金融分野における言語モデル性能評価のための日本語金融ベンチマーク構築

平野 正徳
株式会社 Preferred Networks
research@mhirano.jp

概要

大規模言語モデル (LLM) の発展とともに、分野や言語に特化した言語モデルの構築の必要性が議論されてきている。その中で、現在の大規模言語モデルがどの程度の性能を発揮するかを分野に特化して評価するベンチマークの必要性が高まっている。そこで、本研究では、日本語かつ金融分野に特化した複数タスクからなるベンチマークの構築を行い、主要なモデルに対するベンチマーク計測を行った。その結果、現時点では GPT-4 が突出していることと、構築したベンチマークが有効に機能していることを確認できた。

1 はじめに

大規模言語モデル (LLM) は、近年、著しい性能を発揮している。特に、ChatGPT[1] や GPT-4[2] をはじめとした最新の言語モデルは、性能向上と汎化が著しい。その基本技術は Transformer[3] から始まっており、BERT や [4] や、GPT シリーズ [5, 6, 7] などが続いた。ほかにも、Bard[8] や LLaMA[9, 10]、Dolly[11]、BLOOM[12]、Vicuna[13]、PaLM[14, 15] などのモデルが提案されている。

しかしながら、これらの大規模言語モデルは、様々なタスクでどの程度の性能を発揮するかは未知数であり、それらを評価する取り組みが進められている。たとえば、Language Model Evaluation Harness (lm_eval) [16] と呼ばれる、LLM 用の様々なタスクによるベンチマーク計測プラットフォームが提案されている。また、GPT-4[2] においても、様々なタスクにおける性能を評価している。さらに、会計士試験の達成度 [17] や医学分野における応用 [18]、法律分野への応用 [19, 20] を検証する研究などが存在する。

特に、金融分野への応用は、様々なタスクへの応用可能性から、研究開発が進んでいる。金融に特化した非公開のモデルとして、BloombergGPT[21] が存在するほか、公開されているモデルとしては、LLaMA[9] をチューニングした FinLLAMA[22] や、FinGPT[23]、Instruct-FinGPT[24] などがある。しかしながら、金融にフォーカスして、性能を評価するベンチマークはまだない。

加えて、日本語に特化した LLM という点でも、様々な開発が進んでいる。CyberAgent の CALM シリーズや rinna 社のモデル、stabilityai 社の stablelm シリーズ、Elyza 社のモデル、Preferred Networks 社の Plamo™ や LLM-jp-13B など、様々なモデルが乱立しているが、論文化され、性能評価までしっかりと行われているモデルは少ない。また、日本語に特化させるために、既存の英語ベースのモデルをチューニングした研究も存在する [25, 26, 27]。

こうした特化型 LLM の台頭に際して、これらのモデルの性能を正しく評価する特化型のベンチマークの構築も必要であると考えられる。日本語に特化したベンチマークはすでに存在している [28] が、さらに金融に特化したものは存在しない。

そこで、本研究においては、日本語かつ金融分野に特化したベンチマークの構築を行い、主要なモデルに対するベンチマーク計測を行うことで、日本語金融分野における LLM の性能評価を行う。

構築したベンチマークおよび、各モデルの性能評価結果は、https://github.com/pfnet-research/japanese-lm-fin-harness において公開している。

2 日本語金融ベンチマークデータセット

構築したベンチマークにおいては、現時点で全部で 5 つのベンチマークタスクを用いている。

- chabsa: 金融分野における感情分析タスク
- cma_basics: 証券分析における基礎知識タスク
- cpa_audit: 公認会計士試験における監査に関するタスク

- fp2: ファイナンシャルプランナー試験の選択肢問題のタスク
- security_sales_1: 証券外務員試験の模擬試験タスク

これらのうち、chabsa、cpa_audit については、既存のコーパス等を用いてタスクを構築した。残りのタスクについては、インターネット上で公開されているものをクローリングにより取得し、成形することで作成を行った。以下では、それらのタスクについて詳細を述べる。プロンプトの詳細と例については、前述の公開レポジトリを参照されたい。

