0.00
Llama2-Chat 13B	19	16	-3	0.10	0.11	+0.01	4	10	+6	0.17	0.16	-0.01
Llama-3-8B-Instruct	**7**	2	-5	**0.14**	0.14	0.00	**3**	2	-1	**0.20**	0.20	0.00
BLOOMZ-7B1-P3	72	25	-47	0.02	0.05	+0.03	85	46	-39	0.02	0.08	+0.06
中国語												
Llama2-Chat 7B	43	30	-13	0.13	0.15	+0.02	100	99	-1	0.01	0.02	+0.01
Llama2-Chat 13B	44	36	-8	0.13	0.15	+0.02	18	31	+13	0.18	0.16	-0.02
Llama-3-8B-Instruct	14	7	-7	**0.19**	0.21	+0.02	**3**	8	+5	**0.24**	0.23	-0.01
BLOOMZ-7B1-P3	**6**	10	+4	0.12	0.11	-0.01	84	83	-1	0.03	0.03	0.00
スペイン語												
Llama2-Chat 7B	1	4	+3	0.20	0.20	0.00	56	51	-5	0.09	0.10	+0.01
Llama2-Chat 13B	0	1	+1	0.20	0.20	0	6	5	0.00	0.15	0.16	+0.01
Llama-3-8B-Instruct	0	0	0	0.20	0.20	0.00	0	0	0	0.18	0.18	0.00
BLOOMZ-7B1-P3	10	6	-4	0.10	0.12	+0.02	10	16	+6	0.17	0.16	-0.01
フランス語												
Llama2-Chat 7B	12	10	-2	0.19	0.18	0.00	70	69	-1	0.09	0.09	0.00
Llama2-Chat 13B	3	4	+1	0.20	0.19	0	7	6	0.00	0.17	0.18	+0.01
Llama-3-8B-Instruct	1	0	-1	0.20	0.20	0.00	0	0	0	0.21	0.20	-0.01
BLOOMZ-7B1-P3	9	5	-4	0.11	0.12	+0.01	25	11	-14	0.17	0.18	+0.01
ヒンディー語												
Llama2-Chat 7B	28	24	-4	0.10	0.10	0.00	100	100	0	0.00	0.00	0.00
Llama2-Chat 13B	13	10	-3	0.11	0.12	+0.01	96	97	+1	0.01	0.01	0.00
Llama-3-8B-Instruct	1	0	-1	0.18	0.17	-0.01	0	0	0	0.21	0.20	-0.01
BLOOMZ-7B1-P3	39	45	+6	0.04	0.03	0.00	57	63	+6	0.07	0.05	-0.02

モデルタイプによる影響　ニューロン制御の効果はモデルの種類に強く依存する．日本語タスクにおいて，多言語モデルの Bloomz は大幅な NT Ratio の削減（Dolly で 47%減）を達成したが，Llama シリーズでは 1〜6%の微増減に留まり，大きな格差が見られた．

タスクタイプによる影響　言語ニューロン介入の効果には強いタスク依存性が観察された．Dolly では多くのモデルで改善が見られた一方で，XL-Sum では半数以上の設定で性能の劣化が確認された．

5　分析

言語ニューロン介入の限界を詳細に把握するため，介入の失敗に対して分析を行う．

5.1　言語と内容の相互作用

言語ニューロン介入が非目標言語出力を削減しても，タスク性能が向上せず，むしろ低下する事例が確認された（例：Dolly タスクにおける Llama3-8B）．もし，介入が言語選択のみを制御しているのであれば，出力言語を参考ラベルと一致させることは ROUGE スコアの向上に繋がるはずである．性能の低下は 介入が内容生成を妨害し，情報の損失や内容の歪み，あるいは曖昧さを引き起こしている可能性を示唆している．これは，介入が単なる言語スイッチとして機能するだけでなく，より深い内容生成メカニズムとも相互作用していることを示している．

言語と内容の両面における一貫性への影響を調査するため，介入前後のモデル出力を比較するサン

表 2　Dolly タスク（日本語）における 介入効果のサンプル別分析．言語ニューロン介入は出力言語（Correct, Wrong, Unchanged）と内容の意味（Positive, Negative, Neutral）の両方に変化をもたらしている．

Model	Language (%)			Content (%)		
	Corr.	Wro.	Unch.	Pos.	Neg.	Neutral
Llama2-7B	20	16	64	31	7	62
Llama3-8B	10	6	84	19	18	63
Bloomz-7B	58	5	37	22	11	67

プルごとの分析を行う．言語の一貫性に関しては，fastText を用いて言語の変化を検出し，それらを「正解（correct）」および「不正解（incorrect）」に分類する．内容の一貫性に関しては，介入前後の参照ラベルラベルに対する BLEURT [25] スコア[2)]を算出し，その差分 (Δ = after – before) を計算する．この差分を用いて，各サンプルをポジティブ（Positive），ネガティブ（Negative），またはニュートラル（Neutral）のいずれかに分類し，各カテゴリの分布を報告する．

