LLaMA-2-Econ: Enhancing Title Generation, Abstract Classification, and Academic Q&A in Economic Research

Onur Keleş, Ömer Turan Bayraklı
Boğaziçi University, Istanbul University
Department of Linguistics, Department of Econometrics
onur.keles1@bogazici.edu.tr, omerturanbayrakli@ogr.iu.edu.tr

Abstract

Using Quantized Low Rank Adaptation and Parameter Efficient Fine Tuning, we fine-tuned Meta AI's LLaMA-2-7B large language model as a research assistant in the field of economics for three different types of tasks: title generation, abstract classification, and question and answer. The model was fine-tuned on economics paper abstracts and syntheticically created question-answer dialogues based on the abstracts. For the title generation, the results of the experiment demonstrated that LLaMA-2-Econ (the fine-tuned model) surpassed the base model (7B and 13B) with few shot learning, and comparable models of similar size like Mistral-7B and Bloom-7B in the BLEU and ROUGE metrics. For abstract categorization, LLaMA-2-Econ outperformed different machine and deep learning algorithms in addition to state-of-the-art models like GPT 3.5 and GPT 4 with both single and representative few shot learning. We tested the fine-tuned Q&A model by comparing its output with the base LLaMA-2-7B-chat with a Retrieval Augmented Generation (RAG) pipeline with semantic search and dense vector indexing, and found that LLaMA-2 performed on a par with the base model with RAG.

Keywords: LLaMA-2, economics, SFT, QLoRA, PEFT

1. Introduction

The evolution of neural networks like RNN (Rumelhart et al., 1986), and LSTM (Hochreiter and Schmidhuber, 1997) architectures and later the invention of the transformer architecture (Vaswani et al., 2017) paved the way for the development of the state-of-the-art Large Language Models (LLMs) such as GPT 3.5 (Ouyang et al., 2022), ChatGPT-4[1] by OpenAI, Gemini (Team et al., 2023) by Google or popular open-source LLMs such as LLaMA-2 (Touvron et al., 2023) by Meta AI, Bloom (Scao et al., 2022), Mistral (Jiang et al., 2023), OPT (Zhang et al.), GPT Neo (Black et al., 2021) and Bart (Lewis et al., 2019) especially for text generation tasks with causal language modeling. Earlier models were trained largely on general corpora (e.g., Wikipedia and books) but now there are a myriad of attempts at injecting open-source LLMs with domain-specific knowledge, including transformers pre-trained on medical and biomedical (Lee et al., 2020), financial (Peng et al., 2021), and scientific text (Beltagy et al., 2019).

Fine-tuning is a process where a pre-trained language model, like BERT (Devlin et al., 2018) or GPT, is specialized for a specific task by further training it on a related dataset, enhancing its performance in the target domain. Furthermore, with the advent of newly emerging methodologies such as Retrieval Augmented Generation (RAG) (Lewis et al., 2020), language models can easily retrieve information and use external data sources. Also, the introduction of Low Rank Adaptation (LoRA) (Hu et al., 2021) and more recently Quantized Low Rank Adaptation (QLoRA) (Dettmers et al., 2023) significantly reduced parameters with less memory needed, enabling training on smaller hardware and faster training times and helped with scalability. Likewise, Parameter Efficient Fine Tuning (PEFT) techniques helped optimize LLMs in terms of applicability across various domains/tasks by fine-tuning only a subset of base model parameters. Although there are a few number of attempts made for large-scale domain or task adaptation purposes (Gema et al., 2023), we acknowledge that there is a scarcity of work dedicated to fine-tuning an open-source LLM assistant in economics for specific research tasks like title generation, abstract classification, and open-ended question & answer (Q&A). To address this gap, this work introduces LLaMA-2-Econ, a PEFT adapted version of the open-source LLaMA-2-7B model by Meta AI, fine-tuned on economics paper abstracts and synthetically created question-answer data for research tasks, specifically for title generation, abstract classification and academic open-ended Q&A.

