

# Optimizing Language Models for Real-Time Financial Conversations

ANDREI, GHIURȚU  
UNIVERSITATEA TRANSILVANIA BRAȘOV  
Facultatea de: MATEMATICĂ-INFORMATICĂ  
Specializarea: INFORMATICĂ APLICATĂ  
Email: andrei.ghiurtu@student.unitbv.ro

KARINA, OLARU  
UNIVERSITATEA TRANSILVANIA BRAȘOV  
Facultatea de MATEMATICĂ-INFORMATICĂ  
Specializarea: INFORMATICĂ APLICATĂ  
Email: karina.olaru@student.unitbv.ro

## Abstract

This research focuses on building a custom pipeline centered around a financial large language model that runs well on resource-constrained devices. Some of the most important aspects of our implementation are the use of a synthetic dataset and the text organization method. While the reranker has a teacher-student architecture focused on the financial domain, the generative model is finetuned from a baseline model.

In the subsequent sections of this study, we will thoroughly examine the external services that facilitated the development of our solution, alongside a comprehensive analysis of every significant component within the entire pipeline.

**Keywords:** Small Language Model, fine-tuning, financial domain, Retrieval-Augmented, Generation reranking context-aware



## Introducere

Modelele de limbaj au câștigat popularitate în ultimii ani, cu progrese semnificative în cercetare și optimizări care contribuie la eficacitatea lor în diferite domenii. În timp ce modelele LLM de frunte precum ChatGPT de la OpenAI [1] și Gemini de la Google [2] demonstrează competență în a răspunde la întrebări în mai multe domenii, dependența lor de parametri extinși și resurse computaționale prezintă provocări, în special în contextele financiare. În ciuda eforturilor remarcabile precum FinBert [3], Instruct-FinGPT [4] și BloombergGPT [5], concepute pentru analiza sentimentelor, niciunul nu excelează în răspunsuri la întrebări din domeniul financiar. Aceste modele, cu până la 50 de miliarde de parametri, sunt costisitoare de antrenat și nepotrivite pentru inferență pe dispozitive cu resurse limitate, chiar și quantizate [6].

În contrast cu acestea, pipeline-ul nostru personalizat abordează aceste limitări, prioritizând accesibilitatea și ușurința de utilizare pentru persoanele care doresc să-și îmbunătățească educația financiară. Adaptată pentru gestionarea finanțelor personale, pipeline-ul nostru integrează tehnici avansate de Procesare a Limbajului Natural cu baze de cunoștințe financiare selectate pentru a livra utilizatorilor informații clare și utile.

Acest studiu explorează intersecția dintre modelele de limbaj, datele financiare și inteligența artificială pe dispozitive mobile, valorificând modele precum RoBERTa de la Meta [7] și Qwen1.5 0.5B de la Alibaba [8]. Ne ocupăm de metodologii precum Clusteringul Secvențelor Adiacente și

arhitectura RAG (Retrieval-Augmented Generation), având ca scop îmbunătățirea performanței conversaționale în domeniul financiar.

## Resurse utilizate

Studiul nostru folosește diverse materiale și metode pentru a dezvolta o soluție eficientă. Folosim platforme precum Hugging Face [9] pentru stocarea și explorarea modelelor. Prin intermediul MergeKit [10], o tehnică nouă pentru combinarea modelelor, experimentăm utilizarea algoritmilor precum Passthrough și DARE pentru a îmbunătăți capacitățile modelului.

Bazele de date vectoriale joacă un rol crucial, permițând stocarea și recuperarea eficientă a datelor relevante [11]. Aceste baze de date facilitează căutările rapide de similaritate semantică, îmbunătățind performanța conversațională.

Tehnicile de fine-tuning precum LoRA (Low-Rank Adaptation of LLMs) [12] și QLoRA (Quantized Low-Rank Adaptation of LLMs) [13] optimizează parametrii modelului, îmbunătățind performanța de timp și reduc memoria necesară. În plus, folosim ONNX pentru interoperabilitatea modelului și explorăm tehnici de quantizare pentru a optimiza inferența pe dispozitive cu resurse limitate.

