LLM の多言語内部表現における意味空間と言語成分の構造

山本 泰成 [1,2]　九門 涼真 [1,2]　Danushka Bollegala[3]　谷中 瞳 [1,2,4]
[1] 東京大学　[2] 理化学研究所　[3]University of Liverpool　[4] 東北大学
{yamamo96, kumoryo9, hyanaka}@is.s.u-tokyo.ac.jp, danushka@liverpool.ac.uk

概要

LLM は多言語処理の際、中間層で言語間共通の意味空間で処理を行うことが知られており、言語間で内部表現を揃えることでモデルの多言語能力を向上させる手法が提案されてきた。しかし、これらの手法は言語に依存する推論を妨げることが指摘されている。本研究では、LLM の内部表現を意味成分と言語成分に分け、意味成分が表現される意味空間と言語成分の構造を分析する。中間層では言語成分は入力内容によらない定数であり、また、低資源言語では意味空間と言語成分が干渉し、同一の意味の英文との内部表現の言語間対応が不十分であることが示唆された。これらの知見は、言語情報を保持したまま意味成分を揃える手法の設計に資する。

1　はじめに

大規模言語モデル (LLM) は英語などの高資源言語で様々なタスクにおいて高い性能を示す一方で、用いる言語によって能力に差があることが指摘されている [1, 2]。これまで、LLM の多言語能力向上に向けて、多言語処理における LLM の内部機序が分析されてきた。先行研究では、LLM は入力を言語間で共通の意味空間に写像し、中間層ではその意味空間で処理を行い、終盤層で出力の言語に変換すると報告されている [3, 4, 5]。さらに、中間層における言語間の内部表現の類似度とモデルの多言語能力が相関することも確認されている [6]。これらの性質を用いて、内部表現を言語間で近づけることで多言語能力を改善する手法が提案されてきた [7, 8]。一方で、それらの手法は内部表現が持つ言語情報を損ない、文化的知識を伴う推論などの言語に依存する振る舞いを妨げるという課題も指摘されている [9]。

多言語における LLM の内部表現の構造の理解は、推論に必要な言語情報を損なわずに意味成分を揃え、モデルの多言語能力を改善するために不可欠である。本研究では、LLM の内部表現として各層の出力ベクトルを対象とする。LLM の内部表現を意味を表す意味成分と、処理中の言語の種類に関する情報を担う言語成分に分け、それらを特定し、意味成分が表現される意味空間と言語成分の構造を分析する。そのために、翻訳データセットの文を LLM に入力として与え、文埋め込みの手法を用いてモデルから入力文を表すベクトル表現を抽出する。そして、英語と言語 l の対訳文ペアに対するベクトル表現の差分から共通成分を抽出し、それを LLM の推論時に内部表現に足すことで、英語の入力に対し出力言語を l にするよう操作できることを示す。これは、ある定数ベクトルによって入力内容によらずモデルの出力言語を操作できることを表し、言語成分は意味成分によらない定数であることを意味する。このベクトルを**言語ベクトル**と呼ぶ。

次に、英文のベクトル表現に言語 l の言語ベクトルを加えて並行移動することで得られた表現が、データセット内において、言語 l の対応する翻訳文のベクトル表現に最も近くになることを示す。このことは、ベクトル表現から言語ベクトルを除いた意味成分は言語間で共通することを意味する。また、とくに低資源言語において、並行移動前のベクトル表現は英語との間で揃っておらず、これは言語ベクトルが意味空間と直交性が低く、かつノルムが比較的大きいことに起因することを示す。

最後に、LLM の推論時における各成分の寄与を分析するため、各層のモデル重みと、意味成分及び言語ベクトルとのコサイン類似度の平均をそれぞれ計算する。その結果、モデル重みとの類似度は、中間層では意味成分の方が高く、終盤層では言語ベクトルの方が高くなる傾向がみられた。言語間で比較すると、中間層では低資源言語の言語ベクトルの方が高資源言語よりも重みとの類似度が高いが、終盤層では逆になる。これは、意味処理を担う中間層では低資源言語の言語ベクトルが意味空間により強く干渉していて、また、言語固有の操作を担う終盤層では低資源言語に対応する言語固有パラメータがよ