1.1. Related Work

There are a number of applications of LLMs on related tasks like news headline generation (Gavrilov et al., 2019) or summary generation (Xiao and Chen, 2023). Previous classification methods using transformers include few-shot financial text classification with LLMs like ChatGPT (Loukas et al., 2023b,a). In particular, Loukas et al. demonstrated

[1] http://chat.openai.com

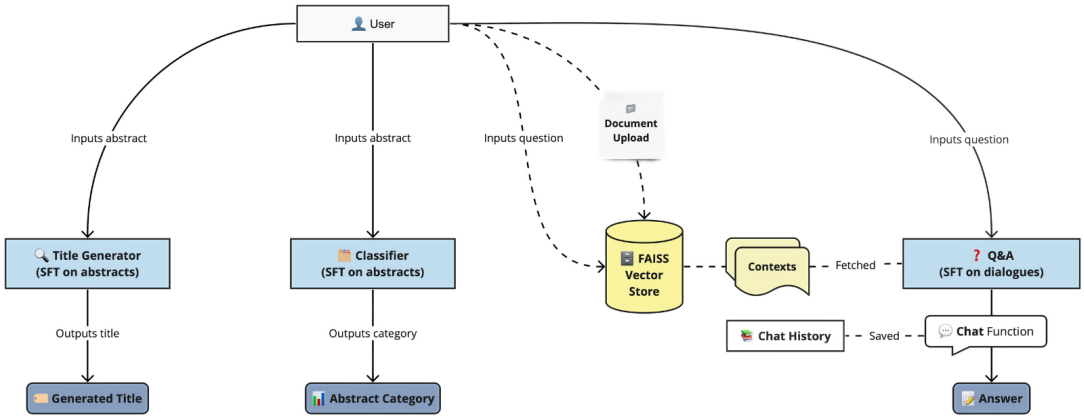

Figure 1: Proposed Workflow for LLaMA-2-Econ. A user inputs an abstract to generate a title and categorize the content using our specialized supervised fine-tuned models. Questions asked by the user are answered through an interactive QA chatbot system that optionally retrieves information from a document vector store, with interactions saved in the chat history.

autoregressive models like GPT 4 can surpass MLM models in text classification. However, their own fine-tuned MP-Net model achieved comparable results in such domain specific tasks. Despite the paucity of research in adapting LLMs in economics, one important contribution is the FinBERT models (Araci, 2019; Yang et al., 2020), BERT-based models trained for financial NLP tasks to tackle financial sentiment analysis and classification problems, outperforming previous state-of-the-art models.

There has been further exploration into financial sentiment analysis by analyzing sentiments in cryptocurrency-related social media posts (Kulakowski and Frasincar, 2023). The authors introduced CryptoBERT, a model fine-tuned on the cryptocurrency domain from BERTweet, and LUKE, a language-universal cryptocurrency emoji sentiment lexicon, to address the challenges in sentiment analysis across languages in social media, and providing tools for enhancing quantitative trading models with sentiment analysis of social media.

As for Q&A, a significant domain adaptation work is PaperPersiChat, which is an open chat-bot designed for discussing scientific papers for computer science (Chernyavskiy et al., 2023). The authors incorporated summarization and Q&A within a single end-to-end online chat-bot pipeline. They trained a dialogue system with scientific grounding. Finally and more relevantly, a recent work employed a PEFT/LoRA based approach for LLaMA-2 fine-tuning in a multitask financial news analysis, and the experimental results showed that the fine-tuned model performs various tasks like main point highlighting, text summarization, and named-entity extraction with sentiments (Pavlyshenko, 2023). Overall, It is clear that LLMs can prove to be helpful agents in (economic) research, performing tasks ranging from paper summaries, generating headlines and text classification to synthesizing information and editing (Korinek, 2023; Dowling and Lucey, 2023; Horton, 2023). However, most available applications for such tasks are not open-source, and there is a lack of research integrating especially decoder-only and open-source LLMs and economics.

2. Methodology

To this end, this paper will attempt at the following: (i) fine-tune LLaMA-2-7B, an open-source and decoder-only model, for the tasks of paper title generation and abstract classification (econometrics, general economics, and theoretical economics) and LLaMA-2-7B Chat for open-ended academic Q&A with QLoRA and PEFT; (ii) perform experiments on metrics to test fine tuned model against the baseline and other language models for these tasks; (iii) propose a Web application acting as a research assistant in economics, utilizing the fine-tuned models with these tasks and an end-to-end chatbot with RAG integration (Figure 1).

2.1. Data

We obtained the data with the arXiv API[2] and searched for economics papers in the following categories/classes: ec.EM (econometrics), ec.GN (general economics), and ec.TH (theoretical economics). In addition to the category, title, abstract, and other metadata were added to our dataset. We preprocessed the data and filtered out low-quality samples following a manual inspection. In the end, we obtained 6362 samples for the train dataset and 707 for the test dataset (Figure 2).

