離散構造レポート

tax_free

2024 年 7 月 28 日

1 概要

本レポートでは，5 つの大規模言語モデル (Meta-Llama-3.1-8B-Instruct，gemma-2-9b-it，Mistral-7B-Instruct-v0.2，Swallow-7b-instruct-v0.1，Qwen2-7B-Instruct) のパラメータの最初の桁の分布がベンフォードの法則に従うかを分析した．分析の結果，各モデルのパラメータ分布はおおむねベンフォードの法則に近い形を示し，特定の桁での偏りが観察された．

2 背景・モチベーション

講義でベンフォードの法則が紹介され，その興味深さに惹かれた．特に，LLM(大規模言語モデル) のパラメータにおいてもベンフォードの法則が成り立つのかどうかが気になり，これを調べることにした．

3 分析を行った環境

- 実行環境: Google Colab `https://colab.google/`, 2024 年 7 月 27 日に動作確認
- ソースコード: `https://github.com/taxfree-python/24f_2Q_Discrete_Mathematics/tree/master/tshun/notebooks`

4 対象のモデル

本レポートでは，以下のモデルを使用してパラメータの最初の桁の分布を分析した．

4.1 Meta-Llama-3.1-8B-Instruct

- 開発元: Meta
- 説明: Meta-Llama-3.1-8B-Instruct は Meta Llama 3.1 コレクションの 1 つで，多言語対応の LLM で，事前学習と指示ファインチューニングが施されている．特に多言語対話のユースケースに最適化されている．[1].
- パラメータ数: 8B
- リンク: `https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct`

4.2 gemma-2-9b-it

- 開発元: Google
- 説明: gemma-2-9b-it は，Gemini モデルの研究および技術を基に開発された，英語で利用可能なテキストからテキストへのデコーダ専用の LLM である [2].
- パラメータ数: 9B
- リンク: `https://huggingface.co/google/gemma-2-9b-it`

4.3 Mistral-7B-Instruct-v0.2

- 開発元: Mistral AI
- 説明: Mistral-7B-Instruct-v0.2 は，Mistral-7B-v0.2 の指示ファインチューニング版である．Mistral-7B-v0.2 は，Mistral-7B-v0.1 にコンテキストウィンドウを 8k から 32k に拡張し，Rope-theta の値が 1e6 に設定するという変更を加えたモデルである [3].
- パラメータ数: 7B
- リンク: `https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2`

4.4 Swallow-7b-instruct-v0.1

- 開発元: TokyoTech-LLM
- 説明: Swallow-7b-instruct-v0.1 は，主に日本語データの追加を伴う形で，Llama 2 ファミリーから継続的に事前学習が行われたモデルの 1 つである．Instruction は，教師ありファインチューニング (SFT) を使用している [4].

- パラメータ数: 7B
- リンク: https://huggingface.co/tokyotech-llm/Swallow-7b-instruct-v0.1

4.5 Qwen2-7B-Instruct

- 開発元: Alibaba
- 説明: Qwen2-7B-Instruct は，最大 131,072 トークンのコンテキスト長をサポートし，広範な入力処理が可能である．また，モデルは Transformer アーキテクチャに基づき，SwiGLU 活性化やグループクエリアテンションなどを備えており，大量のデータで事前学習と指示ファインチューニングが行われている [5].
- パラメータ数: 7B
- リンク: https://huggingface.co/Qwen/Qwen2-7B-Instruct

5 結果

今回調べた 5 つのモデルは，おおむねベンフォードの法則に従っていることが分かった．それぞれのモデルについて個別に見ていく．以下の図は，モデルのパラメータの最初の桁の分布を示している．図の説明は次である．

- 横軸: 各数字の全体に対する割合をパーセントで示している．
- 縦軸: 最初の桁として現れる数字 (1 から 9) を示している．
- 棒グラフ: 各数字の実際の出現割合をパーセントで表示している．数字が多いほど棒が高くなる．棒グラフの先端に ** が付いている桁は，ベンフォードの法則から得られる対応する桁の理論値よりも多いものである．
- 折れ線グラフ: ベンフォードの法則に基づく理論的な出現確率を示している．

5.1 Meta-Llama-3.1-8B-Instruct

Meta-Llama-3.1-8B-Instruct のパラメータの最初の桁の分布は，図 1 になった．1, 7-9 の割合がベンフォードの法則よりも多く，逆に 3-6 が理論値よりも小さくなっていた．

