クロスコーダーを用いた脳と言語モデルにおける内部表現の特徴量比較

青木洸士郎　濵田偉月　折田奈甫　河原大輔　酒井弘
早稲田大学
aokikoshiro@akane.waseda.jp　itsukiku@ruri.waseda.jp
{orita,dkw,hsakai}@waseda.jp

概要

脳と言語モデルが言語処理においてどのような表現を共有し，どのように異なるかを明らかにすることは，神経科学と自然言語処理の双方にとって重要な課題である．従来の研究では，言語モデル（LM）の内部表現から脳応答をある程度予測できることから両者は共通した情報を符号化していると主張されてきた．しかし，それはどのような共通点なのか，逆にどのような点で異なるのかは十分に明らかにできていなかった．本研究では，LM の解釈可能性研究で用いられているクロスコーダーを脳応答と LM 表現に拡張することで，解釈可能な共通特徴量，脳優位特徴量，LM 優位特徴量を抽出することを目的とする．fMRI データを用いた実験の結果，場所に関する特徴量は脳と LM に共通して表現されている一方，負の情動は脳優位に，口語的な表現は LM 優位に表現されていることを特定した．本研究は，脳と LM の類似点と相違点を調べる枠組みを提供し，両分野の相互の発展に貢献する．

1　はじめに

脳と言語モデル（LM）がどのような情報処理を共有し，どこで異なるかを明らかにすることは，神経科学と自然言語処理の両分野にとって重要な問いである．言語刺激に対する脳応答を予測するエンコーディングモデルは，言語刺激が脳内でどのように符号化されているかを検証する枠組みとして発展してきた [1, 2]．文脈を考慮した LM の内部表現（LM 表現）が文脈非依存の埋め込みより高い予測性能を示すことが報告されて以降 [3]，LM 表現を用いたエンコーディング研究が数多く行われ，脳と LM の表現の類似性が主張されている [4, 5, 6, 7]．しかし，LM 表現を使った従来のエンコーディングモデルは，類似性の背後にある要因を解釈可能な特徴量として抽出することが難しい．

本研究では，解釈可能性研究において複数のモデルや層の共通／固有特徴量を同定するために提案されたクロスコーダー [8, 9] を脳応答と LM 表現の比較に拡張した **Brain-LM クロスコーダー**を提案する．この枠組みにより，従来のエンコーディングモデルでは捉えられなかった，脳応答と LM 表現の間の解釈可能な共通特徴量を同定することを試みる．さらにこの枠組みは，主に脳と LM の類似性を主張していたエンコーディングモデルとは異なり，両者の相違点として脳応答と LM 表現のどちらかに優位な特徴量も同定できることを示す．

fMRI データセットを用いた実験の結果，場所に関する特徴量が両者に共通し，負の情動に関する特徴量が脳側で優位に表現され，口語的な表現に関する特徴量が LM 側で優位に表現されていることを特定した．さらに，脳デコーダーの重みを皮質マップ上に投影することで，各特徴量がどの脳領域において表現されているかを可視化し解釈した．

2　関連研究

2.1　脳と言語モデルの相違

我々の研究の目的の一部である脳応答と LM 表現の相違点の分析を行っている研究として Zhou ら [10] の研究がある．この研究は，エンコーディングモデルの予測誤差が大きい単語を分析することで，LM は社会的・感情的および物理的知識を捉えにくいことを示した．しかし，この研究は LM が持っていない脳優位な特徴量は分析できるが，脳応答が符号化していない LM 優位の特徴量は検出できない．これは，エンコーディングモデルが LM 表現から脳応答への一方向の写像を学習することに起因