## Soluția propusă

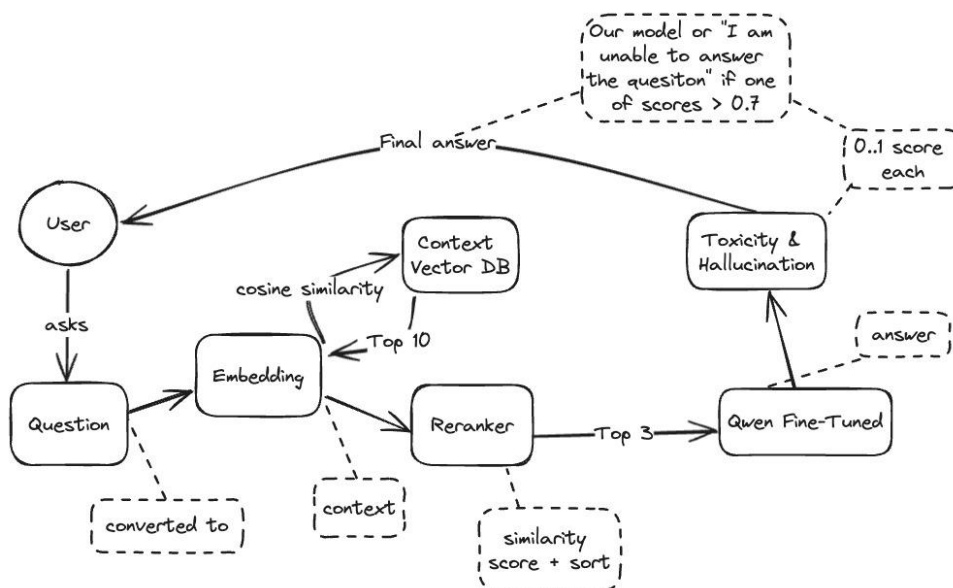


Fig. 1. Pipeline.

Autor: Andrei Ghiurțu

Soluția propusă este un pipeline personalizat, proiectat pentru a îmbunătăți educația financiară pentru utilizatorii obișnuiți. Spre deosebire de soluțiile existente din comunitatea AI, pipeline-ul nostru prioritizează accesibilitatea, fără a avea probleme legate de securitatea datelor trimise în cloud. Integrează tehnici optimizate de procesare a limbajului natural (NLP) cu baze de cunoștințe financiare selectate, adaptate pentru inteligența conversațională. Abordarea noastră include metodologii precum Abordarea de Clustering a Secvențelor Adiacente pentru organizarea datelor textuale și arhitectura Retrieval-Augmented Generation (RAG) pentru performanțe îmbunătățite în conversațiile financiare.

Pipeline-ul extrage date din API-uri precum FinnHub și NewsAPI zilnic, stocându-le într-o instanță Qdrant pentru căutare eficientă. Utilizăm seturi de date sintetice pentru antrenamentul conversațional datorită lipsei seturilor de date fundamentale adaptate explicit pentru domeniul financiar.

Soluția noastră folosește modele precum TinyRoBERTa și Qwen 0.5B, adaptate folosind tehnici precum QLoRA, pentru sarcinile de extragere și generare a răspunsurilor la întrebări. Modelele de detectare a halucinațiilor și toxicității asigură furnizarea de răspunsuri potrivite. În cele din urmă, implementarea modelelor mici și quantizarea permit implementarea pe dispozitive mobile/de mici dimensiuni, precum telefoanele mobile, asigurând accesibilitate și ușurință de utilizare pentru utilizatori.

