ニューロン経験勾配による モデル出力の制御と言語知識表現の統合

趙信 [1]　江澤輝 [1]　吉永直樹 [2]
[1] 東京大学大学院　[2] 東京大学 生産技術研究所
{xzhao,zjiang}@tkl.iis.u-tokyo.ac.jp ynaga@iis.u-tokyo.ac.jp

概要

事前学習済み言語モデルの Feed Forward 層に含まれるニューロンは知識や言語スキルを捉えることが知られている．本研究では，ニューロンの活性値に注目し，ニューロン活性値と出力トークン確率との間に線形関係が存在することを明らかにする．次に，この線形関係の勾配（以降，経験勾配）を効率的に計算する手法 NeurGrad を提案する．さらに，新たに構築した MCEval8K データセットを用いた skill neuron probing を通して NeurGrad で計算した経験勾配に基づく分類器を評価し，ニューロン経験勾配が多様な言語タスクを解ける情報量を有することを示す．

1　はじめに

事前学習済み言語モデル (PLM) は，大規模学習により高い言語生成能力を有する一方，誤った知識を生成する幻覚 [1] の問題が深刻であり，モデルの内部機序を理解することが重要となっている．既存研究 [2, 3, 4, 5]．では，知識や言語スキルとモデルのパラメタとの関係をニューロンへの介入実験により調査し，Transformer [6] における Feed Forward (FF) 層のニューロンが知識の記憶領域として機能することや，特定の事実知識 [7, 8, 9] や言語スキル [10, 11] に対応する “knowledge/skill neuron” が存在することを明らかにした．しかし，これらの研究はニューロンの役割を定性的に評価するのみで [12, 10, 13, 14, 11]，ニューロンがモデル出力にもたらす定量的な影響については明らかになっていない．

そこで本研究ではまず，事実知識の評価データセット MyriadLAMA [15] を用いて，PLM に対するニューロン単位の介入実験を行い，ニューロンの活性値を変化させた際に正解の知識に対応するトークンの確率（以下「出力確率」）がどのように変化するかを分析した．その結果，一定範囲内でニューロン活性値の変化量が出力確率と線形関係を示すことを確認した（以降，この線形関係から得られる勾配を**ニューロン経験勾配**と呼ぶ）．ニューロン経験勾配の計算には介入実験による高い計算コストが必要となるため，本研究では経験勾配に基づく分析を促進すべく，経験勾配の効率的な推定手法 NeurGrad を提案する．複数の PLM を用いて真の経験勾配と比較した結果，NeurGrad はベースライン手法 [7] を効率性・精度の両面で上回ることを確認した．

最後に，NeurGrad を活用した skill neuron probing [10] により，ニューロン経験勾配が言語知識を捉えているかを検証する．具体的に，多様な言語スキルを含むデータセットを基に多肢選択形式のベンチマーク (MCEval8K) を新たに構築して，NeurGrad によって計算した経験勾配に基づく分類器を学習・評価し，各タスクにおける skill neuron が，分類器の構築に有益な情報を提供することを示した（図 1）．

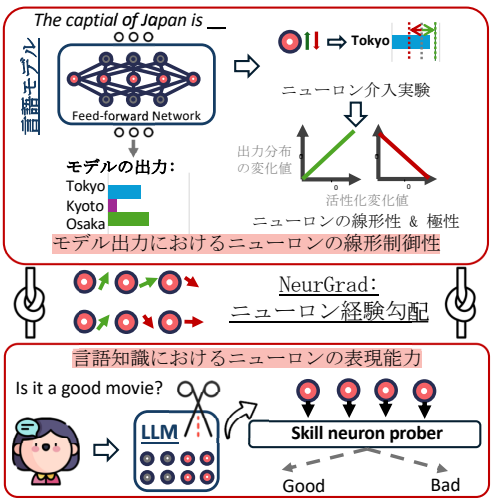

図 1　本研究の貢献：(i) ニューロンの活性値変化による LM 出力の線形制御性の発見．(ii) 経験勾配を効率的に計算する手法 NeurGrad の提案．(iii) 新たに構築した MCEval8K データセットに基づく Skill neuron probing による，経験勾配が多様な言語スキルを捉えることの確認．