[2] https://arxiv.org/help/api/index

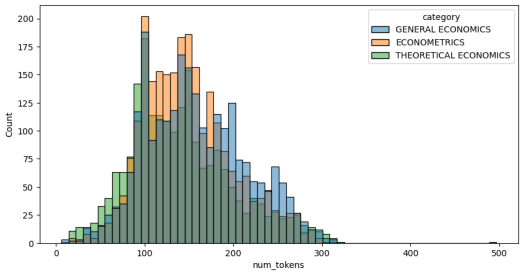

Figure 2: Number of tokens in the training data and their categorical distribution

2.1.1. Creation of Q&A Synthetic Data

Following research that creates synthetic data with state-of-the-art closed source models like ChatGPT (Askari et al., 2023), we have fine-tuned LLaMA-2-7B Chat with a syntheticically created question-and-answer dialogue dataset from academic paper abstracts (7079 in total), employing an approach that utilizes GPT 3.5 Turbo model (costs $0.002 per 1K tokens) from OpenAI that to train the question and answer component of the research assistant. We generated contextual dialogues, where the model both acts as an assistant and user, posing questions and providing answers relevant to a given abstract. Per abstract, we generated 2 to 3 questions and answers. We then filtered out low quality samples, short and incorrectly parsed dialogues, and ended up with 3340 pairs. A sample Q&A pair is given below:

> **What distinguishes revenue management systems? A:** Key characteristics include *fixed capacities*, *homogeneous products*, and sensitivity to *customer pricing decisions*.
>
> **What's problematic with current policy-making indicators? A:** They often subjectively combine a limited number of indicators, overlooking crucial *inter-indicator relationships*.

2.2. Fine Tuning

We fine-tuned Meta AI's LLaMA-2-7B model[3] for the title generation and abstract classification tasks, and LLaMA-2-7B Chat[4] (reinforcement learning with human feedback) using the `transformers` library (Wolf et al., 2020) on a NVIDIA A100 GPU. For the fine-tuning, we used Quantized Low Rank Adaption (QLoRA) with a `lora_r` of 64 and a `lora_dropout` of 0.1. To enhance computational efficiency, we utilized 4-bit precision with a computation dtype of `float16` and the quantization type was set to `nf4`, with nested quantization enabled. The models were scheduled to train for 8 epochs with an early stopping patience of 2 epochs with `bf16` training. Gradient checkpointing and a maximum gradient norm of 0.3 was used. The learning rate was initialized at $2e-4$, using a cosine learning rate scheduler and a warmup ratio of 0.03. Sequences were grouped by length for efficiency and we employed paged AdamW (Loshchilov and Hutter, 2017) with 32-bit precision as the optimizer. The batch size, maximum input and target length were selectively optimized for each task and model.

The PEFT technique we use here integrates fine-tuned components, specifically LoRA weights, into a baseline model, conserving computational resources while keeping the model's task-specific performance. After reloading the model in FP16 for better efficiency and setting up the tokenizer with precision, we then merge these enhancements with the baseline model. This crucial step ensures that our fine-tuning efforts are fully integrated, enhancing the model's overall efficiency and effectiveness. Combined with QLoRA, PEFT allowes for optimized fine-tuning performance and scalability. Our fine-tuned models and dataset are openly available on Huggingface[5].

3. Results

In this section, we report LLaMA-2-Econ's performance on BLEU and ROUGE metrics for the title generation task and compare the results with the baseline LLaMA-2-7B as well as LLaMA-2-13B, Mistal-7B, Bloom-7B and smaller open-source models like GPT Neo and OPT with few shot (5 for this task) learning. As for the classification, we computed the performance metrics and compared the results with those of GPT 3.5 and GPT 4 with one shot and representative few shot (one for each class) learning. We also trained and evaluated different machine learning (ML) and neural network (NN) classifiers. Finally, to evaluate our Q&A model, we measure similarity between LLaMA-2-Econ's generated answers and reference answers obtained through RAG with human verification.

3.1. Experiment 1: Title Generation

As can be seen from the results in Table 1, the fine-tuned model surpasses the baseline and other open-source LLMs of different sizes that use few shot learning. LLaMA-2-13B performs second best in these metrics, followed by other smaller size models.

[3] https://huggingface.co/meta-llama/Llama-2-7b

[4] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

[5] https://huggingface.co/onurkeles/llama-2-7b-econ-abstract-classifier
https://huggingface.co/onurkeles/llama-2-7b-econ-title-generator
https://huggingface.co/onurkeles/llama-2-7b-econ-chat-qa

Model	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
LLaMA-2-Econ (ours)	**0.16**	**0.45**	**0.24**	**0.41**
LLaMA-2-7B (few shot)	0.10	0.41	0.18	0.36
LLaMA-2-13B (few shot)	0.12	0.40	0.19	0.36
Mistral-7B (few shot)	0.11	0.37	0.18	0.33
Bloom-7B (few shot)	0.10	0.37	0.16	0.33
GPT Neo-2.7B (few shot)	0.03	0.19	0.05	0.17
OPT-2.7B (few shot)	0.06	0.25	0.10	0.22