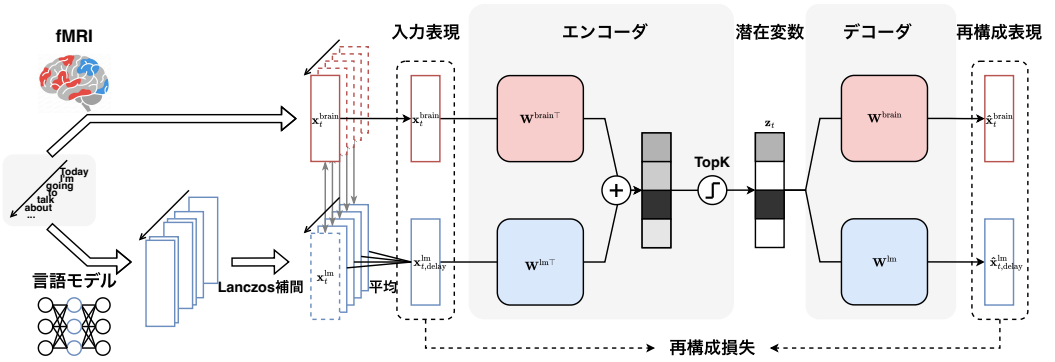


図 1　手法の概要．クロスコーダーを用いて脳応答と LM 表現から共通のスパースな潜在変数を学習する．

する．本研究では，脳応答と LM 表現を対照的に扱うことで，脳応答に優位な相違点だけでなく，LM に優位な相違点も同時に特定することを目指す．

2.2　辞書学習

辞書学習は，観測データをいくつかの基底の疎な線形結合として表現する手法である [11, 12]．神経科学においては，fMRI データから安静時機能的ネットワークや脳領域をデータ駆動で抽出するために応用されてきた [13, 14, 15, 16, 17]．LM の解釈可能性に関する研究では，LM の内部表現から単義的で解釈可能な特徴量を得る手法として，辞書学習の一種であるスパースオートエンコーダー（SAE）が使われている [18, 19]．近年は，クロスコーダーと呼ばれる，複数層や複数モデルの差分を特定するための SAE [8, 9, 20] など，手法の拡張も進んでいる．これらの技術はこれまで脳データまたは LM のいずれか一方の領域のみで適用されてきたが，本研究では両者の間に適用し，表現間の対応関係を調査する．

3　手法

図 1 に示すように，本手法では，脳応答と LM 表現を同一のスパースな潜在空間にエンコードすることで，両者に共通する特徴量と，それぞれに優位な特徴量を同一の枠組みで抽出する．

3.1　表現の取得と前処理

脳応答　脳応答には，被験者が文章を聴取している際に fMRI で計測された BOLD 信号を用いる．

言語モデル表現　LM 表現には，同じ文章を入力したときの LM の特定の層の活性値を用いる．

リサンプリングによる時間方向の整合　脳応答は一定時間の Repetition Time（TR）ごとに得られるのに対し，LM 表現はトークン単位で得られるため，そのままでは同一の時間軸で比較することが難しい．そこで先行研究 [2] と同様に，Lanczos 補間によりトークン列から連続時間の表現系列を構成し，TR ごとの LM 表現をサンプリングする．これにより，時刻 t における脳応答 $\mathbf{x}_t^{\text{brain}} \in \mathbb{R}^{d_{\text{brain}}}$ に対応する LM 表現 $\mathbf{x}_t^{\text{lm}} \in \mathbb{R}^{d_{\text{lm}}}$ が得られる．ここで，d_{brain} は fMRI のボクセル数であり，d_{lm} は LM の埋め込みの次元数である．

遅延の考慮　fMRI で計測される BOLD 信号は神経活動の即時的な反応ではなく，数秒にわたって広がる血流動態応答関数の畳み込みとして観測される．この遅延を考慮するために，時刻 t の脳応答 $\mathbf{x}_t^{\text{brain}}$ に対して，いくつかの t より前の時刻における LM 表現を平均した遅延 LM 表現 $\mathbf{x}_{t,\text{delay}}^{\text{lm}}$ を用いてモデルを学習する．

正規化　脳応答と遅延 LM 表現を対等に扱うため，それぞれ訓練データの平均 ℓ_2 ノルムで割ることで，訓練データにおける平均 ℓ_2 ノルムが 1 になるようにスケーリングして正規化する．検証データとテストデータに対しても，訓練データから求めた平均値を元にスケーリングする．

