Small Language Models for the Democratization of Financial Literacy: A Case Study in Fine-Tuning on Open Financial QA Data

Tagore Rao Kosireddy and **Jeffrey Wall** and **Evan Lucas**
Michigan Technological University / 1400 Townsend Drive
Houghton, Michigan, United States of America
trkosire, jdwall ,eglucas, @mtu.edu

Abstract

This study seeks to test whether low-cost inference and efficient *Small Language Models* (SLMs) fine-tuned on existing open-source question answering datasets are capable of creating financial literacy chat bots that can answer financial questions for those with limited financial knowledge. The use of SLMs is growing in popularity across many domains, but SLMs are not thoroughly explored in the finance sector. This study offers an exploration of challenges and opportunities that exist in the finance sector to utilize SLMs for open-source financial question answering applications. In particular, this study examines the outputs of several open-source SLMs fine-tuned on the open-source FinGPT FiQA_QA financial question answering dataset. We fine-tuned two versions of each model, one with an instruction prompt and one without an instruction prompt and compared the model outputs with ground truth human responses from the dataset. Further qualitative rating and analysis are provided for model outputs and the dataset. The exploration highlighted challenges with available open data and the fine-tuned SLMs. Existing open data sets in the financial AI research community are not sufficient to produce high-quality outputs with SLMs. Successful fine-tuning of SLMs has occurred in other domains with high quality data sets. We thus issue a call for new and better open financial question answering datasets that could result in higher-quality small language models.

1 Introduction

Recent advances in *natural language processing* (NLP) show that *language models* (LMs) produce impressive results on various general-domain NLP tasks (Li et al., 2023a). A language model is a statistical model that predicts the likelihood of missing words given the context of sequences of words, based on the conditional probabilities observed in some large corpora of training text (Bengio et al., 2000).

Specialized LMs have been tuned for a number of targeted domains, such as in health and finance (Ranjit et al., 2024; Araci, 2019; Fu et al., 2024). This study explores the application of LMs to the finance domain. Individuals must learn to successfully navigate financial systems to improve their socio-economic status. However, many in society lack access to quality financial tools and knowledge to improve their economic situation. Advancements in *financial technology* (FinTech) have facilitated further democratization of financial markets, such as providing free trades to *retail investors* (i.e., individual investors) through online and mobile brokerage applications. FinTech and social trends have prompted a surge in retail investor engagement in financial markets (Fisch, 2022).

Although access to markets is further democratized with FinTech, access to financial knowledge has not kept pace. Low levels of financial literacy are well documented across the general population, and financial illiteracy is associated with poorer financial decision making (Lusardi and Mitchell, 2014). This gap matters in the FinTech era: the recent surge in retail investor activity, combined with limited financial knowledge among some investors, has led to risky investing behavior. Some retail investors treat investing as gambling (Gao and Lin, 2015). Others rely on social influencers to make financial decisions or simply follow social trends in fear of missing out (Dalimunthe et al., 2023; Pedersen, 2022). Retail investors, particularly those who have limited access to financial knowledge or means, deserve better open, low-cost, and privacy preserving resources to develop financial literacy through FinTech applications. Open-source and low-cost LMs trained to answer financial questions could support financial literacy in a privacy preserving manner. We use *privacy preserving* in the practical sense of local, on-device inference: because a sufficiently small model can run on a user's own consumer-grade hardware, financial questions need not be sent to a third-party API or provider. We do not claim a formal guarantee such as differential privacy.

In this study we ask: are existing open-source financial question answering datasets and low-cost, efficient, and privacy preserving LMs capable of generating financial tools to answer retail investors' financial questions? To answer this question, we identify open-source financial question answering datasets and state-of-the-art LMs that can run at low-cost on consumer-grade technology to preserve privacy. Open-source datasets and smaller LMs could be the key to democratizing access to advanced financial tools for those in society who most lack financial knowledge and means.

We define *small language models* (SLMs) as LMs with three billion parameters or fewer, since such models are the most likely to run on consumer-grade hardware.

2 Related Work

This section highlights important existing work related to the use of LMs for financial applications, financial question answering datasets, and SLMs.

2.1 Financial Language Models

Many proprietary and open-source financial LMs exist. For example, Bloomberg developed an LM trained on their vast library of financial data, BloombergGPT (Wu et al., 2023). In response, researchers have developed open-source LMs with similar capabilities to democratize access to financial LMs.

