事前学習言語モデルのドメイン適応能力に関する分析：ドメイン特有ニューロンの検出と分析

鈴木雅弘 [1,2]　高柳剛弘 [1]　坂地泰紀 [3]　和泉潔 [1]
[1] 東京大学大学院　[2] 日興アセットマネジメント株式会社　[3] 北海道大学
research@msuzuki.me　takayanagi-takehiro590@g.ecc.u-tokyo.ac.jp
sakaji@ist.hokudai.ac.jp　izumi@sys.t.u-tokyo.ac.jp

概要

本研究では，事前学習言語モデル（PLM）における専門ドメインに特化したニューロンの内部挙動を分析する．具体的には，金融ドメインと一般ドメインのテキストに対する，日本語の Encoder または Decoder のアーキテクチャをもつ複数の PLM のニューロンの挙動を比較し，ドメイン特有のニューロンを検出した．分析の結果，Encoder と Decoder の両アーキテクチャに共通して，ドメイン特有のニューロンは初期層に多く存在することがわかった．次に，Decoder モデルの MLP にはドメイン特有の性質が複数の層に分散して存在することが示唆された．また，金融ドメインで追加事前学習したモデルでは，中間層でドメイン特有の表現を獲得していることがわかった．

1　はじめに

Transformer 構造を持つ事前学習言語モデル（Pre-trained Language Model, PLM）の実用的な応用においては，特定の専門ドメインへの適応が不可欠となる場面が多い．例えば，金融，医療，法律といった専門分野への PLM の適応のためには，専門ドメインの大規模テキストを用いた事前学習 [1, 2, 3] や，専門ドメインのインストラクションデータを用いたチューニング [4, 5, 6] が近年主流となっている．

しかし，これらの手法を用いてモデル全体をチューニングするのは計算コストが非常に高い [7]．ドメイン適応に必要な PLM の構成要素を絞ってチューニングすることで，計算コストを削減し効率的にドメインに適応できると考えられる．そのためには，PLM がドメイン特有の特徴を獲得する過程や適応に寄与する PLM の成分の把握が重要である．

PLM の内部構造を理解する試みとして，ニューロンの活動に着目した研究が行われている [8, 9]．これらの研究は，PLM が異なる性質を持つテキストを，どのように区別し内部で処理しているのかを理解する上で重要な知見を与えているものの，専門ドメインにおける議論は未だ十分とは言えない．

本研究では，PLM がドメインの特徴をどのように表現しているのかを解明することを目的とする．先行研究における言語 [10] やバイアス [11] に特有のニューロンを分析する手法をベースとしつつ，専門ドメインの分析に適用することを試みる．まず，金融ドメインと一般ドメインのテキストに対するモデルの出力から，テキストのドメインの差がニューロンに及ぼす反応の違いを分析する．次に，金融ドメインへの適応に用いられる，事前学習とインストラクションチューニングの 2 つの異なるアプローチが PLM に与える影響を，これらを施されたモデルとベースモデルの比較から検証する．日本語の Encoder または Decoder のアーキテクチャをもつ複数の PLM についての分析の結果，Encoder と Decoder の両アーキテクチャに共通して，ドメイン特有のニューロンは初期層に多く存在することがわかった．また，Decoder モデルの MLP にはドメイン特有の性質が複数の層に分散して存在することが示唆された．金融ドメインで事前学習したモデルでは，中間層でドメイン特有の表現を獲得していることがわかった．

2　手法

2.1　ドメイン特有ニューロンの検出

本研究では，[11] の手法に基づき，金融ドメイン特有のニューロンを検出する．当該研究では，文 $\{x_i\}_{i=1}^{N}$ に対するニューロン m の出力を $\{z_{m,i}\}_{i=1}^{N}$ とする．$\{z_{m,i}\}_{i=1}^{N}$ をドメインラベル $\{b_i\}_{i=1}^{N}$ の予測スコアとして扱い，Precision-Recall 曲線の下の面積であ

