BrainLM を用いた多言語学習での転移学習性能の検証

羅 桜　　小林 一郎
お茶の水女子大学大学院 人間文化創成科学研究科 理学専攻
{ying.luo, koba}@is.ocha.ac.jp

概要

近年，言語による刺激とそれに誘起された脳活動との対応関係を捉えた事前学習済みマルチモーダル言語モデルである BrainLM が提案された．本研究では、このモデルをさらに発展させ、英語だけであった言語刺激をフランス語および中国語へと拡張し多言語からなる BrainLM の開発を行った．言語刺激の拡張には，転移学習を利用することで多言語タスクでのモデルによる脳内状態予測能力を向上させた．特に、英仏間の整合性判別のための二値分類タスクにおいて多言語対応した BrainLM は 51.75%という最高精度を達成した．また，脳内状態予測タスクにおいて，転移学習の前後で相関係数が約 3%から 15%向上した．さらに、大脳皮質全体の脳内状態予測タスクにおいて，他のモデルと比べて BrainLM が最も高い相関を示した．本研究は、BrainLM の応用範囲を広げるだけでなく，さまざまな言語における大規模言語モデルとヒト脳機能との関係性についての理解を深めるものである．

1　はじめに

大規模言語モデルの広範な利用により，機能的磁気共鳴画像法（fMRI）や脳波計測（EEG）などの非侵襲的手法で取得される信号として計測されるヒト脳活動とヒトが操る複雑な言語システムとの関連性を探る研究が注目されている．言語システムは，脳内の意味情報の表現と実際の意味知覚との間に強い相関があることを示している [1, 2, 3, 4]．例えば，Luo らは Brain Language Model（BrainLM）を提案した [5]．これは，自然言語の意味表現と脳の特徴を統合する共通の埋め込み空間を組み込んだ大規模言語モデルである．BrainLM の有効性は，脳内状態のデコーディングとエンコーディングの両方向からのタスクにおいて検証されている．しかし，解決すべき課題は依然として多い．しかし，解決すべき課題は依然として多い．例えば，Luo らの研究では，BOLD5000 データセットを用いて視覚刺激に関連する脳の関心領域（ROIs）を対象とした脳内状態エンコーディングタスクが実施され，平均ピアソン相関係数 50.35%を達成した．この結果は視覚関連の局所的な有効性を示しているものの，言語刺激を含む大脳皮質全体における BrainLM の性能は未だ明らかではない．

さらに，単一言語モデルは多言語使用者の脳内処理を十分に再現できないことが明らかになりつつある．例えば，複数言語使用者は，言語切り替えや統合において、脳内で異なるネットワークが活性化されることが報告されている [6]．一方で，単一言語モデルは特定の言語に特化しており，多言語処理における脳の柔軟性や相互作用を考慮できない．

しかし，大規模言語モデルにおける多言語学習プロセスが人間の脳とどのように対応するかについての定量的分析は，まだ十分に行われていない．また，研究が進むにつれて，利用可能な多言語データセットが増加している [7, 8]．このギャップを埋めるために，BrainLM を多言語学習タスクに適用する新たなアプローチを提案する．特に転移学習 [9] を利用して，多言語データセット上でモデルを検証する．このアプローチにより，BrainLM を通じて脳が複数の言語をどのように処理するかを理解する手がかりを得ることができる．

以上を踏まえ，BrainLM を用いて以下の 3 つの検証実験を実施した．新たな多言語データセットを導入し，可視化ツールと組み合わせることで，転移学習による言語システム間のモデルの適応性を検証した．本研究の貢献は以下の 3 点に要約される：

- 選択された脳の機能的領域から大脳皮質全体への脳活動予測能力を拡張し，異なる被験者およびデータセット間での BrainLM の脳全体予測への汎化能力を検証した．
- BrainLM の有効性を英語以外のシステムにも拡張し，フランス語および中国語を母語とする話

表 1　本実験で使用した脳データセットとテキストコーパスの説明.