The FinGPT project consists of several open-source financial datasets and fine-tuned LMs, such as a Llama-7B and Falcon-7B (ai4finance.org). The models provide classification outputs, such as sentiment analysis and financial named entity recognition (Yang et al., 2023a). Currently, FinGPT does not support question answering capability that could facilitate the development of financial literacy. Similarly, FinBERT is a fine-tuned LM that supports several classification tasks, such as sentiment analysis. FinBERT was tuned on a variety of open data (Araci, 2019). Another model, FinMA was tuned for similar classification tasks. Like FinGPT, FinMA is a Llama-based LM with 7B and 30B versions (Xie et al., 2023). Despite the open nature of these models that could support democratized access to financial LMs, their question answering capabilities are limited to highly specific questions with often single value answers (e.g., yes/no). These models are not ideal for supporting the development of financial literacy, as they assume users already possess a certain level of financial literacy.

Although low-cost and privacy-preserving LMs, namely SLMs, have demonstrated state-of-the-art results on general tasks (Mehta et al., 2024; Abdin et al., 2024; Team et al., 2024) and several domain-specific tasks (Ranjit et al., 2024; Fu et al., 2024; Sen et al., 2024) when fine-tuned on quality data sets, our work shows that SLM quality is not enough to achieve quality outputs with available financial question answering datasets. We show that domains like finance need better open-source datasets to democratize access to financial LMs. The financial open-source community does not yet possess the datasets necessary to train high-quality LMs that can democratize access to financial literacy tools. High-quality, open datasets are crucial to create efficient and quality large and small LMs (Gunasekar et al., 2023; Li et al., 2023b).

2.2 Financial Question Answering Datasets

Several open datasets support finance-related tasks such as named entity recognition, sentiment analysis, and headline classification, but few target long-form financial question answering. Table 1 summarizes the question answering datasets most relevant to this study.

The FiQA datasets, scraped from social media sites including StackExchange, Reddit, and StockTwits, are among the most widely used and were published for an open research challenge in 2018 (Maia et al.); the original challenge link now returns a 404 error. Of the two FiQA subsets, the sentiment dataset (FiQA_SA) is a classification task and is not relevant here, while the textual opinion dataset (FiQA_QA) is. A version of FiQA_QA is available through the FinGPT HuggingFace page and consists of more than 17,000 rows, each with a question prompt, a human response, and an instruction prompt.

Other long-form datasets exist but are not openly usable. The FinMA team's dataset (Huang et al.) provides primarily short-form answers (e.g., yes/no or numeric values) that are unsuitable for supporting financial literacy. FinTextQA (Chen et al., 2024) and the InvestLM dataset (Yang et al., 2023b) are both drawn from high-quality sources such as textbooks and government or academic material and would suit our objectives, but neither is linked in its respective paper or repository (AbciNLP), plausibly because textbook-derived data raises copyright concerns. High-quality financial corpora therefore exist, including proprietary resources such as Bloomberg's, but are not shared openly.

Given the limited number of openly available long-form financial question answering datasets, we selected the FinGPT FiQA_QA dataset for fine-tuning. Truly democratizing access to financial knowledge will require open LMs together with better open financial question answering datasets.

Table 1: Financial question answering datasets most relevant to this study.

Dataset	Primary source	Long-form	Open
FiQA_QA	Social media	Yes	Yes
FinMA QA	Social media / mixed	No	Yes
FinTextQA	Textbooks, gov.	Yes	No
InvestLM	Textbooks, academic	Yes	No

2.3 Small Language Models

SLMs offer a meaningful opportunity to democratize access to low-cost, efficient LMs that can run on consumer-grade hardware. As defined above, we treat SLMs as transformer-based LMs (Vaswani et al., 2017) with three billion parameters or fewer. Several such models are openly available and have shown strong results despite their size, including Google's Gemma-2B (Team et al., 2024), Microsoft's Phi series (Abdin et al., 2024), Apple's OpenELM family (270M to 3B) (Mehta et al., 2024), and TinyLlama (Zhang et al., 2024). We fine-tune and evaluate models from these families in this study.