Table 1: Comparison of LLaMA-2-Econ with Other Models in Title Generation

Model	Accuracy	Precision	Recall	F1 Score
LLaMA-2-Econ (ours)	**0.88**	**0.88**	**0.88**	**0.88**
GPT 3.5 (one shot)	0.43	0.63	0.43	0.40
GPT 3.5 (few shot)	0.59	0.72	0.59	0.53
GPT 4 (one shot)	0.70	0.73	0.70	0.64
GPT 4 (few shot)	0.84	0.85	0.84	0.83
Decision Tree Classifier	0.77	0.72	0.77	0.71
K-Nearest Neighbors Classifier	0.79	0.79	0.79	0.79
Logistic Regression	0.85	0.86	0.85	0.85
Random Forest Classifier	0.85	0.86	0.85	0.85
SVC	0.86	0.86	0.86	0.86
XGB Classifier	0.83	0.83	0.83	0.83
RNN	0.81	0.81	0.80	0.81
LSTM	0.81	0.82	0.81	0.82

Table 2: Comparison of LLaMA-2-Econ with Other Models in Abstract Classification

3.2. Experiment 2: Classification

Table 2 shows that LLaMA-2-Econ outperformed other classifiers, having an F1 score of 0.88. Logistic Regression, Random Forest Classifier and SVC achieve comparable scores to our fine-tuned model, followed by GPT 4 with representative few shot (one for each class) and one shot learning, and other ML and neural models. GPT 3.5 both one and few shot (one for each class) performs worst in this abstract classification task.

3.3. Experiment 3: Q&A

As for the neural evaluation for our Q&A Model, we obtained reference open-ended answers to a subset of our synthetically created questions from the base chat model with RAG integration. Following human verification of the answers and inspection, we compared them with LLaMA-2-Econ's generated answers without RAG. We use BERT-Score (Zhang et al., 2019) as our evaluation metric, which calculates the cosine similarity between the embeddings of tokens in our generated answers and those in the reference answers. The formulas to calculate the precision (P), recall (R), and F1-score (F1) where S_{ij} is the similarity score between token i from the candidate answers and token j from the reference answers, $|C|$ is the total number of tokens in the candidate answers, and $|R|$ is the total number of tokens in the reference answers are:

$$P = \frac{1}{|C|} \sum_{i \in C} \max_{j \in R} S_{ij} \tag{1}$$

$$R = \frac{1}{|R|} \sum_{j \in R} \max_{i \in C} S_{ij} \tag{2}$$

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{3}$$

The generated answers by our LLaMA-2-Econ model without RAG (to the questions in the test dataset) received an average precision value of 0.90, recall value of 0.89, and F1 value of 0.90. This means that it achieved commendable similarity with human verified reference responses provided by a RAG implemented base chat model to academic open-ended questions in the domain of economics.

4. Proposed Workflow

Finally, we propose an open application (Figure 1) that can act as an online research assistant which will be openly available to researchers in economics by using open-source fine-tuned models with QLoRA and PEFT. For the chat module of the system, RAG and Facebook AI Similarity Search are employed as well as Langchain[6]'s loader libraries to allow users to load their own economics paper of their own choice or choose one from the provided database.

[6] http://langchain.com

5. Conclusion

In conclusion, we introduced the LLaMA-2-Econ model a QLoRA and PEFT-based model fine-tuned for specific research tasks in the domain economics. Our fine-tuned model performed well in executing different research related tasks, as supported by the metrics achieved against baseline and other state-of-the-art model across various metrics. Our model was also successful in generating reference-like answers to academic questions related to economics research. Overall, we conclude that smaller adapted models with PEFT can be trained on small set of domain specific papers to perform personalized research tasks and obtain comparable results to larger or more advanced models. The integration of QLoRA and PEFT in this study has also shown that scaling large models to new tasks can be more accessible, as it can reduce the need for extensive computational resources. This, of course, further democratizes the use of LLMs in the social sciences, allowing more entities to fine-tune and deploy state-of-the-art models for their specific research needs.

6. Bibliographical References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Arian Askari, Mohammad Aliannejadi, Evangelos Kanoulas, and Suzan Verberne. 2023. A test collection of synthetic documents for training rankers: Chatgpt vs. human experts. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5311–5315.

Salvador Balkus and Donghui Yan. Improving short text classification with augmented data using GPT-3. pages 1–30.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Alexander Chernyavskiy, Max Bregeda, and Maria Nikiforova. 2023. PaperPersiChat: Scientific paper discussion chatbot using transformers and discourse flow management. In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*, pages 584–587.

