Construction of a Japanese Financial Benchmark for Large Language Models

Masanori Hirano
Preferred Networks, Inc.
Tokyo, Japan
research@mhirano.jp

Abstract

With the recent development of large language models (LLMs), models that focus on certain domains and languages have been discussed for their necessity. There is also a growing need for benchmarks to evaluate the performance of current LLMs in each domain. Therefore, in this study, we constructed a benchmark comprising multiple tasks specific to the Japanese and financial domains and performed benchmark measurements on some models. Consequently , we confirmed that GPT-4 is currently outstanding, and that the constructed benchmarks function effectively. According to our analysis, our benchmark can differentiate benchmark scores among models in all performance ranges by combining tasks with different difficulties.

Keywords: Large Language Model, Benchmark, Finance, Japanese

1. Introduction

Recently, Large Language Models (LLMs) have demonstrated excellent performance. In particular, the latest models, such as ChatGPT(OpenAI, 2023a) and GPT-4(OpenAI, 2023b), exhibit high performance and significant generalization abilities. The basis of these models begins with the transformer (Vaswani et al., 2017) and BERT(Devlin et al., 2019), and GPT series (Radford et al., 2018, 2019; Brown et al., 2020) were developed using the transformer. Other LLMs have also been proposed, such as Bard(Google, 2023), LLaMA(Touvron et al., 2023a,b), Dolly(Databricks, 2023), BLOOM(Scao et al., 2022), Vicuna(Vicuna, 2023), PaLM(Chowdhery et al., 2022; Anil et al., 2023), and Gemini (Team, 2023).

The major difference between the latest LLMs and previous language models, such as BERT, is that one model can answer questions in multiple languages and domains and respond to questions by following the instructions. Previously, BERT was trained separately in different languages and domains (SUZUKI et al., 2023). However, the latest LLMs, such as GPT4, can freely process multiple languages. Moreover, whereas BERT can only fill in incomplete sentences, the latest LLMs can answer questions in the same manner as humans.

Because of these improvements, the evaluation tasks should be reconstructed. The latest LLM performances far exceed those of previous language models regarding the variety and accuracy of questions they can answer. Therefore, a greater variety of questions is necessary to evaluate LLMs more accurately. Thus, evaluation tasks are important for developing high-performance LLMs.

Currently, some evaluation tasks for LLMs have already been prepared, but are insufficient as concerns domain-specified tasks and those for languages other than English. For instance, a language model evaluation harness (lm_eval) (Gao et al., 2021) was proposed for LLM evaluation using several English tasks. Moreover, several domain-specified tasks have been evaluated using GPT-4(OpenAI, 2023b). Eulerich et al. (2023) evaluated it using certified public accountant (CPA) tests, Nori et al. (2023) tested it in the medical domain, and its applications to legal services were also tested (Iu and Wong, 2023; Choi et al., 2023). However, only a small number of domain-specified tasks have been tested, and the response of LLMs to other tasks is still being investigated comprehensively.

This study focuses on evaluations of the Japanese financial domain. Financial services are relatively large as concerns money spendings. Moreover, according to World Bank data[1], Japan has the third-largest listed capital market in the world as of 2020. Therefore, the usability of LLMs in Japanese and financial domains is a crucial issue.

Several studies have been conducted on Japanese LLMs. Various models such as CyberAgent's CALM series, Rinna's model, stabilityai's stablelm series, Elyza's model, Preferred Networks' Plamo™, and LLM-jp-13B have been proposed. However, few models have been published in academic research papers, and their performances have not been thoroughly evaluated. Other studies have tuned existing English-based models to specialize in Japanese-language use(HIRANO et al., 2023; Sukeda et al., 2023; Suzuki et al., 2023). As for the Japanese task evaluation for LLMs, several benchmarks are available, including the jlm_eval(StabilityAI, 2023), llm-jp-eval(LLM-jp,

[1] https://data.worldbank.org/indicator/CM.MKT.LCAP.CD

2024), and Rakuda benchmarks[2].