り少ないことを示唆する。

2 言語ベクトルの計算

2.1 実験設定

実験に用いる言語は、高資源言語である英語、標準中国語 (cmn)、スペイン語 (spa)、フランス語 (fra)、ドイツ語 (deu)、ロシア語 (rus)、日本語 (jpn)、タイ語 (tha)、中〜低資源言語であるスワヒリ語 (swh)、ベンガル語 (ben)、テルグ語 (tel) とする。これらの言語は複数の語族に属し、複数の語順を持つ。対象とするモデルは、Llama-3.1-8B-Instruct [10]、Qwen3-4B、Qwen3-8B [11] とする。分析には翻訳データセット FLORES+ [12]、OPUS-100 [13] と、英語の指示学習データセット dolly-15k [14] を用いる。

2.2 LLM の内部表現

モデルへの入力文に対する LLM のベクトル表現を計算するため、デコーダーモデルから文埋め込みを計算する手法を用いる。文埋め込みの計算手法は様々なものが提案されているが [6]、事前実験の結果をもとに、以下の式で定義されるトークン位置による重み付き和 [15] を用いる：

$$\mathbf{x} = \sum_{t=1}^{T} w_t \mathbf{h}_t \quad \text{where} \quad w_t = \frac{t}{\sum_{t=1}^{T} t} \tag{1}$$

ここで、$\mathbf{h}_t$ はトークン位置 t におけるモデルの内部表現、T は文中のトークン数を表す。

内部表現の多言語構造を分析するために、内部表現が、意味を表す意味成分と処理中の言語の種類に関する情報を担う言語成分に分けられると仮定する。つまり、言語 l の入力文 i に対する内部表現を

$$\mathbf{x}_i^{(l)} = \mathbf{s}_i^{(l)} + \mathbf{b}_i^{(l)} \tag{2}$$

と分解できるとする。$\mathbf{s}_i^{(l)}$ は入力 i の言語 l における意味成分、$\mathbf{b}_i^{(l)}$ は言語成分である。ここで、LLM の意味空間は、学習言語に最も多く含まれている英語に対応することがこれまでの研究で示されてきた [5, 16]。そのため、以下では英語の言語成分 $\mathbf{b}_i^{(en)}$ を基準（$\mathbf{0}$）として、$\mathbf{x}_i^{(en)} = \mathbf{s}_i^{(en)}$ として扱う。

2.3 言語ベクトルによる出力言語の操作

言語成分は入力の意味内容に依存しないかを分析する。そのために、言語ごとの定数ベクトルを内部表現に足すことで、異なるドメインの入力に対して LLM の出力言語が操作できるかを調べる。

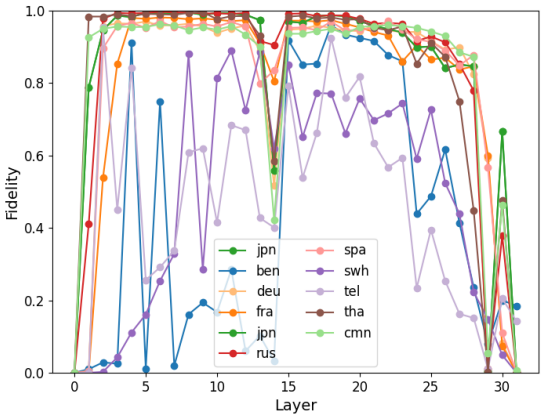
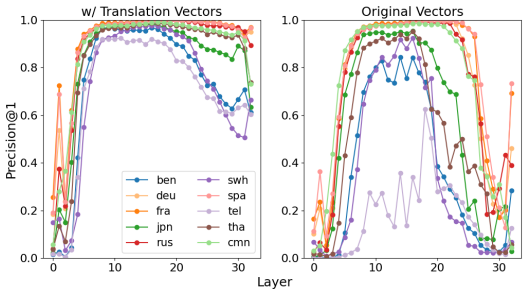


