大規模言語モデルを用いた言語刺激下の脳内意味表象解読

佐藤 杏奈 [1] 小林 一郎 [1]
[1] お茶の水女子大学
{g1920519,koba}@is.ocha.ac.jp

概要

脳情報解読技術は，神経活動を解釈して思考や感情を再現する手法として注目されている．Tang ら (2023) [1] は，fMRI データを使用した，言語モデルを生成モデルとして活用する新たな解読手法を提案した．本研究はその拡張として，先行研究で用いられた GPT モデルに加え 3 種類の言語モデルを導入し，精度比較を行った．結果として，高い解読精度には，使用する言語モデルが脳活動の予測に優れていることだけでなく，言語モデルが生成するテキストの種類も重要である可能性を明らかにした．

1 はじめに

脳活動をリアルタイムで解析し，思考や動作を解読する脳の解読技術は，医療，コミュニケーション支援など様々な分野で革新的な変化をもたらすと期待されている．侵襲的 BMI は ECoG などを用いて高精度なデータ解析が可能だが [2, 3]，脳手術のリスクや制約がある．一方，非侵襲的 BMI は fMRI や EEG を活用し安全でコストも低いが，ノイズや低解像度が課題で，実用化には多くの挑戦が残る [4]．そこで Tang ら [1] は，非侵襲的データから直接刺激へデコーディングを行うのではなく，言語モデル GPT [5] の生成を基に神経データを補助的に使用することで，より応用性の高い脳解読を試みた．この手法は事前に fMRI にて取得したデータを使用するオフラインの脳内情報解読であるが，その革新的なアプローチにより注目を集めている．

2 関連研究

Tang らは，非侵襲的に取得した fMRI データを用いて，被験者が聞いているまたは想像している刺激を自然言語で再構築するデコーダを提案した．概要を図 1 に示す ([1] より引用)．このデコーダは，言語モデルを用いて候補となる単語群を生成し，事前訓練した脳状態をモデリングする符号化モデルを用いて各候補から引き起こされる脳内状態を推定する．そして予測された脳状態と実際の脳状態が類似している単語候補を選択することで，低い時間分解能である fMRI データの弱みを克服し，被験者が聞いている文の再構築を可能にした．

符号化モデルは一般に，深層学習モデルから抽出される刺激を表す特徴量ベクトルから脳内状態を推定する．言語特徴量として近年は GPT-2 [6]，Llama [7] などの言語モデルの中間層の内部状態が文の特徴を表すベクトルとして多く脳内状態の推定に利用される [8, 9, 10]．Antonello ら [11] は，特徴量抽出に使用される言語モデルのパラメタ数が上がるにつれて符号化モデルの精度が対数線形的に向上するスケーリング法則を報告した．符号化モデルの性能は，作業モデルとして使用される言語モデルの能力と密接に関係しており，符号化モデル構築においてどの言語モデルを用いるかは重要な選択である．

本研究では，先行研究で使われた Fine-tuned GPT の追加学習前のモデルを導入し、追加学習によるデコーダの精度の変化を明らかにする．またより高精度な符号化モデルの構築を目指して，新しく強力な言語モデル Llama3，OPT を使用し，従来のデコーディング手法の有効性を確認する．

3 手法

3.1 連続した自然言語の意味再構成

本研究で構築するデコーダは Tang らにより導入された (図 1)．脳活動データは，一人の話し手により語られる複数の物語の音声刺激のもと，fMRI より取得された．まず自然言語刺激下の脳反応パターンを学習するために，言語モデルを用いて抽出された特徴量から言語刺激下の脳状態を推定する符号化モデルを構築する (図 1a)．デコーディングでは，大規模なテキストデータで訓練された言語モデルを使用して次に来る可能性のある単語を制限することで，被験者が聞いたり想像したりしていると考えら