2.1 chabsa: 感情分析タスク

chabsa とは、金融文書の一種である、有価証券報告書に含まれる文章に関して、特定の単語に対するセンチメントを判定するタスクである。https://github.com/chakki-works/chABSA-dataset で公開されている。このセンチメントには、positive, negative, neutral の 3 種類が存在する。しかしながら、neutral は非常に個数が少なく、安定的な性能評価を妨げる可能性があることから、positive か negative の 2 値分類として扱うこととした。そのため、neutral とタグのついているデータは、positive と negative のどちらを回答しても不正解となることに注意されたい。すべての問題が 2 択問題であるため、ランダムで回答しても、おおよそ 50%の正解を得ることができる。最終評価値としては、positive と negative に対する macro-f1 値を用いる。positive が 4334 件、negative が 3131 件、neurtal が 258 件存在する。そのため、ランダムで回答した場合には、49.15 ポイントの f1 値を得ることができる。

2.2 cma_basics: 証券分析における基礎知識タスク

cma_basics とは、証券分析における基礎知識を問うタスクである。証券アナリスト試験のサンプル問題をクローリングにより取得し、成形することで作成を行った。そのため、一般に言う公益財団法人日本証券アナリスト協会が実施している証券アナリスト試験の 1 次試験や 2 次試験とは異なるものであるが、選択肢問題である点も含め、その 1 次試験と同様の性質を持つ。なお、図を含む問題は削除し、表はマークダウン形式により問題を成形した。すべての問題が 4 択であるため、ランダムで回答しても、25.00%の正解を得ることができる。

2.3 cpa_audit: 公認会計士試験における監査に関するタスク

cpa_audit とは、公認会計士試験における短答式試験監査論の問題を収録したものであり、先行研究 [29] のデータを使用した。6 択問題を 360 問、5 択問題を 38 問収録している。そのため、ランダムに回答した場合には、16.98%の正解を得ることができる。

2.4 fp2: ファイナンシャルプランナー試験の選択肢問題のタスク

fp2 とは、ファイナンシャルプランナー 2 級の選択問題である。日本 FP 協会 2 級ファイナンシャル・プランニング技能検定学科試験の 2021 年 5 月から 2023 年 9 月の過去問題を公式 HP[1] より取得し、加工して作成した。なお、図を含む問題は削除し、表はマークダウン形式により問題を成形した。すべての問題が 4 択であるため、ランダムで回答しても、25.00%の正解を得ることができる。

2.5 security_sales_1: 証券外務員試験の模擬試験タスク

security_sales_1 とは、証券外務員資格試験 1 級に相当する模擬試験のタスクである。外務員資格試験 1 級の模擬試験や対策問題例をクローリングにより取得し、成形することで作成を行った。そのため、一般的な外務員資格試験とは問題構成や難易度に若干の相違があることに注意されたい。4 択問題が 29 問、2 択問題が 28 問収録されている。そのため、ランダムに回答しても、37.28%の正解を得ることができる。

3 各モデルを用いたベンチマーク値の測定

前節で説明したベンチマークを用いて、様々なモデルに対してベンチマークを計測する。

プロンプトによる性能への影響が大きいことから、前節で示したプロンプトのほか、日本語に特化したベンチマークの先行研究 [28] で採用されているすべてのプロンプトと同様のプロンプトを各タスクに対して用意した。これらのプロンプトを用いて、0-4 shot での事前実験を行い、もっともよい性能を発揮したプロンプトと shot 数を採用して最終的な実験を行った。この手順は、ある種の in-sample training にも見えるかもしれないが、実際には、プロンプトの種類を限っていることや、モデル作成者であれば最も適切なプロンプトを選択することは難し

1) https://www.jafp.or.jp/exam/mohan/

表 1 ベンチマーク一覧 (一部省略。詳細は前述の公開レポジトリにて。)