る平均適合率 $\mathrm{AP}_m = \mathrm{AP}(z_m, b) \in [0, 1]$ を計算することで，ニューロン m のドメイン検出における性能を測定する．[10] と同様に，入力トークン列に対しては [PAD] トークンの出力を省いたうえで $z_{m,i}$ を平均する．すべてのニューロンに対して AP_m を計算し，降順に並べたうえで，上位 (Top)，中位 (Middle)，下位 (Bottom) の各 1,000 ニューロンを選択する．Top と Bottom のニューロンによる出力は，テキストの属するドメインのラベルに対してそれぞれ強い正と負の相関を持つことから，これらのニューロンはドメイン特有の情報を表現していると考えられる．

2.2 ドメイン特有の語彙の検出

本研究は，先行研究 [10] と異なり，同一言語のテキスト間での比較のため，日本語の表現に共通する部分ではドメインによる差異はほとんどなく，ドメイン特有のニューロンの検出が難しくなることが考えられる．そこで本研究では，AP_m を算出するトークンを，金融ドメインと一般ドメインのそれぞれにおいて重要な語彙のみに絞る．文書における単語の重要度を算出する指標としては，単語の出現頻度と稀さの積によって算出される TF-IDF がある．しかしながら，TF-IDF では金融ドメインの文書では重要な一方で一般ドメインの文書では重要ではない単語を抽出することができない．そこで，本研究では簡易的に以下の条件を満たす語彙（トークン）を，金融ドメインと一般ドメインの文書のそれぞれにおいて重要な語彙として扱う．

- 金融または一般ドメインの 2%以上の文書に出現する
- トークンの文字数が 2 文字以上である
- ひらがなのみで構成されない
- トークン t が出現する金融ドメインと一般ドメインの文書数をそれぞれ $N_{\mathrm{fin}}(t)$ と $N_{\mathrm{gen}}(t)$ としたときに次の式を満たす [1)]:

$$0.7 < \frac{|N_{\mathrm{fin}}(t) - N_{\mathrm{gen}}(t)|}{N_{\mathrm{fin}}(t) + N_{\mathrm{gen}}(t)}$$

抽出された金融ドメインと一般ドメインのトークン集合を結合し，結合したトークン集合に含まれるトークンについてのみ AP_m を算出する．

1) 0.7 の閾値は，両ドメインの文書におけるトークンの出現頻度を詳細に分析した結果，ドメイン特有の語彙とそうでない語彙を効果的に分離できる値として経験的に得られた．

2.3 モデル間の比較

2.1 節で述べた手法は，ドメインの異なるテキスト間でのニューロンの比較を目的としている．我々はさらに，金融モデルやインストラクションモデルがベースモデルとどのように異なるかを計測するために，同じテキスト集合の入力に対するベースモデルとその派生モデルを比較する．2.1 節では，2 つの異なるテキスト集合を同じモデルに入力することで得られた 2 つの出力の集合に対して AP_m を算出する．本節では同じテキスト集合を異なるモデルに入力することで得られた 2 つの出力の集合に対して AP_m を算出することで，モデル間の差異を計測する．金融モデルと汎用モデルの比較では，2.2 節で述べたトークンのフィルタリングを行う．インストラクションモデルと汎用モデルの比較では，トークンのフィルタリングを行わない．

2.4 モデルとデータセット

本研究では，日本語の代表的な事前学習言語モデルの中で，Encoder と Decoder の 2 種類について，それぞれ異なるアーキテクチャをもつ 2 つのモデルからなる 4 つのモデルを分析対象とする．Encoder モデルとして BERT [12] の日本語 Large モデル [13] と DeBERTaV2 [14] の日本語 Base モデル [15]，Decoder モデルとして llm-jp-3-13b [16] と Sarashina1-7B [17] を用いる．llm-jp-3-13b と Sarashina1-7B はそれぞれ LLaMA [18] と GPT-NeoX [19] をアーキテクチャのベースとした日本語モデルである．