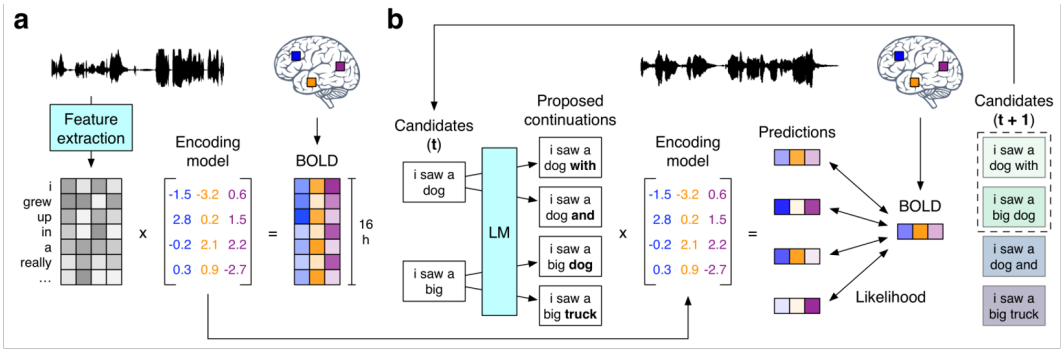

図 1: 言語刺激下の脳データによる文の再構築 ([1] より引用)．(a) 音声刺激として被験者に与えた単語列から，MRI 実験で取得された BOLD 反応を予測する符号化モデルを構築．訓練には 16 時間分のデータが使用された．(b) 言語モデルで次に来る可能性のある単語を出力，符号化モデル (3.4 節参照) を用いて候補の文から誘起される脳反応を予測し，実測脳反応と近い k 個の候補が次のタイムステップへ保持される．

れる刺激を効率よく特定することができる．デコーダではビームサーチが採用されており，それぞれの候補から推定された脳反応と，実際に計測された脳反応が最も近かった k 個の候補が次のタイムステップへ保持される（図 1b）．

3.2 MRI データ

本研究では，先行研究 [1] と同じ公開データセット [12] を使用する．MRI データは，シーメンス社 3T MRI を使用し，23 歳〜36 歳の健康な大人 3 名 (内女性 1 名) より取得された．刺激データセットは，*The Moth Radio Hour* と *Modern Love* から抽出された 5〜15 分の 82 ストーリーで，各ストーリーでは一人の話し手が自伝的な物語を語る音声刺激である．本研究では，LeBel ら [13] により前処理が施された fMRI データを使用する．テストデータは，訓練データと同じ条件で取得された “Where There’s Smoke”(10 分) を聴いたときの fMRI データを用いる．

3.3 言語モデル

本研究では，先行研究で使用された Fine-tuned GPT をベースラインとして用い，さらに他言語モデルでのデコーダ性能を確認するために，Pre-trained GPT，Llama3-8B，OPT-6.7B モデルを使用した (表 1，付録表 3)．Fine-tuned GPT は，Reddit のコメント (200 億語以上) と MRI 実験で使用されなかった *The Moth Radio Hour* と *Mordern Love* から抽出された自叙伝の物語 (40 万語以上) のコーパスで訓練された．本研究で追加した 3 つのモデルは，Hugging Face Hub で公開されている事前学習済モデルである．

表 1: 使用した言語モデル．FT は Fine-tuned，PT は Pre-trained を表す．

Model	Size	Training Data
FT GPT	120M	Reddit posts, autobiographical stories
PT GPT	120M	Unpublished books in various genres
PT Llama3	8B	Large public text datasets
PT OPT	6.7B	Books, story-like data, news, web text

3.4 符号化モデル

ヒト脳への刺激単語から抽出された特徴から，正則化線形回帰を使用して，各特徴が各ボクセル内の BOLD 信号にどのような影響を与えるかを予測する重みを学習する [14]．各トークンの特徴量は，前 5 トークンと対象トークンからなる単語列を言語モデルに与えたときの，対象トークンの隠れ状態が，それぞれ符号化モデルの特徴量として使用された．その後，得られたトークンごとの特徴量を Lanczos フィルタを用いて MRI 繰り返し時間 (TR) にダウンサンプリングする．また，刺激に対する BOLD 反応の遅延を考慮するために，1〜4TR[1)] 前の特徴量を結合して回帰する．符号化モデルの線形回帰には一般的に用いられるリッジ回帰を使用し，正則化項の係数 α は 10^1〜10^3 の中の 10 個の値から 50-fold 交差検証により各ボクセルごとに採用された．

3.5 トークン数予測モデル

各被験者についてトークン数予測モデルを推定し，単語を知覚または想像するタイミングを予測する．聴覚皮質に対応するボクセルの BOLD 信号から，時間 $t-1$ から t までに与えられたトークン数を

1) 1TR=2.0 秒

予測するモデルである．符号化モデルと同様に，刺激に対する BOLD 反応の遅延を考慮するため，1〜4TR 後の特徴量を結合して回帰する．各被験者における聴覚皮質は，20 秒間の音楽，スピーチ音声，自然音を含む 1 分間の聴覚刺激を 10 回繰り返し聴く，聴覚ローカライザタスクにより定義された [13].