3.2　Brain-LM クロスコーダー

LM の解釈可能性においてモデル間の差分を解明する手法としてクロスコーダー [8, 9] が提案されているが，これまでは言語モデル間の実験しか行われていなかった．本研究では，これを脳応答と言語モデル表現に拡張した **Brain-LM クロスコーダー**を訓練する．

Brain-LM クロスコーダーは，時刻 t における脳応答 $\mathbf{x}_t^{\text{brain}}$ と遅延 LM 表現 $\mathbf{x}_{t,\text{delay}}^{\text{lm}}$ を共通の d_{latent} 次元の潜在空間に写像する．この表現に TopK 活性化関数を適用することで，スパースな潜在変数 $\mathbf{z}_t \in \mathbb{R}^{d_{\text{latent}}}$ を得る．この潜在変数 $\mathbf{z}_t$ の i 番目の次元の値は特徴

量 i の活性値に対応する．デコーダーは各表現に固有の重みを持ち，潜在変数から脳応答と遅延 LM 表現を再構成する．数式で表すと次のようになる．

$$\mathbf{z}_t = \mathrm{TopK}(\mathbf{W}^{\mathrm{brain}\top}\mathbf{x}_t^{\mathrm{brain}} + \mathbf{W}^{\mathrm{lm}\top}\mathbf{x}_{t,\mathrm{delay}}^{\mathrm{lm}} + \mathbf{b}^{\mathrm{enc}}) \quad (1)$$

$$\hat{\mathbf{x}}_t^{\mathrm{brain}} = \mathbf{W}^{\mathrm{brain}}\mathbf{z}_t = \sum_{i=1}^{d_{\mathrm{latent}}} z_{t,i}\mathbf{d}_i^{\mathrm{brain}} \quad (2)$$

$$\hat{\mathbf{x}}_t^{\mathrm{lm}} = \mathbf{W}^{\mathrm{lm}}\mathbf{z}_t = \sum_{i=1}^{d_{\mathrm{latent}}} z_{t,i}\mathbf{d}_i^{\mathrm{lm}} \quad (3)$$

ここで，$\mathbf{W}^{\mathrm{brain}} \in \mathbb{R}^{d_{\mathrm{brain}} \times d_{\mathrm{latent}}}$ と $\mathbf{W}^{\mathrm{lm}} \in \mathbb{R}^{d_{\mathrm{lm}} \times d_{\mathrm{latent}}}$ は，特徴量ベクトル $\mathbf{d}_i^{\mathrm{brain}} \in \mathbb{R}^{d_{\mathrm{brain}}}$ と $\mathbf{d}_i^{\mathrm{lm}} \in \mathbb{R}^{d_{\mathrm{lm}}}$ をそれぞれ列ベクトルとするデコーダーの重み行列である．エンコーダーの重みは，デコーダーの重みの転置を用いる（tied weights）．$\mathbf{b}^{\mathrm{enc}} \in \mathbb{R}^{d_{\mathrm{latent}}}$ はバイアス項である．TopK 活性化関数は，活性値の上位 k 個のみを残し，それ以外を 0 にする関数である．

損失関数は次のように定義する．

$$\mathcal{L} = \|\mathbf{x}_t^{\mathrm{brain}} - \hat{\mathbf{x}}_t^{\mathrm{brain}}\|_2^2 + \|\mathbf{x}_t^{\mathrm{lm}} - \hat{\mathbf{x}}_t^{\mathrm{lm}}\|_2^2 + \alpha\mathcal{L}_{\mathrm{aux}} \quad (4)$$

$\mathcal{L}_{\mathrm{aux}}$ は活性化しない特徴量（dead feature）を抑制するための補助損失であり [21]，α はその係数である（詳細は付録 B を参照）．

3.3 脳-LM 優位性の指標

特徴量 i が脳応答と LM 表現のどちらにより強く寄与するかを定量化するため，次の相対差分スコア r_i を定義する．

$$r_i = \frac{\|\mathbf{d}_i^{\mathrm{brain}}\|_2 - \|\mathbf{d}_i^{\mathrm{lm}}\|_2}{\max(\|\mathbf{d}_i^{\mathrm{brain}}\|_2, \|\mathbf{d}_i^{\mathrm{lm}}\|_2)} \quad (5)$$

