突出次元の特性および発生要因に関する考察

藤原寛隆 [1]　新納浩幸 [2]
[1] 茨城大学大学院理工学研究科情報工学専攻　[2] 茨城大学大学院理工学研究科情報科学領域
{24nm757h, hiroyuki.shinnou.0828}@vc.ibaraki.ac.jp

概要

Transformer ベースの言語モデルでは，ごく少数の次元が他と比べて極端に大きな値を示す現象が確認される．しかしながら，その発生要因は十分に理解されていない．本研究では，言語モデルの隠れ状態に現れる突出次元に着目し，その形成過程と発生要因を分析する．具体的には，複数の言語モデルを対象に，突出次元がどのように発生・推移するのかを調査するとともに，学習過程の解析を通じて，突出次元が学習初期から出現することを示す．また，Layer Normalization を他のモジュールに置き換えた実験により突出次元が消失することが確認され，Layer Normalization が突出次元の直接的な発生要因である可能性が示唆された．

1　はじめに

近年，大規模言語モデルは様々なサービスにおいて広く活用されており，それに伴って多様な言語モデルやその応用手法に関する研究が活発に行われている．一方で，言語モデルの内部表現や計算過程といった内部メカニズムについては依然として十分に解明されていない．

Transformer [1] においては，ごく少数の次元が他と比べて極端に大きな値を取る現象が確認されており [2, 3]，その発生要因については不明な点も多い．このような突出次元は，量子化やアーキテクチャ設計，最適化手法を考える上で無視できない現象であり，その性質や発生要因を明らかにすることは重要である．

本研究では，Transformer ベースの言語モデルに見られる隠れ状態の外れ値について，その形成過程および発生要因を分析する．まず，複数の言語モデルを対象に，突出次元がどのように発生・推移するのかを調査する．次に，学習過程を追跡することで，突出次元が学習初期から出現することを確認する．さらに，Attention [1]，MLP，Layer Normalization [4] といった各モジュール別の分析を通して，Layer Normalization が突出次元の形成に強く関与している可能性を示す．最後に，Layer Normalization を Dynamic Tanh [5] に置き換える実験を行い，Layer Normalization が突出次元の直接的な発生要因であることを示す．

2　関連研究

2.1　Massive Activations

Massive Activations [6] とは，大規模言語モデルにおいて一部のトークンのごく少数の活性値が他と比較して極端に大きな値を示す現象であり，Sun らによって報告された．彼らは，Massive Activations が入力に依存しない実質的なバイアスとして機能していることを示し，特に Self-Attention の計算において一部のトークンに Attention を集中させる働きを持つことを明らかにした．さらに，次式のように Attention に明示的なバイアスとしてパラメータ k', v' を導入することで，Massive Activations の発生を抑制できることを示した．

$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}\left(\frac{Q[K^\top\ k']}{\sqrt{d}}\right)\begin{bmatrix} V \\ v'^\top \end{bmatrix} \quad (1)$$

2.2　Outliers

Transformer における外れ値（outlier）に関する研究は複数報告されている．ここで，外れ値とはすべてのトークンにおいて恒常的に見られるものであり，Massive Activations とは異なる．

Kovaleva ら [2] は，Layer Normalization のスケールおよびバイアスにおいて少数の次元が異常に大きな値を持つことを指摘し，これらの次元を無効化することでモデル性能が著しく低下することを示した．一方，Luo ら [3] は，BERT や RoBERTa の隠れ状態の一部の次元に恒常的な外れ値が存在し，それらが，トークンの位置的な情報と強く関係している可能性を示した．