However, no benchmarks or LLMs are specified for both Japanese and financial domain.

Thus, this study proposes a new benchmark for the Japanese financial domain and evaluates several models specified for Japanese. The benchmark and performance results of the models are publicly available at https://github.com/pfnet-research/japanese-lm-fin-harness.

2. Related Works

Studies on specialized language models in finance and Japanese have been conducted for a long time. The classic vector embedding technique used in language processing is word2vec (Mikolov et al., 2013). Word2vec has also been used in the financial domain HIRANO et al. (2019). After word2vec, ELMo (Peters et al., 2018), which uses a bidirectional long short-term memory (LSTM) (Schuster and Paliwal, 1997) to pre-train a distributed representation, appeared, along with transformer (Vaswani et al., 2017), which is a good alternative to LSTM in time-series processing, and transformer-based BERT (Devlin et al., 2019).

In contrast, the methodologies to fit language models to specific languages or domains are also pursued. For instance, Howard and Ruder (2018) proposed universal language model fine-tuning. Following this study, some domain- or language-specific language models were developed, such as SciBERT (Beltagy et al., 2019), MedBERT (Rasmy et al., 2021), Japanese BERT[3], and Japanese financial BERT (SUZUKI et al., 2022). Moreover, the methodologies and effects of domain-specified fine-tuning were discussed in(Gururangan et al., 2020; SUZUKI et al., 2023).

In the era of LLMs, although several transformer-based language models have been proposed, as described in the Introduction section, several unknown mechanisms of LLMs exist and numerous trials have been performed.

Several proposed LLMs that focus specifically on finance exist. For instance, BloombergGPT(Wu et al., 2023) is a private LLM focused on finance. In addition, publicly available models, such as FinLLAMA(William Todt, 2023), which is a tuned version of LLaMA(Touvron et al., 2023a), FinGPT(Yang et al., 2023), and Instruct-FinGPT(Zhang et al., 2023), exist.

Japanese-focused LLMs and benchmarks have also been developed, as mentioned in the Introduction section.

However, currently, no LLMs and benchmarks focused on the Japanese financial domain exist. Therefore, in this study, we construct a benchmark.

3. Japanese Financial Benchmark Dataset

We construct a new Japanese financial benchmark for LLMs, comprising the following five benchmark tasks:

- chabsa: Sentiment analysis task in the financial field.
- cma_basics: Fundamental knowledge questions in securities analysis.
- cpa_audit: Tasks on auditing in the Japanese Certified Public Accountant (CPA) exam.
- fp2: Multiple choice questions for 2nd grade Japanese financial planner exam.
- security_sales_1: Practice exam for the 1st grade Japanese securities broker representative test.

For chabsa and cpa_audit, we constructed a dataset using corpora from previous studies. We constructed the remaining tasks by crawling and cleansing the documents available on the Internet. In the following section, we describe these tasks in detail. For each task, an example prompt is shown below, but this is only for illustrative purposes. Several other types of prompts were also prepared, and those prompts were originally written in Japanese. For details of the prompts, please refer to the aforementioned public repository.

3.1. chabsa: Sentiment Analysis Task in the Financial Field

chabsa (Kubo et al., 2018) is a task to determine the sentiments of specific words with respect to sentences contained in securities reports. In Japan, listed companies publish securities reports annually. These data are available from https://github.com/chakki-works/chABSA-dataset. Three types of sentiments exist: positive, negative, and neutral. However, the number of neutral words is extremely small, which may hinder a stable performance evaluation. Therefore, we decided to treat it as a binary classification task, that is, positive or negative classification. This implies that data tagged as "neutral" will be regarded as incorrect regardless of whether the output is positive or negative. Because all the questions were two-choice questions, a random response would yield approximately 50% correct answers. For the final evaluation values, we employed the macro-f1 value.