分析の結果（表 2），言語ニューロンの介入は言語制御の成否にかかわらず，出力内容の意味に一貫して変化を与えていることが判明した．このことは，言語特有のニューロンと内容生成に関わるニューロンが密接に結合していることを裏付けている．

5.2　ニューロン介入のメカニズム理解

実験の結果，言語ニューロン介入による非目標言語出力の削減効果は限定的であり，大半のサンプルで言語の切り替えに失敗した．本節では，介入後も元の言語が保持される要因を解明するため，言語ニューロン介入のメカニズムを詳細に分析する．

分析設定　介入前後における目標言語トークンの生成確率の推移を定量化するため，LogitFlow 分析を導入する．本分析手法では，(1) top-k Logit 内の目標トークンを出現頻度（x 軸），および (2) 目標トークンの最高順位（y 軸）という 2 つの指標に基づき，各トークンの遷移を 2 次元ベクトルで表現する．これらの推移を，上昇（Upward），下落（Downward），混合（Mixed）の 3 パターンに分類し可視化を行う．

分析結果　図 1 に示す通り，言語ニューロン介入は多くの場合で目標言語トークンの確率を上昇させていることが明らかとなった．しかし，これらの改善は競合するトークンを追い越して top-1 に到達するには不十分なケースが大半であった．

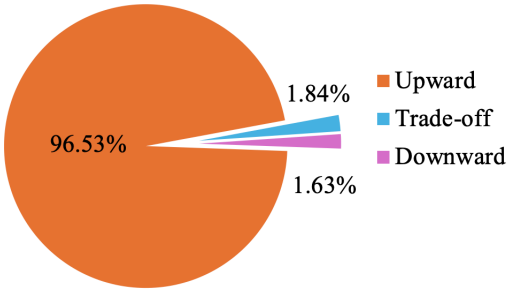

図 1　Llama2-7B を用いた XL-Sum（日本語）における LNI 介入後のターゲット言語トークンの Logit 変化．上昇（Upward）が支配的である一方、トレードオフや下落（Downward）は稀である．

本実験では，高いサンプリング温度が言語混同を悪化させるという知見 [7] に基づき，モデルが常に最高確率のトークンを選択する決定論的生成（temperature = 0.0）を用いている．この設定では，最高確率のトークンのみが生成されるため，言語ニューロン介入が目標トークンの確率を上昇させたとしても，それが最高確率のトークンにならない限り，言語の切り替えは実現されず効果を発揮しないことを意味している．

この知見は，言語ニューロン介入単体による制御の限界を示すとともに，適切なデコーディング戦略，例えば，温度スケーリングやビームサーチと組み合わせることで，言語制御能力を強化できる可能性を示唆している．以上の結果は，言語ニューロン介入 単体では不十分であることを示しており，多言語生成におけるより堅牢な制御メカニズムの必要性を示している．

6　おわりに

本研究では，より実用的な設定における言語ニューロン介入の潜在能力と限界を調査した．実験の結果，言語ニューロン介入による非目標言語出力の抑制効果は限定的であり，モデルやタスク間で不安定であることを示した．また，言語と内容の生成メカニズムの結合，および決定論的生成におけるトークン確率の上昇幅の不足が，言語ニューロン介入の主要な限界要因であることを明らかにした．これらの一連の知見は，LLM の多言語制御における課題を浮き彫りにするものであり，今後はより精緻な制御手法の確立が重要な課題であることを示している．

2)　言語が異なる場合でも，テキスト間の意味的類似度を測定可能な指標．

謝辞

本研究は，JSPS 科研費 JP24H00727，JST ムーンショット型研究開発事業 JPMJMS2011-35（fundamental research），JST 次世代研究者挑戦的研究プログラム JPMJSP2114 の支援を受けたものです．また，本研究の実施にあたっては，産総研および AIST Solutions が提供する ABCI 2.0，ならびにデータ活用社会創成プラットフォーム mdx を利用しました．最後に，有益な助言を下さった東北大学 Tohoku NLP Group の皆様に感謝致します．

参考文献

[1]Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and et al. The llama 3 herd of models, 2024.

[2]OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, and et al. Gpt-4 technical report, 2024.

[3]Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, and Alham Fikri Aji. Multilingual large language models are not (yet) code-switchers. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, 2023.

[4]Nadezhda Chirkova and Vassilina Nikoulina. Zero-shot cross-lingual transfer in instruction tuning of large language models. In **Proceedings of the 17th International Natural Language Generation Conference**, 2024.

[5]Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. Monolingual or multilingual instruction tuning: Which makes a better alpaca. In **Findings of the Association for Computational Linguistics: EACL 2024**, 2024.

[6]Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. Llama beyond english: An empirical study on language capability transfer, 2024.