図 1 Llama-3.1-8B-Instruct に言語ベクトルで介入した時の dolly-15k における fidelity。

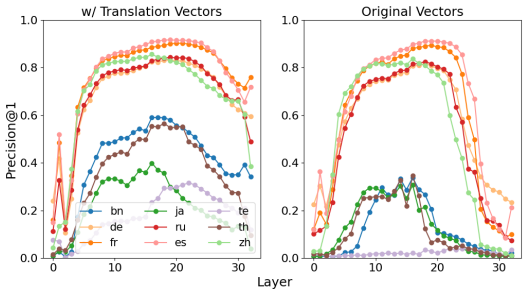
まず、FLORES+開発セット (dev) を用いて英語と言語 l の対訳文に対するベクトル表現 $\mathbf{x}_i^{(en)}, \mathbf{x}_i^{(l)}$ を層ごとに計算する。これらの差分 $\mathbf{x}_i^{(l)} - \mathbf{x}_i^{(en)}$ に共通する成分を抽出することで、言語を表す定数成分、つまり言語ベクトルを計算することを考える。そのために、まず、$\mathbf{x}_i^{(l)} - \mathbf{x}_i^{(en)}$ を列ベクトルとして並べた行列 $X \in \mathbb{R}^{d \times n}$ を作る（d は次元数、n はインスタンス数）。X を主成分分析し、分散が中央値の 10 倍より大きい p 個の方向を並べて行列 $W \in \mathbb{R}^{d \times p}$ とする。そして、差分ベクトルの平均 μ から W の成分を $\frac{1}{2}$ だけ除いたベクトル $\mathbf{f} = (I - \frac{1}{2}WW^T)\mu$ に、μ を射影したベクトル $\frac{\mathbf{f}\mathbf{f}^\mathbf{T}}{||\mathbf{f}||^2}\mu$ を言語ベクトル $\mathbf{b}^{(l)}$ とする。この方法により、差分ベクトルの平均方向を維持しながら、分散を軽減することができる。

英語の入力を与えた際に、この $\mathbf{b}^{(l)}$ を各層に足すことで介入を行う。入力は、dolly-15k の 8 カテゴリから 100 件ずつランダムにサンプリングした、合計 800 件のデータを用いる。出力された文章の言語を GlotLID [17] により特定し、出力言語が l となっている割合 (fidelity) を計算する。

Llama-3.1-8B-Instruct の結果を図 1（Qwen3-4B、8B は図 6、 7）に示す。とくに中間以降の層で fidelity が高かった。さらに、テルグ語などの低資源言語では fidelity が低いが、それでも最大約 90% の fidelity を示した。また、μ をそのまま用いて介入した結果を図 8 に示すが、$\mathbf{b}^{(l)}$ を用いた場合よりも fidelity は低かった。これらの結果から、$\mathbf{b}^{(l)}$ を足すという介入により、入力ドメインに関わらず出力言語を操作できることがわかる。このことは、内部表現において意味成分と言語成分は式 2 のように和の形で表せ、かつ言語ベクトルは入力 i によらない定数であることを示唆する。これ以降、言語ベクトルは



(a) FLORES+評価セット



(b) OPUS-100 テストセット

図 2　Llama-3.1-8B-Instruct の内部表現の距離による言語間での Precision@1。言語ベクトルによる並行移動をした場合（左）としない場合（右）。

FLORES+開発セットで計算した定数 $\mathbf{b}^{(l)}$ を用いる。

3　多言語における内部表現の構造

3.1　言語間で共通する意味成分の特定

次に、意味成分が言語間でどれほど一致しているかを分析する。まず、FLORES+評価セット (devtest) または OPUS-100 テストセットから、英語と言語 l の対訳文をそれぞれ入力として与えたときのベクトル表現 $\{\mathbf{x}_i^{(en)}\}_{i=1}^n, \{\mathbf{x}_i^{(l)}\}_{i=1}^n$ を計算する。そして、英語のベクトル表現を言語ベクトルで並行移動した $\mathbf{x}_i^{(en)} + \mathbf{b}^{(l)}$ との距離が最も近い言語 l のベクトル表現が、対訳文のベクトル表現 $\mathbf{x}_i^{(l)}$ になっている割合 $\mathbb{E}\left[\mathbb{I}\left(\mathrm{argmin}_j(\|\mathbf{x}_i^{(en)} + \mathbf{b}^{(l)} - \mathbf{x}_j^{(l)}\|) = i\right)\right]$（Precision@1）を、層ごとに評価する。さらに、$\mathbf{b}^{(l)}$ による並行移動をしない場合の、元々のベクトル表現の Precision@1 $\mathbb{E}\left[\mathbb{I}\left(\mathrm{argmin}_j(\|\mathbf{x}_i^{(en)} - \mathbf{x}_j^{(l)}\|) = i\right)\right]$ も計算する。なお、OPUS-100 については英文と各言語の対訳文とのアライメントの品質を担保するため、英文のトークン数が 10 未満のペアは除いた。