[2] https://yuzuai.jp/benchmark

[3] https://huggingface.co/tohoku-nlp/bert-base-japanese

In this dataset, 4334 positive, 3131 negative, and 258 neutral responses were observed. Therefore, the random response yields an f1 value of 49.15 points.

An example of chabsa

Please indicate the sentiment of the targeted word in the following sentences, whether positive or negative.

Sentence: The Japanese economy continued to gradually recover during the fiscal year ending March 31, 2012.
Target Word: Japanese economy
Answer: positive

3.2. cma_basics: Fundamental Knowledge Questions in Securities Analysis

cma_basics questions basic knowledge in securities analysis. It was created by crawling and cleansing sample questions from the securities analyst examination. Therefore, it differs from the first and second rounds of the Japanese securities analyst examination administered by the Securities Analysts Association of Japan. However, it has the same characteristics as the first-round test, including a multiple-choice format. In addition, questions containing figures were deleted and the tables were translated into a markdown format. Since all questions had four choices, randomly selecting an answer results in 25.00% accuracy.

An example of cma_basics

Please answer the letter corresponding to the appropriate choice for the following question.

Question:
Which of the following statements about the Japanese economy is incorrect?
A: Real GDP (real gross domestic product) is the level of production activity excluding the effects of price fluctuations.
B: Inflation implies a sustained increase in the general price level.
C: Indirect finance is a form of financial intermediation in which banks and other financial intermediaries play a central role in mediating money lending and borrowing.
D: The fiscal policy of the Bank of Japan adjusts the price level through an increase or decrease in money supply.

Answer:
D

3.3. cpa_audit: Tasks on Auditing in the Japanese CPA Exam

cpa_audit is a collection of short-answer questions on audit theory from the Japanese CPA examination, and data from a previous study (Masuda et al., 2023) were used. It contains 360 questions with six choices and 38 questions with five choices. Therefore, 16.98% of the questions could be answered correctly if they are answered randomly.

An example of cpa_audit

Please answer the letter corresponding to the appropriate combination of symbols to answer the following questions:

Question:
Choose the most appropriate combination of the following statements regarding CPA audits.
(i) In a stock company, the management has a fiduciary responsibility to properly manage and invest the capital contributed by shareholders and provide an accounting report to shareholders regarding the results of this management responsibility. CPA audits of these financial reports contribute to proper management accountability.
(ii) CPA audit not only plays a role in ensuring the reliability of financial information but also supports corporate governance because it encourages the correction of internal control deficiencies and fraudulent acts discovered in the process.
(iii) As listed companies have a significant influence on society, special provisions are placed on CPAs who audit listed companies, such as the prohibition of independent audits, prohibition of certain non-audit attestation services, and restrictions on employment.
(iv) Because a listed company can raise funds widely from general investors, several interested parties arise, and protection against them is necessary. Therefore, establishing a management system for timely and appropriate disclosure of information to stakeholders is necessary. Therefore, CPAs must perform an internal control audit when a company is newly listed.

Choices:
A: (i) and (ii)
B: (i) and (iii)
C: (i) and (iv)
D: (ii) and (iii)
E: (ii) and (iv)
F: (iii) and (iv)

Answer:
A

3.4. fp2: Multiple Choice Questions for 2nd Grade Japanese Financial Planner exam

fp2 is the choice question for a 2nd grade Japanese financial planner exam. The past questions from the Japan FP Association's 2nd grade financial planning skills examination from May 2021 to September 2023 were obtained from the official HP[4] and processed. Questions containing figures were removed, and the tables were translated into a markdown format. Because all the questions had four choices, a random answer yielded 25.00% correct answers.