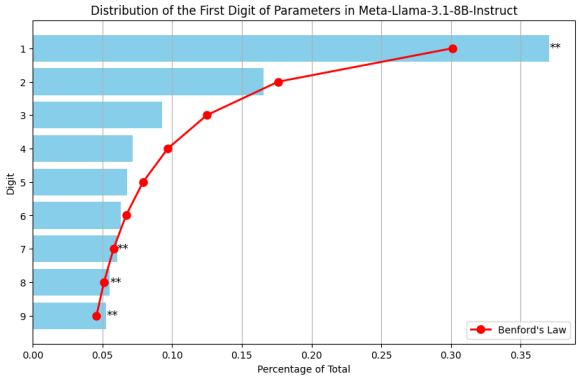

図 1 Meta-Llama-3.1-8B-Instruct のパラメータの最初の桁の分布

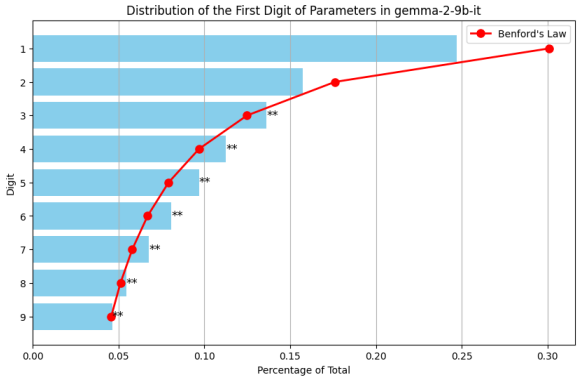

図 2 gemma-2-9b-it のパラメータの最初の桁の分布

5.2 gemma-2-9b-it

gemma-2-9b-it のパラメータの最初の桁の分布は，図 2 になった．3-9 の割合がベンフォードの法則よりも多く，逆に 1-2 が理論値よりも小さくなっていた．

5.3 Mistral-7B-Instruct-v0.2

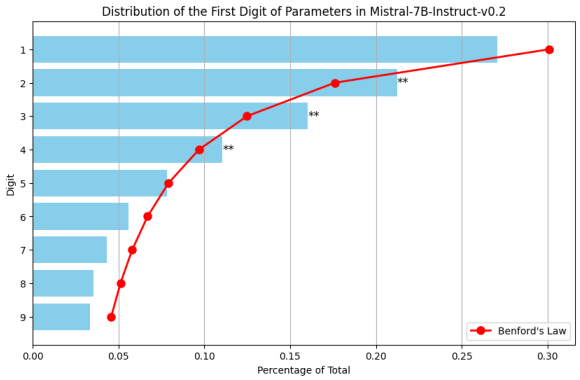

図 3 Mistral-7B-Instruct-v0.2 のパラメータの最初の桁の分布

Mistral-7B-Instruct-v0.2 のパラメータの最初の桁の分布は，図 3 になった．2-5 の割合がベンフォードの法則よりも多く，逆に 1, 6-9 が理論値よりも小さくなっていた．

5.4 Swallow-7b-instruct-v0.1

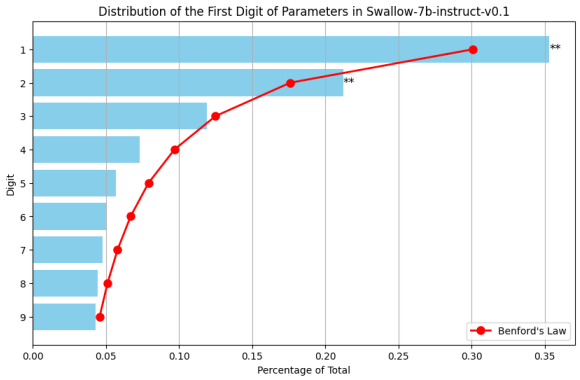

図 4 Swallow-7b-instruct-v0.1 のパラメータの最初の桁の分布

Swallow-7b-instruct-v0.1 のパラメータの最初の桁の分布は，図 4 になった．1-2 の割合がベンフォードの法則よりも多く，逆に 3-9 が理論値よりも小さくなっていた．