Llama-3.1-8B-Instruct の結果を図 2（Qwen3-8B は図 9）に示す。FLORES+に関して、言語ベクトルによる並行移動を行った場合は、全ての言語で中間層では 90% 以上の精度を示した。これは、各インスタンス i について、以下が成り立つことを表す：

$$\forall j\ (i \neq j),\ \|\mathbf{x}_i^{(en)} + \mathbf{b}^{(l)} - \mathbf{x}_i^{(l)}\| < \|\mathbf{x}_i^{(en)} + \mathbf{b}^{(l)} - \mathbf{x}_j^{(l)}\| \tag{3}$$

$$\Leftrightarrow \|\mathbf{s}_i^{(en)} - \mathbf{s}_i^{(l)}\| < \|\mathbf{s}_i^{(en)} - \mathbf{s}_j^{(l)}\| \tag{4}$$

つまり、データセット内で意味成分は言語間で一致している。一方、OPUS-100 では、とくに低資源言語で精度が低い。これは、OPUS-100 は短い文を多く含んでおり、対訳文間で意味が完全には一致しないケースが存在することや、意味の分散が小さいことなどの原因が考えられる。

並行移動を行わない場合は、行う場合よりも特に低資源言語で精度が低かった。ここで、あるインスタンス $i, j\ (i \neq j)$ に対しそれぞれ言語間でアライメントが取れる場合、以下の関係式が成立する：

$$\|\mathbf{x}_i^{(en)} - \mathbf{x}_i^{(l)}\| < \|\mathbf{x}_i^{(en)} - \mathbf{x}_j^{(l)}\| \tag{5}$$

$$\wedge\ \|\mathbf{x}_j^{(en)} - \mathbf{x}_j^{(l)}\| < \|\mathbf{x}_j^{(en)} - \mathbf{x}_i^{(l)}\| \tag{6}$$

ここで、前述の結果より $\mathbf{s}_i^{(en)} \approx \mathbf{s}_i^{(l)}$ を仮定すると、以下が導出できる：

$$|\cos(\mathbf{x}_i^{(en)} - \mathbf{x}_j^{(en)}, \mathbf{b}^{(l)})| < \frac{1}{2}\frac{\|\mathbf{x}_i^{(en)} - \mathbf{x}_j^{(en)}\|}{\|\mathbf{b}^{(l)}\|} \tag{7}$$

この式は、意味空間と言語が直交している場合、または言語ベクトルのノルムが意味成分の違いに比べて小さい場合に、アライメントが取れることを表している。これらに関して 3.2 章で詳しく分析する。

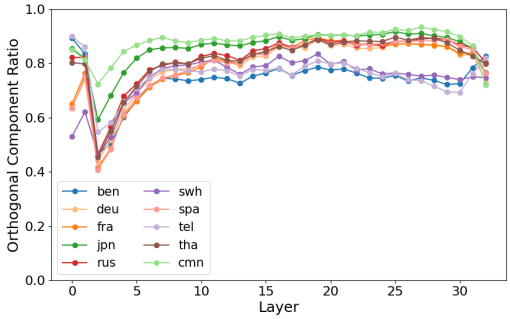


図 3　Llama-3.1-8B-Instruct の言語ベクトルのノルムのうち、意味空間に直交する成分の割合。

3.2　意味空間と言語成分の関係

モデルの内部表現における意味空間と言語ベクトルの直交性を分析する。意味空間 $\mathcal{S}$ は、FLORES+評価セットの英文を入力として与えたときのベクトル表現に対し主成分分析を行い、寄与率 90% までの要素によって張られる空間とする。言語 l につい