An example of fp2

Please select the appropriate answer to the following question using numbers from 1 to 4:

Question:
Which of the following statements regarding the conduct of financial planners ("FP") toward their clients is most inappropriate as concerns the relevant laws and regulations?
1. Mr. A, an FP who is not qualified as a lawyer, was consulted by a client about adult guardianship and provided a general explanation on the difference between legal and voluntary guardianship.
2. Ms. B, who is not a licensed tax accountant, received a client's consultation regarding the deduction of medical expenses for income tax purposes and explained that the amount of medical expenses paid, which is compensated for by insurance proceeds, is not deductible as a medical expense deduction.
3. Mr. C, an FP who is not a licensed social insurance consultant, received consultation from a client regarding the deferral of receipt of the basic old-age pension and estimated the pension amount in the case of deferral based on the estimated amount of pension receipt in the client's pension benefit report.
4. Mr. D, an FP who is not registered as a financial instruments business operator, concluded an investment advisory contract regarding asset management with the client and recommended the purchase of individual stocks that were expected to rise in value.

Answer:
4

[4] https://www.jafp.or.jp/exam/mohan/

3.5. security_sales_1: Practice Exam for the 1st Grade Japanese Securities Broker Representative Test

security_sales_1 is a practice exam task that corresponds to the first level of the Japanese securities broker representative test. It was created by crawling and cleansing to obtain practice examinations and sample questions for the 1st-grade Japanese securities broker representative test. Consequently, some differences in the question structure and difficulty levels from official Japanese securities broker representative tests exist. It contains 29 questions with four choices and 28 questions with two choices. Therefore, even if the questions were answered randomly, 37.28% of correct answers could be obtained.

An example of security_sales_1

Please answer the letter corresponding to the appropriate choice for the following question.

Question:
Please answer if the following statement is correct or incorrect:
A securities broker representative is deemed to have the authority to perform all judicial acts on behalf of the financial instrument firm to which they belong with respect to acts prescribed by law, such as the purchase and sale of securities.

Choices:
A: Correct
B: Wrong

Answer:
B

4. Experiments: Benchmark Calculation for LLMs

We measured the benchmarks for various models using the benchmarks described in the previous section.

Given the significant impact of prompts on performance, we prepared prompts for each task in addition to the prompts presented in the previous section. These prompts were similar to those employed in previous Japanese-specific benchmark studies (StabilityAI, 2023). Preliminary experiments with 0–4 shots were conducted using these prompts, and the best-performing prompts and numbers of shots were employed for the final experiment. Although this procedure may seem to be a type of in-sample training, in practice, we believe that such an evaluation procedure would provide a fair comparison. This is because the number of prompts was limited,

Table 1: All Benchmark Results. Some low-performance models are omitted. See full results at the repository as previously mentioned