Fine-tuning adapts these models to a domain at far lower cost than training from scratch (Wei et al., 2021), and fine-tuned SLMs have performed well across domains such as meeting summarization (Fu et al., 2024) and hate speech detection (Sen et al., 2024). For example, Ranjit et al. (2024) fine-tuned an SLM on radiology-specific data and obtained quality outputs with fewer hallucinations than the base model, and Li et al. (2023b) showed that the Phi-1.5 SLM, when trained on high-quality data, matched or exceeded much larger models on reasoning tasks, underscoring that data quality matters as much as scale. To our knowledge, no prior study has explored fine-tuning SLMs for financial question answering.

3 Method

This research study consisted of selecting a dataset and SLM for fine-tuning, and assessing the outputs of the model.

3.1 Data Selection and Preparation

As outlined in Section 2.2, the FinGPT FiQA_QA dataset on HuggingFace was the only open financial question answering dataset we identified that met the criteria for this study, and we therefore selected it for fine-tuning. We randomly partitioned the 17,110 data points into a fine-tuning set (13,688 samples) and a test set (3,422 samples). The human responses in the dataset have a mean length of 258.2 tokens and a median of 195.0 tokens, indicating a positively skewed length distribution. This distribution informed our generation settings: we capped generation at 250 new tokens (Section 3.4), a value close to the mean human response length, so that model outputs were comparable in length to the ground-truth answers while remaining bounded for the majority of examples.

3.2 Small Language Model Selection

Prior work has begun to examine whether SLMs are suitable for financial question answering. Kosireddy et al. (2024) studied several base SLMs without fine-tuning to assess whether they possessed qualities appropriate for the task. They examined Gemma-2B, Phi-2, OpenELM-270M, OpenELM-540M, OpenELM-1B, OpenELM-3B, and TinyLlama for characteristics that would make the models suitable for democratizing LMs, and found that several were promising and warranted further fine-tuning. We build on their findings by fine-tuning these models to determine whether fine-tuning further improves their ability to answer financial questions. We also include ChatGPT-4o as a benchmark, since it is a widely used LLM that some individuals have turned to for financial guidance.

3.3 Model Fine-tuning

As outlined in Section 2.2, each FinGPT FiQA_QA example contains an input question, an output containing the human response, and an instruction prompt. The instruction prompt is not authored by us: the dataset supplies one of five fixed instruction paraphrases per example (for example, *"Utilize your financial knowledge, give your answer or opinion to the input question or subject. Answer format is not limited."*). All five instructions are listed in Appendix A. To determine the benefit of including an instruction prompt when fine-tuning SLMs, we fine-tuned each model twice, once with the dataset-provided instruction prompt and once without it. Instruction fine-tuning has been used in multiple studies to improve model outputs (Mishra et al., 2022; Sanh et al., 2022; Ouyang et al., 2022; Marvin et al., 2023; Wei et al., 2022).

The fine-tuning of each model was completed with the SFTTrainer class from the HuggingFace transformers library on a NVIDIA RTX 3090 or A6000 GPU for 2 epochs. The AdamW optimizer (Loshchilov and Hutter, 2017) was utilized with bf16 mixed precision (Burgess et al., 2019). The models were evaluated every 250 steps during fine-tuning and the best model was saved based on evaluation loss.

The initial learning rate was set to 5e-5 with a weight decay of 0.01 to 0 using a linear learning rate schedule with warmup steps set to 500, train batch and eval batch sizes set to 2, and gradient accumulation steps set to 1 with gradient clipping at 1.0. These hyperparameters, including the train and evaluation batch sizes and the gradient accumulation steps, were held identical across both the RTX 3090 and A6000 GPUs, so the effective batch size was the same for every run regardless of which GPU was used.

3.4 Model Output Evaluation

After fine-tuning the models, we generated outputs from the various fine-tuned and their base models to evaluate the responses. Hereafter, we refer to the fine-tuned models with FT and the base zero-shot models as ZS. The models with an instruction prompt are identified with an I, such as FT+I or ZS+I. Outputs for each model were generated from the test dataset using the same generation parameters for all models (max_new_tokens = 250, top_k = 30, top_p = 0.8, no_repeat_ngram_size = 5, num_return_sequences = 1).