Model	Ave.	chabsa	cma_basics	cpa_audit	fp2	security_sales_1
openai/gpt-4-32k	66.27	93.16	81.58	37.44	50.74	68.42
openai/gpt-4	66.07	93.20	78.95	37.69	50.32	70.18
openai/gpt-4-turbo	64.59	92.86	76.32	36.18	50.95	66.67
openai/gpt-35-turbo	50.27	89.98	52.63	18.09	29.26	61.40
meta-llama/Llama-2-70b-hf	50.21	89.37	57.89	20.85	30.32	52.63
meta-llama/Llama-2-70b-chat-hf	49.53	90.29	52.63	18.84	28.00	57.89
Xwin-LM/Xwin-LM-13B-V0.2	47.53	88.11	52.63	22.11	25.68	49.12
meta-llama/Llama-2-13b-chat-hf	46.98	87.95	52.63	19.60	27.37	47.37
elyza/ELYZA-japanese-Llama-2-7b-fast	46.04	82.52	44.74	17.84	30.74	54.39
lmsys/vicuna-13b-v1.5-16k	45.57	85.81	52.63	19.10	28.21	42.11
mosaicml/mpt-30b-instruct	45.18	83.27	42.11	21.36	26.53	52.63
meta-llama/Llama-2-7b-chat-hf	44.86	83.70	39.47	20.35	29.89	50.88
llm-jp/llm-jp-13b-instruct-full-jaster-v1.0	44.66	85.91	39.47	20.10	26.95	50.88
meta-llama/Llama-2-13b-hf	44.19	82.04	36.84	20.85	30.32	50.88
rinna/youri-7b-instruction	43.84	86.88	34.21	21.61	27.37	49.12
llm-jp/llm-jp-13b-instruct-full-dolly-oasst-v1.0	43.76	83.23	39.47	19.60	27.37	49.12
rinna/youri-7b-chat	43.67	86.67	36.84	19.60	26.11	49.12
cyberagent/calm2-7b-chat	43.67	81.09	36.84	18.09	29.68	52.63
llm-jp/llm-jp-13b-instruct-full-jaster-dolly-oasst-v1.0	43.60	86.83	39.47	18.59	24.00	49.12
lmsys/vicuna-33b-v1.3	43.44	87.81	34.21	19.60	28.21	47.37
lmsys/vicuna-7b-v1.5-16k	43.21	84.78	39.47	19.60	24.84	47.37
mosaicml/mpt-30b-chat	43.10	86.40	39.47	21.36	24.42	43.86
elyza/ELYZA-japanese-Llama-2-7b	42.99	83.48	42.11	19.60	25.89	43.86
pfnet/plamo-13b	42.87	76.97	39.47	21.61	27.16	49.12
mosaicml/mpt-30b	42.80	83.44	36.84	19.60	26.74	47.37
stabilityai/japanese-stablelm-base-alpha-7b	42.73	78.74	34.21	19.10	30.74	50.88
Xwin-LM/Xwin-LM-7B-V0.2	42.73	82.79	42.11	19.85	25.05	43.86
llm-jp/llm-jp-13b-v1.0	42.39	81.24	39.47	19.10	26.53	45.61
openai/text-davinci-003	37.68	53.92	44.74	17.59	26.53	45.61
ランダム	30.68	49.15	25.00	16.98	25.00	37.28

くないということから、このような評価手順をとる方が、公平な比較ができると考えた。

ただし、Open AI 社が API を通じて提供するモデルに関しては、コスト面を考慮し、標準的な 1 つのプロンプトのみを用いることとし、shot 数も 0-shot のみの採用とした。また、Open AI 社の API は、Azure を用いて使用することとし、Content filter が適用されて回答が得られなかった場合には不正解と判定することとした。

また、選択肢問題の解答にあたっては、最も生成尤度の高いものをモデルの出力として利用することとした。ただし、GPT3.5 および GPT-4 シリーズについては、温度パラメータを 0 に指定した状態での出力を API 経由で獲得し、その出力内で最も早く出現する選択肢を出力とした。

結果は表 1 に示す。

4 考察

まず、GPT-4 シリーズのスコアが圧倒的に高いことが明確である。表 1 によれば、ベンチマーク平均スコアが 60 を上回っているのは、GPT-4 シリーズのみである。スコアが 50 を上回るモデルが非常に限られる中で圧倒的なスコアである。GPT-4 は、GPT3.5 や他の 70B クラスのモデルよりもモデルパラメータ数が 1 桁か 2 桁多いとされており、金融タスクにおいても、モデルサイズが大きいほど性能が高い可能性が示唆される。

一方で、スコアが 35〜45 程度のモデルはその性能差は顕著ではなく、モデルパラメータ数の影響が軽微であるようにも見える。これは、モデルの学習の際に、金融に関連したコーパスが非常に限られていることが原因である可能性があり、今後、金融文書を重点的に追加して学習した場合などの効果を確認することで、コーパスの影響を評価できると考えられる。

今回構築したベンチマークの有効性について考える。chabsa に関しては、GPT-4 シリーズはほぼ理論上の上限に近づいてきている。今回のタスクの設計上、95 程度が達成しうる現実的な上限であると考えており、ほぼ限界値であると言っても過言ではない。一方で、それ以外のタスクに関しては、まだまだ伸びしろがある。特に cpa_audit については、現状では正答率が低く、先行研究 [29] でも、GPT-4 と Retrieval-Augmented Generation を組み合わせて、初