特徴量ベクトルのノルム $\|\mathbf{d}_i^{\mathrm{brain}}\|_2$ および $\|\mathbf{d}_i^{\mathrm{lm}}\|_2$ を，それぞれ脳応答と遅延 LM 表現の再構成における潜在変数 i の寄与度とみなし，r_i はそれらの相対的な差異を表す．r_i は $[-1, 1]$ の範囲の値を取り，$r_i \approx 0$ のとき両者が同程度の寄与を示す．本研究では，$r_i > 0.5$ の潜在変数を脳優位，$r_i < -0.5$ の場合を LM 優位，$-0.2 < r_i < 0.2$ の場合を共通と分類する．

4 実験

4.1 実験設定

データセット　実験には，LeBel ら [22] が公開した fMRI データセットを使用した．このデータセットは，8 人の被験者が英語のポッドキャストを聴取している時の fMRI（TR=2s）で計測された BOLD 信号を収録したものである．この BOLD 信号を脳応答として使用した．本研究では，LeBel ら [22] の実験結果に基づき，最も品質の高い [1] 被験者（UTS03）のデータを使用した．ポッドキャストにはいくつかのストーリーがあり，本研究では訓練データとして 58 本，検証データとして 8 本，テストデータとして 17 本のストーリーを用いた．

言語モデル　Llama-3.1-8B-Instruct [23] を使用して実験を行った．この LM は 32 層の Transformer Block を持ち，中間の 16 層目の Transformer Block の出力（residual stream）を LM 表現として使用した．

ハイパーパラメータ　潜在変数の次元数 d_{latent} は 1,024，遅延の平均に用いる時間は-2, -4, -6, -8 秒，TopK 活性化の k は 64，学習エポック数は 300 エポック，バッチサイズは 2,048，学習率は 1×10^{-4} とし，Adam [24] を用いて最適化した．

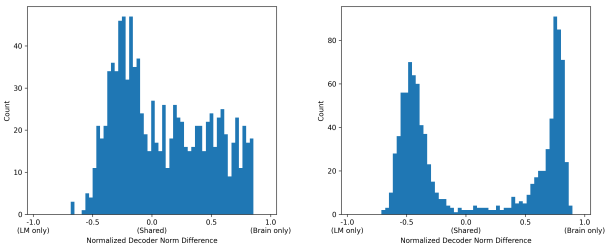
4.2 再構成性能

本手法で得られる潜在変数が脳応答と遅延 LM 表現をどの程度正確に再構成できているかを評価する．評価指標には，脳応答と遅延 LM 表現のボクセルまたは次元ごとに計算した，実測値と再構成値の間のピアソン相関係数を用いる．全次元を平均した脳応答と遅延 LM 表現の平均相関係数はそれぞれ 0.44 と 0.52 であった．この脳応答の平均相関係数は，入力に LM 表現のみを用いるエンコーディングモデルの先行研究 [22, 25] で報告されている平均相関係数（~0.1）よりも高い．これは，LM 表現のみでは説明できない脳応答の特徴量が存在することを示唆する．ボクセルごとに脳応答の相関係数を大脳皮質上に投影した結果は付録 C に示す．

4.3 相対差分スコアの分布

図 2(a) に相対差分スコア r_i のヒストグラムを示す．共通特徴量（$-0.2 < r_i < 0.2$）に分類される特徴量の割合は 24%であった．これらの共通特徴量が，同一の言語刺激に対する脳と LM の共通した表現を捉えているかを検証するため，脳応答と LM 表現の対応関係を除去するアブレーション実験を行った．具体的には，脳応答の時刻をシャッフルすることで得られる，異なる言語刺激に対する脳応答と LM 表現のペアからなるデータセットを用いてモデルを学習した．データセット以外の設定は変更しない．その結果，図 2(b) に示すように，r_i の分布は二峰性を

1) ここで，品質が高いとは，頭の動きが小さく，再現性（同じ刺激に対する同じボクセルの反応の類似性）が高く，エンコーディングモデルの性能が高いことを意味する [22].