Colin B. Clement, Matthew Bierbaum, Kevin P. O'Keeffe, and Alexander A. Alemi. On the use of ArXiv as a dataset.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Michael Dowling and Brian Lucey. 2023. Chatgpt for (finance) research: The bananarama conjecture. *Finance Research Letters*, 53:103662.

Raymond Fok, Joseph Chee Chang, Tal August, Amy X. Zhang, and Daniel S. Weld. 2023. Qlarify: Bridging scholarly abstracts and papers with recursively expandable summaries.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Daniil Gavrilov, Pavel Kalaidin, and Valentin Malykh. 2019. Self-attentive model for headline generation. In Leif Azzopardi, Benno Stein, Norbert Fuhr, Philipp Mayr, Claudia Hauff, and Djoerd Hiemstra, editors, *Advances in Information Retrieval*, volume 11438, pages 87–93. Springer International Publishing. Series Title: Lecture Notes in Computer Science.

Aryo Gema, Luke Daines, Pasquale Minervini, and Beatrice Alex. 2023. Parameter-efficient fine-tuning of llama for the clinical domain. *arXiv preprint arXiv:2307.03042*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

John J Horton. 2023. Large language models as simulated economic agents: What can we learn

from homo silicus? Technical report, National Bureau of Economic Research.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Zhiqiang Hu, Yihuai Lan, Lei Wang, Wanyu Xu, Ee-Peng Lim, Roy Ka-Wei Lee, Lidong Bing, and Soujanya Poria. 2023. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. *arXiv preprint arXiv:2304.01933*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. Challenges and applications of large language models.

Anton Korinek. Language models and cognitive automation for economic research.

Anton Korinek. 2023. Generative ai for economic research: Use cases and implications for economists. *Journal of Economic Literature*, 61(4):1281–1317.

Mikolaj Kulakowski and Flavius Frasincar. 2023. Sentiment classification of cryptocurrency-related social media posts. *IEEE Intelligent Systems*, 38(4):5–9.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436–444.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Zongxi Li, Xianming Li, Yuzhang Liu, Haoran Xie, Jing Li, Fu-lee Wang, Qing Li, and Xiaoqin Zhong. Label supervised LLaMA finetuning.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lefteris Loukas, Ilias Stogiannidis, Odysseas Diamantopoulos, Prodromos Malakasiotis, and Stavros Vassos. 2023a. Making llms worth every penny: Resource-limited text classification in banking. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 392–400.

Lefteris Loukas, Ilias Stogiannidis, Prodromos Malakasiotis, and Stavros Vassos. 2023b. Breaking the bank with ChatGPT: Few-shot text classification for finance.

Thanh Thi Nguyen, Campbell Wilson, and Janis Dalins. Fine-tuning llama 2 large language models for detecting online sexual predatory chats and abusive texts.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback, 2022. *URL https://arxiv. org/abs/2203.02155*, 13.

Bohdan M. Pavlyshenko. Financial news analytics using fine-tuned llama 2 GPT model.

Bohdan M Pavlyshenko. 2023. Financial news analytics using fine-tuned llama 2 gpt model. *arXiv preprint arXiv:2308.13032*.

Chandrashekhar S. Pawar and Ashwin Makwana. Comparison of BERT-base and GPT-3 for marathi text classification. In Pradeep Kumar Singh, Sławomir T. Wierzchoń, Jitender Kumar Chhabra, and Sudeep Tanwar, editors, *Futuristic Trends in Networks and Computing Technologies*, volume 936, pages 563–574. Springer Nature Singapore. Series Title: Lecture Notes in Electrical Engineering.

Bo Peng, Emmanuele Chersoni, Yu-Yin Hsu, and Chu-Ren Huang. 2021. Is domain adaptation worth your investment? comparing bert and finbert on financial tasks. In *Proceedings of the Third Workshop on Economics and Natural Language Processing*, pages 37–44.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Issey Sukeda, Masahiro Suzuki, Hiroki Sakaji, and Satoshi Kodera. JMedLoRA:medical domain adaptation on japanese large language models using instruction-tuning.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, {\textbackslash}Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*.

Sharon Whitfield and Melissa A. Hofmann. Elicit: AI literature review research assistant. 19(3):201–207.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. PMC-LLaMA: Towards building open-source language models for medicine.

Le Xiao and Xiaolin Chen. 2023. Enhancing llm with evolutionary fine tuning for news summary generation. *arXiv preprint arXiv:2307.02839*.

Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models, 2022. *URL https://arxiv.org/abs/2205.01068*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.