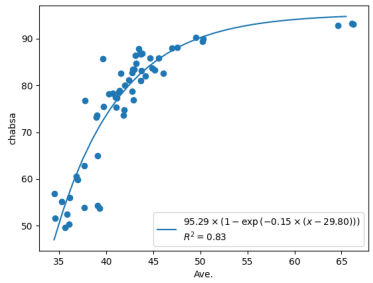

図 1　平均スコアと chabsa の関係

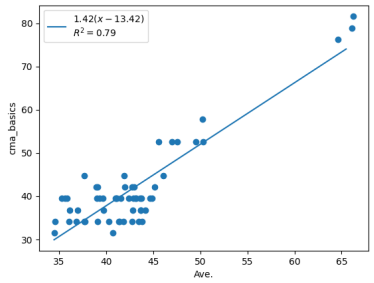

図 2　平均スコアと cma_basics の関係

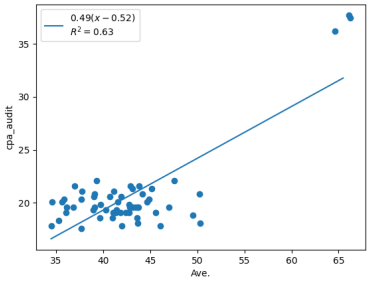

図 3　平均スコアと cpa_audit の関係

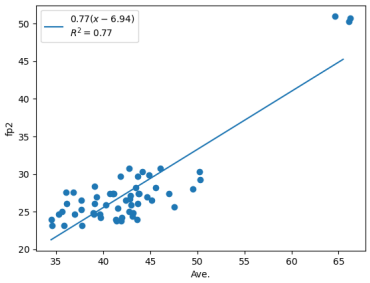

図 4　平均スコアと fp2 の関係

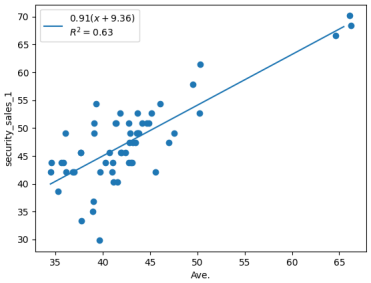

図 5　平均スコアと security_sales_1 の関係

めて合格レベルの性能を発揮しているとの報告が出ており、モデル単体での性能という点では今後の改善の余地があると考えられる。

ここで、さらなるベンチマークの有効性の確認のために、図 1 – 5 に、最終的な平均スコアと各タスクのスコアの関係を示した。無論、平均スコアの 1/5 は各タスクのスコアなので、一定の相関性が認められるのは自明であるものの、あらためて各タスクのスコア分布と平均スコアの関係をプロットした。

図 1 では、プロットの関係が $1 - \exp(x)$ の相似拡大平行移動関係のグラフに見えたため、その関数でフィッティングを行った。この意味としては、タスクとして簡単である傾向があるため、ある程度まで行くと、飽和してしまうという仮説を置いたということである。実際に、フィッティング関数もかなりきれいに当てはまっている。

これらのプロットを見ると、chabsa はやはり比較的簡単なタスクであり、今回のベンチマークにおいては中堅層のスコアの差を拡大させる良い指標になっていると言える。また、cma_basics や security_sales_1 に関しても、下位層は差があまり見られないものの、中堅層でスコアの差が拡大している。一方で、それ以外の指標に関しては、下位・中堅層ともに性能差が観測しにくく、GPT-4 だけが圧倒的に高いという結果になっており、これらに関しては、今後のより性能の高いモデルが出てきた場合に、うまく差を生み出す指標になる可能性が期待される。

今後の課題としては、さらなるタスクの追加や、より妥当なプロンプトチューニング手法の導入、さらには、金融に特化した言語モデルが性能を発揮できるか、という点などがあげられる。

5　結論

本研究では、日本語の金融タスクに特化した LLM 用のベンチマークを構築し、様々なモデルで実際にベンチマークを計測した。その結果、GPT-4 シリーズが圧倒的な性能を発揮することが明らかになった。一方で、ベンチマークとしての有用性も確認することができた。今回構築したベンチマークは、LLM の性能の上位・中位・下位で性能差の出やすさがタスクによって異なっており、結果として、それらの平均は、どの性能領域でも一定の差が出るようなベンチマークになっていることがわかった。今後、より多くのタスクを追加することで、より妥当なベンチマークを構築することができると考えられる。

Declarations

著者は、pfnet/plamo-13b の開発元である、株式会社 Preferred Networks に所属しているが、本研究での実験においては、他のモデルと公平な評価を行っており、透明性の確保のために、すべてのコードを公開している。

参考文献

[1] OpenAI. ChatGPT, 2023. https://openai.com/blog/chatgpt/.