Name	Type	Author	Stimuli	N Subjects	Description
Little Prince Dataset	Brain Dataset	littlePrince	Auditory	112	This dataset collected fMRI data from native speakers of Chinese, English, and French asked to listen to the audio of the novel The Little Prince in their native language during a behavioral experiment. The stimulus audio was divided into nine parts and each part was played three times. In total, the data consisted of approximately 15,000 words. Individual parts were approximately 6,000 seconds long.
The XNLI Corpus	Text Corpus	XNLICorpus	-	-	A multilingual text corpus derived from translated sentences across diverse languages, designed for evaluating cross-lingual understanding and transfer capabilities in natural language processing models.

者を対象とした実験で成功を収めた．また，多言語学習タスクにおける転移学習の効果に関する新しい実証例を追加した．

- 可視化ツールを使用し，モデル構造内の異なる層と対応する脳の機能的領域間の定量的関係を詳細に示した．

2　方法

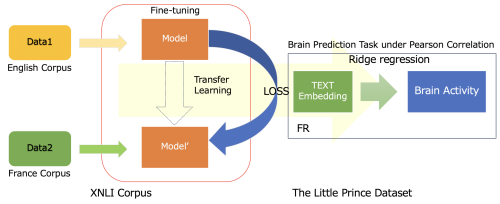

図 1　Flowchart of transfer learning for the models.

データセット

モデルの学習および検証において，2 つの新しい多言語データセット（表 1）を使用した. Luo らの研究 [5] と同様に，脳データセットには fMRI データを使用した. 脳データセット収集のための行動実験における刺激ソースは異なるものの，これらはすべて人間の脳による自然言語処理能力を探ることを目的としている.

モデル

本研究では，脳活動と言語データの接続を確立するモデルとして BrainLM [1] を使用した. このモデルは BERT [10] を基盤とし，大規模なテキストコーパス（英語 Wikipedia コーパス [2]）および比較的小規模な脳データセット（Alice Dataset [11]）でモデルが訓練された.

BrainLM は，24 層の Transformer ブロックで構成されており，BERT-Large-Uncased [3] のパラメータを基盤としている. モデルの入力はテキストと脳データを受け取るように設計が改良されている. 具体的には，各時点の fMRI 脳データは事前訓練された Autoencoder モデルによって 1,024 次元の脳特徴量として前処理されている [4].

二値分類タスク

合計 10,000 ペアのデータを用い，二値分類タスクの学習を行った. 実験の段階では，半数のデータをランダムにサンプリングし，現在のペアに対応する英語文を別の英語文に置き換えて不一致ペア（ラベル：‘0’）を生成した. 3 つの異なる最先端（SOTA）モデルを制御グループとして選定し，BrainLM と共にモデルの訓練を実施した. その後，モデル適用前後における大脳皮質全体の予測タスクにおけるピアソン相関係数の比較分析を行った.

$$\begin{aligned} \mathrm{LOSS}_{\mathrm{TL}} &= \mathrm{Loss}_{\mathrm{BinaryClassification}} + \mathrm{Loss}_{\mathrm{CosineEmbedding}} \\ &= \mathrm{BinaryCrossEntropy}(y) + \mathrm{Cosine}(e_{\mathrm{EN}}, e_{\mathrm{FR}}) \\ &= -(y \log(p) + (1-y)\log(1-p)) + y\frac{e_{\mathrm{EN}} \cdot e_{\mathrm{FR}}}{|e_{\mathrm{EN}}||e_{\mathrm{FR}}|} \end{aligned} \quad (1)$$

ここで，y は実際のラベルを表し，p は予測ラベルの確率を表す. e_{EN} は英語文の埋め込みを，e_{FR} は対応するフランス語文の埋め込み表現を示す.

モデル訓練時の損失は 2 つの部分から成る. 1 つは二値分類タスクのラベル一致のためのバイナリ交差エントロピー損失で，モデルの予測精度を確保する. もう 1 つは，2 つの文の埋め込み間のコサイン損失で，モデルの学習過程が 2 つの言語間の関連性を学習できるようにする.