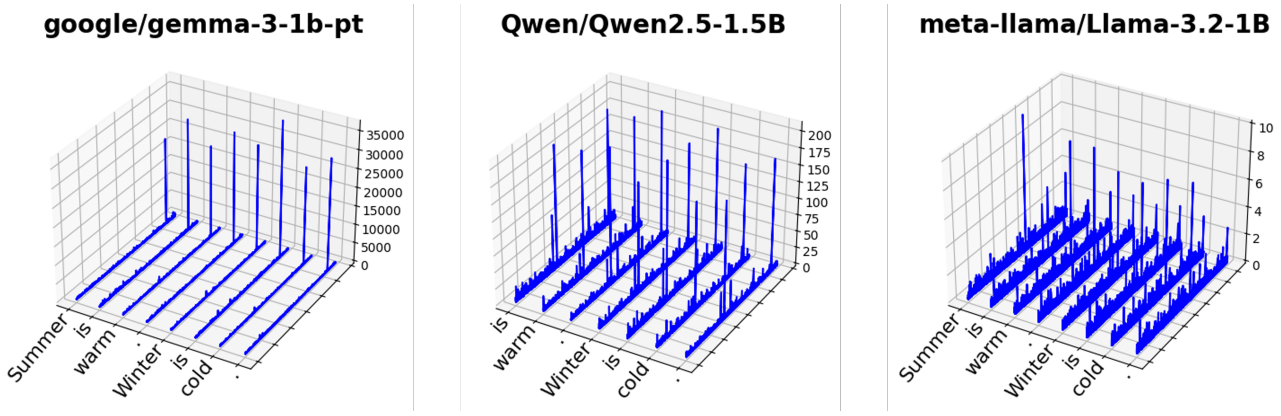

図 1　突出次元の可視化

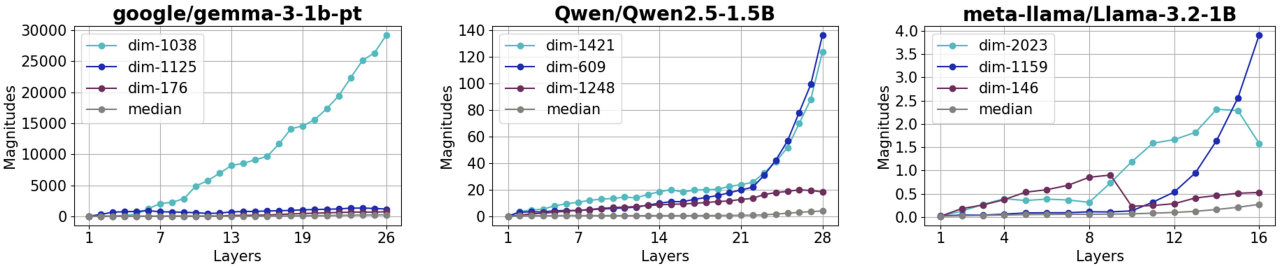

図 2　突出次元の推移

本研究ではこれらを踏まえ，このような突出次元が形成されるより本質的な要因について調査する．

3　突出次元

はじめに，突出次元の存在を視覚的に示す．図 1 は Gemma3-1B[7], Qwen2.5-1.5B[8], Llama-3.2-1B [9] にそれぞれ短い文章を入力し，最終層の出力についてその大きさを可視化したものである．一部の次元が他の次元と比較して極端に突出していることがわかる．これは，入力文章に拠らず任意の系列で確認することができる．本節では，この活性値の突出について，その特性を調査する．

3.1　発生層・発生次元

どの層のどの次元で活性値が突出するのかを調査した．より正確な値を得るため，RedPajama [10] に含まれる Wikipedia 記事から 100 件をサンプリングし，各言語モデルに対して 100 回の生成を行った．得られた隠れ状態について層ごとに平均を取り，その推移を分析する．入力は 1024 トークンに切り詰めている．なお，Massive Activations（活性値が 100 以上，かつ中央値の 1000 倍以上となる次元）が観測されたトークンについては，平均値の算出から除外した．

図 2 は全体を平均して特に大きな値を持つ上位 3 次元と中央値の推移を示したものである．言語モデル間で差異は見られるものの，いずれのモデルにおいても，活性値の突出は比較的浅い層から現れ，最終層に向かって増大する傾向が確認できる．また，突出次元間の大小関係は層によって変化し，一部の次元では途中の層で消失することから，突出する次元は層間で一貫していないことが分かる．

