疎な言語固有の次元の同定およびそれを用いた大規模言語モデルの多言語生成制御

鐘 承志 [1]　程 飛 [1]　劉 倩瑩 [2]　村脇有吾 [1]　Chenhui Chu[1]　黒橋禎夫 [1,2]
[1] 京都大学大学院情報学研究科　[2] 国立情報学研究所
zhong@nlp.ist.i.kyoto-u.ac.jp
{feicheng, murawaki, chu, kuro} @i.kyoto-u.ac.jp　ying@nii.ac.jp

概要

英語中心に事前学習された大規模言語モデルは，限られた非英語データにもかかわらず高い多言語能力を示す．先行研究では，Transformer の中間層で英語の表現が形成され，最終層で目的言語へ投影される段階的な表現遷移が報告されている．本研究では，この言語間遷移が少数かつ疎な言語固有の次元によって制御されるという仮説を立てる．この仮説に基づき，わずか 50 文の単言語または対訳文のみを用いて言語固有の次元を同定し，推論時にそれらの次元のみを操作して言語を制御する学習不要な手法を提案する．多言語生成制御タスクにおける実験により，提案手法は意味を保ったまま出力言語を切り替え，既存のニューロン操作手法による言語生成制御を一貫して上回る性能を示すことを確認した．

1　はじめに

大規模言語モデル（LLM）は，主として英語中心のコーパスで事前学習されているにもかかわらず，翻訳，質問応答など多様な多言語タスクにおいて高い性能を示している．このような多言語能力がモデル内部でどのように実現されているかを理解することは，解釈可能性の向上のみならず，効率的な言語制御手法の設計においても重要な課題である．

Wendler ら [1] は，英語中心の LLM において，入力文の意味表現が Transformer の層を進むにつれて段階的に変化することを示している．特に，中間層では意味内容が英語の表現空間に表され，最終層に近づくにつれて目的言語のトークン空間へと投影される現象が報告されている．この潜在言語遷移は logit lens[2] による解析からも確認されており，図 1 に示すように，“English: Today is hot. – 日本語: 今日は__” といった入力では，中間層で英語の表現が顕在化し，最終層で日本語トークンが生成される．

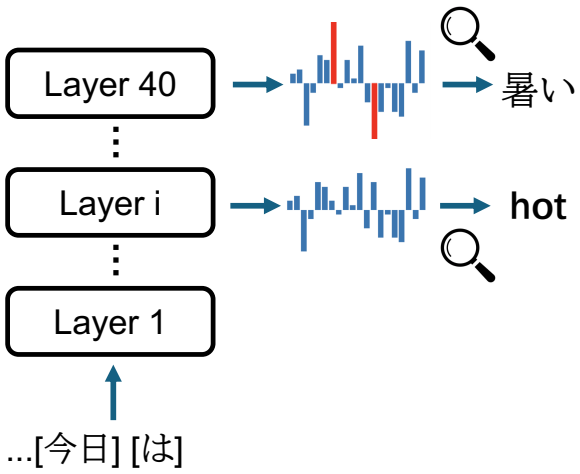


図 1: 言語固有の次元に関する仮説

これまでの研究の多くは，言語選択を司るニューロンの存在に着目し，特定のニューロンや FFN ユニット [3, 4]，あるいはスパースオートエンコーダを用いた特徴抽出により [5]，言語制御を実現してきた．しかし，これらの手法は，大規模なコーパスや計算資源を必要とすることが多い．また，表現が層をまたいでどのように変化するかという表現遷移の構造そのものには十分に踏み込めていない．

本研究では，多言語表現の制御を表現空間上の次元レベルで捉える．言語間の表現遷移が一貫したインデックスを持つ少数かつ疎なの次元に集中して生じているという仮説を立てる．図 1 に示すように，意味内容は主に言語非依存的な次元に保持され，特定の言語固有の次元の変化によって最終的な出力言語が決定されると考えられる．