We evaluated the model outputs in several ways. First, we compared each model's outputs to the human responses from the dataset using quantitative metrics. We treated the human answers as ground truth for similarity comparison metrics. We calculated ROUGE metrics, which are commonly used for summarizing and evaluating the degree of overlap between words, bi-grams, or common substrings between a candidate and reference sentence (or sentences) (Lin, 2004). The ROUGE metrics range between 0 and 1, with higher scores indicating higher similarity between the produced and the reference text. ROUGE scores do not take semantic meaning into account, and have been criticized for this shortcoming (Akter et al., 2022; Schluter, 2017). We also calculated BERTScore, which measures the similarity between a candidate and a reference using contextual embeddings (Zhang et al., 2019), with scores that in principle range from -1 to 1. We used the default English configuration of the `bert_score` library, which uses `roberta-large` with the library's default layer selection and without baseline rescaling. Because raw, non-rescaled BERTScore for fluent English text occupies a narrow band near the top of its range, the absolute scores cluster between roughly 0.80 and 0.84 across all models; the small variation we observe is therefore expected behavior of the metric under this configuration rather than evidence that the outputs are near-identical. BERTScore is also less sensitive to small errors when the candidate text is lexically or stylistically similar to the reference text (Hanna and Bojar, 2021). Further, we calculated the Semantic Textual Similarity (STS), which produces similarity scores from 0 to 1 between a candidate and reference sentence (or sentences) by using a Cross-Encoder which achieves better performance than a Bi-Encoder (Reimers and Gurevych, 2019; Risch et al., 2021).

We also calculated a Flesch readability score (Flesch, 2007) for the outputs of each model. The Flesch readability score assesses the complexity of text on a scale from 0 to 100, with higher values representing text that is more complex and difficult to read. Further, we calculated a diversity score (four distinct n-grams) (Li et al., 2015), which assess the lexical diversity in model outputs on a scale from 0 to 100 with higher values representing greater lexical diversity in model responses.

Second, three domain experts, consisting of the business researcher co-authors, rated the outputs of each model. These individuals include an Accounting researcher, a Finance researcher, and an Information Systems researcher. Since the FiQA_QA dataset is founded on social media data that contains personal opinion, the outputs were not rated for factual accuracy. Instead, the raters evaluated the outputs on whether the output answered the question or not without regard for accuracy.

Following best practice in multiple-rater situations, the three business researchers met to identify and reach consensus on a set of rating behaviors to enhance rating reliability (Sattler et al., 2015). The evaluation proceeded in three stages: a small calibration round to establish shared rating behaviors, a larger reliability round on the outputs of a single model to test whether those behaviors produced consistent ratings at scale, and a final cross-model round on a carefully screened set of questions for the substantive comparison across models. In the calibration round, the authors randomly selected 10 questions, which the three business researchers rated individually. Then, in a video conference, the three researchers shared their ratings and explained why they gave each rating. Where rating discrepancies existed, the researchers discussed the differences and arrived at a set of common rating behaviors for assessing the relevance of the model outputs.

The business researchers then conducted a larger reliability round of 100 outputs from a single model, OpenELM-270M, rated on a 5-point Likert scale (1 = irrelevant to 5 = relevant). Holding the model fixed for this round was deliberate, so that any disagreement reflected the rating protocol rather than differences in model quality. We report inter-rater reliability as pairwise scores, using the Pearson Correlation Coefficient and Cohen's Kappa, rather than a single omnibus statistic such as Fleiss' or Krippendorff's alpha. With only three raters, the pairwise breakdown is diagnostic: it shows that raters 1 and 2 agreed reasonably well while agreement involving rater 3 was substantially lower, a pattern that a single aggregate coefficient would obscure. While one correlation was above 0.8 for raters 1 and 2, the other correlation scores were less than 0.55. Similarly, the highest Kappa was close to 0.65 for raters 1 and 2, but lower than 0.32 for the other rater comparisons. The authors met and discussed the results. The raters noted that many of the responses had expansive hallucinated narratives around the core answer or lack thereof, which made it difficult to rate the relevance. Rater 1 had a mean rating of 3.22 ± 1.0971. Rater 2 had a mean rating of 3.22 ± 1.2519. Rater 3 had a mean rating of 3.89 ± 1.4695. Based on the concerns over hallucination and the means of two of the raters scoring close to the center of the 5-point scale, the authors determined to simplify the scale to a 3-point Likert scale from irrelevant to relevant for the final round. This was done to simplify the issue of interpreting the relevance of hallucinated narratives and to push ratings away from the center of the scale.