2　ニューロンの線形制御性

本節では，PLM の FF 層に含まれるニューロンがモデルの出力にどのように影響を与えるかを，定量的な観点から考察する．具体的には，関係知識評価を対象に，同一プロンプトに対してニューロンの活性値を調整する介入実験を行い，正解トークンの出力確率がどのように変化するかを観測した．

2.1　実験設定

モデル: 本研究では，マスク言語モデルと因果言語モデルの複数のモデルを用いて実験を行った．マスク言語モデルとしては，BERT [6, 16] 系列 ($\text{BERT}_{\text{base}}$，$\text{BERT}_{\text{large}}$，$\text{BERT}_{\text{wwm}}$) を用い，マスク付きプロンプトを作成してモデルにマスクトークンの予測を行わせる．一方，因果言語モデルとしては，Llama2-7B/13B/70B [17] を利用し，先行研究 [15] に従い単一トークンの解答を生成するよう指示する．

データセット: 介入実験には，事実知識評価データセット MyriadLAMA [15] を用いる．本研究では，回答を単一トークンで出力する probing に着目する．具体的には，正解トークンが一つのみのプロンプトを候補にして，モデルが正解を予測できたプロンプトを各 PLM ごとにランダムに 1000 件抽出し，これらを対象として以下のニューロン介入実験を行う．

評価手順: 各ニューロンの活性値を $[-10, 10]$ の範囲でステップ幅 0.2 ずつ変化させ，正解トークンの出力確率がどのように変動するかを観測した．プロンプト・ニューロン・トークンの各対に対し 100 回の推論が必要となり，全ニューロンを対象とすると計算コストが問題となるため，ランダムにサンプリングしたニューロンを対象に介入実験を行う．

2.2　結果とその分析

$\text{BERT}_{\text{base}}$ と Llama2-7B を対象に少数のニューロンの介入実験を行ったところ，ニューロンの活性値の変化と出力確率の変化に強い線形相関が確認された．これを踏まえ，ニューロンの活性値が出力確率に一貫した線形的な影響を与えるかを検証する．

相関の分析: まず，正解トークンの出力確率と活性値変化量の間におけるピアソン相関係数の絶対値を計算する．各活性値変化範囲（例えば± 10）に対して，10 個のプロンプト（各プロンプトで 1000 個のニューロンを対象）を平均した値をこの変化量に対する平均相関係数とする．その結果，図 2 に示すように，全てのモデルで高い相関が得られることがわかった．特に ±2 以内では相関が 0.95 以上となり，強い線形性を示すことを確認した．そのため，以下の実験では活性値変化量を ±2 とし，分析を行う．

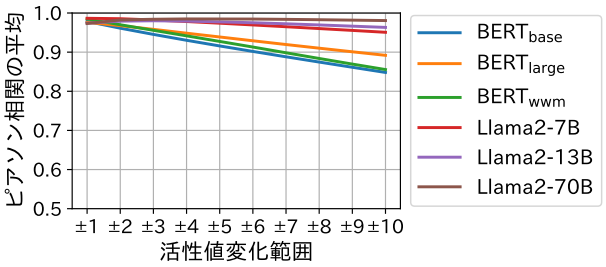

図 2　対象トークンの出力確率とニューロン活性値変化量間のピアソン相関係数の絶対値の平均．

ニューロンの線形性: 次に，異なるプロンプトや Transformer 層においてニューロンが線形性を示す割合と，それらニューロンの汎用性を定量的に分析する．厳密な線形性の定義は存在しないため，以降の議論では，活性値と正解トークンの出力確率の相関係数が活性値の変化量 ±2 の範囲で 0.95 以上となるときに，ニューロンが線形性を示すと判定することとした．計算結果の網羅性を高めるため，各 PLM に対して 1000 個のプロンプトとランダムに 100 個のニューロンをサンプリング[1)]し，合計で 10 万件のニューロン介入実験を実施した．$\text{BERT}_{\text{base/large/wwm}}$ における “線形” ニューロンの割合は，それぞれ 0.9565/0.8756/0.9564 であり，Llama2-7B/13B/70B では 0.9387/0.9677/0.9208 となった．いずれも高い割合となっており，多くのニューロンが線形性を保持することが確認できた．また，線形ニューロンは Transformer 層やプロンプトに依存せず一般的に存在することもわかった（詳細は付録 § A を参照）．