この仮説に基づき，本研究では，ごく少量（50 文）のデータから言語固有の次元を同定し，推論時にそれらの次元のみを操作して言語を制御する手法を提

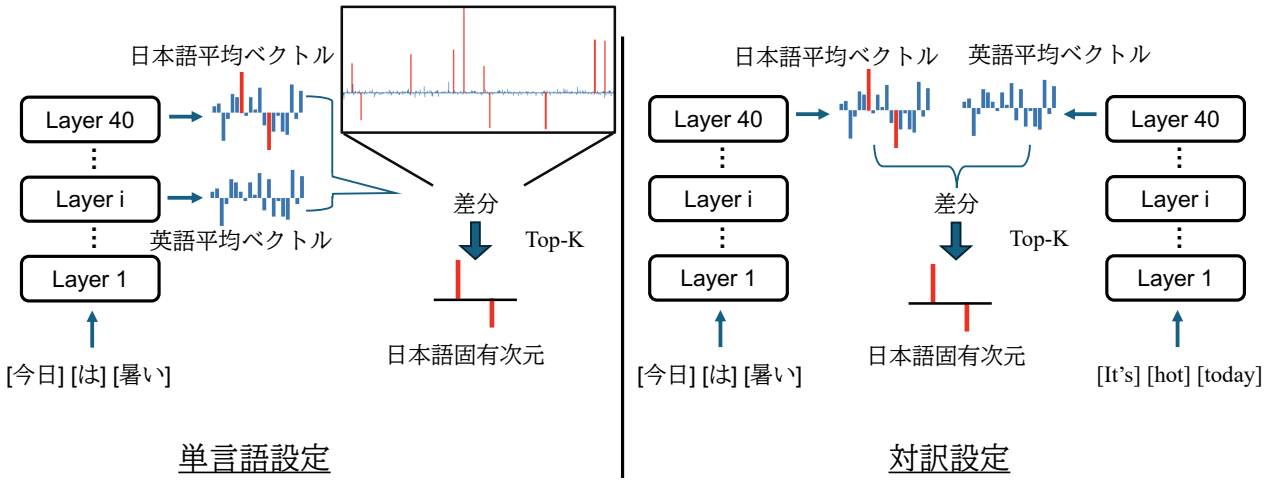


図 2: 言語固有の次元を同定する 2 つの設定

案する. 提案手法は，大規模なコーパス作成やモデルの学習を必要とせず，多言語生成制御タスクにおいて，意味内容を保ったままで出力言語を切り替えられることを実験的に確認した．これは，言語固有の次元が LLM における言語制御の中核的な役割を果たしていることを示唆している.

2 関連研究

多言語処理は，mBERT[6] や mBART[7] など初期の Transformer モデル以降，事前学習の拡張によって大きく発展してきた. 近年の LLM は英語中心で学習されでも，多数の言語に対して汎化性能を示すものの，英語との差は依然として残ることが指摘されている。この課題に対し，対訳コーパスを用いて英語表現空間から他言語空間への投影行列を学習し，推論時に写像を適用することで多言語能力を向上させる手法が提案されている [8, 9]。これらの手法は追加学習を必要とし，表現空間間の写像に依存して言語切替を実現する点に特徴がある。

本研究は，表現空間上の次元レベルで捉え，層を通じて一貫して現れる言語固有の次元に着目する点で，既存研究とは異なる視点を提供する.

3 提案手法

言語固有の次元を同定し，推論時にそれらの次元のみを操作することで出力言語を制御する手法を提案する. 提案手法は，単言語データのみを用いる設定と，対訳データを用いる設定の両方に対応しており，全体の流れを図 2 に示す.

3.1 言語固有の次元の同定

言語固有の次元は，同一意味内容に対応する表現の平均ベクトルを比較することで同定する．入力文に対し，中間層 ℓ と最終層 L の文レベル平均表現を

$$\boldsymbol{\mu}_{\ell}^{(\mathrm{lang})},\ \boldsymbol{\mu}_{L}^{(\mathrm{lang})} \in \mathbb{R}^{d} \tag{1}$$

と定義する．単言語設定では単言語の層間，対訳設定では英語対訳文との最終層の間で次元ごとの差分の絶対値を次式で計算する：

$$\begin{cases} \boldsymbol{\delta}_{monoligual}^{(\mathrm{lang})} = \left|\boldsymbol{\mu}_{L}^{(\mathrm{lang})} - \boldsymbol{\mu}_{\ell}^{(\mathrm{lang})}\right| \\ \boldsymbol{\delta}_{parallel}^{(\mathrm{lang})} = \left|\boldsymbol{\mu}_{L}^{(\mathrm{lang})} - \boldsymbol{\mu}_{L}^{(\mathrm{en})}\right| \end{cases} \tag{2}$$

差分の絶対値が大きい Top-K 次元を言語固有の次元とし，そのインデックス集合を

$$\mathcal{I}_{K}^{(\mathrm{lang})} = \mathrm{TopK}\left(\boldsymbol{\delta}^{(\mathrm{lang})}, K\right) \tag{3}$$

と定義する．この集合 $\mathcal{I}_{K}^{(\mathrm{lang})}$ は後述の言語制御のための操作に用いられる.