その後，新たに訓練されたモデルを取得し，転移学習の有無によるモデルの脳活動予測性能を比較

1）　ベースラインモデルは以下よりダウンロード可能：https://github.com/luoying050601/BrainLM
2）　https://www.corpusdata.org/
3）　https://huggingface.co/google-bert/bert-large-uncased
4）　事前トレーニングモデルは以下よりダウンロード可能：https://github.com/luoying050601/BrainLM

した．この予測タスクでは，モデルを用いてテキスト埋め込みを取得し，その埋め込みを使用してリッジ回帰により対応する実際の脳活動データを予測した．そして，予測データと実際のデータの間の相関性をピアソン相関係数を計算することにより求めた．

音声-テキスト整合アプローチ

実験には，オープンソースの Little Prince Dataset [5] を使用した．このデータセットでは，音声ファイルが刺激源として聴覚神経を活性化し，時間的連続性を示す．一方で，fMRI 取得装置は 2 秒ごとに脳活動を 3 次元浮動小数点配列として記録する．連続データと離散データの整合性を取ることは結果の品質に大きく影響するため，重要な課題である．

実験検証の過程で，固定長アプローチ（FLA）と可変長アプローチ（VLA）の 2 つのデータ整合アプローチを設計した．

3 実験

表 2 英語話者の被験者における大脳皮質各領域のモデル予測課題の PC 結果．

Model	PC(%)			
	Cerebral Cortex	Visual	Auditory	Linguistic
albert-xlarge-v1 with VLA	11.22	13.00	13.23	13.15
albert-xlarge-v1 with FLA	10.57	10.37	10.64	10.48
albert-xlarge-v2 with VLA	8.77	9.98	9.97	9.70
albert-xlarge-v2 with FLA	1.71	10.39	9.96	9.85
BrainLM1.0 with VLA	8.99	9.23	8.44	9.42
BrainLM1.0 with FLA	10.04	10.43	10.15	9.99
BrainLM2.0 with VLA	**13.44**	**13.94**	**13.51**	**13.30**
BrainLM2.0 with FLA	11.78	11.56	11.49	11.24

BrainLM の転移可能性を調査し，特に Little Prince データセットを使用した脳信号予測に焦点を当てた．英語話者の被験者から脳活動データを選択し，3 回実施された実験のうち最初の 2 回を 4:1 の比率で訓練と検証用に分割した．その後，5 分割交差検証を使用してリッジ回帰モデルを訓練した．さらに，最後の実験からランダムに 10%のデータをテストセットとして割り当て，脳データの予測性能を評価した．検証手続きにはピアソン相関係数を使用し，False Discovery Rate（FDR）アルゴリズムによる統計的有意性が P 値 ≤ 0.05 であることを確認した．

すべてのコード実装は Pytorch フレームワーク[6] を使用し，分散並列計算のために 8 台の Quadro RTX 8000 GPU を使用した．

ファインチューニング結果

4 つの選択されたモデルそれぞれに対して，2 つのデータペアリングアプローチを適用し，合計 8 つの制御グループを予測タスクに用いた（表 3 の第 1 列参照）．この分析を容易にするため，Freesurfer ツール[7] を使用して被験者の脳の皮質モデルを構築し，Pycortex を用いてデータマッピングを実施した．その結果，約 250,000 ~ 300,000 のデータポイントが得られた．その後，医療文献 [12, 13, 14] に基づき，視覚，聴覚，言語処理の各ドメイン内の脳の関心領域（ROIs）を選択的に特定した．

結果は表 3 に示されており，BrainLM2.0（VLA）が全体的に最も高い予測関連性を示していることが分かる．この観察結果は，BrainLM が異なるデータセット間で脳活動の関連性を予測する能力を有しており，データセットの境界を越えたテキスト入力から関連する脳特徴を抽出する能力を持つことを裏付けるものである．この結果から，VLA が FLA を包括的に上回っていることが分かる．不均等なデータセグメンテーションアプローチによって得られるデータペアリングが，実際の脳活動および刺激の基礎的パターンとより密接に一致していることを示している．これらの結果に基づき，後続の検証タスクではデフォルトで注釈に基づいてテキストと脳データを整合させる方法を採用した．