ニューロンの極性: ニューロンの活性値を増加させると，出力確率の変動方向にばらつきが見られた．この変動に基づき，ニューロンの活性値増加によって対象トークンの出力確率が上昇する場合を正極ニューロン，逆に出力確率が低下する場合を負極ニューロンと呼ぶこととする．

3　ニューロン経験勾配とその推定

ニューロンが対象トークンの出力確率に及ぼす影響の強さは，ニューロン介入実験で得られる線形

1)　計算量の配慮から，Llama2-70B では 200 個のプロンプトと 100 個のニューロンのみで介入実験を行った．

	G_C	IG.	NeurGrad
$\text{BERT}_{\text{base}}$	-.9216	.8098	.9999
$\text{BERT}_{\text{large}}$	-.8986	.7317	1.000
BERT_{wwm}	-.9006	.8694	.9999
Llama2-7B	.3020	.5383	.8136
Llama2-13B	-.1973	.7261	.9965
Llama2-70B	.0283	n/a	1.000

表 1 サンプルしたニューロンに対する経験勾配の実測値と推定値のピアソン相関係数．Llama2-70B では GPU のメモリ制限により，IG の結果を得ることができなかった．

関係の勾配（以降，**ニューロン経験勾配**と呼ぶ）により定量化できる．この経験勾配は，ニューロンによる出力の制御能力を示し，その絶対値は制御の強度，符号は制御の方向を示す．本研究では，活性値変化量と出力確率変化量との関係を線形回帰で近似し，得られた回帰係数をニューロン経験勾配の実測値とする．このニューロン介入実験に基づく経験勾配の計算では多数の推論が必要となるため，その計算コストが問題となる．

そこで本研究では，計算勾配[2)]の絶対値が経験勾配の絶対値をおおよそ近似する一方，ニューロンの極性を正確に反映できず，特にニューロン活性値の符号と負の相関が見られることに着目し，効率的かつ高精度に経験勾配を推定する手法 **NeurGrad** を提案する．**NeurGrad** の計算式は以下に示す．

$$\bar{G_E} = G_C \times -\text{sign}(A), \tag{1}$$

ここで，$\bar{G_E}$ は推定される経験勾配，A は活性値，G_C は計算勾配，$\text{sign}(A)$ は A の符号を表す．

経験勾配の計算精度の評価：NeurGrad の有効性を検証するため，ニューロン介入実験により経験勾配の実測値を計算し，NeurGrad によって計算した推定値とのピアソン相関係数を計算する．具体的には 1000 件のプロンプトに対して各 100 個のニューロンをサンプリングし，活性値変化量の範囲を ±2 に設定して介入実験を実施した．経験勾配の推定方法として，(i) 計算勾配 (G_c)，(ii) Integrated gradients (IG) [7]（小ステップごとにニューロンに介入して勾配を近似的に積分する手法），(iii) NeurGrad の 3 手法を比較したところ，表 1 に示すように NeurGrad が経験勾配を最も正確に推定できることを確認した．

計算効率の評価：また，各手法の効率を評価するため，1000 個のプロンプトごとに 1 つのニューロンを用いて介入実験を行った．その結果，計算勾配の計算時間を 1 として，$\text{BERT}_{\text{base}}$ では IG と NeurGrad の計算時間はそれぞれ 25.73，1.04，Llama2-7B では 134.24，1.04 となり[3)]，NeurGrad が効率よく経験勾配を推定できることがわかった．

2) 計算勾配とは，逆伝播を用いて計算グラフから算出される勾配を指す．

4 Skill Neuron Probing

本節では，skill neuron probing [10] において，経験勾配を用いて多様な言語スキルに対応するニューロンを特定し，経験勾配が言語スキルを持つニューロンを特定する情報を有するかを検証する．