3.2 推論時操作

同定された言語固有の次元を用い，推論時に表現を部分的に操作する. 英語入力に対して推論を行い，中間層 j における表現 $\mathbf{h}_j \in \mathbb{R}^d$ に対し，言語固有の次元集合 $\mathcal{I}_{K}^{(\mathrm{lang})}$ のみを上書きする.

具体的には，目的言語に対応する最終層の平均表現 $\boldsymbol{\mu}_{L}^{(\mathrm{lang})} \in \mathbb{R}^d$ を用い，次式により操作を行う：

$$\mathbf{h}'_j[i] = \begin{cases} \alpha\, \boldsymbol{\mu}_{L}^{(\mathrm{lang})}[i], & i \in \mathcal{I}_{K}^{(\mathrm{lang})}, \\ \mathbf{h}_j[i], & \text{otherwise.} \end{cases} \tag{4}$$

ここで α は操作強度を制御するスケーリング係数である．この操作により，意味内容を概ね保持したまま，出力言語のみを切り替えることが可能となる．

4 実験設定

4.1 モデル

Llama2–7B，Llama2–13B，Aya23–8B[10, 11] の 4 つの LLM を評価対象とする．Llama2 系列は主に英語中心のデータで学習されている一方，Aya23 はより多言語性の高いモデルであり，異なる学習データ分布を持つモデル間での一般性を検証する．

4.2 次元同定に用いるデータ

英語からフランス語，ドイツ語，スペイン語，中国語，日本語への 5 言語方向を対象とする．言語固有の次元の同定には，FLORES-200[12] の開発セットから各方向につき 50 文の対訳文対をランダムに抽出して用いる．

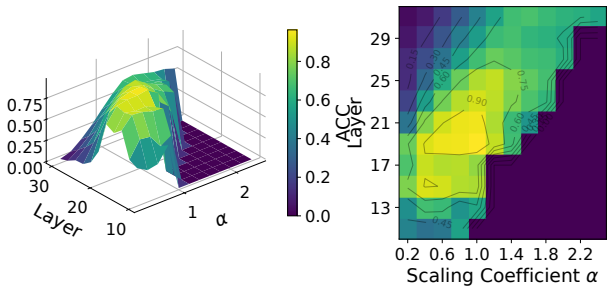
4.3 多言語生成制御タスク

既存研究 [3] に従い，多言語生成制御タスクを用いて評価を行う．モデルには “Translate an English sentence into a target language. English: {source} Target language:” という翻訳指示形式のプロンプトを与えるが，目的言語は明示的に指定しない．この設定により，同定した言語固有の次元が出力言語の制御に寄与するかを検証する．

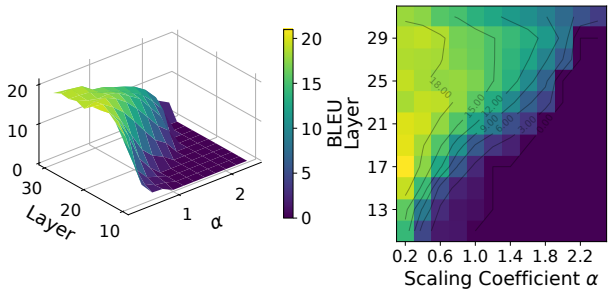
評価には，FLORES-200[12]，IWSLT2017[13]，WMT[14] の各データセットから英語→各言語方向につき 100 文対を用い，平均性能を報告する．指標として，出力言語が目的言語となる割合を測る Accuracy（ACC），目的言語となっている成功例に対する翻訳品質を測る BLEU，およびその積である ACC×BLEU を用いる．出力言語の判定には fastText[15] による言語識別器を用い，識別スコアが 0.5[3] を超える場合を成功と判定する．