3.6 類似度評価指標

被験者が想起する文をどれだけ再構築できているか評価するため，デコーダで再構築された文と実際の刺激文の類似度を測る．先行研究では，word error rate(WER)，BLEU，METEOR，BERTScore [15] で評価されたが，単語レベルではなく高次な意味的類似の評価をするために BERTScore[2] のみ継承した (その他指標の結果は付録図 5 参照)．また異なる評価指標として，近年クラスタリングや検索拡張生成 [17] などで幅広く使われる [18]，Embedding モデルを使用した文章間類似評価を導入する[3]．実際の刺激文とデコードされた文の埋め込みベクトルのピアソン相関係数を図り類似度を計算した．

文章間類似度の比較は，先行研究に倣って Window similarity と Story similarity で計測する．Window similarity は 20 秒のウィンドウ内にある単語列から類似度が計算され，Story similarity は Window similarity の平均によって計算される．

4 実験設定

4.1 デコーダ設定

生成モデルの候補単語生成方法として top-p サンプリングを用い，$k = 5$ のビームサーチでデコーディングを行う．デコーダにより生成された文章は，“He”，“I”，“It”，“She”，“They” のいずれかで始まるように設定した．デコーダには，各被験者において符号化モデルの交差検証で最も精度の良かった 10,000 個のボクセルを使用する．これら設定は先行研究と同様である．

内部モデルとして用いる，符号化モデルとトークン数予測モデルの精度を付録図 3, 図 4，表 4 に示す．デコーダで使用された符号化モデルは，テストデータを用いない事前実験において最も予測精度が高かった層 (FT GPT：9 層，PT GPT：10 層，Llama3：13 層，OPT：22 層) の隠れ状態を使用した．

2) BERTScore には，DeBERTa [16] xlarge が使用された．
3) 本研究では OpenAI の text-embedding-3-small が使用された．

4.2 統計的検定

生成された文が有意に高いスコアを持つのかを確かめるため，デコーダで使用した同じ言語モデルから，脳活動を用いずに出力させた 300 文を用いて同様にスコアを計測した．各 300 文と実際の刺激文の類似度を測ることで帰無分布を構築し，デコーダは脳内状態を反映する文を再構築できないという帰無仮説のもと検定を行う．ここで p 値は，脳活動を使用しない 300 文の内デコーダにより生成された文以上のスコアを持つ文の割合と定義され，false discovery rate(FDR) を使用して多重補正された．

4.3 実験結果

Story similarity(図 2a) は，デコードされた文全体が実際の刺激文と有意に似ているかを示す．**Chance** として描かれた帰無分布は各言語モデルにより生成された文から構成されており，言語モデルにより異なる分布を持つ．全言語モデル，全被験者において偶然レベルより有意に実際の刺激と類似した文を再構成できた (q(FDR)<0.05)．Window simlarity(図 2b) は，各タイムポイントでデコードされた文が実際の刺激文と有意に似ているかを示す．BERTScore，Embedding モデルによる評価共に，全言語モデル多くのタイムポイントで有意なスコアを確認できた．

実際に被験者が聞いた文と，それぞれのデコーダにより生成された文の一部を表 2 に示す．Llama3 や OPT を用いたデコーダでは、大文字小文字が区別されていたり、記号が含まれたり，よりリッチな文章が生成されるが，評価ではテストデータセットの表記法に準じすべてのテキストを小文字に統一し，記号 (アポストロフィーを除く) を除去する前処理を行った．どの言語モデルにおいても，強調された一部分で実際の刺激文と似た意味を持つ単語列が再構成されているを確認できた．

5 考察

Story similarity の結果 (図 2(a)) より，BERTScore と Embedding モデルの評価共に，先行研究で使用された Fine-tuned GPT のデコーダが，僅かに他 3 つの言語モデルのデコーダよりも高いスコアを持つことがわかる．これは，Fine-tuned GPT の訓練データセットに実際の刺激文と同じ (実験では使用されていない) データセットが含まれており，ボキャブラリが他の言語モデルに比べて限定されている (付録表 3 参照)

