Triple-Phase Transition: 脳との関係から捉える大規模言語モデルの学習ダイナミクス

中木 裕子 [1,2*‡]　多田 圭吾 [1,2*]　吉野 草太 [1,2*]　西本 伸志 [1,2†]　高木 優 [1,2,3†]

[1] 大阪大学 大学院生命機能研究科　[2] 情報通信研究機構 脳情報通信融合研究センター
[3] 国立情報学研究所 大規模言語モデル研究開発センター
*Equal first author.　†Equal corresponding author.　‡Team lead.

概要

大規模言語モデル（LLMs）の学習時，ある能力を学習過程で突然獲得する現象が知られている．このような LLMs の変化は相転移（Phase Transition）現象と呼ばれ，その原因として LLMs の内部状態における相転移が示唆されているが，その実態は未解明の点が多い．本研究で我々は LLMs の相転移について LLMs の内部表現とヒト脳活動を対比させることにより解釈することを試みる．具体的には，学習過程における，LLMs と脳の類似度，LLMs の内部状態，下流タスク精度の変化という 3 つの観点を統合した解析を行い，LLMs の学習ダイナミクスに新たな解釈を与える．我々は，LLMs が下流タスクの能力を獲得する過程で，脳との対応や内部状態に 3 段階の相転移が起きることを示す．

1　はじめに

大規模言語モデル（LLMs）は，従来の言語モデルに対して常識問題や推論問題，機械翻訳などの下流タスクにおける性能が大幅に向上し，ヒトのように自然言語を扱えるようになってきた．一般的に，LLMs は学習が進むにつれて損失が減少していく [1, 2] が，LLMs の学習ダイナミクスを分析する研究からは，LLMs はすべての能力を同時には獲得せず，学習の異なる段階で異なる能力を獲得していくことが示されている．ある能力は連続的な変化を通して獲得される [3, 4] 一方，ある能力は非連続的に突然獲得される．後者は言い換えると，モデルサイズや計算量，学習データ量に応じて突然これまで扱えなかった能力を扱えるようになる現象であり，近年注目されている．この現象は下流タスクを用いた Benchmark 解析を通じて発見され，Phase Transition[5] や Emergent Abilities[6] などと呼ばれている．また，近年の Mechanistic Interpretability の研究を通じて，LLMs の内部状態でもこのような相転移的現象が報告されている [7, 8, 9].

しかし，LLMs 内部で報告される相転移現象が学習過程でどのように発現するのかについては，未だ不明な点が多く残されている．ここで，近年 LLMs の内部を人間中心に解釈する方法として，LLMs の内部表現とヒト脳活動とを対比するアプローチが注目されている [10, 11, 12, 13]．具体的には，LLMs の潜在表現に単純な線形変換を適用するだけでヒト脳活動をモデル化できることがわかってきており，LLMs のヒト脳活動との類似性が示されただけでなく，ヒト脳活動と対比することにより LLMs の内部を解釈する方法として注目されている．これらの研究の代表的なものとして，LLMs と人間の両者の学習特性から議論した研究 [14, 15, 16, 17] や LLMs の内部状態の変化から議論した研究 [18] が挙げられる．しかし，多くの先行研究では学習済みの LLMs を用いてヒト脳活動との対応が探られており，学習過程における相転移現象に注目していなかった．

これらを踏まえ，本研究では，学習過程における，LLMs と脳の類似度を調べる Encoding 解析，LLMs の内部状態の変化を調べる Probing 解析，下流タスク精度の変化を調べる従来の Benchmark 解析という 3 つの観点を統合した解析を行う．これによって，LLMs の学習ダイナミクスに新たな解釈を与える．