5.5 Qwen2-7B-Instruct

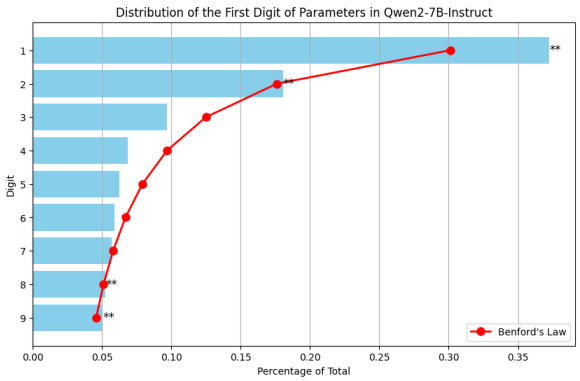

図 5 Qwen2-7B-Instruct のパラメータの最初の桁の分布

Qwen2-7B-Instruct のパラメータの最初の桁の分布は，図 5 になった．1-2, 8-9 の割合がベンフォードの法則よりも多く，逆に 3-7 が理論値よりも小さくなっていた．

5.6 各モデルの理論値との MAE

各モデルの理論値との MAE は表 2 になった．

表 1 各モデルの MAE

Model	MAE
Meta-Llama-3.1-8B-Instruct	0.0185
gemma-2-9b-it	0.0161
Mistral-7B-Instruct-v0.2	0.0189
Swallow-7b-instruct-v0.1	0.0196
Qwen2-7B-Instruct	0.0182

6 議論

6.1 MAE とベンチマークスコアの関係性

表 2 に Open LLM Leaderboard の Average スコア [6, 7, 8, 9, 10, 11, 12, 13, 14](以下スコアと書く) を追加して，そのスコアが高い順に並び変えた．ただし，Swallow-7b-instruct-v0.1 はデータが掲載されていなかったので表にスコアを掲載してない．スコアが計算されている 4 つのモデルに関して，

表 2 各モデルの MAE

Model	MAE	スコア
Meta-Llama-3.1-8B-Instruct	0.0185	26.59
Qwen2-7B-Instruct	0.0182	24.76
gemma-2-9b-it	0.0161	23.18
Mistral-7B-Instruct-v0.2	0.0189	18.44
Swallow-7b-instruct-v0.1	0.0196	-

MAE とスコアのピアソンの積率相関係数を計算すると，-0.1700 となった．しかし，p 値は 0.8300 となったので，これらのデータだけから統計的に有意なことがあるか分からない．

今回は基本的な調査であったが，今後，以下のような仮説について検証したいと考えている．

- パラメータの分布がベンフォードの法則にまったく従わない場合，llm の性能は下がるのか
- より多くのモデルを分析すると MAE と (Open LLM Leaderboard の Average スコア以外を含めた) スコアで有意な相関があるか
- アーキテクチャや最適化アルゴリズムはパラメータの分布に影響を与えるのか
- 対応する言語の種類や数とパラメータの偏りはあるのか

- ファインチューニングを行った場合に分布にどういった変化があるのか (例えば，swallow は Llama 2 を日本語で継続事前学習したモデルである．)
- 特に instruct の前後で大きく分布は変化するのか

参考文献

[1] Meta, *Meta-Llama 3.1 8B Instruct*, https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct, 2024.

[2] Gemma Team, *Gemma*, https://www.kaggle.com/m/3301, Kaggle, 2024.

[3] Mistral AI, *Mistral 7B Instruct v0.2*, https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2, 2024.

[4] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki, *Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities*, arXiv preprint https://arxiv.org/abs/2404.17790, 2024.

[5] Qwen2 Team, *Qwen2 Technical Report*, 2024.

[6] Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf, *Open LLM Leaderboard v2*, https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard, Hugging Face, 2024.

[7] Gao, Leo, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou, *A framework for few-shot language model evaluation*, Zenodo, https://doi.org/10.5281/zenodo.5371628, 2021.

[8] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou, *Instruction-Following Evaluation for Large Language Models*, arXiv preprint https://arxiv.org/abs/2311.07911, 2023.

[9] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei, *Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them*, arXiv preprint https://arxiv.org/abs/2210.09261, 2022.

[10] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt, *Measuring Mathematical Problem Solving With the MATH Dataset*, arXiv preprint https://arxiv.org/abs/2103.03874, 2021.

[11] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman, *GPQA: A Graduate-Level Google-Proof Q&A Benchmark*, arXiv preprint https://arxiv.org/abs/2311.12022, 2023.

[12] Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett, *MuSR: Testing the Limits of Chain-of-thought with Multistep Soft Reasoning*, arXiv preprint https://arxiv.org/abs/2310.16049, 2024.

[13] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen, *MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark*, arXiv preprint https://arxiv.org/abs/2406.01574, 2024.

[14] Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar San-

seviero, Lewis Tunstall, and Thomas Wolf, *Open LLM Leaderboard (2023-2024)*, `https://huggingface.co/spaces/open-llm-leaderboard-old/open_llm_leaderboard`, Hugging Face, 2023.