(a) 刺激対応データ　　(b) 刺激非対応データ

図 2　相対差分スコア r_i の分布.

表 1　代表的な特徴量とその解釈.

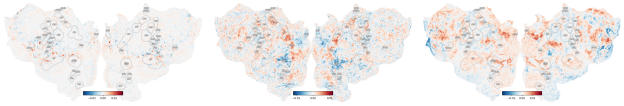
インデックス	r_i	分類	解釈
694	−0.01	共通	場所
1013	0.80	脳優位	負の情動
462	−0.67	LM 優位	口語的な表現

示し，共通特徴量の割合は 4%に減少した．この結果は，刺激対応データで得られた共通特徴量が，同一の言語刺激に対して脳と LM が共通して表現する情報を反映していることを示唆する．

4.4　発見された解釈可能な特徴量

各特徴量の意味を解釈するため，以下の手順で特徴量の代表文章を取得する．まず，テストデータのストーリーごとに特徴量の活性値が最大となる時点を特定し，その時刻の 8 秒前からその時刻までに現れた文章を取得する．この文章を最大活性化文章と呼ぶ．次に，すべてのストーリーから得られたこれらの文章の中で活性値の上位のものを用いて特徴量がどのような意味を表現するかを解釈した．表 1 に共通特徴量，脳優位特徴量，LM 優位特徴量の具体例を示す（代表文章の具体例は付録 D を参照）．これらの特徴量は，相対差分スコアがそれぞれ 0 付近，特に大きい，特に小さい特徴量の中から解釈可能な特徴量を選んだ．ただし，特に脳優位特徴量において，代表文章からの解釈が難しい特徴量も多く存在したことを明記しておく．これらの特徴量の解釈については § 4.5 で述べる．

共通特徴量の例として，特徴量 694（$r_i = -0.01$）は，場所や情景の描写に関する文章で活性化した．脳優位特徴量では，特徴量 1013（$r_i = 0.80$）が，葛藤や抑うつ，復讐心といった負の情動に関わる文章で活性化した．LM 優位特徴量の例として，特徴量 462（$r_i = -0.67$）は，口語的な定型表現，言いよどみや修正が多い文章で活性化した．これは，LM の学習コーパスの多くは書き言葉であり，口語は LM にとって特異的に表現されているからと考えられる．



(a) 代表文章からは解釈困難であった脳優位特徴量（特徴量 288）　(b) 場所に関連する共通特徴量（特徴量 694）　(c) 負の情動に関連する脳優位特徴量（特徴量 1013）

図 3　皮質マップ上の脳デコーダーの重み.

4.5　皮質マップ上の重みの可視化と解釈

脳デコーダーの重み $\mathbf{d}_i^{\text{brain}}$ はボクセル空間上のベクトルであり，特徴量 i がどの脳領域に寄与するかを可視化できる．

代表文章からは解釈困難であった脳優位特徴量の重みは図 3(a) のような縞模様を示した．これは，マルチバンド fMRI でみられるアーティファクトに由来すると考えられる [26].

場所に関する共通特徴量（特徴量 694）は，頭頂間溝（IPS）および横後頭溝（TOS）で相対的に大きな重みが観察された（図 3(b)）．これらの脳領域は情景処理に関与する領域であることが知られており [27, 28]，特徴量の解釈と先行研究の知見は整合的であることがわかる．

負の情動に関する脳優位特徴量（特徴量 1013）は，脳梁膨大後部皮質（RSC）と頭頂間溝（IPS）の間の領域や，前頭前皮質（PFC）の一部で相対的に大きな重みが観察された（図 3(c)）．RSC は感情的な刺激によって活性化することが報告されている [29]．また，PFC は感情調節において重要な役割を果たし，負の感情の処理において活性化することが知られている [30, 31, 32]．本研究で負の情動に関する特徴量がこれらの領域と関連していたことは，これらの先行研究の知見を支持するものである．さらに，この特徴量が脳優位であったことは，LM が負の情動に関する情報を脳と同程度には表現していない可能性を示唆し，LM が社会的・感情的内容の脳応答を十分に説明できないことを報告した Zhou ら [10] の結果とも整合的である．