本研究の主な貢献は以下である．

1. LLMs の学習ダイナミクス解析を，従来の解析に加えてヒト脳とのアライメントを含めた 3 つの統合的な観点から行う．
2. 学習データが異なる複数の LLMs を用いた解析

連絡先: nishimoto.shinji.fbs@osaka-u.ac.jp,　yu-takagi@nii.ac.jp

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

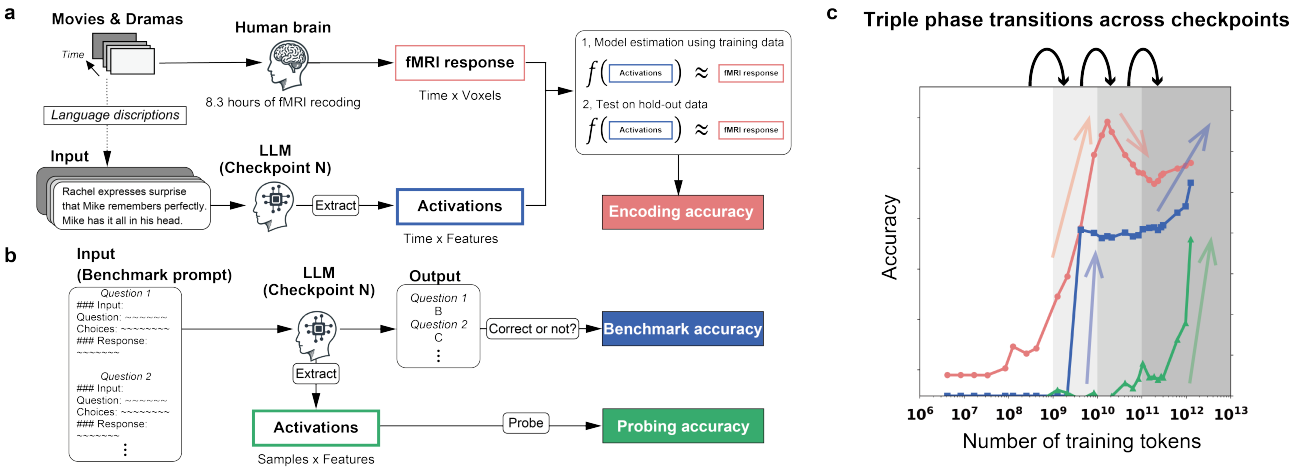

図 1　研究概要図．a. Encoding 解析の概要．b. Benchmark 解析 (上) と Probing 解析 (下) の概要．c. Encoding，Benchmark，Probing の 3 つの解析結果から確認された LLMs の学習過程における相転移現象の概要．

を通して，学習データ依存の学習ダイナミクスを示す．

3. LLMs の学習ダイナミクスに，以下の 3 段階の相転移現象としての解釈を与える：
   (a) 脳整合・指示追従期
   (b) 脳特化期
   (c) 理解定着期

2　方法

2.1　fMRI データセット

本研究では，母語が日本語の健常な被験者 6 名が 9 本の映画やドラマ（10 エピソード，合計 8.3 時間）を 3 テスラの fMRI 内で視聴しているときの脳活動データセット [19, 13] を用いた．映画とドラマは様々なジャンルを含み，8 本が海外の映画またはドラマ，1 本が日本のアニメーションであった．全ての海外作品は日本語吹き替えで再生されており，被験者は日本語でそれらを理解している．また，本データセットには，複数の自然言語アノテーションが付けられており，我々はその内のシーンごとの物語背景に関するアノテーションを用いた．アノテーションは予め用意されている日本語版を DeepL で英語に翻訳したものと、それをさらに日本語に翻訳したものを使用した．

本研究では 2.4 節の Encoding 解析で，29,993 秒分の本データを用いた．このうち，各エピソードの最後の分割動画に対応する 7,737 秒分をテストデータに，残りの 22,262 秒分を訓練データに用いた．

2.2　下流タスクデータセット

本研究では，Massive Multitask Language Understanding（MMLU）[20] と CommonsenseQA（CSQA）[21] を用いた．MMLU は幅広い知識と問題解決能力を評価するためのデータセットで，CSQA は常識推論能力を評価するためのデータセットである．