Model	Pearson correlation coefficient(%)			
	Cerebral Cortex	Visual	Auditory	Linguistic
albert-xlarge-v1 with VLA	11.22	13.00	13.23	13.15
albert-xlarge-v1 with FLA	10.57	10.37	10.64	10.48
albert-xlarge-v2 with VLA	8.77	9.98	9.97	9.70
albert-xlarge-v2 with FLA	1.71	10.39	9.96	9.85
BrainLM1.0 with VLA	8.99	9.23	8.44	9.42
BrainLM1.0 with FLA	10.04	10.43	10.15	9.99
BrainLM2.0 with VLA	**13.44**	**13.94**	**13.51**	**13.30**
BrainLM2.0 with FLA	11.78	11.56	11.49	11.24

表 3 Pearson correlation coefficient results for model prediction tasks for an English-speaking subject in various functional regions of the cerebral cortex.

さらに、図 2 に示されているように、転移学習を適用する前後の予測結果を折れ線グラフとして可視化した。転移学習を適用した後、英語データセットにおける SOTA モデルの予測パフォーマンスが全体的に向上していることが確認された。この観察結果は、多言語学習タスクに従事するモデルにおいて、

5） https://openneuro.org/datasets/ds003643/versions/2.0.5
6） https://pytorch.org/
7） https://surfer.nmr.mgh.harvard.edu/

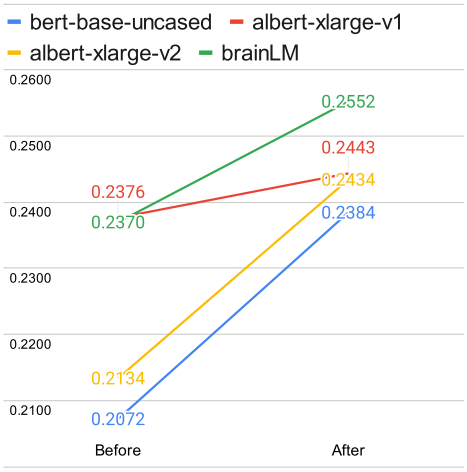

図 2　転移学習の前後における脳活動予測のための脳言語領域 ROI 周辺のピアソンの相関係数の結果.

転移学習が有益な影響を及ぼすことを強調している。要約すると、転移学習の適用は、さまざまな指標にわたって肯定的な結果をもたらし、モデル性能の向上においてその有効性を示している。

4　結果分析

多言語転移学習タスク

転移学習の結果は，150 の脳領域（ROIs）の頂点に適用したリッジ回帰の相関値として示されている (付録の図表 3 を参照). データは，視覚皮質の頂点から機能的 ROIs を統合し，2 次元の皮質平面上に投影して異なる被験者間の比較を可能にした．赤はより強い相関を，青はより弱い相関を示している．また，黄色と緑の矢印は英語（EN）から中国語（ZH）への転移学習を表している．比較は，SUB_FR057（フランス語ユーザー）と SUB_CN003（中国語ユーザー）に焦点を当て，英語からフランス語（FR）および中国語（ZH）への転移学習の予測結果を分析している．英語から中国語への転移学習も実施した．その結果，BrainLM が予測した脳活動信号と実際の信号には正の相関が見られた．枕葉皮質は 3 つの言語グループ全体で一貫して強い相関を示し，特に Pole_occipital-lh 領域で顕著だった．しかし，英語とフランス語のグループと比較して，中国語グループはこれらの領域で有意に高い相関を示した．この結果は，中国語の文字の複雑さがその視覚処理の要求を高め，これがこれらの領域におけるより強い脳応答を引き起こしたことを示唆する．言語処理は純粋に聴覚的または言語的なものではなく，さまざまな感覚および認知モダリティから情報を統合する必要がある．このモデルは，視覚皮質がマルチリンガル情報の処理において重要な役割を果たしていることを示している．