5 結果

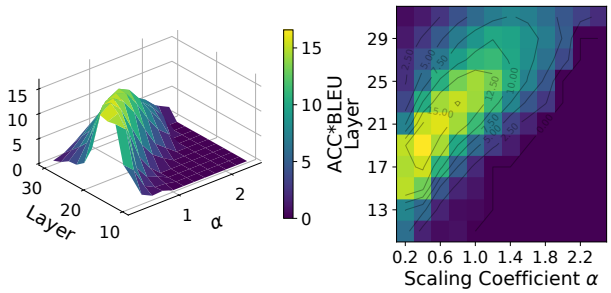
最適なハイパーパラメータをチューニングするため，一連の検証実験を行った．詳細は付録 A に示す．また，同定した言語固有の次元の分析を付録 B に示す．



(a) Overall ACC



(b) Overall BLEU



(c) Overall ACC*BLEU

図 3: 単言語設定における Llama2-7B の結果

5.1 言語固有の次元操作の検証結果

本節では，言語固有の次元が層をまたいで一貫して機能するかを検証するため，各中間層に対して操作を適用し，スケーリング係数 α を変化させた検証を行った．図 3 は，Llama2–7B の単言語設定における結果を示す．

図より，同一の言語固有の次元インデックス集合に対して，適切な α を選択することで，多くの中間層において出力言語を安定して制御できることが分かる．このことは，言語固有の次元が特定の層に局在するのではなく，層を通じて一貫したインデックスとして整列して存在し，共通の機能的役割を担っていることを示唆している．

また，層が深くなるにつれて内部表現のスケールが増大するため，後段の層ほど出力言語を制御するにはより大きなスケーリング係数 α が必要となる．

表 1: 多言語生成制御の結果

Model	Method	Fr		De		Zh		Ja		Es		Overall		
		ACC	BLEU	ACC	BLEU	ACC	BLEU	ACC	BLEU	ACC	BLEU	ACC	BLEU	A*B
Llama2-7b	**Neuron Kojima**	42.0	15.3	55.7	15.1	81.3	10.2	57.5	7.5	77.0	16.6	60.8	12.5	7.56
	Neuron Tang	83.7	**28.2**	72.7	**21.1**	86.7	**17.9**	20.5	**14.4**	98.0	**20.8**	72.3	**21.8**	15.80
	Ours (Monolingual)	98.3	23.4	97.3	19.4	84.7	16.4	**76.5**	9.9	98.7	17.30	91.1	18.3	**16.63**
	Ours (Parallel)	**99.1**	22.7	**98.8**	18.3	**88.3**	17.2	**76.5**	9.4	**99.3**	17.4	**92.6**	17.9	16.57
Llama2-13b	**Neuron Kojima**	15.7	7.3	30.3	9.5	96.0	15.9	64.5	8.6	14.0	4.3	47.4	12.2	5.80
	Neuron Tang	83.7	**32.5**	14.7	**23.5**	**99.3**	**19.1**	64.5	8.6	42.0	**25.8**	63.7	**22.4**	14.23
	Ours (Monolingual)	**96.2**	23.3	**99.2**	17.5	97.6	16.1	91.5	8.9	97.0	11.2	**96.9**	16.7	16.14
	Ours (Parallel)	93.1	26.0	98.6	17.9	99.1	17.6	**92.5**	**9.4**	**98.7**	13.4	96.3	18.1	**17.40**
Aya23-8b	**Ours (Monolingual)**	68.0	**21.7**	**69.4**	**14.7**	**22.8**	20.3	69.8	10.8	58.7	14.0	55.6	16.5	9.34
	Ours (Parallel)	**88.0**	20.1	58.8	14.4	**22.8**	**25.6**	**76.0**	**12.8**	**80.0**	**16.1**	**61.7**	**17.3**	**10.69**

一方で，α の増加に伴い BLEU が低下する傾向も観察されており，操作強度と生成品質の間には明確なトレードオフが存在することが分かる．さらに，13 層以前の浅い層では本手法はほとんど効果を示さなかった．これは，この段階ではモデルが出力言語をまだ確定しておらず，後続層での処理によって操作の効果が上書きされるためと考えられる．これは，出力言語決定が中間層付近で行われるとする Dumas ら [16] の報告とも整合的である．

5.2 多言語生成制御の結果

前節の解析に基づき，最適な層と操作強度を選択した上で，多言語生成制御タスクにおける性能を評価した．4 つの LLM と 5 言語方向における多言語生成制御の結果を表 1 に示す。結果は，言語固有の次元の同定に用いる文を異なるランダムシードでサンプリングした 3 回の独立した実行の平均である。