金融ドメイン特化モデルと汎用モデルの比較では，DeBERTaV2 の日本語金融 Base モデル [20] と上述の日本語汎用 Base モデルを用いる．インストラクションモデルと汎用モデルの比較では，llm-jp-3-13b-instruct [21] と上述の llm-jp-3-13b を用いる．金融ドメイン特化モデルとインストラクションモデルは，それぞれ汎用モデルをベースとして金融ドメインのテキストによる追加事前学習とインストラクションチューニングを行ったモデルである．

モデルに入力するテキストとして，金融ドメインと一般ドメインのコーパス [2)] を用いる．金融ドメインのデータとして，決算短信等が公開される TDnet と有価証券報告書などが公開される EDINET からそれぞれ取得した．一般ドメインのデータには，

2) これらのテキストデータの収集期間は [20] の事前学習で用いられたコーパスと同一である

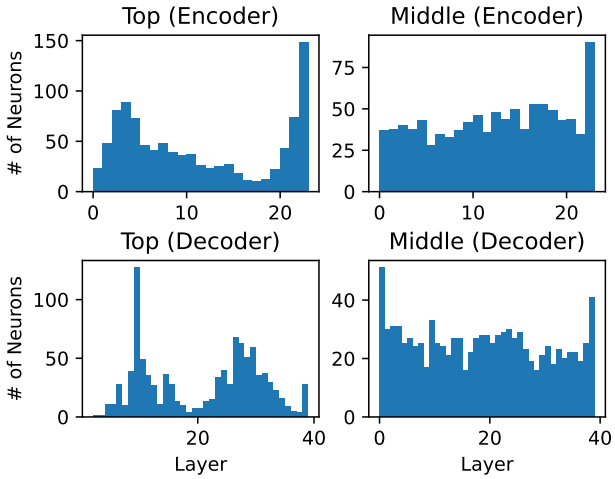


図 1: 平均適合率で降順に並べたときの上位 (Top)・中位 (Middle) のニューロンの，各層における分布．

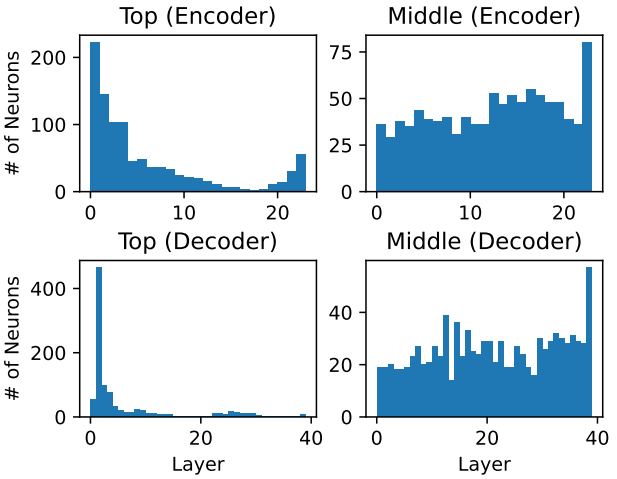


図 2: ドメイン特有のトークンのみに対して AP_m の算出を行った際のニューロンの各層における分布．

Wikipedia の日本語版を用いる．インストラクションモデルと汎用モデルの比較では，一般的なドメインのインストラクションデータ [22] を用いる．金融ドメイン，一般ドメイン，インストラクションデータのそれぞれから，ランダムに 1,000 文を抽出しそれぞれの代表データとして用いる．

3 結果と考察

3.1 金融テキストと一般テキストの比較

本節で述べる結果について，Encoder の 2 モデルと Decoder の 2 モデルにはそれぞれ同様の傾向が見られたため，本稿では Encoder モデルとして BERT の日本語 Large モデル，Decoder モデルとして llm-jp-3-13b を用いた実験結果のみを示す．DeBERTaV2 の日本語 Base モデルと Sarashina1-7B についての結果は Appendix A に記載する．