左右半球の機能差異の詳細検証

モデルの層間関係と脳 ROIs の研究結果は付録の図 4 および図 5 に示される．EN（英語）のベースラインは，各機能領域の異なる層で多様な主導効果を示している．転移学習（FR および ZH）後，この多様性はより集中し，主導効果は主に浅層と最終層にシフトした．EN ベースラインでは，異なる層が情報処理で異なる役割を果たしているが，転移学習はモデルを微調整し，中間層のタスク特化表現を削減する．この結果，浅層の一般的な特徴と最終層の高次セマンティック出力により多く依存するようになる．浅層，特に第一層は 3 つの言語全体で主導的な役割を果たしており，多くの脳領域が情報流の初期段階で基本的な信号を処理していることを示している．これは，初期言語信号のデコードや音声と語彙の認識など，脳の低次知覚プロセスを反映する．視覚関連領域はモデルの最終層と強く関連しており，視覚的な知覚とセマンティック統合における役割を示している．この結果は，視覚皮質が多言語情報処理で重要な役割を果たしていることを示唆する．

この分析から，転移学習によって BrainLM を多言語フレームワークに適応させ，英語，フランス語，中国語における脳活動の予測能力が大幅に向上したことがわかる．特に非英語言語において，予測結果と実際の脳応答との相関が顕著に改善された．中国語とフランス語の予測精度は重要な機能脳領域で顕著に向上し，転移学習が多言語神経信号をよりよく捉えるためのモデルの感度を向上させることを示している．

5　おわりに

転移学習の実施を通じて, マルチモーダル言語モデルを英語フレームワークから多言語環境へ効果的に変換することに成功した. 3 つの異なる実験を通じて, BrainLM モデルがさまざまな被験者やデータセットに対して持つ堅牢性を実証し, 大規模言語モデルの層と脳応答信号との関係や傾向を定量的に探究した. 本研究が進展する中で, BrainLM は脳のエンコーディングおよびデコーディングプロセスの背後にある複雑なメカニズムを解明し続けると期待される.

謝辞

本研究は、2022 年度内閣府国立大学イノベーション創出環境強化プロジェクトおよび日本学術振興会（JSPS）の支援を受けて行われた．これらの組織に対し、心より感謝の意を表します．（https://research-er.jp/projects/view/1226696）

参考文献

[1] Alfredo Ardila, Byron Bernal, and Monica Rosselli. How localized are language brain areas? a review of brodmann areas involvement in oral language. **Archives of clinical neuropsychology : the official journal of the National Academy of Neuropsychologists**, Vol. 31, , 12 2015.

[2] Satoshi Nishida, Antoine Blanc, Naoya Maeda, Masataka Kado, and Shinji Nishimoto. Behavioral correlates of cortical semantic representations modeled by word vectors. **PLOS Computational Biology**, Vol. 17, No. 6, pp. 1–35, 06 2021.

[3] Jiaxin Wang, Kiichi Kawahata, Antoine Blanc, Naoya Maeda, Shinji Nishimoto, and Satoshi Nishida. Asymmetric representation of symmetric semantic information in the human brain. **bioRxiv**, 2024.

[4] Yuko Nakagi, Takuya Matsuyama, Naoko Koide-Majima, Hiroto Yamaguchi, Rieko Kubo, Shinji Nishimoto, and Yu Takagi. The brain tells a story: Unveiling distinct representations of semantic content in speech, objects, and stories in the human brain with large language models. **bioRxiv**, 2024.

[5] Ying Luo and Ichiro Kobayashi. Brainlm: Estimation of brain activity evoked linguistic stimuli utilizing large language models. In **Proceedings of the The 2023 IEEE Conference on Systems, Man, and Cybernetics**, pp. 1904–1909, Hawaii, U.S.A., 10 2023.

[6] Eve Higby, Jungna Kim, and Loraine Obler. Multilingualism and the brain. **Annual Review of Applied Linguistics**, Vol. 33, , 03 2013.

[7] Sabrina Stehwien, Lena Henke, John Hale, Jonathan Brennan, and Lars Meyer. The little prince in 26 languages: Towards a multilingual neuro-cognitive corpus. In Emmanuele Chersoni, Barry Devereux, and Chu-Ren Huang, editors, **Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources**, pp. 43–49, Marseille, France, May 2020. European Language Resources Association.