本研究では LLMs の Benchmark 解析と Probing 解析のために，英語版と日本語版の 5-shot のプロンプトを用いた．LLMs の最大コンテキスト長を超えるプロンプトはサンプルから除去した．MMLU の英語版には MMLU[20]，日本語版には MMMLU[22] の日本語翻訳されたものを用いた．サンプル数は日英版ともに 13,571 問であった．CSQA の英語版には CSQA[21]，日本語版には JCommonsenseQA[23] を用いた．サンプル数は英語版が 10,957 問，日本語版が 8,934 問であった．また，Probing 解析ではサンプルを訓練/テストデータに分割した．訓練データは日英 MMLU が 10,856，英 CSQA が 8,765，日 CSQA が 7,147 サンプルで，残りをテストデータとした．

2.3　解析で使用したモデル

LLMs の学習ダイナミクスを 3 つの観点から調査するために，学習チェックポイントが用意されている OLMo-2[24] と LLM-jp[25]，Amber[26] を用いた．チェックポイントは，OLMo-2 を 28 個，LLM-jp を 27 個，Amber を 18 個用いた．パラメータ数はそれぞれ 7.3B，7.28B，6.74B であった．全てのモデルはレイヤー数が 32，潜在次元数が 4096 であった．使用チェックポイントの詳細は A 章に記述する．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

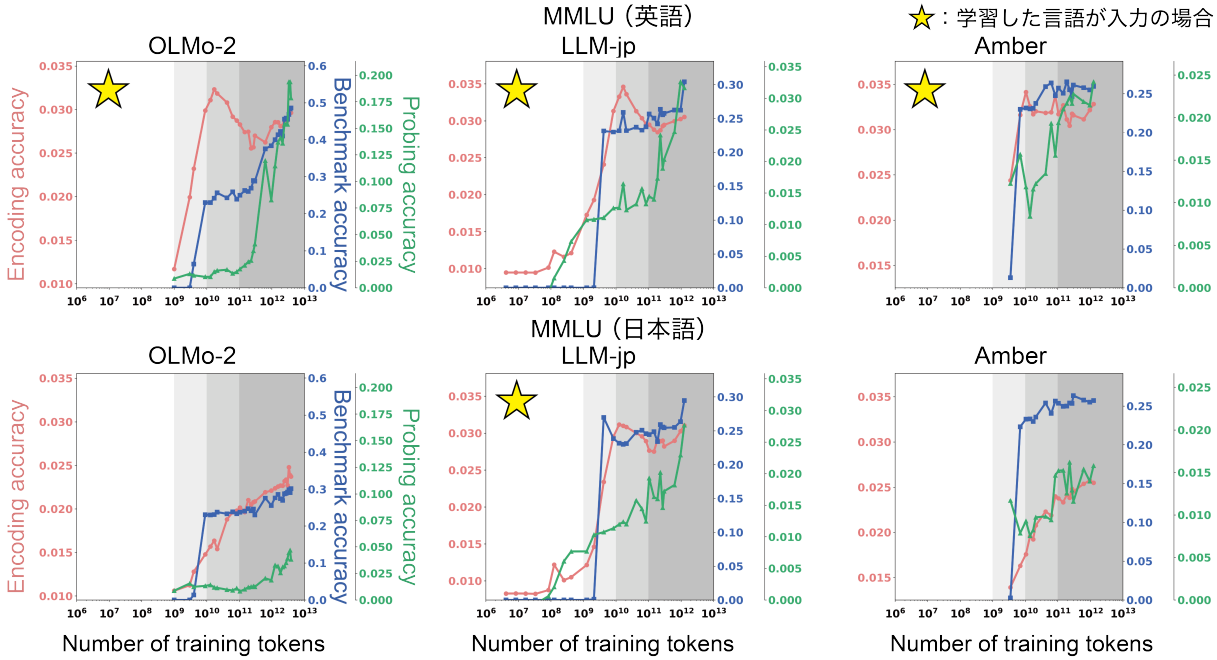

図 2　3 回の相転移がある学習ダイナミクス．横軸は学習トークン数．縦軸は第 26 層の潜在表現と単一被験者の脳活動を用いた際の平均 Encoding 精度 (赤)，Benchmark 精度 (青)，第 26 層での平均 Probing 精度 (緑)．背景色は LLMs の状態．