5　おわりに

本研究では，クロスコーダーを脳応答と言語モデルの内部表現の比較に拡張し，いくつかの解釈可能かつ神経科学的に妥当な共通特徴量と優位特徴量を特定した．複数被験者での検証や，表現への介入実験等を通じた因果関係の解明が今後の課題である．

謝辞

本研究は日本学術振興会科研費 JP23H05493 および JST CREST JPMJCR2565 の助成を受けたものである．また，本研究は東京科学大学のスーパーコンピュータ TSUBAME4.0 を利用して実施した．

参考文献

[1] Tom M. Mitchell, Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L. Malave, Robert A. Mason, and Marcel Adam Just. Predicting human brain activity associated with the meanings of nouns. **Science**, Vol. 320, No. 5880, pp. 1191–1195, 2008.

[2] Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. **Nature**, Vol. 532, No. 7600, pp. 453–458, 2016.

[3] Shailee Jain and Alexander Huth. Incorporating context into language encoding models for fmri. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 31. Curran Associates, Inc., 2018.

[4] Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. **Proceedings of the National Academy of Sciences**, Vol. 118, No. 45, p. e2105646118, 2021.

[5] Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. **Communications biology**, Vol. 5, No. 1, p. 134, 2022.

[6] Mariya Toneva and Leila Wehbe. **Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)**. Curran Associates Inc., Red Hook, NY, USA, 2019.

[7] Eghbal A. Hosseini, Martin Schrimpf, Yian Zhang, Samuel Bowman, Noga Zaslavsky, and Evelina Fedorenko. Artificial neural network language models predict human brain responses to language even after a developmentally realistic amount of training. **Neurobiology of Language**, Vol. 5, No. 1, pp. 43–63, 04 2024.

[8] Jack Lindsey, Adly Templeton, Jonathan Marcus, Thomas Conerly, Joshua Batson, and Christopher Olah. Sparse Crosscoders for Cross-Layer Features and Model Diffing. **Transformer Circuits Thread**, 2024.

[9] Julian Minder, Clément Dumas, Caden Juang, Bilal Chughtai, and Neel Nanda. Overcoming sparsity artifacts in crosscoders to interpret chat-tuning. In **The Thirty-ninth Annual Conference on Neural Information Processing Systems**, 2025.

[10] Yuchen Zhou, Emmy Liu, Graham Neubig, Michael J. Tarr, and Leila Wehbe. Divergences between language models and human brains. In **The Thirty-eighth Annual Conference on Neural Information Processing Systems**, 2024.

[11] Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. **Nature**, Vol. 381, No. 6583, pp. 607–609, 1996.

[12] M. Aharon, M. Elad, and A. Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. **IEEE Transactions on Signal Processing**, Vol. 54, No. 11, pp. 4311–4322, 2006.

[13] Harini Eavani, Roman Filipovych, Christos Davatzikos, Theodore D. Satterthwaite, Raquel E. Gur, and Ruben C. Gur. Sparse dictionary learning of resting state fmri networks. In **2012 Second International Workshop on Pattern Recognition in NeuroImaging**, pp. 73–76, 2012.

[14] Gael Varoquaux, Alexandre Gramfort, Fabian Pedregosa, Vincent Michel, and Bertrand Thirion. Multi-subject dictionary learning to segment an atlas of brain spontaneous activity. In Gábor Székely and Horst K. Hahn, editors, **Information Processing in Medical Imaging**, pp. 562–573, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[15] Alexandre Abraham, Elvis Dohmatob, Bertrand Thirion, Dimitris Samaras, and Gael Varoquaux. Extracting brain regions from rest fmri with total-variation constrained dictionary learning. In Kensaku Mori, Ichiro Sakuma, Yoshinobu Sato, Christian Barillot, and Nassir Navab, editors, **Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013**, pp. 607–615, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[16] Kangjoo Lee, Sungho Tak, and Jong Chul Ye. A data-driven sparse glm for fmri analysis using sparse dictionary learning with mdl criterion. **IEEE Transactions on Medical Imaging**, Vol. 30, No. 5, pp. 1076–1089, 2011.