[8] Jixing Li, Shohini Bhattasali, Shulin Zhang, Berta Franzluebbers, Wen-Ming Luh, R. Nathan Spreng, Jonathan R. Brennan, Yiming Yang, Christophe Pallier, and John Hale. Le petit prince: A multilingual fmri corpus using ecological stimuli. **bioRxiv**, 2021.

[9] Maddineni Bhargava, Karthika Vijayan, Oshin Anand, and Gaurav Raina. Exploration of transfer learning capability of multilingual models for text classification. In **Proceedings of the 2023 5th International Conference on Pattern Recognition and Intelligent Systems**, PRIS '23, p. 45–50, New York, NY, USA, 2023. Association for Computing Machinery.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[11] Shohini Bhattasali, Jonathan Brennan, Wen-Ming Luh, Berta Franzluebbers, and John Hale. The alice datasets: fMRI & EEG observations of natural language comprehension. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Twelfth Language Resources and Evaluation Conference**, pp. 120–125, Marseille, France, May 2020. European Language Resources Association.

[12] Zhaowen Liu, Edmund T Rolls, Zhi Liu, Kai Zhang, Ming Yang, Jingnan Du, Weikang Gong, Wei Cheng, Fei Dai, He Wang, Kamil Ugurbil, Jie Zhang, and Jianfeng Feng. Brain annotation toolbox: exploring the functional and genetic associations of neuroimaging results. **Bioinformatics**, Vol. 35, No. 19, pp. 3771–3778, 03 2019.

[13] Nancy Kanwisher. Functional specificity in the human brain: A window into the functional architecture of the mind. **Proceedings of the National Academy of Sciences**, Vol. 107, No. 25, pp. 11163–11170, 2010.

[14] Jeffrey R. Binder, Julie A. Frost, Thomas A. Hammeke, Robert W. Cox, Stephen M. Rao, and Thomas Prieto. Human brain language areas identified by functional magnetic resonance imaging. **Journal of Neuroscience**, Vol. 17, No. 1, pp. 353–362, 1997.

5.1 付録 (Appendix)

3 で、選択された ROIs は以下の通り：

1. 視覚 ROIs: Cuneus, Fusiform Gyrus, Inferior Parietal Lobule (IPL), Inferior Temporal Gyrus/Cortex (IT), Lateral Occipital Cortex (LOC), Lingual/Limbic Gyrus (Medullary Gyrus), Middle Temporal Gyrus, Middle Temporal Visual area, Orbital part of Inferior Frontal Gyrus (pars Orbitalis Pericalcarine Cortex), Superior Temporal Gyrus (STG), Supramarginal Gyrus.
2. 聴覚 ROIs: Inferior Parietal Lobule (IPL), Middle Temporal Gyrus, Superior Temporal Gyrus (STG), Supramarginal Gyrus, Transverse Temporal Gyrus (TTG).
3. 言語 ROIs: Entorhinal Cortex (EC), Inferior Parietal Lobule (IPL), Parahippocampal Gyrus, Pars Opercularis, Pars Triangularis, Superior Temporal Gyrus (STG), Supramarginal Gyrus.