4.1 タスク設定

本研究で実施する Skill neuron probing は以下のように定式化される．個別の言語スキルを表すデータセット $\mathcal{D}$ はプロンプト q_i と回答 a_i のペア $\mathcal{D} = \{(q_1, a_1), ..., (q_{|\mathcal{D}|}, a_{|\mathcal{D}|})\}$ から構成される．例えば，評判分析タスクでは q_i は文書，a_i は正解の評価極性ラベルとなる．ニューロン部分集合 $\mathcal{N}_s \subseteq \mathcal{N}$ の経験勾配を特徴量として用い，プロンプト q_i に対する正解回答 a_i を予測する分類器を構築する．ここで，$\mathcal{N}$ はモデルの FF 層が含む全ニューロンの集合を指す．Skill neuron prober は，データセット $\mathcal{D}$ に対して最高精度を達成する $\mathcal{N}_s^*$ を探索する．

$$\mathcal{N}_s^* = \arg\max_{\mathcal{N}_s \subseteq \mathcal{N}} \text{Acc}(f(\mathcal{N}_s), \mathcal{D}) \tag{2}$$

$$\text{Acc}(f(\text{N}_s), D) = \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \mathbb{1}[f(\mathcal{N}_s, q_i) = a_i]. \tag{3}$$

ここで，$f(\mathcal{N}_s, q_i)$ はプロンプト q_i に対して選定したニューロン部分集合 $\mathcal{N}_s$ を用いた分類器 f の出力である．$\mathbb{1}[X = Y]$ は X が Y と一致する場合に 1，それ以外の場合に 0 となる指示関数である．

4.2 経験勾配による Skill Neuron Prober

本節では NeurGrad によって推定された経験勾配を用いて Skill Neuron Prober を学習し，経験勾配が多様な言語スキルを有するニューロンの特定に役立つかを確認する．

経験勾配の極性に基づく Prober (Polar-prober): まず，経験勾配の極性を用いて正解の回答に反応するニューロン (skill neuron) を検出し，多数決分類器を構成する．データセット $\mathcal{D} = \{(q_i, a_i)\}_{i=1}^{|\mathcal{D}|}$ とニューロン $n_k \in \mathcal{N}$ が与えられたとき，各 n_k について，$\mathcal{D}$ 中の全ペアに対してニューロンの極性 $\mathbf{x}_{q_i,a_i}^{n_k}$（§ 2.2）

3) IG は FF 層ごとにニューロン活性値を変更し複数回の推論が必要であるが，NeurGrad は単一の推論で計算可能である．

を計算し，多数となる極性をそのニューロンのスキル極性 $\bar{\mathbf{x}}^{n_k}$ とする．次に，前述の多数決において極性一致度が高いニューロンの上位 $|\mathcal{N}_s|$ 個を skill neuron とみなし，多数決分類器を構成する．テストプロンプト q に対する予測は次式で行う:

$$f(\mathcal{N}_s, q) = \arg\max_{a \in \hat{\mathcal{A}}_{\text{cands}}} \sum_{n_k \in \mathcal{N}_s} \mathbb{1}[\mathbf{x}^{n_k}_{q,a} = \bar{\mathbf{x}}^{n_k}] \qquad (4)$$

ここで，$\mathcal{A}_{\text{cands}}$ は回答候補ラベルの集合である．$|\mathcal{N}_s|$ は開発データを用いて決定する．

経験勾配の大きさに基づく Prober (Magn-prober): 次に，経験勾配の大きさを用いて skill neuron を検出し，多数決分類器を構成する．この Prober では，各 q_i およびニューロン n_k について，全回答候補 $a \in \mathcal{A}_{\text{cands}}$ の経験勾配を取得し比較する．正解の回答候補 a_i に対する経験勾配が他の全ての回答候補の経験勾配より大きければ $\mathbf{x}^{n_k}_{q_i,a_j} = 1$，逆に小さければ $\mathbf{x}^{n_k}_{q_i,a_j} = -1$ と記録する（Polar-prober に倣って，これらを極性と呼ぶこととする）．$\mathcal{D}$ 中の全てのペアに対して極性の多数決を取り，ニューロン n_k のスキル極性 $\bar{\mathbf{x}}^{n_k}$ を求める．このようにして求めたスキル極性を用いて，式 4 と同様に，多数決を行う．