[7]Kelly Marchisio, Wei-Yin Ko, Alexandre Berard, Théo Dehaze, and Sebastian Ruder. Understanding and mitigating language confusion in LLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, November 2024.

[8]Weichuan Wang, Zhaoyi Li, Defu Lian, Chen Ma, Linqi Song, and Ying Wei. Mitigating the language mismatch and repetition issues in LLM-based machine translation via model editing. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, 2024.

[9]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 **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, 2024.

[10]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 Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics**, 2024.

[11]Xie Yunfan, Lixin Zou, Dan Luo, Min Tang, Chenliang Li, Xiangyang Luo, and Liming Dong. Mitigating language confusion through inference-time intervention. In **Proceedings of the 31st International Conference on Computational Linguistics**, 2025.

[12]Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. How do large language models handle multilingualism?, 2024.

[13]Shaomu Tan, Di Wu, and Christof Monz. Neuron specialization: Leveraging intrinsic task modularity for multilingual machine translation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, 2024.

[14]Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models, 2024.

[15]Zhihao Zhang, Jun Zhao, Qi Zhang, Tao Gui, and Xuanjing Huang. Unveiling linguistic regions in large language models. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics**, 2024.

[16]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 Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, 2019.

[17]NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, and et al. No language left behind: Scaling human-centered machine translation, 2022.

[18]Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, and et al. Llama 2: Open foundation and fine-tuned chat models, 2023.

[19]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, 2022.

[20]Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**, 2021.

[21]Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023.

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

[23]Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics**, 2017.

[24]Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, 2004.

[25]Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, 2020.

[26]Qiyuan Chen Ziang Leng and Cheng Li. Luotuo: An instruction-following chinese language model, lora tuning on llama. https://github.com/LC1332/Luotuo-Chinese-LLM, 2023.

[27]Stephen H. Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, and et al. Promptsource: An integrated development environment and repository for natural language prompts, 2022.

A　データセットと生成条件

A.1　データセット

XL-Sum　新聞要約タスクである XL-Sum は，与えられた記事の要旨を生成するタスクである．本実験では，テストデータから各言語につき，888 件のサンプルを無作為に抽出して使用した．

databricks-dolly-15k　databricks-dolly-15k は，Wikipedia をベースとした人手作成によるデータセットであり，閉じた QA（Closed QA），自由記述 QA（Open QA），要約などの多様な生成タスクを含んでいる．中国語および日本語については，オリジナル版が存在しないため，機械翻訳版である chinese-dolly-15k [26][3] および databricks-dolly-15k-ja[4] を使用した．

本研究では主に質問生成タスクを対象とし，シード値 42 で，`open_qa` および `general_qa` の各カテゴリから 200 サンプルずつを無作為に抽出した．なお，chinese-dolly-15k においてリファレンスの回答が英語のみであったため，GPT-4o mini（ChatGPT API）を用いて中国語へ翻訳した上で実験に供した．

A.2　生成条件

Marchisio ら [7] の研究では，temperature が高いほど言語の混同が悪化することが報告されている．これに基づき，本実験では temperature を 0.0，Top-p を 0.9 に設定した．また，Max new tokens は 100 とした．

プロンプトは英語に設定し，その中で出力言語を明示的に指定した．XL-Sum タスクについては，PromptSource [27] が提供するテンプレートを基にプロンプトを作成した．各データセットにおけるプロンプトの詳細は，表 3 に示す通りである．

表 3　プロンプト設定．

Task	Prompts
XL-Sum	Write one sentence to summarize the given document. The document is: {*Input*} Summarize in language:
Dolly	Answer the following question in language.{*Input*}

B　予備実験

B.1　異なる言語ニューロンの介入手法

言語ニューロンの特定手法として，主に 2 つの代表的なアプローチが提案されている [9, 10]．両者はともに言語特有性の高いニューロンを対象とする点は共通しているが，選択の細部において異なる．これら 2 つの手法を同一条件で比較した先行研究は存在しないため，本研究のタスクに最適な手法を決定するための予備実験を行った．

実験設定（生成時の推論パラメータおよびニューロン制御の構成）は本実験と同一とし，Llama2-Chat 7B を用いて XL-Sum タスク（中国語・日本語）で評価した．表 4 に示す通り，実験の結果，Kojima ら [9] による AP ベースの手法が他方を上回る性能を示した．したがって，本研究では同手法を採用する．

表 4　Llama2-Chat 7B を用いた日本語（JA）および中国語（ZH）データセットにおけるニューロン選択手法の比較結果．

Method	Lang. Change		Content Change	
	Correct	Wrong	Positive	Negative
JA				
AP	1.6	0.0	3.0	3.2
LAPE	0.1	0.2	2.1	2.8
ZH				
AP	1.0	0.0	6.5	4.3
LAPE	0.0	0.2	2.9	3.4

3）　silk-road/chinese-dolly-15k
4）　llm-jp/databricks-dolly-15k-ja