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
Qwen/Qwen-72B	62.18	92.36	78.95	32.91	40.00	66.67
Qwen/Qwen-72B-Chat	57.89	92.52	78.95	29.90	28.42	59.65
rinna/nekomata-14b	56.03	89.70	63.16	25.13	42.53	59.65
Qwen/Qwen-14B	55.95	90.73	63.16	22.61	38.32	64.91
Qwen/Qwen-14B-Chat	54.71	91.56	65.79	22.36	32.42	61.40
rinna/nekomata-14b-instruction	54.43	91.27	63.16	24.12	37.47	56.14
stabilityai/japanese-stablelm-base-beta-70b	53.07	90.87	60.53	22.36	33.68	57.89
stabilityai/japanese-stablelm-instruct-beta-70b	52.77	91.85	60.53	22.86	36.00	52.63
tokyotech-llm/Swallow-13b-instruct-hf	52.32	87.79	60.53	19.60	35.79	57.89
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
lightblue/qarasu-14B-chat-plus-unleashed	50.04	89.69	57.89	20.35	31.37	50.88
rinna/nekomata-7b-instruction	49.90	90.34	47.37	22.61	27.79	61.40
Qwen/Qwen-7B-Chat	49.86	86.38	50.00	20.85	32.42	59.65
meta-llama/Llama-2-70b-chat-hf	49.53	90.29	52.63	18.84	28.00	57.89
Qwen/Qwen-7B	48.67	85.11	57.89	19.35	30.11	50.88
elyza/ELYZA-japanese-Llama-2-13b	48.37	88.37	47.37	19.35	28.84	57.89
tokyotech-llm/Swallow-13b-hf	48.31	87.59	52.63	19.60	32.63	49.12
Xwin-LM/Xwin-LM-13B-V0.2	47.53	88.11	52.63	22.11	25.68	49.12
rinna/nekomata-7b	47.12	79.18	42.11	21.61	33.05	59.65
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
elyza/ELYZA-japanese-Llama-2-13b-fast	45.70	86.37	39.47	20.60	31.16	50.88
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
elyza/ELYZA-japanese-Llama-2-13b-instruct	44.27	89.40	44.74	18.59	26.53	42.11
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
elyza/ELYZA-japanese-Llama-2-13b-fast-instruct	43.59	87.27	42.11	18.59	26.11	43.86
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
tokyotech-llm/Swallow-7b-hf	42.91	72.27	39.47	19.60	28.84	54.39
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
cyberagent/calm2-7b	41.96	80.02	42.11	17.84	24.21	45.61
rinna/japanese-gpt-neox-3.6b-instruction-ppo	41.89	74.71	44.74	20.60	23.79	45.61
rinna/youri-7b	41.84	73.60	34.21	19.10	29.68	52.63
elyza/ELYZA-japanese-Llama-2-7b-fast-instruct	41.59	82.53	39.47	20.10	25.47	40.35
stabilityai/japanese-stablelm-instruct-alpha-7b	41.43	78.94	34.21	19.35	23.79	50.88
tokyotech-llm/Swallow-7b-instruct-hf	41.36	83.61	31.58	18.09	24.42	49.12
stabilityai/japanese-stablelm-instruct-alpha-7b-v2	41.36	78.62	34.21	19.10	24.00	50.88
pfnet/plamo-13b-instruct	41.13	77.33	39.47	21.11	27.37	40.35
rinna/japanese-gpt-neox-3.6b-instruction-sft-v2	41.03	75.36	39.47	19.10	27.37	43.86
meta-llama/Llama-2-7b-hf	40.99	77.41	39.47	18.59	27.37	42.11
rinna/bilingual-gpt-neox-4b-instruction-ppo	40.71	78.38	31.58	20.60	27.37	45.61
rinna/bilingual-gpt-neox-4b-instruction-sft	40.31	78.23	34.21	19.35	25.89	43.86
llm-jp/llm-jp-1.3b-v1.0	39.70	75.48	36.84	19.85	24.21	42.11
At Random	30.68	49.15	25.00	16.98	25.00	37.28

and it was easy for a human to train the model to select the most appropriate prompts.

However, for the models provided by Open AI through its API, we decided to use only one standard prompt and only 0-shots for the number of shots because of the cost. The Open AI API was used with Azure; if a content filter was applied and no answer was obtained, it was determined to be incorrect.

To answer the multiple-choice questions, the likelihoods of the choices in the context were calculated and the choice with the highest likelihood was employed as the output. For GPT3.5 and GPT-4 series, the outputs with the temperature parameter set to 0 were obtained via API, and the choice that appeared earliest in the outputs was used as the output.

The results are summarized in Table 1.

5. Discussion

According to the results, the GPT-4 series exhibited a significantly high performance. Although the number of parameters in GPT-4 has not been determined, it is estimated to be more than 500 billion. Compared with other models, which have approximately 70 billion or fewer parameters, the number of parameters in GPT-4 is significantly larger, at least a few times. Considering that Qwen-72B exhibited the second-best results, the effect of the number of parameters in the models was important for achieving the highest results.

Compared to the existing Japanese leaderboard, Nejumi[5], our benchmark results for Japanese financial tasks almost correspond to the general Japanese task performance, but an exception exists. Nekomata-14b exhibits a high performance in financial tasks, which differs from that of the Nejumi leaderboard. Nekomata-14b is a tuned model of Qwen-14b that has not yet been evaluated on the Nejumi leaderboard. Moreover, the training corpora for the Qwen series were not revealed, but corpora of professional fields were included according to the official website. Therefore, the corpora used in the training of Qwen may include financial-related texts in their pre-training, and the performance of nekomata-14b is owing to this. However, models other than the nekomata, Qwen, and GPT series are already known to not include financial-related texts in their pre-training.