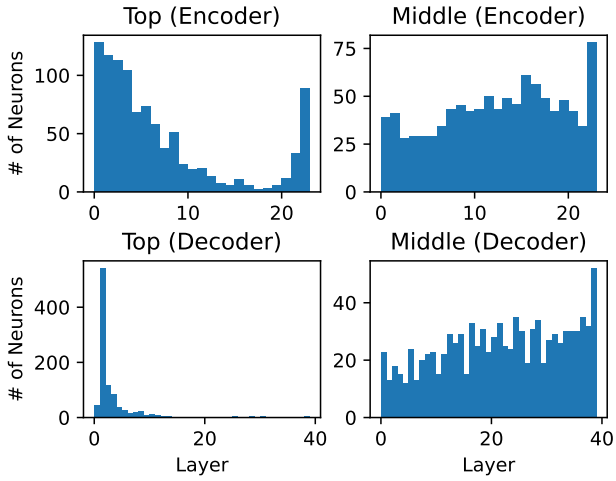
図 1 に [10] と同様の手法で検出されたニューロンの各層における分布を表すヒストグラムを示す．Top と Middle の分布を比較すると，Encoder モデルの Top では，前方の層で緩やかな山型の分布を形成している．Decoder モデルでは大きく 2 つのピークを持つ分布となっている．

2.2 節の手法によって AP_m を算出するトークンを金融ドメインと一般ドメインのそれぞれにおいて重要な語彙に絞った際のヒストグラムを図 2 に示す．図 1 と異なり，Encoder モデルと Decoder モデルの両方で特に前方の層に急な山型の分布を形成しており，ドメイン特有の語彙に対するニューロンの活動が初期層で行われていることが示唆される．

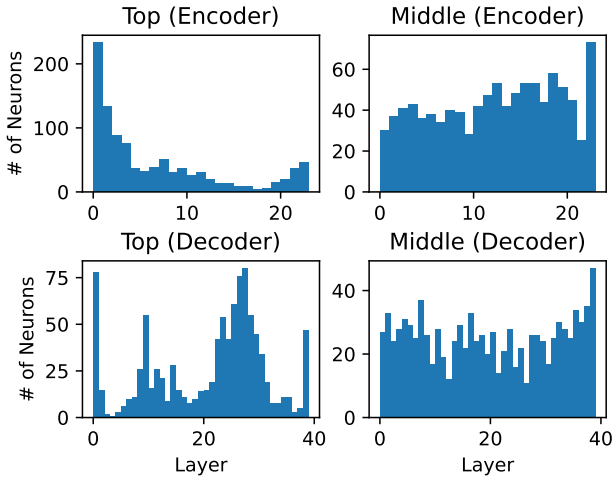
図 2 の分析を，Attention 層と MLP 層に分けて行った結果を図 3 に示す．図 3 (a) で見られる傾向は図 2 と概ね一致しており，Attention 層においてはドメイン特有の語彙に対するニューロンの活動が初期層で行われていることが確認された．図 3 (b) において，Encoder モデルの MLP は Attention 層と同様に前方の層にドメイン特有ニューロンが見られる一方，Decoder モデルではドメイン特有ニューロンが複数のピークに分散していることがわかる．ドメイン知識は MLP 層に蓄積されている [23] ことと合わせると，ドメイン知識はいくつかの層の範囲に分散して蓄積されている可能性が示唆される．Encoder が前方の特定の層に固まってドメイン特有のニューロンを持つことから，ドメイン適応のための学習を行う際には，Encoder の初期層に重点を置くことで効率的にチューニングが可能であると考えられる．Decoder は複数の層に分散してドメイン特有のニューロンを持つと考えられることから，ドメイン適応のための学習を行う際には，Encoder のように初期層のみをチューニングするだけでは不十分である可能性がある．

3.2 金融モデルと汎用モデルの比較

汎用 DeBERTaV2 モデルとそれをベースに追加事前学習を行った金融 DeBERTaV2 モデルに対し，同じ金融テキストを入力し 2.3 節の分析を行った結果を図 4 に示す．3.1 節で見られた Encodwer モデルの傾向とはやや異なり，Attention 層と MLP 層の両方で 2 層目から真ん中の層にかけて山型の分布を形成している．これは，金融モデルと汎用モデルの差異