## Rezultate

În cazul în care este utilizată tehnica fine-tuning-ului, pornind de la un model de bază, succesul acestuia nu este garantat. Pentru a se verifica dacă performanțele în timpul inferenței au fost îmbunătățite și pentru a testa dacă obiectivul inițial, acela de a crește eficacitatea și corectitudinea, mai ales în domeniul țintă, cel financiar, a fost atins, se pot aplica diverse teste și se pot calcula metrici specifice. În cazul studiului prezentat în această lucrare, pentru crearea metricilor s-a utilizat o resursă externă [14] care automatizează procesul de evaluare a modelelor de limbaj. Utilizând metrica MMLU (Massive Multitask Language Understanding), fiecare model a fost testat pe cele șapte domenii care se pot observa în graficele de mai jos. Pentru fiecare domeniu, modelul trebuie să răspundă la un set de întrebări cu răspuns multiplu. Modelele care au fost analizate sunt: Qwen1.5-0.5B, modelul Qwen de bază, care nu este specializat pe domeniul financiar, Qwen1.5-0.5B-FinancialQA, modelul nostru fine-tuned pe conversații financiare, iar ultimele două au la bază acest ultim model, dar optimizat utilizând tehnica DARE, respectiv Passthrough (PT). Se poate observa că în majoritatea domeniilor s-a adus o îmbunătățire față de modelul de bază, cu precădere în domeniul contabilității (*Accounting*). De asemenea, modelul optimizat folosind Passthrough are cele mai slabe performanțe, lucru ce s-a observat și în diversele teste de inferență. Acest tip de metrici sunt utile, deoarece pot fi indicatori a unor erori produse în timpul fine-tuning-ului sau a altor optimizări. Fiind metrici bine-cunoscute în domeniul modelelor de limbaj, pot fi utilizate, de asemenea, ca punct de plecare pentru îmbunătățirea performanței modelelor.



Fig. 2. Rezultate.  
 Autor: Karina Olaru

Table 2. prezintă un exemplu de discuție între utilizator și asistentul financiar, modelul de limbaj obținut în urma fine-tuning-ului. Deși în majoritatea testelor de inferență răspunsurile modelului sunt relativ consistente, așa cum se poate observa și în exemplul oferit, informația este prezentată pe larg, încercând să ofere cât mai multe detalii despre subiectul abordat.

User:	Hi
Qwen Assistant:	Hello! How can I assist you today?
User:	SUA has 80% increase in stock prices.
Qwen Assistant:	This statement is quite vague and there isn't much information to provide. If the company is an exchange-traded fund (ETF) then SUA is the ETF name and its value has increased by 80%. However, the reason for this increase might not be clear. In general, it's possible that the ETF is gaining market share or capital appreciation because the shares have become more

	expensive and investors are willing to pay more for them. This is a common phenomenon as long as companies continue to expand.
User:	What USA stands for?
Qwen Assistant:	The USA is the official country code top-level domain (ccTLD) for the United States of America. It is used for websites related to the American people, business and organization. The first American registered domain was called "america.com" which was given away to private individuals on May 14, 1963. The current most popular term is "usacom" which means "United States of America". There is not standard acronym for "USA" as the U.S.
User:	What is AAPL?
Qwen Assistant:	Apple Inc. is a multinational technology company headquartered in Cupertino, California. It is one of the world's largest tech companies with a market capitalization of over \$2 trillion. Apple is known for producing iPhones, iPads, Mac computers, and other products that include software applications and accessories. It is also known for developing and manufacturing some of the latest and most innovative technologies in the field of consumer electronics. Apple operates through a variety of businesses including the AppStore for iPhone and iPad, the Mac App.

Table 1. Example of a conversation with fine-tuned Qwen.  
Autor: Andrei Ghiurțu

## Concluzii

În concluzie, modelul pe care l-am antrenat promite a fi o soluție viabilă pentru mai multe aspecte cheie ale aplicațiilor viitoare. Natura compactă îl face scalabil, facilitând integrarea pe diverse platforme și dispozitive. În plus, implementarea de asistenți personalizați cu capacități de antrenament pe dispozitiv oferă experiențe îmbunătățite utilizatorilor și oferă avantaje aplicațiilor ce restricții dure privind confidențialitatea datelor utilizatorilor, aliniindu-se cu regulamentele GDPR. Mai mult, rentabilitatea antrenării acestui model îl face mai accesibil și fezabil pentru întreprinderi de diferite dimensiuni. În plus, specializarea modelului în sarcini financiare subliniază potențialul său de rafinare și optimizare. Prin aplicarea acestei tehnici, am reușit să creștem capacitatea modelului de bază cu 13% în evaluarea pe contabilitate, aspect deosebit de important într-un caz de utilizare pentru analiza financiară și înțelegerea obiceiurilor de cheltuieli în istoricul de achiziții al utilizatorilor. Alți indicatori importanți pentru un asistent financiar au fost de asemenea îmbunătățiți cu 4%-8%, indicând că modelul și-a crescut capacitatea de înțelegere și asistență a utilizatorilor [ Table 2. Table 2. . Creșterea în aceste metrici poate fi considerată substanțială, având în vedere constrângerile de resurse și obiectivul de performanță pe dispozitive cu limite hardware. Într-adevăr, există oportunitate pentru progrese și îmbunătățiri ulterioare, poziționând acest model ca o perspectivă captivantă pentru eforturile viitoare în procesarea limbajului natural și inteligența conversațională.