4.3 実験設定

NeurGrad は単一トークンの出力にのみ対応しているため，本研究では 22 のタスクを含む多肢選択評価ベンチマーク MCEval8K を新たに構築して skill neuron probing の実験を行った．MCEval8K[4] は，基礎解析，自然言語推論，文書分類，事実性判定，モデルの自己認知，多言語理解の 6 カテゴリにわたる 22 タスクを含む多肢選択型知識評価ベンチマークであり，多様な言語スキルを包括的に評価する枠組みを提供する（誌面の都合のため構築の詳細については [18] を参照されたい）．

各タスクには，学習用に 6K，開発用に 1K，評価用に 1K の事例がそれぞれ含まれており，Llama2-7B を用いた few-shot 設定で実験を行う．MCEval8K の全データセットに対して，人手で指示を作成した．Skill neuron を用いた多数決分類器の性能は，二つのベースライン（ランダム予測 (Rand) と，正解トークンの出力確率に基づく分類 (LM-Prob)）と比較する．LM-Prob は，最も高い確率を持つ候補トークンを出力し，文脈内学習による性能を評価する．詳細な実験設定は付録 § B を参照されたい．

4) https://huggingface.co/datasets/iszhaoxin/MCEval8K

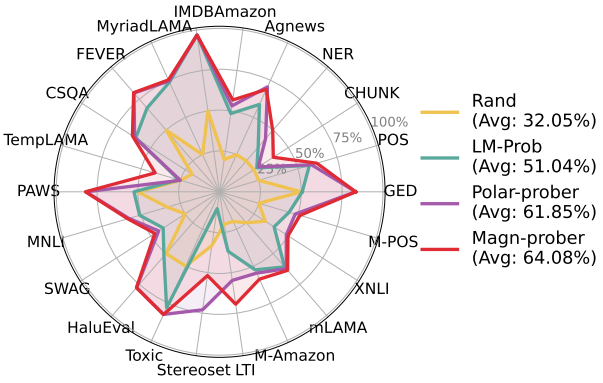

図 3 Llama2-7B における skill Neuron Prober を用いた MCEval8K での精度.

4.4 実験結果

図 3 に，Llama2-7B を用いた MCEval8K の全タスクに対する精度を示す．LM-Prob が Rand を上回っていることから，Llama2-7B が指示を理解して各言語タスクを解く能力を有することが確認できる．各タスクの分類精度を見ると，本研究で用いた単純な多数決に基づく skill neuron prober でも平均して LM-Prob を 10%以上上回る結果が得られた．また，経験勾配の大きさまで考慮して多数決分類器のための skill neuron を選ぶ Magn-Prober が経験勾配の極性のみを考慮して skill neuron を選ぶ Polar-Prober の性能を平均で 2%上回っていることから，ニューロンの経験勾配が言語タスクを解くための情報を保持することが明らかとなった．

5 おわりに

本研究では，事前学習済み言語モデル (PLM) が内包するニューロンの活性値への介入実験を通じて，ニューロンがモデル出力を線形的に制御できることを明らかにした．さらにその線形関係の勾配（ニューロン経験勾配）を効率的かつ正確に推定する手法 NeurGrad を提案し，その有効性を経験勾配との実測値との相関により確認した．この NeurGrad を活用し，新たに構築した広範な言語理解タスクを対象とする MCEval8K データセットを用いた skill neuron probing により，経験勾配が言語スキルを捉える上で有用であることを示した．本研究は，モデルの内部表現と出力の定量的関係を明らかにした．将来，経験勾配を用いて PLM のニューロンの活性値を調整することで，モデル出力を制御し言語タスクの性能向上を実現する可能性を探求する．

謝辞

本研究は，東京大学生産技術研究所特別研究経費，JSPS 科研費 JP21H03494, JP21H03445 および JST, CREST, JPMJCR19A4 の支援を受けたものである．

参考文献

[1] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's song in the ai ocean: A survey on hallucination in large language models. **arXiv preprint arXiv:2309.01219**, 2023.