2.4　Encoding 解析

LLMs の学習ダイナミクス解析の 1 つ目として，LLMs の潜在表現と脳活動の類似度が LLMs の学習進度によりどのように変化していくのかを調べた．そのために，2.3 節の LLMs の各チェックポイントについて，LLMs の各層の潜在表現を脳活動に線形変換し，その予測精度を評価する Encoding 解析を行った [27, 28, 29]（図 1a 右）．

まず，2.1 節のアノテーションを LLMs に入力し，各層の潜在表現を抽出した．潜在表現には，各層の特徴をよく表現するとされるために LLMs の内部表現解析でよく用いられる MLP 層の出力とし，トークン間で平均化して使用した．次に，訓練データを用いて潜在表現から脳活動を L2 正則化線形回帰により予測し，その後テストデータで評価を行った．評価には予測 fMRI 信号と実際の fMRI 信号のピアソン相関係数を使用した．統計的有意性は，Blockwise permutation をもとに予測 fMRI 信号とシャッフルした実際の fMRI 信号間の相関を比較して計算した．統計的閾値は $p<0.05$ とし，FDR 法により多重比較補正を行うことで有意な予測精度のボクセルを選んだ．また，神経活動から BOLD 信号への血行動態遅延を 8〜10 秒と仮定してモデル化を行った．

2.5　Benchmark 解析

LLMs の学習ダイナミクス解析の 2 つ目として，下流タスクを行う能力が LLMs の学習進度によりどのように獲得されていくのかを調べた．そのために，2.3 節の LLMs の各チェックポイントについて，2.2 節の下流タスクデータセットの精度評価を行った（図 1b 上段）．このときに使用した評価指標は，LLMs の最終層の出力が正解と完全に一致した問題数を全問題数で割った正解比率とした．

2.6　Probing 解析

LLMs の学習ダイナミクス解析の 3 つ目として，LLMs 内部で下流タスクに必要な情報表現が LLMs の学習進度によりどのように獲得されていくのかを調べた．そのために，2.3 節の LLMs の各チェックポイントについて，下流タスクの解答ラベルから LLMs の全層の潜在表現を予測し，その予測精度を評価する Probing 解析を行った（図 1b 下段）．

まず，2.2 節の下流タスクデータセットの問題文を LLMs に入力し，各層の最終トークンに対する潜在表現（MLP 層の出力）を抽出した．次に訓練データを用いて，正解を 1，不正解を 0 とした解答行列（サンプル数×選択肢数）から LLMs の全層の潜在表現を L2 正則化線形回帰で予測したあと，テストデータで評価を行った．評価には予測潜在表現と実

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

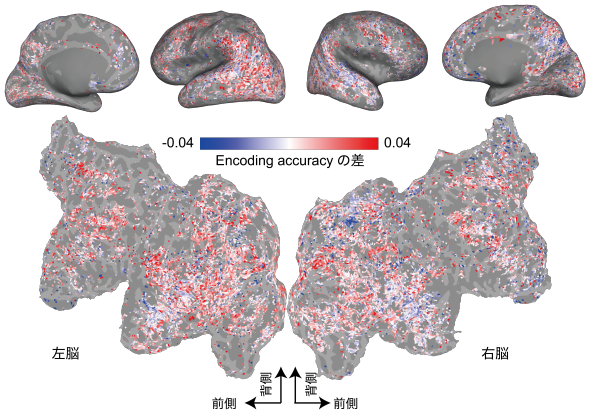

図 3　単一被験者の大脳皮質における脳特化期前後での Encoding 精度の差．日本語入力/LLM-jp の第 26 層での結果．図は左右大脳皮質 (上) とその平面図 (下) を表し，有意な予測を示したボクセルに色付けした．脳特化期前で精度が高いほど赤く，後で精度が高いほど青い．