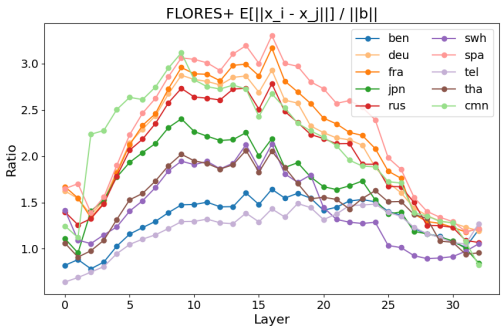


図 4 Llama-3.1-8B-Instruct の英語入力をした時の内部表現の差分のノルムと、言語ベクトルのノルムの比。

て、$\mathbf{b}^{(l)}$ のうち $\mathcal{S}$ に直交する成分 $\mathbf{b}_o^{(l)}$ が、$\mathbf{b}^{(l)}$ のノルムに占める割合 $\frac{||\mathbf{b}_o^{(l)}||}{||\mathbf{b}^{(l)}||}$（直交率）を計算する。

Llama-3.1-8B-Instruct の結果を図 3（Qwen3-8B は図 10）に示す。低資源言語では特に終盤層で直交率が低く、言語ベクトルが意味空間に干渉することが示唆される。また、高資源言語の中で見ると、英語と同じラテン文字を使う言語では直交率が低い。これは、同じ文字体系を用いる言語間では語彙などに共通要素があるため、言語ベクトルが意味空間に干渉しやすくなることが原因だと考えられる。

次に、FLORES+評価セットにおける英語インスタンス間のベクトル表現差分のノルム平均と、言語ベクトルのノルム比 $\frac{\mathbb{E}_{i\neq j}[||\mathbf{x}_i^{(en)}-\mathbf{x}_j^{(en)}||]}{||\mathbf{b}^{(l)}||}$ を図 4（Qwen3-8B は図 11）に示す。低資源言語ではとくに前半層でこの比が小さく、意味成分の差に対して言語ベクトルの影響が大きいことがわかる。これらの結果から、高資源言語では中間層までは言語ベクトルのノルムが比較的小さく、終盤層では意味空間と言語ベクトルがより直交することで、言語間のベクトル表現のアライメントが達成されていることが示唆される。一方で、低資源言語はどの層でも言語ベクトルのノルムが大きく、終盤層での直交率の向上も見られないため、ベクトル表現の整合が不十分であると考えられる。

4 推論時における意味空間の分析

最後に、LLM の推論において、意味成分と言語ベクトルがモデルの重みにどのように作用するのかを調べる。LLM の内部表現 $\mathbf{h}$ に対し、MLP のゲートモジュールの活性値は以下のように計算される：

$$\mathbf{g} = W_{\text{gate}} R \frac{\mathbf{h}}{\sqrt{||\mathbf{h}||^2 + \epsilon}} \tag{8}$$

W_{gate} はゲートモジュールの重み行列、R は RMSNorm の重み（対角行列）を表す。これは実際は非線形変換であるが、分散による正規化を無視し $W_{\text{gate}}R$ の各行と内部表現のコサイン類似度を考える。各言語について、言語ベクトル $\mathbf{b}^{(l)}$ と、FLORES+開発セットで計算したベクトル表現の平均 $\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i^{(l)}$ から $\mathbf{b}^{(l)}$ を引いた平均意味成分のそれぞれに対し、$W_{\text{gate}}R$ の各行とのコサイン類似度の平均を計算する。

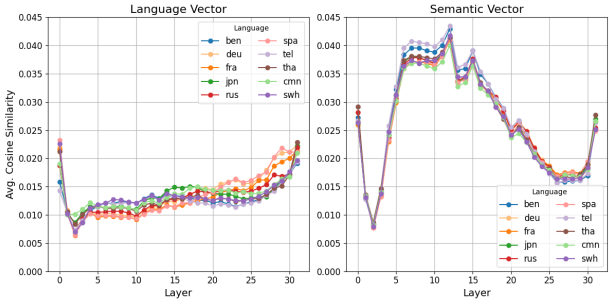


図 5 Llama-3.1-8B-Instruct の言語成分、意味成分と、MLP ゲートモジュールの重みとのコサイン類似度の平均。

Llama-3.1-8B-Instruct の結果を図 5（Qwen3-8B は図 12）に示す。まず、中間層では意味成分の類似度が高く、終盤層では言語ベクトルの類似度が高かった。これは、中間層では意味処理を行い、終盤層は言語固有の操作を担うためだと考えられる。また、重みと言語ベクトルとの類似度に関して、前半層では低資源言語の方が高資源言語よりも高く、終盤層では逆になった。これは、低資源言語では意味処理を担う中間層で言語ベクトルが意味空間に干渉する一方、終盤層において言語固有の操作を担うパラメータが少ないことを示唆しており、これが言語性能の低さと関係している可能性がある。