(a) Attention



(b) MLP

図 3: ドメイン特有のトークンのみに対して AP_m の算出を行い，さらに Attention 層と MLP 層に分けた場合のニューロンの各層における分布．

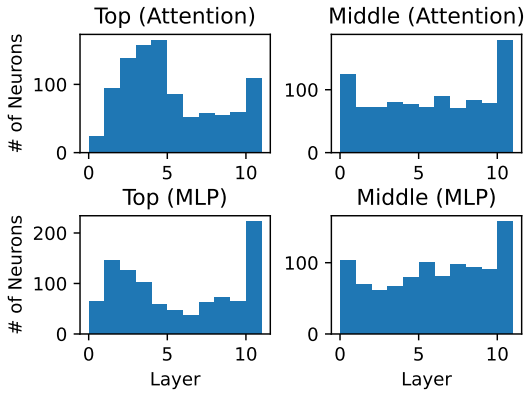


図 4: 金融テキストを入力した金融 DeBERTaV2 とベースモデルの汎用 DeBERTaV2 の出力についてのニューロンの各層における分布．

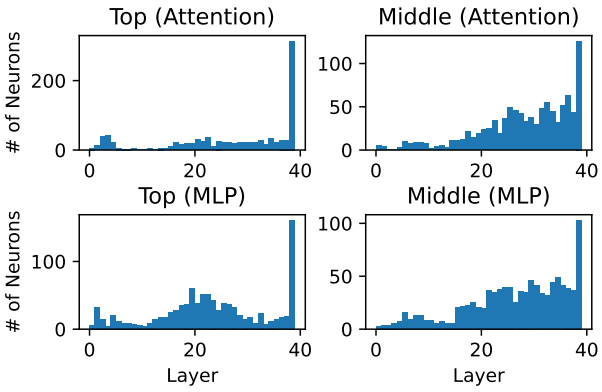


図 5: インストラクションデータを入力したインストラクションモデルとベースモデル（llm-jp-3-13b）の出力についてのニューロンの各層における分布．

は初期のドメイン特有のトークンの解釈ではなく，中間層での意味理解にあらわれていると解釈することができる．

3.3 インストラクションモデルとの比較

llm-jp-3-13b のベースモデルと当該モデルにインストラクションチューニングを行ったインストラクションモデルに対し，同じインストラクションデータを入力し 2.3 節の分析を行った結果を図 5 に示す．Attention 層と MLP 層の両方で，モデル間の大きな差異は観測できなかった．この要因としては，インストラクションデータによるチューニングがモデル内部のニューロン活動に与える影響は比較的小さいことが考えられる．

4 おわりに

本研究では，PLM が金融ドメインと一般ドメインに示す反応の違いを，ニューロン活動の観点から分析した．ドメイン特有の語彙に着目した分析により，金融ドメインのテキストは前方の層のニューロンがより反応した．Decoder モデルの MLP ではドメインの知識がいくつかのピークを形成した．また追加事前学習によるドメイン適応は，モデルの中間層のニューロンの出力に影響を与えていた．一方で，インストラクションチューニングはニューロン活動の観点からは限定的な変化しか与えないことが示唆された．今後の課題としては，ドメイン特有のニューロンがある層を重視したチューニングによる，ドメイン適応の効率性の検証が挙げられる．

参考文献

[1] Masahiro Suzuki, Hiroki Sakaji, Masanori Hirano, and Kiyoshi Izumi. Constructing and analyzing domain-specific language model for financial text mining. **Information Processing & Management**, Vol. 60, No. 2, p. 103194, 2023.

[2] Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. BioMistral: A collection of open-source pretrained large language models for medical domains. In **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 5848–5864, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[3] Masanori Hirano and Kentaro Imajo. Construction of domain-specified japanese large language model for finance through continual pre-training. In **2024 16th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)**, pp. 273–279, 2024.

[4] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. **Nature**, Vol. 620, No. 7972, pp. 172–180, 2023.