4　発生過程

次に，突出次元が学習の進行に伴ってどのように形成されるかを調査するため，GPT-2 [11] の事前学習を行い，その過程を分析した．訓練設定は Massive Activations の設定を踏襲した[1)]．ただし，計算資源の制約から batch size は 4 とし，また，突出次元の形成に着目するため，Massive Activations の発生を抑制する attention bias (式 1) を使用した．以上の設定の下，訓練を 50k ステップ行った．

4.1　発生ステップ

図 3 はそれぞれの学習ステップごとに短い文章を入力し，最終層の出力についてその大きさを可視化したものである．学習初期には見られなかった活性値の突出が 3000 ステップ以降から確認され，突出次元が学習の比較的早い段階で出現することが分

1)　https://github.com/locuslab/massive-activations

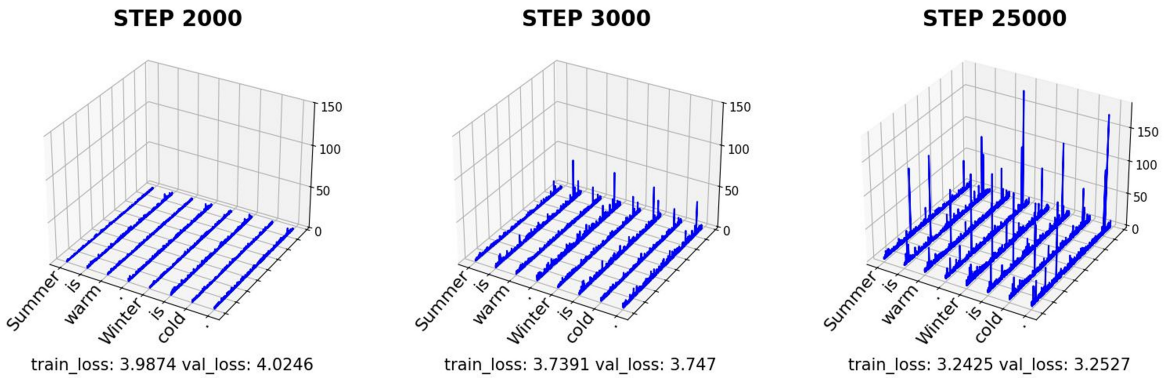

図 3　突出次元の発生過程

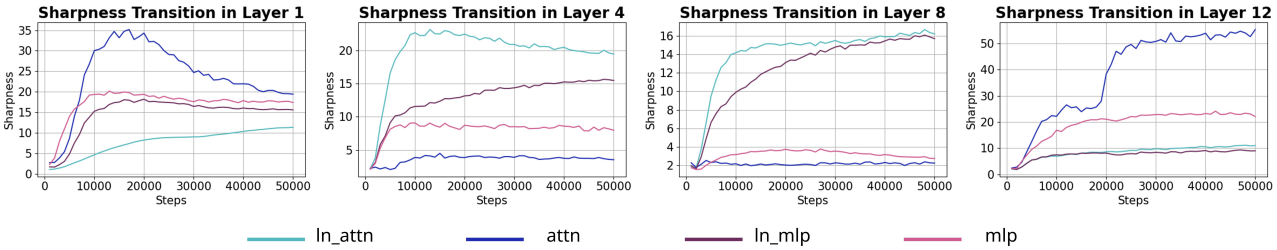

図 4　モジュール別に見た突出度合いの推移

かる．

4.2　発生モジュール

GPT-2 [11] の Decoder Layer は，Attention [1]，MLP，およびそれぞれの入力を正規化する Layer Normalization [4] から構成される．これらのうちどのモジュールにおいて活性値の突出が生じるのかを調査した．具体的には，各モジュールの出力に対して 3.1 と同様に 100 回の生成を行い，得られた隠れ状態の平均について，式 2 で定義される突出度合い Sharpness を算出し，学習過程におけるその推移を分析した．

$$Sharpness_k = \frac{\sum_{k \in \text{top}k} |x_k|}{\sum_{i=1}^{d} |x_i|} \tag{2}$$