5 まとめ

本論文では、LLM の多言語内部表現における意味空間と言語成分の構造を分析した。まず、モデルの言語成分は定数ベクトルによって近似でき、それを用いて介入することで入力内容によらず出力言語を操作できることを実験的に示した。次に、意味成分は言語間で一致していることが多いが、特に低資源言語では、モデルの前半層で言語ベクトルのノルムが大きく、終盤層で言語ベクトルと意味空間が直交していないために、言語間で内部表現が整合していないことが示唆された。また、モデルの重みに関して、低資源言語では前半層で意味処理に言語ベクトルが干渉し、終盤層で言語情報の処理を担うパラメータが少ないことがわかった。今後、これらの知見をもとに、言語情報を保持しながら LLM の内部表現を揃える手法について検討する。

謝辞

本研究は JST CREST　JPMJCR2565，JST BOOST JPMJBY24H5 の支援を受けたものである．

参考文献

[1] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners. In **The Eleventh International Conference on Learning Representations**, 2023.

[2] Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. 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. 749–775, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[3] Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in English? on the latent language of multilingual transformers. 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. 15366–15394, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[4] 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 (Volume 1: Long Papers)**, pp. 5701–5715, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[5] Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. In **The Thirteenth International Conference on Learning Representations**, 2025.

[6] Amir Hossein Kargaran, Ali Modarressi, Nafiseh Nikeghbal, Jana Diesner, François Yvon, and Hinrich Schuetze. MEXA: Multilingual evaluation of English-centric LLMs via cross-lingual alignment. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Findings of the Association for Computational Linguistics: ACL 2025**, pp. 27001–27023, Vienna, Austria, July 2025. Association for Computational Linguistics.

[7] Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. Improving in-context learning of multilingual generative language models with cross-lingual alignment. 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. 8058–8076, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[8] Danni Liu and Jan Niehues. Middle-layer representation alignment for cross-lingual transfer in fine-tuned LLMs. 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. 15979–15996, Vienna, Austria, July 2025. Association for Computational Linguistics.

[9] HyoJung Han, Sweta Agrawal, and Eleftheria Briakou. Rethinking cross-lingual alignment: Balancing transfer and cultural erasure in multilingual llms, 2025.

[10] Llama Team, AI @ Meta. The llama 3 herd of models. **arXiv preprint arXiv:2407.21783**, 2024. version 3.

[11] Qwen Team. Qwen3 technical report. **arXiv preprint arXiv:2505.09388**, 2025.

[12] NLLB Team, Marta R. Costa-jussà, James Cross, et al. Scaling neural machine translation to 200 languages. **Nature**, Vol. 630, No. 8018, pp. 841–846, Jun 2024.

[13] Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 1628–1639, Online, July 2020. Association for Computational Linguistics.

[14] 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.

[15] Niklas Muennighoff. Sgpt: Gpt sentence embeddings for semantic search. **arXiv preprint arXiv:2202.08904**, 2022.

[16] Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schuetze. Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models. 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. 5075–5094, Vienna, Austria, July 2025. Association for Computational Linguistics.

[17] Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. GlotLID: Language identification for low-resource languages. In **The 2023 Conference on Empirical Methods in Natural Language Processing**, 2023.