[5] Issey Sukeda, Masahiro Suzuki, Hiroki Sakaji, and Satoshi Kodera. JMedLoRA:Medical Domain Adaptation on Japanese Large Language Models using Instruction-tuning. In **Deep Generative Models for Health Workshop NeurIPS 2023**, 2023.

[6] Kota Tanabe, Masahiro Suzuki, Hiroki Sakaji, and Itsuki Noda. JaFIn: Japanese Financial Instruction Dataset. In **2024 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CIFEr)**, pp. 1–10, 2024.

[7] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance, 2023.

[8] Hassan Sajjad, Nadir Durrani, and Fahim Dalvi. Neuron-level interpretation of deep NLP models: A survey. **Transactions of the Association for Computational Linguistics**, Vol. 10, pp. 1285–1303, 2022.

[9] Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In **Advances in Neural Information Processing Systems**, Vol. 36, pp. 16318–16352. Curran Associates, Inc., 2023.

[10] Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. In **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 6919–6971, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[11] Xavier Suau Cuadros, Luca Zappella, and Nicholas Apostoloff. Self-conditioning pre-trained language models. In **International Conference on Machine Learning**, pp. 4455–4473. PMLR, 2022.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics**, pp. 4171–4186. Association for Computational Linguistics, 2019.

[13] BERT large Japanese, （2025-01 閲覧）. https://huggingface.co/tohoku-nlp/bert-large-japanese-v2.

[14] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In **International Conference on Learning Representations**, 2021.

[15] DeBERTa V2 base Japanese, （2025-01 閲覧）. https://huggingface.co/izumi-lab/deberta-v2-base-japanese.

[16] llm-jp-3-13b,（2025-01 閲覧）. https://huggingface.co/llm-jp/llm-jp-3-13b.

[17] Sarashina1-7b,（2025-01 閲覧）. https://huggingface.co/sbintuitions/sarashina1-7b.

[18] 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.

[19] Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. GPT-NeoX-20B: An open-source autoregressive language model. In **Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models**, pp. 95–136. Association for Computational Linguistics, 2022.

[20] 鈴木雅弘, 坂地泰紀, 平野正徳, 和泉潔. Findebertav2: 単語分割フリーな金融事前学習言語モデル. 人工知能学会論文誌, Vol. 39, No. 4, pp. FIN23–G_1–14, 2024.

[21] llm-jp-3-13b-instruct, （2025-01 閲覧）. https://huggingface.co/llm-jp/llm-jp-3-13b-instruct.

[22] Masahiro Suzuki, Masanori Hirano, and Hiroki Sakaji. From Base to Conversational: Japanese Instruction Dataset and Tuning Large Language Models. In **2023 IEEE International Conference on Big Data (BigData)**, pp. 5684–5693, 2023.

[23] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 8493–8502. Association for Computational Linguistics, 2022.