In the middle score of the benchmarks, around the model exhibiting an overall score of 35–40, no significant differences in their performances or the effect of the number of parameters in the models were present. We believe that this is also related to the corpora used in the training of the models. Currently, several LLMs do not learn financial documents. Therefore, in the future, the impact of financial texts on training should be evaluated, and developing models trained with financial documents is also important.

From the overall summary of the results, the benchmarks that we constructed exhibited considerable variation in difficulty from task to task, and it is possible that we were making an effective assessment. With respect to Chabsa, the highest-performing models approached the theoretical upper limit. For the design of this task, we believe that 95 is a realistic upper limit that can be achieved and is almost at this limit. However, room for further improvement in other tasks still exists, specifically regarding the performance of cpa_audit. A previous study (Masuda et al., 2023) reported that a combination of GPT-4 and retrieval-augmented generation is necessary to achieve a passing level of performance. The model's performance in solving the cpa_audit task without any external information sources can still be improved.

To investigate the effectivity of our benchmark, we analyzed the results, and the plots shown in Figures 1 – 5 were created. The relationships between the overall benchmark score and the individual scores for each task are plotted in Figures 1 – 5. Because 1/5 of the mean score is obtained from each task, a certain degree of correlation can be observed. In Figure 1, the scatter plot appears to be similar to that of $1 - \exp(x)$; therefore, fitting was performed using that function. This implies that the task tended to be easy and saturated for higher-performing models. The fitting function was found to fit well.

According to the plots, each task has its own difficulties. Chabsa is a relatively easy task and a good indicator that the difference in scores widens in lower-performing tiers. In addition, for cma_basics and security_sales_1, there is little difference in the scores of the lower-performing tiers, but the difference in the scores of the mid-performing tiers is increasing. In contrast, for the other indicators, that is, cpa_audit and fp2, observing differences in performance for both the lower and middle-performing tiers is difficult, and only some of the models exhibit overwhelmingly high performance. Because of the inclusion of these tasks with varying difficulty levels, our constructed benchmarks seem to be suitable for evaluating the Japanese financial performance of LLMs.

In future studies, we need to add more tasks, introduce more reasonable prompt-tuning methods, and determine whether a finance-specific language model can perform well.

[5] https://wandb.ai/wandb-japan/llm-leaderboard/reports/Nejumi-LLM-Neo--Vmlldzo2MTkyMTU0