既存のニューロン操作手法 Neuron-Kojima[3], Neuron-Tang[4] と比較すると，提案手法は一貫して高い性能を示した. 特に，ACC×BLEU の観点から，Llama2–7B では最良設定において Neuron-Kojima を 8.87 ポイント，Neuron-Tang を 1.49 ポイント上回り，Llama2–13B ではそれぞれ 11.43 ポイントおよび 4.61 ポイントの改善が確認された．これにより，提案手法は出力言語制御の成功率と意味保持の両立において優れていることが分かる. 単言語設定および対訳設定のいずれにおいても，安定して目的言語への制御が可能であり，設定間の性能差は小さい.

さらに，Aya23–8B といった新しいモデルに対しても評価を行った結果，提案手法はモデル系列を跨いで有効であることが示された.

6 おわりに

本研究では，LLM における言語表現の層方向遷移を分析し，英語の表現から目的言語トークン空間への変換が，層を通じて一貫した少数かつ疎な言語固有の次元によって制御されていることを示した. さらに，50 文のデータのみを用いてこれらの次元を同定・操作する学習不要な手法を提案し，意味内容を保ったままで出力言語を制御できることを確認した. 本手法は，多言語生成制御タスクにおいて既存のニューロン操作手法を上回る性能を示し，言語固有の次元が LLM における言語制御の中核的機構であることを示唆する.

本研究の評価は特定の生成制御タスクと設定に限定されており，他タスクやモデルごとの最適設定への一般化については今後の課題である.

謝辞

本研究は文部科学省の補助事業「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」および JSPS 科研費 JP23K28144 の助成を受けたものです.

参考文献

[1] Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in english? on the latent language of multilingual transformers. **ArXiv**, Vol. abs/2402.10588, , 2024.

[2] Nostalgebraist. Interpreting gpt: The logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, 2020. Accessed: 2024-07-28.

[3] 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 Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 6919–6971, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[4] Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, et al. 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 (Volume 1: Long Papers)**, pp. 5701–5715, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[5] Boyi Deng, Yu Wan, Baosong Yang, Yidan Zhang, and Fuli Feng. Unveiling language-specific features in large language models via sparse autoencoders. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 4563–4608, Vienna, Austria, July 2025. Association for Computational Linguistics.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In **North American Chapter of the Association for Computational Linguistics**, 2019.

[7] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, et al. Multilingual denoising pre-training for neural machine translation. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 726–742, 2020.

[8] Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. Bridging the language gaps in large language models with inference-time cross-lingual intervention. In **Annual Meeting of the Association for Computational Linguistics**, 2024.

[9] Omar Mahmoud, Buddhika Laknath Semage, Thommen George Karimpanal, and Santu Rana. Improving multilingual language models by aligning representations through steering. **ArXiv**, Vol. abs/2505.12584, , 2025.

[10] Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, et al. Llama 2: Open foundation and fine-tuned chat models. **ArXiv**, Vol. abs/2307.09288, , 2023.

[11] Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, et al. Aya 23: Open weight releases to further multilingual progress. **ArXiv**, Vol. abs/2405.15032, , 2024.

[12] Nllb team, Marta Ruiz Costa-jussà, James Cross, Onur cCelebi, Maha Elbayad, et al. No language left behind: Scaling human-centered machine translation. **ArXiv**, Vol. abs/2207.04672, , 2022.

[13] Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, et al. Overview of the IWSLT 2017 evaluation campaign. In Sakriani Sakti and Masao Utiyama, editors, **Proceedings of the 14th International Conference on Spoken Language Translation**, pp. 2–14, Tokyo, Japan, December 14-15 2017. International Workshop on Spoken Language Translation.

[14] Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, et al. Findings of the 2018 conference on machine translation (WMT18). In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, **Proceedings of the Third Conference on Machine Translation: Shared Task Papers**, pp. 272–303, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[15] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers**, pp. 427–431, Valencia, Spain, April 2017. Association for Computational Linguistics.

[16] Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 31822–31841, Vienna, Austria, July 2025. Association for Computational Linguistics.