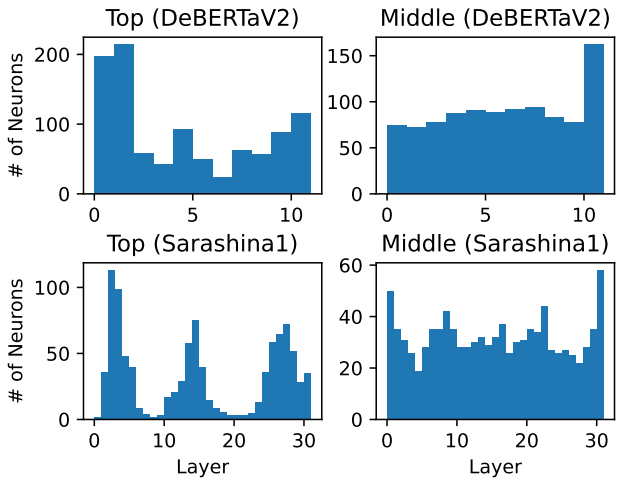


図 6: 平均適合率で降順に並べたときの上位 (Top)・中位 (Middle) のニューロンの，各層における分布．

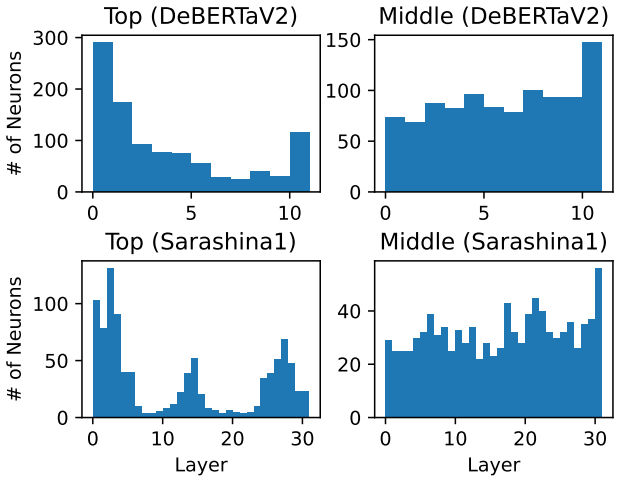
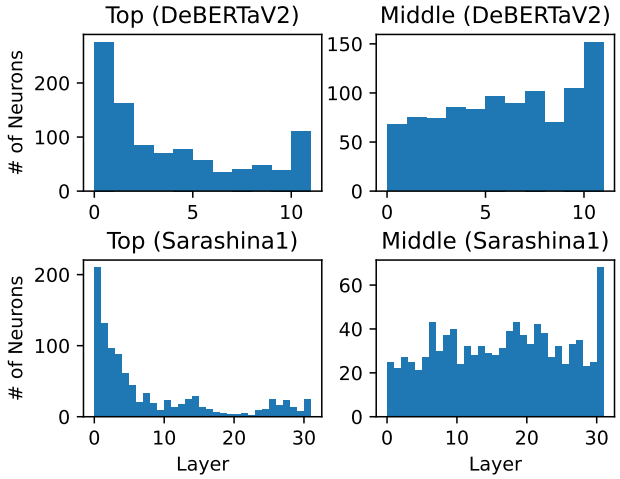
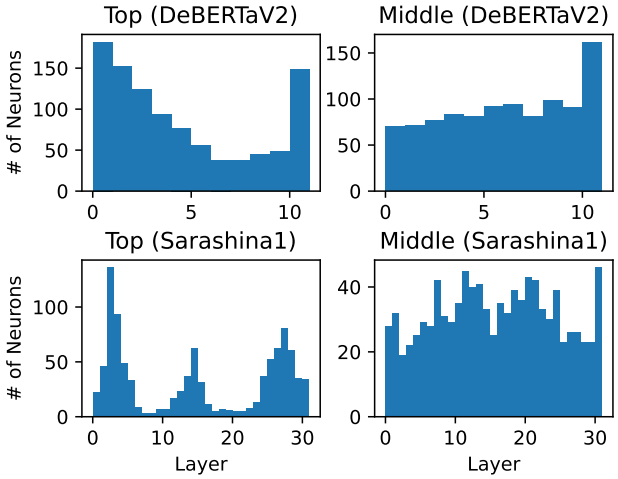


図 7: ドメイン特有のトークンのみに対して AP_m の算出を行い，さらに Attention 層と MLP 層に分けた場合のニューロンの各層における分布．



(a) Attention



(b) MLP

図 8: ドメイン特有のトークンのみに対して AP_m の算出を行い，さらに Attention 層と MLP 層に分けた場合のニューロンの各層における分布．

A　ヒストグラム

Encoder モデルである DeBERTaV2 の日本語 Base モデルと Decoder モデルである Sarashina1-7B についての実験結果を，図 1 から図 3 までに示す．図 6，7，8 は，それぞれ本稿中の図 1，2，3 に対応する．3.1 節において述べた結果は，DeBERTaV2 の日本語 Base モデルと Sarashina1-7B についても観測されたことが確認できる．これは，本研究で得られた知見がモデルのアーキテクチャに依存しない普遍的な性質であることを示唆している．