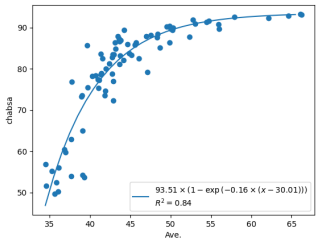

Figure 1: Relationship between Benchmark and chabsa scores

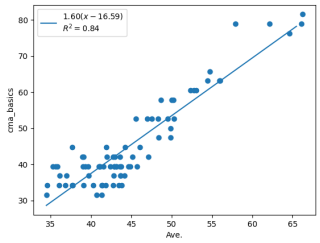

Figure 2: Relationship between Benchmark and cma_basics scores

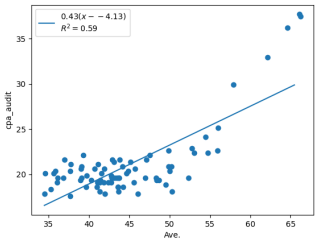

Figure 3: Relationship between Benchmark and cpa_audit scores

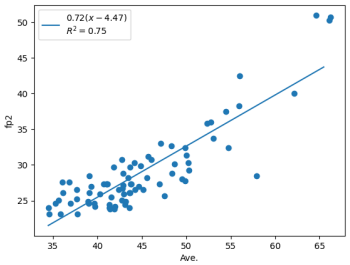

Figure 4: Relationship between Benchmark and fp2 scores

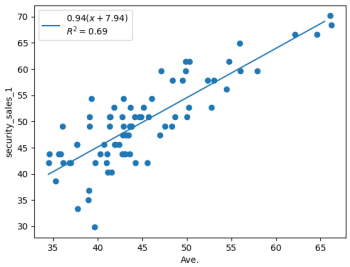

Figure 5: Relationship between Benchmark and security_sales_1 scores

6. Conclusion

In this study, we constructed a new LLM benchmark specialized for Japanese financial tasks and measured the actual benchmarks for various models. The results demonstrated that the GPT-4 series exhibited overwhelming performance. In contrast, we were also able to confirm the usefulness of our benchmark. We confirmed that our benchmark could differentiate the benchmark scores among models in all performance ranges by combining tasks with different difficulties. Future studies should also include more tasks for benchmarking to ensure a more accurate performance evaluation of LLMs.

Declarations

The author is affiliated with Preferred Networks, Inc., the developer of pfnet/plamo-13b, pfnet/plamo-13b-instruct, and pfnet/plamo-13b-instruct-nc. However, in the experiments conducted in this study, all codes were made publicly available for transparency and fair evaluation with other models.

7. Bibliographical References

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

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620.

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

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

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

Databricks. 2023. Dolly. https://github.com/databrickslabs/dolly.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, pages 4171–4186. Association for Computational Linguistics.

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

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

Google. 2023. Bard. https://bard.google.com/.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Masanori HIRANO, Hiroki SAKAJI, Shoko KIMURA, Kiyoshi IZUMI, Hiroyasu MATSUSHIMA, Shintaro NAGAO, and Atsuo KATO. 2019. Related Stocks Selection with Data Collaboration Using Text Mining.

Masanori HIRANO, Masahiro SUZUKI, and Hiroki SAKAJI. 2023. 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*, pages 442–454.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339. Association for Computational Linguistics.

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

LLM-jp. 2024. llm-jp-eval.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 26, pages 3111–3119.

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

OpenAI. 2023a. ChatGPT. https://openai.com/blog/chatgpt/.

OpenAI. 2023b. GPT-4 Technical Report.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 2227–2237. Association for Computational Linguistics.

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

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

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86.

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

M. Schuster and K.K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

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

Issey Sukeda, Masahiro Suzuki, Hiroki Sakaji, and Satoshi Kodera. 2023. JMedLoRA: Medical Domain Adaptation on Japanese Large Language

Models using Instruction-tuning. *arXiv*. https://arxiv.org/abs/2310.10083.

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

Masahiro SUZUKI, Hiroki SAKAJI, Masanori HIRANO, and Kiyoshi IZUMI. 2022. Construction and Validation of a Pre-Training and Additional Pre-Training Financial Language Model [in Japanese]. In *The 28th meeting of Special Interest Group on Financial Informatics of Japanese Society for Artificial Intelligence*, pages 132–137.

Masahiro SUZUKI, Hiroki SAKAJI, Masanori HIRANO, and Kiyoshi IZUMI. 2023. Constructing and Analyzing Domain-Specific Language Model for Financial Text Mining.

Gemini Team. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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

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

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

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

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

Shijie Wu, Ozan ˙ Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. BloombergGPT: A Large Language Model for Finance. *arXiv*. https://arxiv.org/abs/2303.17564v2.

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

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

8. Language Resource References

Kubo, Takahiro and Nakayama, Hiroki and Kamura, Junya. 2018. *chABSA: Aspect Based Sentiment Analysis dataset in Japanese*. PID https://github.com/chakki-works/chABSA-dataset.

Tatsuki Masuda, Kei Nakagawa, and Takahiro Hoshino. 2023. Can chatgpt pass the jcpa exam?: Challenge for the short-answer method test on auditing. In *The 31st meeting of Special Interest Group on Financial Informatics of Japanese Society for Artificial Intelligence*, pages 81–88.