際の潜在表現のピアソン相関係数を使用した．

3　結果

3.1　LLMs の学習ダイナミクス

2.4 節，2.5 節，2.6 節の解析を通して学習進度による各精度の変化を調べた結果，LLMs の学習過程において後期層で 3 回の相転移を確認した (図 2)．具体的には，1 回目の相転移で Encoding 精度と Benchmark 精度が急上昇する状態（学習トークン数: 約 10^9〜10^{10}），2 回目の相転移で Encoding 精度が低下する状態（約 10^{10}〜10^{11}），3 回目の相転移で Benchmark 精度と Probing 精度が上昇する状態（約 10^{11}〜）になることがわかった．特に，Benchmark 精度から下流タスクを行う能力に注目すると，1 回目の相転移後の状態を経て LLMs はタスクの指示に従い始め，3 回目の後ではタスクを解けるようになっていたため，それぞれを脳整合・指示追従期と理解定着期と解釈した．本現象はモデルが学習した言語を用いた時のみ確認された．図 2 では MMLU を使用し，Encoding 解析および Probing 解析の結果は各 LLMs の学習後期での Probing 精度が第 26 層付近で高かったことから，第 26 層を用いた際の全脳ボクセルおよび LLMs の層内の全ニューロンでの平均予測精度を示した．CSQA での結果は B.1 節に示す．

3.2　LLMs と対応する脳領域の変化

次に，2, 3 回目の相転移の間の Encoding 精度が下がる状態において，LLMs の潜在表現と類似した脳活動を持つ脳領域の変化を調べた．図 3 に，この期間で最も高い精度と低下後の精度の差分を示す．興味深い点として，単一被験者での結果において潜在表現と脳活動は 2 回目の相転移時の方が平均的には類似しているにも関わらず，一部の楔前部や高次視覚野では精度低下後の方がより類似することが確認できた．そこで，この状態を脳特化期と解釈した．

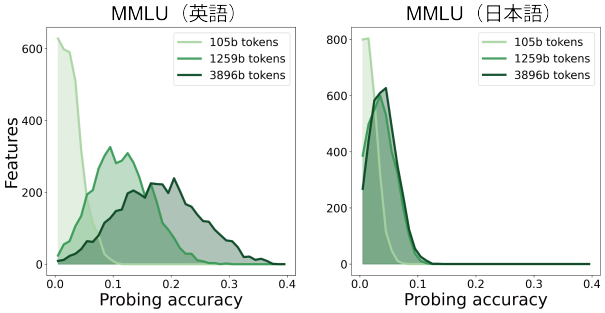

図 4　日英 MMLU の Probing 精度の理解定着期での変化．OLMo-2 の第 26 層の結果．横軸は Probing 精度，縦軸は bin 幅 0.01 の精度に該当するニューロン数．

この結果はある内容に関する学習済みの LLMs の潜在表現が特定の領域の脳活動と固有に対応すること [13] と関連していると考えられる．脳特化期前後で，潜在表現は広範な脳領域を薄く説明する一般的な表現から，より内容に関連した情報を表現するようになり，対応する脳領域が特化した可能性がある．今後はこの点に注目し，全被験者で検証することが重要になる．他モデルの結果は B.2 節に示す．

3.3　LLMs 内部の情報表現の変化

最後に，LLMs のニューロンがタスクに特化した表現を理解定着期に獲得しているのかを検証するため，この期間における Probing 精度の変化を調べた．図 4 に，OLMo-2 の第 26 層における Probing 精度の変化を示す．理解定着期に英語の MMLU で一部のニューロンの精度が上昇していく様子が読み取れる．この結果から，LLMs がこの期間で下流タスクを解けるようになる裏で，内部でもタスクに特化した表現が獲得されていたことがわかる．

4　結論

本研究では，脳と LLMs のアライメント，LLMs の内部表現解析，下流タスク精度の 3 つの観点から LLMs の学習ダイナミクスを調べた．我々は LLMs の後期層で 3 回の相転移があることを示し，LLMs は脳整合・指示追従期と脳特化期，理解定着期を経て学習していくと解釈した．また，この傾向は学習した言語が入力であるかに依存することも示した．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

謝辞

本研究は，JSPS 科研費 JP24H00619，JST JPMJCR24U2 の助成を受けたものです．

参考文献

[1] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. **arXiv [cs.LG]**, 22 January 2020.