ここで，top$k$ は絶対値の大きさが上位 k の次元を表し，d は隠れ状態の次元数である．

図 4 は各層における $Sharpness_3$ の学習ステップに伴う変化を示したものである．浅い層および深い層では Attention や MLP の出力において高い Sharpness が見られる一方で，中間層では Layer Normalization の出力において高い突出度合いを示すことが分かる．このことから，突出次元の形成には層の深さに応じて寄与するモジュールが異なり，特に中間層においては Layer Normalization が重要な役割を果たしている可能性が示唆される．

5　発生要因

前節では，Layer Normalization [4] が突出次元の形成に関与している可能性を示した．また，突出した隠れ状態は，最終層以降のトークン予測や各層の Attention [1] および MLP の計算に先立って Layer Normalization に入力される．これらのことから，活性値の突出は Layer Normalization の処理と密接に関係していると考えられる．

そこで，Layer Normalization を代替可能なモジュールである Dynamic Tanh [5] に置き換え，突出次元の形成にどのような影響が生じるのかを検証した．Dynamic Tanh は学習済みモデルにおける Layer Normalization の入出力関係を近似するように設計されたモジュールであり，次式で定義される．

$$\text{DyT}(\boldsymbol{x}) = \boldsymbol{\gamma} \cdot \tanh(\alpha \boldsymbol{x}) + \boldsymbol{\beta} \tag{3}$$

ここで，$\boldsymbol{\gamma}$，α および $\boldsymbol{\beta}$ は学習可能なパラメータである．本実験では，GPT-2 [11] のすべての Layer Normalization を Dynamic Tanh に置き換え，事前学習を行った．

図 5 はそれぞれの学習ステップにおいて短い文章を入力し，最終層の出力についてその大きさを可視化したものである．同程度の訓練損失を示す学習ステップ同士を比較すると，Dynamic Tanh に置き換えたモデルでは，Layer Normalization を用いたモデル

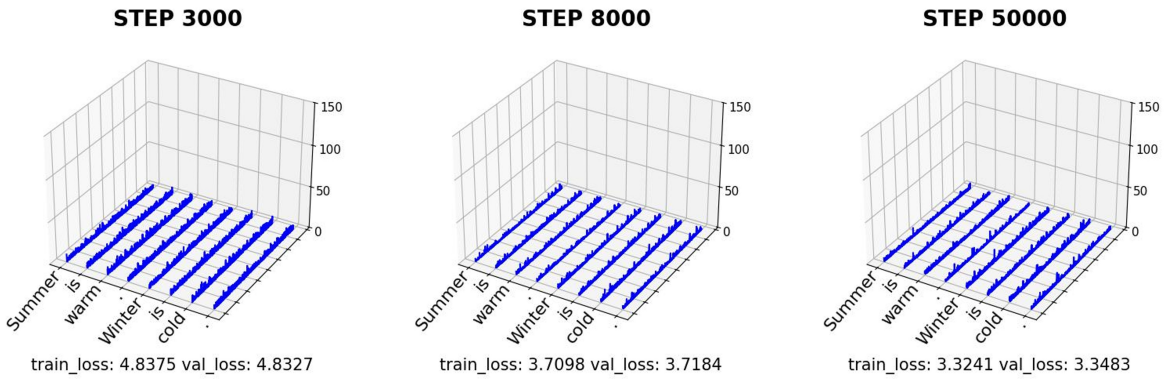

図 5　LayerNorm を DyT に置換した場合，突出次元は消失する

で見られた突出次元が見られないことが分かる．以上の結果から，Layer Normalization が突出次元の直接的な発生要因であることが示唆される．

6　考察

これまでの実験から，突出次元の形成において Layer Normalization [4] が重要な役割を果たしていることが明らかになった．一方で，Layer Normalization が存在するとなぜ突出次元が発生するのかという，より本質的な発生要因については依然として十分に説明されていない．本節では，これまでの実験結果を踏まえ，Layer Normalization の性質に着目してその要因を考察する．