A　言語固有の次元操作のハイパーパラメータチューニング

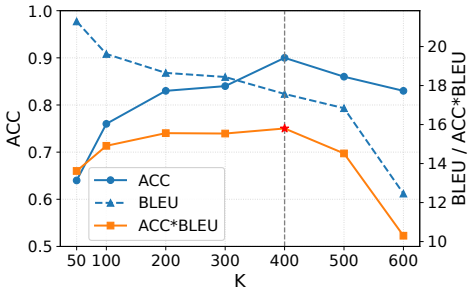


図 4: 言語固有の次元同定における Top-K の選択

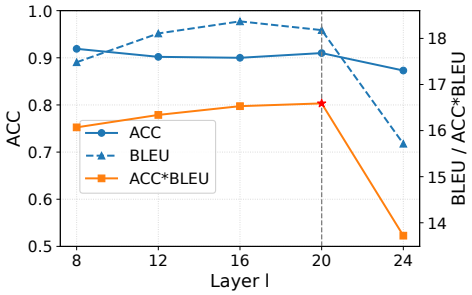


図 5: 単言語設定における中間層の選択

本節では，言語固有次元の同定および推論時操作に関わる主要ハイパーパラメータの影響を検証する。まず，言語固有の次元の同定における Top-K の影響を検証する．Llama2–7B を対象とし，対訳設定において層 19 に操作（$\alpha = 0.4$）を適用した上で，K を変化させて多言語生成制御タスクを評価した．

図 4 に示すように，K の増加に伴い ACC は向上する一方，BLEU は低下する傾向が確認された．その結果，言語制御において本質的な次元は比較的小さな K でも概ね捉えられることが分かる．一方で，性能を最大化する目的から，全実験において $K = 400$ を採用した．この値は，Llama2–13B では表現次元の約 7.8%，その他のモデルでは約 9.8% に相当する．

次に，単言語設定では，最終層と比較するための中間層の選択が必要となる．そこで，$K = 400$ とし，層 19 に操作（$\alpha = 0.4$）を適用した上で，次元同定に用いる中間層 l を変化させ，同モデルで性能を評価した．

図 5 より，ACC および BLEU は広い範囲の l にわたって安定しており，ACC×BLEU は $l = 20$ 付近で最大となった．一方で，非常に深い層を用いた場合には性能が低下する傾向が見られた．これは，この段階で表現空間がすでに目的言語へと遷移しつつあり，言語固有次元の識別が困難になるためと考えられる．これらの結果から，提案手法は中間層の選択に対して高い頑健性を有しており，本研究では $l = 20$ を採用した．

B　言語固有の次元の分析

表 2: 言語固有の次元の言語間重複

	zh	ja	fr	es	de
zh	400	**193**	105	100	118
ja		400	88	98	113
fr			400	**152**	143
es				400	140
de					400

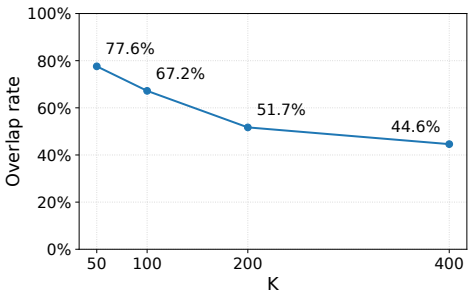


図 6: 単言語と対訳設定の言語固有の次元の一致率

同定された言語固有次元が言語間でどの程度共有されるかを分析した．表 2 には，Llama2–7B の単言語設定下における言語ペア間の次元重複数を示す．

結果から，言語類型的に近い言語ほど多くの次元を共有する傾向が確認された．例えば，中国語と日本語は 400 次元中 193 次元を共有しており，ロマンス語族およびゲルマン語族内でも比較的高い重なりが観察された．これらの結果は，言語固有の次元の多くが単一言語に固有というよりも，言語間で部分的に共有されていることを示唆している．

最後に，Llama2–7B における単言語設定と対訳設定において同定される言語固有の次元の一致性を評価する．図 6 に示すように，$K = 400$ の場合，言語平均での一致率は 44.6% であった．一方，$K = 50$ にすると，一致率は 77.6% に向上した．すなわち，中核となる言語次元は両設定で概ね一致する一方，K の拡大に伴い，純粋な言語次元に加えて，意味的・文脈的要因を含む周辺次元が混入する傾向が見られる．この傾向は，図 4 に示したように，K の増加とともに BLEU が低下する現象とも整合的である．