[2] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In Tal Linzen, Grzegorz Chrupa la, Yonatan Belinkov, and Dieuwke Hupkes, editors, **Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP**, pp. 276–286, Florence, Italy, August 2019. Association for Computational Linguistics.

[3] 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, 2022.

[4] Samet Oymak, Ankit Singh Rawat, Mahdi Soltanolkotabi, and Christos Thrampoulidis. On the role of attention in prompt-tuning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, **Proceedings of the 40th International Conference on Machine Learning**, Vol. 202 of **Proceedings of Machine Learning Research**, pp. 26724–26768. PMLR, 23–29 Jul 2023.

[5] Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 30–45, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

[7] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 8493–8502, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[8] Yifei Wang, Yuheng Chen, Wanting Wen, Yu Sheng, Linjing Li, and Daniel Dajun Zeng. Unveiling factual recall behaviors of large language models through knowledge neurons. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 7388–7402, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[9] Zeping Yu and Sophia Ananiadou. Neuron-level knowledge attribution in 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. 3267–3280, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[10] Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. Finding skill neurons in pre-trained transformer-based language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 11132–11152, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[11] Shaomu Tan, Di Wu, and Christof Monz. Neuron specialization: Leveraging intrinsic task modularity for multilingual machine translation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 6506–6527, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[12] Elena Voita, Javier Ferrando, and Christoforos Nalmpantis. Neurons in large language models: Dead, n-gram, positional. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 1288–1301, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[13] Ran Song, Shizhu He, Shuting Jiang, Yantuan Xian, Shengxiang Gao, Kang Liu, and Zhengtao Yu. Does large language model contain task-specific neurons? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 7101–7113, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[14] Wen Lai, Viktor Hangya, and Alexander Fraser. Style-specific neurons for steering LLMs in text style transfer. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 13427–13443, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[15] Xin Zhao, Naoki Yoshinaga, and Daisuke Oba. What matters in memorizing and recalling facts? multifaceted benchmarks for knowledge probing in language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Findings of the Association for Computational Linguistics: EMNLP 2024**, pp. 13186–13214, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[16] 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.

[17] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models, 2023.

[18] Xin Zhao, Zehui Jiang, and Naoki Yoshinaga. Neuron empirical gradient: Connecting neurons' linear controllability and representational capacity, 2024.

[19] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In Marina Meila and Tong Zhang, editors, **Proceedings of the 38th International Conference on Machine Learning**, Vol. 139 of **Proceedings of Machine Learning Research**, pp. 12697–12706. PMLR, 18–24 Jul 2021.

	Linear neuron ratio	LG	PG
$\text{BERT}_{\text{base}}$	.9999	.9844	.9999
$\text{BERT}_{\text{large}}$	1.000	.9492	.9999
BERT_{wwm}	.9999	.9494	.9999
Llama2-7B	.8136	.9518	.9999
Llama2-13B	.9965	.9833	.9999
Llama2-70B	1.000	.9780	.9999

表 2　線形ニューロンの一般性の評価結果.

A　線形ニューロンの一般性

本節では，PLM に内包されるニューロンの線形性が特定の FF 層やプロンプトに依存せず成り立つかを確認する．具体的に，FF 層における一般性 (**LG**) およびプロンプトに対する一般性 (**PG**) の二つ指標を提案する．二つ指標は直感的には次のような性質を捉えるものである．線形性という観点で性質の異なるニューロンと N 個のビン（FF 層（**LG**）またはプロンプト（**PG**））があり，ニューロンの線形性が一般性を持つかを確認したい場合，以下の 2 点を考えられる．**高いカバレッジ**：線型ニューロンがほとんどのビンに存在する．**均等な分布**：線形ニューロンの数がビンによって大きく異ならない．具体的に，**LG** と **PG** を次のように定義する．

$$\textbf{LG} \triangleq \text{coverage}_{\text{layer}} \times \text{distribution}_{\text{layer}}, \tag{5}$$

$$\textbf{PG} \triangleq \text{coverage}_{\text{prompt}} \times \text{distribution}_{\text{prompt}}, \tag{6}$$

coverage と distribution は以下のように定義される．

$$\text{coverage}_{\text{x}} = \frac{\sum_i \mathbb{1}(\text{linear neuron exists in } x_i)}{\#x}, \tag{7}$$

$$\text{distribution}_{x} = 1 - \frac{\text{Var}(\#\text{neurons in } x)}{\text{maxVar}(\#\text{neurons in } x)}, \tag{8}$$