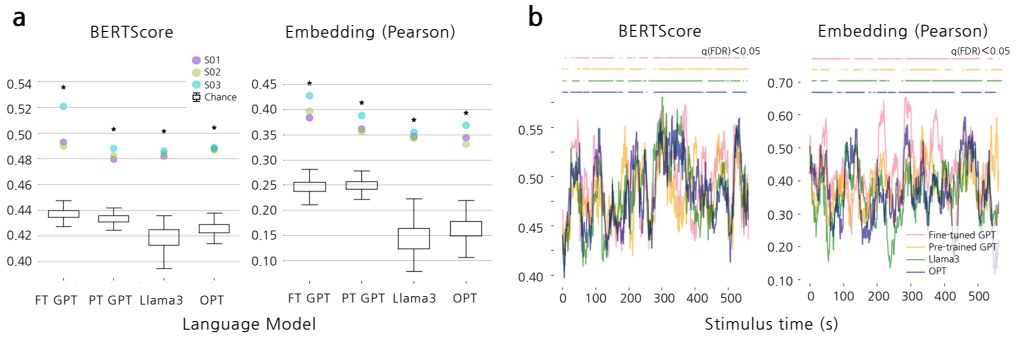

図 2: デコーダによる再構成文の結果．(a) 文全体の類似度を示す Story similarity．箱ひげ図は帰無分布，*は有意に高い (q(FDR)<0.05) スコアであることを示す．(b)20 秒の window 内における類似度を示す Window similarity．上部の線はそのタイムポイントにおいて有意に高い (q(FDR)<0.05) スコアであることを示す．

表 2: 2 つのタイムポイントにおける実際の刺激文とデコードされた文．似た表現を持つ部分を太字で示す．

	Example 1	Example 2
Actual	**in that little crack of light** and i hear the man and he says where were you and she says never mind i'm back and **he says you alright**	**the roads** are getting wider and wider and there's more **cars** and i see um lots of stores you know laundromats and
FT GPT	the windshield a minute later and **the guy said to me are you okay** and i replied well i'm fine and he says ok	little trail and then **the main road** and the trees and there are houses and some kind of town hall and a gas station
PT GPT	**candle** in the foyer burning bright is it time to leave yet no i'll be back soon	i'll rent a **car** and **drive** my first step is to find a **car** rental agency a small town a bank and
Llama3	**my phone's screen was brighter** than the sun it's time to sleep i'll see you soon okay i love you	as we **drive** i explain what we'll do when we arrive the warehouse is an old military surplus store now a gun shop
OPT	dozen different calls how long are you here i have to go i'm sorry i'll see	i **drove** i **drove** to the only place i knew of a diner a greasy spoon a diner in a strip

ことが理由で，刺激文に含まれる単語や言い回しが文中に現れやすく，デコード文，帰無分布ともに高いスコアを維持していると考える．

また注目すべきなのは，特に Embedding モデルによる評価で．2 つの GPT モデルの帰無分布が，より大きな言語モデル Llama3，OPT の帰無分布よりも高く位置していることである．これは，テストデータと似たボキャブラリを持つ Fine-tuned GPT だけでなく，Pre-trained GPT も実際の刺激文に似た文を生成しやすいことを示しており，著者らはこの原因もモデルの訓練データにあると考える．大きなモデルは一般に大量の訓練データを必要とするため，Web テキストなどの今回の刺激 (自叙伝の物語) と大きく異なるデータが用いられる一方，Pre-trained GPT は物語調のデータで訓練されている (表 1 参照) ため，実際の刺激と似た文が生成されやすいと考える．

より人脳を反映した文を再構築するときに，使用する言語モデルのサイズが大きいことは脳状態の推定の精度が高くなる可能性がある点で重要であるが、それが必ずしも精度の高いデコードに繋がるとは限らないことがわかった．一方で，適用する fMRI データにおける刺激セットが明らかでない場合は，より広い候補を生成できるという点で，より多様なデータセットで訓練された言語モデルを用いることが重要となる可能性がある．