[2] OpenAI. GPT-4 Technical Report, 2023.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In **Advances in Neural Information Processing Systems**, Vol. 30, pp. 5999–6009, 2017.

[4] 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.

[5] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training, 2018. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[6] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners, 2019. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

[7] Tom Brown, Benjamin Mann, et al. Language Models are Few-Shot Learners. **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901, 2020.

[8] Google. Bard, 2023. https://bard.google.com/.

[9] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and Efficient Foundation Language Models. **arXiv**, 2023. https://arxiv.org/abs/2302.13971.

[10] Hugo Touvron, Louis Martin, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. **arXiv**, 2023. https://arxiv.org/abs/2307.09288v2.

[11] Databricks. Dolly, 2023. https://github.com/databrickslabs/dolly.

[12] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. **arXiv**, 2022. https://arxiv.org/abs/2211.05100.

[13] Vicuna. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality, 2023. https://vicuna.lmsys.org/.

[14] Aakanksha Chowdhery, Sharan Narang, et al. PaLM: Scaling Language Modeling with Pathways. **arXiv**, 2022. https://arxiv.org/abs/2204.02311v5.

[15] Rohan Anil, Andrew M. Dai, et al. PaLM 2 Technical Report. **arXiv**, 2023. https://arxiv.org/abs/2305.10403v3.

[16] Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, et al. A framework for few-shot language model evaluation, 2021. https://github.com/EleutherAI/lm-evaluation-harness.

[17] Marc Eulerich, Aida Sanatizadeh, Hamid Vakilzadeh, and David A. Wood. Is it All Hype? ChatGPT's Performance and Disruptive Potential in the Accounting and Auditing Industries. **SSRN Electronic Journal**, 2023. https://papers.ssrn.com/abstract=4452175.

[18] Harsha Nori, Nicholas King, Scott Mayer Mckinney, Dean Carignan, and Eric Horvitz. Capabilities of GPT-4 on Medical Challenge Problems. **arXiv**, 2023. https://arxiv.org/abs/2303.13375v2.

[19] Kwan Yuen Iu and Vanessa Man-Yi Wong. ChatGPT by OpenAI: The End of Litigation Lawyers? **SSRN Electronic Journal**, 2023. https://papers.ssrn.com/abstract=4339839.

[20] Jonathan H. Choi, Kristin E. Hickman, Amy Monahan, and Daniel B. Schwarcz. ChatGPT Goes to Law School. **SSRN Electronic Journal**, 2023. https://papers.ssrn.com/abstract=4335905.

[21] 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. **arXiv**, 2023. https://arxiv.org/abs/2303.17564v2.

[22] Pedram Babaei William Todt, Ramtin Babaei. Fin-LLAMA: Efficient Finetuning of Quantized LLMs for Finance, 2023. https://github.com/Bavest/fin-llama.

[23] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. FinGPT: Open-Source Financial Large Language Models. **arXiv**, 2023. https://arxiv.org/abs/2306.06031.

[24] Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. Instruct-FinGPT: Financial Sentiment Analysis by Instruction Tuning of General-Purpose Large Language Models. **arXiv**, 2023. https://arxiv.org/abs/2306.12659.

[25] llm-japanese-dataset v0: Construction of Japanese Chat Dataset for Large Language Models and its Methodology. In **The 26th International Conference on Network-Based Information Systems**, pp. 442–454, 2023.

[26] Issey Sukeda, Masahiro Suzuki, Hiroki Sakaji, and Satoshi Kodera. JMedLoRA: Medical Domain Adaptation on Japanese Large Language Models using Instruction-tuning. **arXiv**, 2023. https://arxiv.org/abs/2310.10083.

[27] Masahiro Suzuki, Masanori Hirano, and Hiroki Sakaji. From Base to Conversational: Japanese Instruction Dataset and Tuning Large Language Models. **arXiv**, 2023. https://arxiv.org/abs/2309.03412.

[28] StabilityAI. JP Language Model Evaluation Harness, 2023. https://github.com/Stability-AI/lm-evaluation-harness/tree/jp-stable.

[29] Tatsuki Masuda, Kei Nakagawa, and Takahiro Hoshino. Chatgpt は公認会計士試験を突破できるか？: 短答式試験監査論への挑戦. 人工知能学会第 31 回金融情報学研究会, pp. 81–88, 2023.