Model	Business Ethics	Econo-metrics	Macro-economics	Micro-economics	Juri-prudence	Accounting	Professional Law
Qwen 1.5 0.5B Chat	<b>39.00</b>	21.92	25.12	31.93	<b>41.66</b>	26.59	29.33
Qwen 1.5 0.5B Chat	38.00	<b>23.68</b> ↑8%	<b>26.15</b> ↑4%	<b>33.19</b> ↑4%	38.89	<b>30.14</b> ↑13%	29.34

Financial QA							
Our solution + DARE	37.00	<b>23.68</b> ↑8%	25.90	32.77	39.81	<b>30.14</b> ↑13%	<b>29.66</b> ↑1%
Our solution + PT	32.00	<b>23.68</b> ↑8%	21.03	24.79	36.11	25.18	28.36

Table 2. OpenLLM Leaderboard  
Autor: Karina Olaru

--

|

## Bibliografie

- [1] P. P. Ray, „ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope,” *Internet of Things and Cyber-Physical Systems*, vol. 3, pp. 121-154, 2023.
- [2] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, Soricut, Radu, J. Schalkwyk, A. Dai, A. Hauth și e. al., „Gemini: a family of highly capable multimodal models,” *arXiv preprint arXiv:2312.11805*, 2023.
- [3] D. Araci, „Finbert: Financial sentiment analysis with pre-trained language models,” *arXiv preprint arXiv:1908.10063*, 2019.
- [4] H. Yang, X.-Y. Liu și C. D. Wang, „Fingpt: Open-source financial large language models,” *arXiv preprint arXiv:2306.06031*, 2023.
- [5] S. Wu, O. Irsoy, S. Lu, V. Dabrovolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg și G. Mann, „Bloomberggpt: A large language model for finance,” *arXiv preprint arXiv:2303.17564*, 2023.
- [6] Y. Zhao, C.-Y. Lin, K. Zhu, Z. Ye, L. Chen, S. Zheng, L. Ceze, A. Krishnamurthy, T. Chen și B. Kasikci, „Atom: Low-bit quantization for efficient and accurate llm serving,” *arXiv preprint arXiv:2310.19102*, 2023.
- [7] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer și V. Stoyanov, „Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.
- [8] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang și e. al., „Qwen technical report,” *arXiv preprint arXiv:2309.16609*, 2023.
- [9] T. W. Jer, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. v. Platen, C. Ma, Y. Jernite, J. Plu, C. Xu și e. al., „Transformers: State-of-the-Art Natural Language Processing,” în *Association for Computational Linguistics*, 2020.
- [10] M. Labonne, *Merge Large Language Models with mergekit*.
- [11] Z. Jing, Y. Su, Y. Han, B. Yuan, C. Liu, H. Xu și K. Chen, „When Large Language Models Meet Vector Databases: A Survey,” *arXiv preprint arXiv:2402.01763*, 2024.
- [12] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang și W. Chen, „Lora: Low-rank adaptation of large language models,” *arXiv preprint arXiv:2106.09685*, 2021.
- [13] T. D. Zettlemoyer, A. Pagnoni, A. Holtzman și Luke, *QLoRA: Efficient Finetuning of Quantized LLMs*, 2023.
- [14] „Github LLM Course,” [Interactiv]. Available: <https://github.com/mlabonne/llm-course>. [Accesat 13 May 2024].