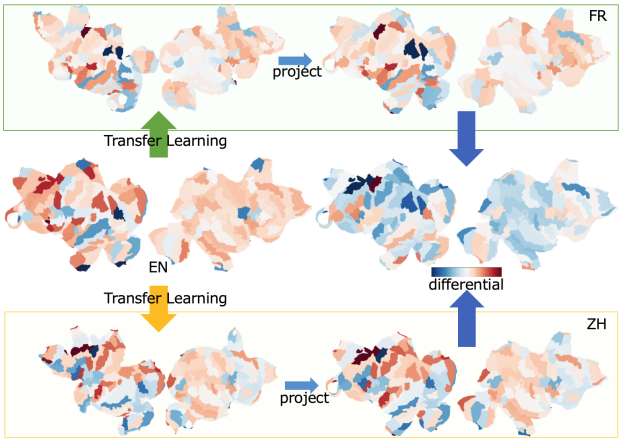

図 3 転移学習における 3 言語の BrainLM の PC 値. 図の詳細：多言語転移学習の結果を示しています. データは，大脳皮質の頂点にある 150 の脳 ROI に適用されたリッジ回帰の PC 値です． 赤は相関が強いことを示し，青は反対を示します． 黄色と緑の矢印は，英語から中国語への転移学習を表しています． 大脳皮質は機能的 ROI に基づいて統合され，異なるユーザー間の領域を比較するために 2 次元の皮質平面に投影されます．比較では，SUB_FR057 と SUB_CN003 の差異に焦点を当て，英語（EN）をベースとしたフランス語（FR）と中国語（ZH）への転移学習の予測結果を分析しています.

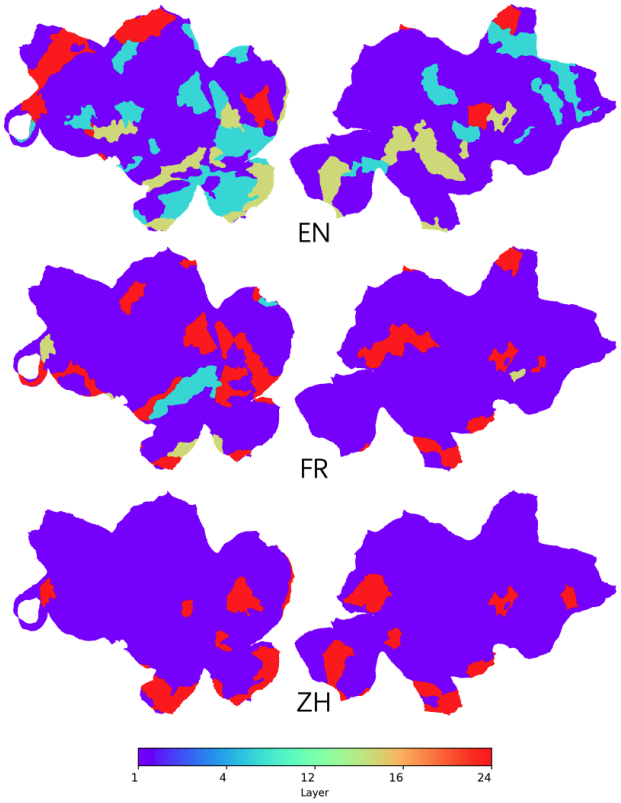

図 4 BrainLM によって異なる層で予測された脳活動の PC 値. 各言語で最も高い相関を示す 3 つの脳活動領域は：英語：G_occipital_sup-lh (0.1739), S_oc_middle&Lunatus-lh(0.1719), Pole_occipital-rh (0.1558). フランス語：S_oc-temp_lat-lh (0.2360), Pole_occipital-lh (0.2094), G_oc-temp_lat-fusifor-lh (0.2087). 中国語：Pole_occipital-lh (0.3575), S_suborbital-lh (0.3505), G_occipital_sup-rh (0.3184).

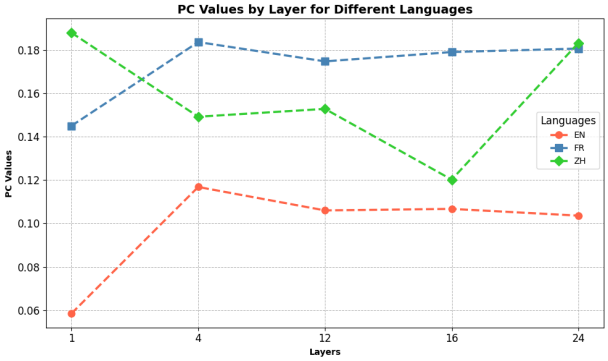

図 5 3 言語（EN（英語）, FR（フランス語）, ZH（中国語））の 5 つの異なるレイヤー（1, 4, 12, 16, 24）における PC 値. バーは言語別にグループ化され, 各色は異なるレイヤーを表す.