[2] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. **arXiv [cs.CL]**, 29 March 2022.

[3] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In **Findings of the Association for Computational Linguistics: EMNLP 2023**, pages 5687–5711, Stroudsburg, PA, USA, December 2023. Association for Computational Linguistics.

[4] Ian R. McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, Andrew Gritsevskiy, Daniel Wurgaft, Derik Kauffman, Gabriel Recchia, Jiacheng Liu, Joe Cavanagh, Max Weiss, Sicong Huang, The Floating Droid, Tom Tseng, Tomasz Korbak, Xudong Shen, Yuhui Zhang, Zhengping Zhou, Najoung Kim, Samuel R. Bowman, and Ethan Perez. Inverse scaling: When bigger isn't better, 2024.

[5] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. **arXiv [cs.LG]**, 23 September 2022.

[6] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. **arXiv [cs.CL]**, 15 June 2022.

[7] Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L Leavitt, and Naomi Saphra. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. **arXiv [cs.CL]**, 13 September 2023.

[8] Core Francisco Park, Maya Okawa, Andrew Lee, Hidenori Tanaka, and Ekdeep Singh Lubana. Emergence of hidden capabilities: Exploring learning dynamics in concept space, 2024.

[9] Emily Cheng, Diego Doimo, Corentin Kervadec, Iuri Macocco, Jade Yu, Alessandro Laio, and Marco Baroni. Emergence of a high-dimensional abstraction phase in language transformers. **arXiv [cs.CL]**, 24 May 2024.

[10] Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A Norman, Orrin Devinsky, and Uri Hasson. Shared computational principles for language processing in humans and deep language models. **Nat. Neurosci.**, 25(3):369–380, March 2022.

[11] Shailee Jain and Alexander G Huth. Incorporating context into language encoding models for fMRI. In **Proceedings of the 32nd International Conference on Neural Information Processing Systems**, NIPS'18, pages 6629–6638, Red Hook, NY, USA, 3 December 2018. Curran Associates Inc.

[12] Subba Reddy Oota, Manish Gupta, and Mariya Toneva. Joint processing of linguistic properties in brains and language models. **arXiv [cs.CL]**, 15 December 2022.

[13] 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**, pages 20313–20338, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[14] 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. **Proc. Natl. Acad. Sci. U. S. A.**, 118(45), 9 November 2021.

[15] Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. Evidence of a predictive coding hierarchy in the human brain listening to speech. **Nat Hum Behav**, 7(3):430–441, March 2023.

[16] Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. Instruction-tuning aligns LLMs to the human brain. **arXiv [cs.CL]**, 1 December 2023.

[17] Richard Antonello, Aditya Vaidya, and Alexander G Huth. Scaling laws for language encoding models in fMRI. **arXiv [cs.CL]**, 19 May 2023.

[18] Emily Cheng and Richard J Antonello. Evidence from fMRI supports a two-phase abstraction process in language models. **arXiv [cs.CL]**, 9 September 2024.

[19] Hiroto Q Yamaguchi, Naoko Koide-Majima, Rieko Kubo, Tomoya Nakai, and Shinji Nishimoto. Narrative movie fMRI dataset, 4 October 2024.

[20] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. **arXiv [cs.CY]**, 7 September 2020.

[21] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In **Proceedings of the 2019 Conference of the North**, pages 4149–4158, Stroudsburg, PA, USA, 2019. Association for Computational Linguistics.

[22] openai/MMMLU · datasets at hugging face. https://huggingface.co/datasets/openai/MMMLU. Accessed: 2025-1-1.

[23] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pages 2957–2966, 2022.

[24] Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious, 2024.