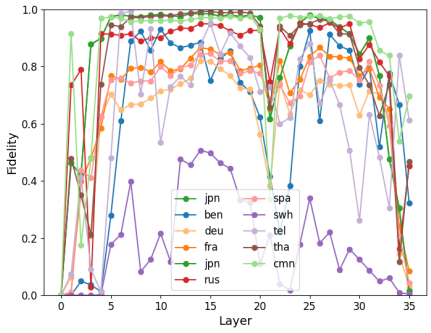


図 6 Qwen3-4B に言語ベクトルで介入した時の dolly-15k における fidelity。

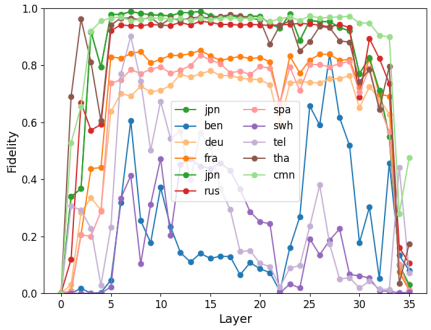


図 7 Qwen3-8B に言語ベクトルで介入した時の dolly-15k における fidelity。

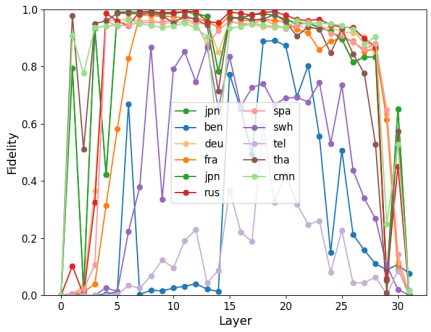


図 8 Llama-3.1-8B-Instruct に平均差分ベクトルで介入した時の dolly-15k における fidelity。

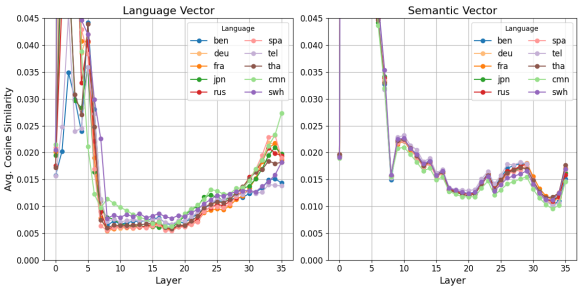
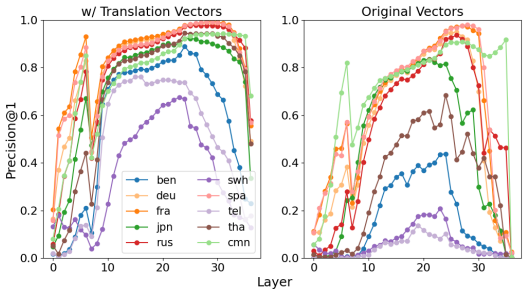
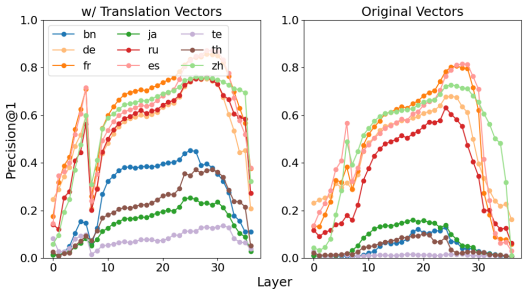


図 12 Qwen3-8B の言語成分、意味成分と、MLP ゲートモジュールの重みとのコサイン類似度の平均。



(a) FLORES+評価セット



(b) OPUS-100 テストセット

図 9 Qwen3-8B の内部表現の距離による言語間での Precision@1。言語ベクトルによる並行移動をした場合（左）としない場合（右）。

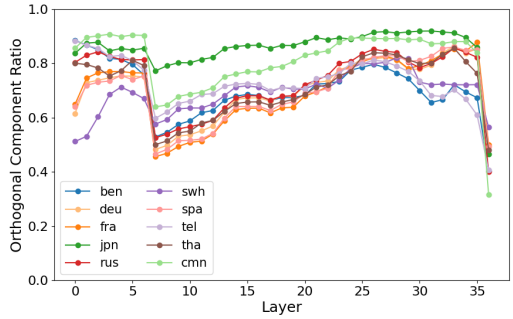


図 10 Qwen3-8B の言語ベクトルのノルムのうち、意味空間に直交する成分の割合。

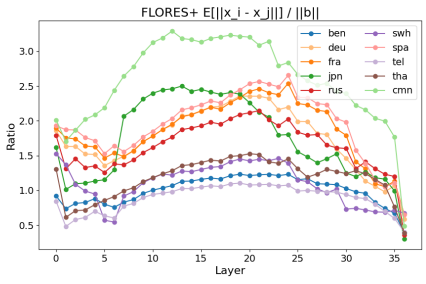


図 11 Qwen3-8B の英語入力をした時の内部表現の差分のノルムと、言語ベクトルのノルムの比。