6 おわりに

本研究では，言語モデルを使用した脳内デコーディングを提案した Tang らの研究の拡張を行った．特に，先行研究で使用されていた GPT モデルに加え 3 つの言語モデルを導入，デコード結果の比較を行なった．どの言語モデルを使用したときも有意に実際に被験者に与えた刺激文と類似した文を再構成できた一方で，Llama3-8B，OPT-6.7B などの大規模モデルよりも，120M の GPT モデルの方が高いスコアを出す傾向を確認した．これは，言語モデルの訓練データが実際の刺激文と似ていたことが原因のひとつだと考える．本研究では，実際の刺激文とデコード文の類似度の評価のみを行ったが，類似度が高い文が必ずしも脳内をより反映しているとは限らない．今回と違って，デコーダを適用する fMRI データの刺激セットが明らかでない場合は，より多様なデータで学習された言語モデルを用いることが脳内状態をより反映した再構築に繋がると考える．

6.1 謝辞

本研究は JSPS 科研費 23K18489 の助成を受けたものです.

参考文献

[1] Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G. Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. **Nature Neuroscience**, Vol. 26, No. 5, pp. 858–866, 2023.

[2] Francis R. Willett, Erin M. Kunz, Chaofei Fan, Donald T. Avansino, Guy H. Wilson, Eun Young Choi, Foram Kamdar, Matthew F. Glasser, Leigh R. Hochberg, Shaul Druckmann, Krishna V. Shenoy, and Jaimie M. Henderson. A high-performance speech neuroprosthesis. **Nature**, Vol. 620, No. 7976, pp. 1031–1036, 2023.

[3] Sean L. Metzger, Jessie R. Liu, David A. Moses, Maximilian E. Dougherty, Margaret P. Seaton, Kaylo T. Littlejohn, Josh Chartier, Gopala K. Anumanchipalli, Adelyn Tu-Chan, Karunesh Ganguly, and Edward F. Chang. Generalizable spelling using a speech neuroprosthesis in an individual with severe limb and vocal paralysis. **Nature Communications**, Vol. 13, p. 6510, 2022.

[4] Diego Lopez-Bernal, David Balderas, Pedro Ponce, and Arturo Molina. A state-of-the-art review of eeg-based imagined speech decoding. **Frontiers in Human Neuroscience**, Vol. 16, p. 867281, 2022.

[5] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[6] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2018.

[7] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[8] 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.

[9] Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. Gpt-2's activations predict the degree of semantic comprehension in the human brain. **bioRxiv**, 2021.

[10] Yuko Nakagi, Takuya Matsuyama, Naoko Koide-Majima, Hiroto Q. Yamaguchi, Rieko Kubo, Shinji Nishimoto, and Yu Takagi. Unveiling multi-level and multi-modal semantic representations in the human brain using large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 20313–20338, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[11] Richard Antonello, Aditya Vaidya, and Alexander G. Huth. Scaling laws for language encoding models in fmri, 2024.

[12] Amanda LeBel, Lauren Wagner, Shailee Jain, Aneesh Adhikari-Desai, Bhavin Gupta, Allyson Morgenthal, Jerry Tang, Lixiang Xu, and Alexander G. Huth. "an fmri dataset during a passive natural language listening task", 2024.

[13] 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.

[14] Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. Encoding and decoding in fmri. **NeuroImage**, Vol. 56, No. 2, pp. 400–410, May 2011.

[15] Tianyi Zhang, Varsha Kishore*, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In **International Conference on Learning Representations**, 2020.

[16] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention, 2021.

[17] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

[18] Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models, 2024.

[19] Yuejiao Wang, Xianmin Gong, Lingwei Meng, Xixin Wu, and Helen Meng. Large language model-based fmri encoding of language functions for subjects with neurocognitive disorder, 2024.