[17] Elvis DOHMATOB, Arthur Mensch, Gael Varoquaux, and Bertrand Thirion. Learning brain regions via large-scale online structured sparse dictionary learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 29. Curran Associates, Inc., 2016.

[18] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. **Transformer Circuits Thread**, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

[19] Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In **The Twelfth International Conference on Learning Representations**, 2024.

[20] Harrish Thasarathan, Julian Forsyth, Thomas Fel, Matthew Kowal, and Konstantinos G. Derpanis. Universal sparse autoencoders: Interpretable cross-model concept alignment. In **Forty-second International Conference on Machine Learning**, 2025.

[21] Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In **The Thirteenth International Conference on Learning Representations**, 2025.

[22] Amanda LeBel, Lauren Wagner, Shailee Jain, Aneesh Adhikari-Desai, Bhavin Gupta, Allyson Morgenthal, Jerry Tang, Lixiang Xu, and Alexander G Huth. A natural language fmri dataset for voxelwise encoding models. **Scientific Data**, Vol. 10, No. 1, p. 555, 2023.

[23] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. The llama 3 herd of models. **arXiv preprint arXiv:2407.21783**, 2024.

[24] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, **3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings**, 2015.

[25] Yu Lei, Xingyang Ge, Yi Zhang, Yiming Yang, and Bolei Ma. Do large language models think like the brain? sentence-level evidences from layer-wise embeddings and fmri. **arXiv preprint arXiv:2505.22563**, 2025.

[26] Ludovica Griffanti, Gwenaëlle Douaud, Janine Bijsterbosch, Stefania Evangelisti, Fidel Alfaro-Almagro, Matthew F. Glasser, Eugene P. Duff, Sean Fitzgibbon, Robert Westphal, Davide Carone, Christian F. Beckmann, and Stephen M. Smith. Hand classification of fmri ica noise components. **NeuroImage**, Vol. 154, pp. 188–205, 2017. Cleaning up the fMRI time series: Mitigating noise with advanced acquisition and correction strategies.

[27] Daniel D Dilks, Joshua B Julian, Alexander M Paunov, and Nancy Kanwisher. The occipital place area is causally and selectively involved in scene perception. **Journal of neuroscience**, Vol. 33, No. 4, pp. 1331–1336, 2013.

[28] Tim J. Preston, Fei Guo, Koel Das, Barry Giesbrecht, and Miguel P. Eckstein. Neural representations of contextual guidance in visual search of real-world scenes. **Journal of Neuroscience**, Vol. 33, No. 18, pp. 7846–7855, 2013.

[29] Richard J Maddock. The retrosplenial cortex and emotion: new insights from functional neuroimaging of the human brain. **Trends in neurosciences**, Vol. 22, No. 7, pp. 310–316, 1999.

[30] Jun Zhao, Licheng Mo, Rong Bi, Zhenhong He, Yuming Chen, Feng Xu, Hui Xie, and Dandan Zhang. The vlpfc versus the dlpfc in downregulating social pain using reappraisal and distraction strategies. **Journal of Neuroscience**, Vol. 41, No. 6, pp. 1331–1339, 2021.

[31] Aaron S. Heller, Tom Johnstone, Michael J. Peterson, Gregory G. Kolden, Ned H. Kalin, and Richard J. Davidson. Increased prefrontal cortex activity during negative emotion regulation as a predictor of depression symptom severity trajectory over 6 months. **JAMA Psychiatry**, Vol. 70, No. 11, pp. 1181–1189, 11 2013.

[32] Mo Yang, Shang-Jui Tsai, and Chiang-Shan R Li. Concurrent amygdalar and ventromedial prefrontal cortical responses during emotion processing: a meta-analysis of the effects of valence of emotion and passive exposure versus active regulation. **Brain Structure and Function**, Vol. 225, No. 1, pp. 345–363, 2020.