[25] LLM-jp, :, Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, Yuto Harada, Chikara Hashimoto, Tatsuya Hiraoka, Shohei Hisada, Sosuke Hosokawa, Lu Jie, Keisuke Kamata, Teruhito Kanazawa, Hiroki Kanezashi, Hiroshi Kataoka, Satoru Katsumata, Daisuke Kawahara, Seiya Kawano, Atsushi Keyaki, Keisuke Kiryu, Hirokazu Kiyomaru, Takashi Kodama, Takahiro Kubo, Yohei Kuga, Ryoma Kumon, Shuhei Kurita, Sadao Kurohashi, Conglong Li, Taiki Maekawa, Hiroshi Matsuda, Yusuke Miyao, Kentaro Mizuki, Sakae Mizuki, Yugo Murawaki, Akim Mousterou, Ryo Nakamura, Taishi Nakamura, Kouta Nakayama, Tomoka Nakazato, Takuro Niitsuma, Jiro Nishitoba, Yusuke Oda, Hayato Ogawa, Takumi Okamoto, Naoaki Okazaki, Yohei Oseki, Shintaro Ozaki, Koki Ryu, Rafal Rzepka, Keisuke Sakaguchi, Shota Sasaki, Satoshi Sekine, Kohei Suda, Saku Sugawara, Issa Sugiura, Hiroaki Sugiyama, Hisami Suzuki, Jun Suzuki, Toyotaro Suzumura, Kensuke Tachibana, Yu Takagi, Kyosuke Takami, Koichi Takeda, Masashi Takeshita, Masahiro Tanaka, Kenjiro Taura, Arseny Tolmachev, Nobuhiro Ueda, Zhen Wan, Shuntaro Yada, Sakiko Yahata, Yuya Yamamoto, Yusuke Yamauchi, Hitomi Yanaka, Rio Yokota, and Koichiro Yoshino. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms, 2024.

[26] Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, Richard Fan, Yi Gu, Victor Miller, Yonghao Zhuang, Guowei He, Haonan Li, Fajri Koto, Liping Tang, Nikhil Ranjan, Zhiqiang Shen, Xuguang Ren, Roberto Iriondo, Cun Mu, Zhiting Hu, Mark Schulze, Preslav Nakov, Tim Baldwin, and Eric P Xing. LLM360: Towards fully transparent open-source LLMs. **arXiv [cs.CL]**, 11 December 2023.

[27] Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fMRI. **Neuroimage**, 56(2):400–410, 15 May 2011.

[28] Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. **Curr. Biol.**, 21(19):1641–1646, 11 October 2011.

[29] Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. **Neuron**, 76(6):1210–1224, 20 December 2012.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

A　LLMs の使用チェックポイント

本解析では，複数の LLMs について公開済みの学習チェックポイントを用いた．具体的には，OLMo-2 を 28 個（1B〜3896B のモデル．150, 600, 900, 2K, 3K, 4K, 5K, 10K, 15K, 20K, 25K, 35K, 45K, 55K, 65K, 72K, 150K, 230K, 300K, 370K, 441K, 513K, 584K, 656K, 727K, 799K, 870K, 928.646K の学習ステップ），LLM-jp を 27 個（4.2M〜1258B のモデル．1, 2, 4, 8, 20, 30, 60, 100, 300, 500, 1K, 2K, 3K, 4K, 5K, 10K, 15K, 20K, 25K, 35K, 45K, 55K, 65K, 72K, 150K, 230K, 300K の学習ステップ），Amber を 18 個（3.5B〜1259B のモデル．1, 2, 3, 4, 5, 6, 12, 18, 24, 30, 42, 54, 66, 78, 86, 179, 275，最終のチェックポイント）用いた．ただし，LLM-jp は手元にあるものを使用した．主な学習データは，OLMo-2 と Amber は英語，LLM-jp は英語と日本語であった．