A　言語モデルの詳細

使用した各言語モデルのボキャブラリサイズと，本研究で追加した 3 つの事前学習済モデルの Hugging Face のモデル ID は以下の通りある．

表 3: 使用した言語モデルの補足情報．

Model	Vocab	ID
FT GPT	17378	-
PT GPT	40478	openai-community/openai-gpt
PT Llama3	128000	meta-llama/Meta-Llama-3-8B
PT OPT	50272	facebook/opt-6.7b

B　内部モデルの精度

B.1　符号化モデル

各言語モデルを用いて構築された，各被験者の符号化モデルのテストデータにおけるピアソン相関での精度を図 3 に示す．全皮質ボクセルにおける実測値と予測値のピアソン相関係数の，FDR 補正 (q(FDR)<0.05) 後の平均値が計算された．各言語モデル最も精度が高かった層では，Llama3，OPT の方が GPT よりも高い精度を出した．これらは先行研究と整合性のある結果である [11, 19]．

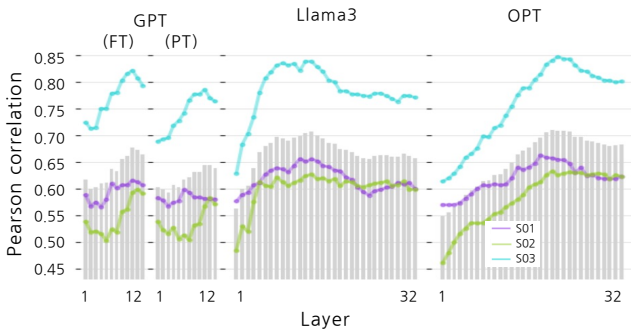

図 3: 各言語モデルの符号化モデルの精度．グレーの棒線は全被験者 (n=3) の平均のスコアを示す．

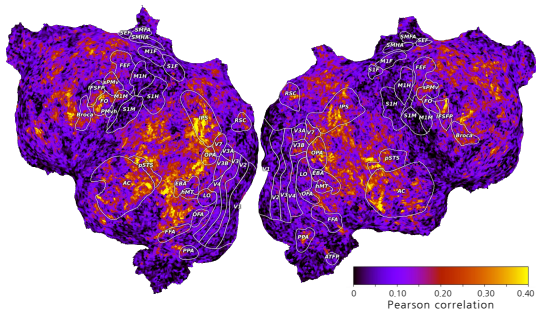

図 4: 皮質マップ上の符号化モデルの精度．

また，被験者 S02 の Fine-tuned GPT による皮質上の符号化モデルの精度を図 4 に示す (q(FDR)<0.05)．同データを使用した先行研究 [13] と同様に，頭頂皮質，側頭皮質，前頭前皮質などが高い精度を持つことがわかる．ページ数の制約上，掲載は困難であるが，他の言語モデル，他の被験者においても非常によく似た傾向が見られた．

B.2　トークン数予測モデルの精度

トークン数予測モデルの，テストデータにおけるピアソン相関での精度は以下の通りであった．

表 4: トークン数予測のデルの精度 (n=3)．

Model	Pearson correlation
FT GPT	0.740 ± 0.012
PT GPT	0.708 ± 0.011
PT Llama3	0.722 ± 0.009
PT OPT	0.729 ± 0.008

C　他の類似度評価指標結果

3.6 節で述べた，先行研究で使用された他の類似度評価指標の Story similarity の結果を図 5 に示す．WER 以外の指標については，Fine-tuned GPT が他の言語モデルより僅かに良いスコアとなった．

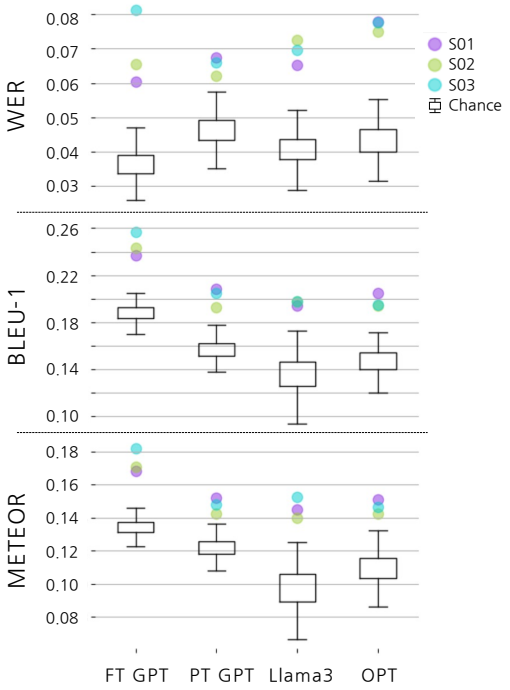

図 5: 単語レベルの評価指標による Story similarity．