A　限界と今後の展望

本研究は一人の被験者のデータを用いた分析に限定されている．個人差の影響を検討するためには，複数の被験者を対象とした分析が必要である．また，本研究で観察された共通特徴量は相関に基づくものであり，因果関係を示すものではない．脳と LM の表現が類似しているからといって，両者が同じ計算処理を行っているとは限らない．因果的な関係を検証するためには，介入実験などが必要である．fMRI のアーティファクトが脳優位特徴量として検出されたが，これらのノイズは全変動（Total Variation）を基準に特定できると考えられる．

B　補助損失の詳細

TopK 活性化関数を用いる場合，一部の特徴量が学習中に一度も活性化しなくなる問題が生じうる．このような特徴量は dead feature と呼ばれ，モデルの表現能力を低下させる．Brain-LM クロスコーダーの訓練では，Gao ら [21] および Minder ら [9] に従い，dead feature を抑制するための補助損失を導入する．

補助損失の計算手順は次のとおりである．1 エポック内で一度も活性化していない特徴量の中から k_{aux} 個をランダムに選択し，選択された dead feature のみを用いて残差誤差を再構成する．この再構成値と残差誤差との間の二乗誤差を補助損失とし，次のように表せる．

$$\mathcal{L}_{\text{aux}} = \|\mathbf{e}_t^{\text{brain}} - \hat{\mathbf{e}}_t^{\text{brain}}\|_2^2 + \|\mathbf{e}_t^{\text{lm}} - \hat{\mathbf{e}}_t^{\text{lm}}\|_2^2 \tag{6}$$

ここで，$\mathbf{e}_t^{\text{brain}} = \mathbf{x}_t^{\text{brain}} - \hat{\mathbf{x}}_t^{\text{brain}}$ および $\mathbf{e}_t^{\text{lm}} = \mathbf{x}_t^{\text{lm}} - \hat{\mathbf{x}}_t^{\text{lm}}$ は，それぞれ脳応答と LM 表現の残差誤差である．$\hat{\mathbf{e}}_t^{\text{brain}}$ および $\hat{\mathbf{e}}_t^{\text{lm}}$ は，選択された dead feature のみを用いた残差誤差の再構成値である．この補助損失により，dead feature が残差を説明できるように学習が促され，特徴量の利用効率が向上する．先行研究 [21, 9] に従い，$k_{\text{aux}} = 512$，補助損失の係数 $\alpha = 1/32$ に設定した．

C　皮質マップ上の相関係数

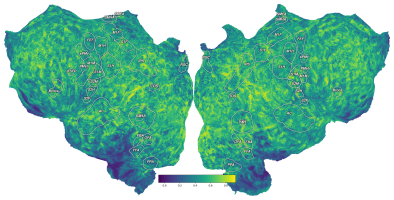


図 4　テストデータにおける各ボクセルの実測値と再構成値の相関係数の分布を皮質マップに投影した結果．

D　特徴量の代表文章の例

表 2 に各特徴量の代表文章の具体例を示す．

表 2　各特徴量の代表文章の具体例．

インデックス	分類	解釈	代表文章の例
694	共通	場所の描写	stop was that woodsy restaurant on the cliff’s edge where i had eaten with my family two decades ago and when i landed on little diomede as a wildlife biologist to work on a walrus study with the local people the good feelings you have when you win a prize then you put it back in storage but it’s different it has
1013	脳優位	負の情動	speed i’d gone from easygoing to confrontational when a friend called my personal lust for revenge against those first grade bullies didn’t last long at least not at that in my brain and also as time went on i was able to start remembering the the only side effect of being depressed that was working for me you know what i mean so
462	LM 優位	口語的な表現	uh oh my god but so i could tell chloe i’m like you know i’m really a little nervous and she goes no no no you’ll have no problem she’s says you know why don’t you make a a left turn and i’m in the far right lane i say mom that’s tha you know i can’t go i gotta lot of people who want to see you i’m like yeah i don’t want to go and you know i want to keep the focus on my dad i don’t want to be