ここで x はビン（FF 層またはプロンプト）を指し，maxVar(·) は分散の最大値を示す．coverage が 1, distribution が 1 に近いほど，線形ニューロンの一般性が成り立つと言える．図 2 は，各 PLM において 1000 プロンプトとランダムにサンプリングした 100 ニューロンを用いた介入実験に基づく LG，PG，および線形ニューロンの割合を報告する．結果として，線形性を持つニューロンが一般的に存在し，特定のプロンプトや FF 層に依存しないことを示した．

B　Skill Neuron Probing 実験の詳細

B.1　実験設定

本研究では，MCEval8K ベンチマークに対し，Llama2-7B を用いた文脈内学習により，多肢選択問題として正解の選択肢のラベル（単一トークン）を生成する形式で評価を行った．具体的には，各タスクごとに，タスク指示，入出力事例，選択肢を含むプロンプトを設計した．この際，ユーザの実際の使用状況を模倣するために，プロンプトの最適化は行わなかった．（誌面の都合のため作成したプロンプトの例については [18] を参照されたい）．なお，多数ラベルバイアス [19] を回避するため，入出力事例内で全ての選択肢トークン (a,b,c,d など) が一度ずつ出現することを保証した．

言語タスク	Rand	LM-Prob	Polar-prober ($\lvert\mathcal{N}_s\rvert$)	Magn-prober ($\lvert\mathcal{N}_s\rvert$)
GED	.5000	.5060	**.8330** (16)	**.8330** (64)
POS	.2500	.5730	.5870 (4)	**.6210** (16)
CHUNK	.2500	.2710	.2820 (8192)	**.3910** (64)
NER	.2500	.3610	.4300 (4)	**.4970** (64)
Agnews	.2500	.5880	**.7060** (64)	.6890 (512)
Amazon	.2000	.4840	.5310 (1)	**.5680** (128)
IMDB	.5000	.9700	**.9700** (64)	.9690 (64)
MyriadLAMA	.2500	.7380	.7450 (256)	**.7530** (4096)
FEVER	.5000	.6780	.8000 (1)	**.8030** (4)
CSQA	.2000	.6100	.6180 (32)	**.6340** (8192)
TempLAMA	.2500	.2600	.2500 (1)	**.4110** (4)
PAWS	.5000	.5240	.8180 (16)	**.8210** (32)
MNLI	.3333	.5100	.5780 (32)	**.5860** (64)
SWAG	.2500	.4100	.4430 (256)	**.4710** (64)
HaluEval	.5000	.5200	.7750 (2048)	**.7770** (256)
Toxic	.5000	.7800	.8250 (8)	**.8260** (4)
Stereoset	.3333	.1040	**.7297** (128)	.5180 (16)
LTI	.2000	.3680	.5480 (64)	**.6950** (8)
M-POS	.2500	.4440	.4830 (4)	**.5130** (8)
M-Amazon	.2000	.5250	.5470 (1024)	**.5880** (128)
mLAMA	.2500	.6080	.6230 (8192)	**.6360** (512)
XNLI	.3333	.3970	.4860 (32)	**.4980** (32)
Macro average	.3205	.5104	.6185	**.6408**

表 3　Llama2-7B を用いた MCEval8K の各タスクの精度.

B.2　実験結果

表 3 に，skill neuron prober とベースライン手法を用いた各タスクの分類精度を示す．多数決分類を用いた Prober の評価では，開発データを用いて最適なニューロン数 (#n_neurons) を $2^N (1 \le N \le 16)$ の中から探索した．結果として得られた各タスクにおける最適な $|\mathcal{N}_s|$ の値も併せて報告する．表 3 の結果から，skill neuron prober が少数のニューロンで高精度を達成可能であることが明らかとなった．特に，ほとんどのタスクにおいて数として 256 以下のニューロンで最適な精度を達成しており，これは経験的勾配が言語スキルの表現において効率的であることを示唆する．