B　結果

B.1　CSQA を用いたときの LLMs の学習ダイナミクス

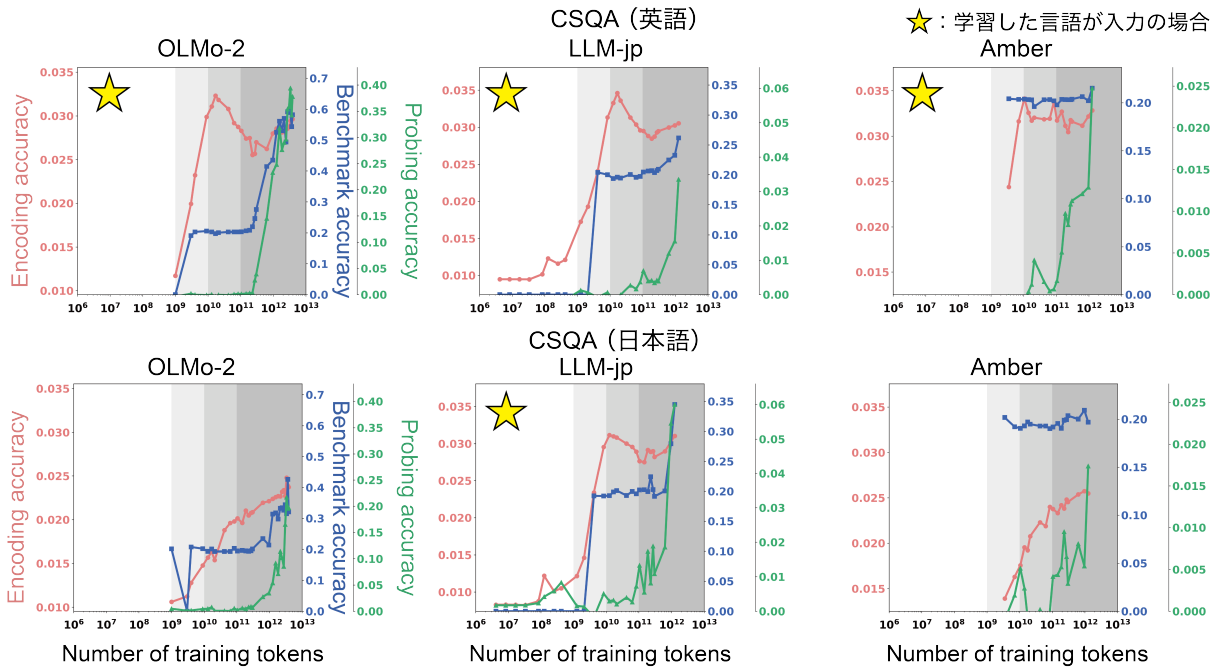

図 B.1　CSQA の場合の学習ダイナミクス．横軸は学習トークン数．縦軸は第 26 層の潜在表現と単一被験者の脳活動を用いたときの平均 Encoding 精度 (赤)，Benchmark 精度 (青)，第 26 層での平均 Probing 精度 (緑)．背景色は LLMs の各状態の範囲．CSQA でも MMLU と同様の相転移が確認できた．

B.2　LLMs と対応する脳領域の変化に関する追加資料

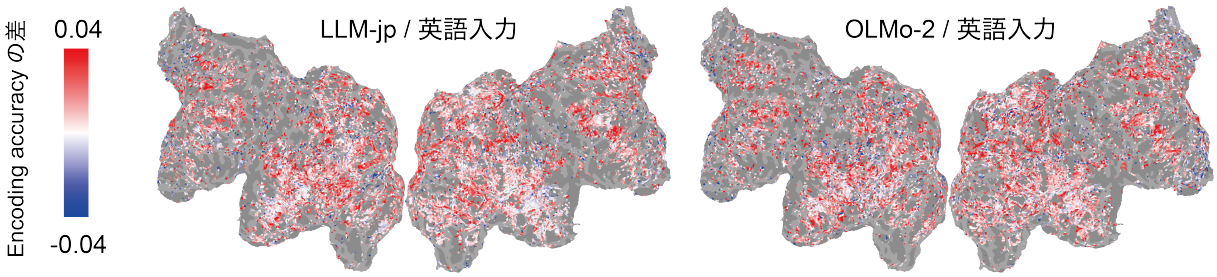

図 B.2　単一被験者の大脳皮質における脳特化期前後での Encoding 精度の差．英語入力/LLM-jp と OLMo-2 の第 26 層での結果．LLMs が学習した言語を入力とした潜在表現の場合に注目すると，本被験者では同様の傾向が確認された．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).