Layer Normalization は次式で定義される．

$$\mathrm{LayerNorm}(\boldsymbol{x}) = \boldsymbol{\gamma} \cdot \frac{\boldsymbol{x} - \mu}{\sigma} + \boldsymbol{\beta} = \boldsymbol{\gamma} \cdot \sqrt{d}\, \frac{\boldsymbol{x} - \mu}{\|\boldsymbol{x} - \mu\|} + \boldsymbol{\beta} \quad (4)$$

ここで正規化処理に着目すると，入力ベクトルをそのノルムで割る操作に等しく，出力ベクトルの大きさが一定に保たれることが分かる．このとき，入力ベクトルの中に特に大きな値を取る次元が存在すると，ノルムは主としてそれらの次元によって決まり，他の次元の影響は相対的に小さくなる．その結果，大きな値を持つ次元では正規化後の出力がほぼ一定となるのに対し，その他の次元では入力の変化がより直接的に反映される (図 6)．このような特性を実現する過程において，突出次元が形成されたと考えられる．実際，大きな入力値に対して出力の変化が緩やかであり，小さな入力値に対しては出力の変化が大きいという性質を持つ Dynamic Tanh [5] では突出次元が見られなかった．このことからも，Dynamic Tanh に類似した入出力特性を実現する過程において，活性値の突出が形成されたと考えることができる．

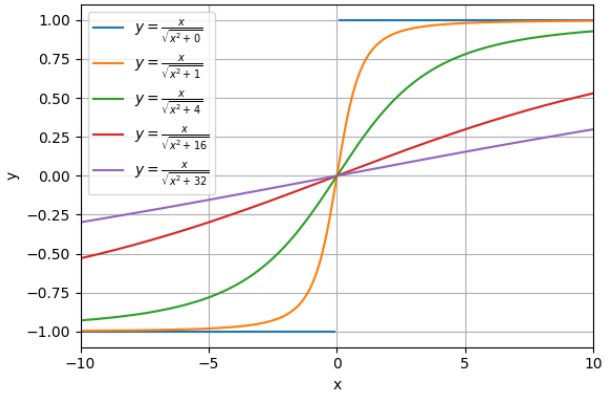

図 6　相対的に小さい次元は線形変換に近づく

7　おわりに

本研究では，言語モデルにおける突出次元の遍在性を示し，その形成過程と発生要因について詳細な分析を行った．

層ごとの解析および学習過程の調査を通じて，突出次元は浅い層から出現し，学習の比較的早い段階で形成されることを示した．また，モジュール単位の分析から，層の深さに応じて寄与するモジュールは異なるものの，Layer Normalization が突出次元の形成と強く関連していることが明らかになった．

さらに，Layer Normalization を Dynamic Tanh に置き換えた実験では，モデル性能が同程度であるにもかかわらず突出次元が消失することを確認した．このことから，Layer Normalization が突出次元の直接的な発生要因である可能性が示唆された．

謝辞

本研究は JSPS 科研費 23K11212 の助成を受けています．

参考文献

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. **CoRR**, Vol. abs/1706.03762, , 2017.

[2] Olga Kovaleva, Saurabh Kulshreshtha, Anna Rogers, and Anna Rumshisky. BERT busters: Outlier layernorm dimensions that disrupt BERT. **CoRR**, Vol. abs/2105.06990, , 2021.

[3] Ziyang Luo, Artur Kulmizev, and Xiaoxi Mao. Positional artefacts propagate through masked language model embeddings. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 5312–5327, Online, August 2021. Association for Computational Linguistics.

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016.

[5] Jiachen Zhu, Xinlei Chen, Kaiming He, Yann LeCun, and Zhuang Liu. Transformers without normalization, 2025.

[6] Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models, 2024.

[7] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025.

[8] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

[9] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models, 2024.

[10] Maurice Weber, Daniel Y. Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. Redpajama: an open dataset for training large language models. **NeurIPS Datasets and Benchmarks Track**, 2024.

[11] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. **OpenAI**, 2019. Accessed: 2024-11-15.