For the final domain expert evaluation, the raters reviewed a sample of the same 10 questions across all of the different models. The questions were more carefully screened for the final round than for the random selection used in the 100-output reliability round. This was done because many of the entries in the dataset were not actually questions. In Appendix B.0.1 examples of this issue are provided with a calculation of the number of potentially problematic questions. The 10 questions for the final evaluation were hand selected from a random sample of 100 questions by one of the domain experts, who selected the first 10 valid questions from that sample. Here, a *valid* question is an entry that is an actual, well-formed question rather than a statement or malformed prompt; as quantified in Appendix B.0.1, a substantial fraction of dataset entries do not meet this bar.

Third, after reviewing the quantitative and qualitative results, one of the business researchers performed a post-hoc qualitative review of the question prompts, the human outputs, and the model outputs on a small random subset of the data to identify patterns that could explain the seemingly odd quantitative results.

4 Results

The quantitative and qualitative results are provided in the following sub-sections.

4.1 Quantitative Output Comparisons

When comparing the model outputs to the human responses designated as the ground truth for models with an instruction prompt, OpenELM-3B FT performed the best on the ROUGE-1 (0.2458), ROUGE-2 (0.0373) and ROUGE-L (0.1340). In comparison, ChatGPT-4o scored slightly higher for ROUGE-1 and ROUGE-2, but lower for ROUGE-L and BERTScore. The TinyLlama-1.1B ZS model performed the best on the BERTScore (0.8334), which for comparison was higher than the ChatGPT-4o score. Phi-2 ZS performed the best on the STS metric (0.4454). For comparison, ChatGPT-4o scored well above all models on STS. Table 2 shows further comparison details. In most cases, the metrics are not substantially different across the models. As depicted in the table, many of the base models (i.e., ZS models) performed better than the fine-tuned models. The implications of these findings are discussed later.

To understand the impact of the fine tuning and instruction prompt, we perform a short ablation study and present the results in Table 3. We compare the automatic metrics for Phi 1.5 with (notated by +) and without (notated by -) fine-tuning and the instruction prompt. We find that the instruction prompt has only a small impact and that the fine-tuning has a larger positive impact.

In addition to similarity comparisons, we also compared the readability of each model using the Flesch readability measure (Flesch, 2007) and the lexical diversity of the text with a diversity score (Li et al., 2015) and report the results in Table 5 and 4. The results suggest that the OpenELM-450M FT with instruction prompt (80.98), OpenELM-270M FT without instruction prompt (79.10), and OpenELM-270M FT with instruction prompt (78.50) performed the best for readability, even better than the readability of the human output (63.68). The top three models with the greatest lexical diverse were Phi-2B FT with instruction prompt (72.58), Phi-2B FT without instruction prompt (72.02) and Gemma-2B FT with instruction prompt (71.79), which are slightly lower than the human output (78.38). Many models showed an increase in diversity after fine-tuning. Fine-tuning also seemed to increase readability for models, likely because the social media fine-tuning data was simpler than the original training data for models, which included scholarly articles in many cases.

4.2 Qualitative Ratings

Despite following best practices to establish consistency in rating (as outlined in Section 3.4), inter-rater agreement was low. Averaged across models, the pairwise Cohen's Kappa values were 0.1213, 0.2236, and 0.1506 (Table 6), which corresponds to only slight-to-fair agreement on the Landis and Koch scale. Per-model pairwise Kappa ranged widely, from -0.273 (below chance) to 0.667 (substantial), and the Pearson correlation scores ranged from -0.333 to 0.843. In other words, the raters agreed well on a minority of models but barely above chance on most, and the aggregate signal is weak agreement rather than moderate agreement.

We do not treat this low agreement as solely a shortcoming of the rating protocol. Three factors specific to the data make consistent relevance judgments difficult, even for trained domain experts. First, many question prompts are truncated, missing context, grammatically incorrect, or otherwise ill-formed (Sections 4.3.1 and 4.3.2), so there is often no single defensible notion of what a relevant answer would be. Second, many outputs embed an answer or non-answer inside an expansive hallucinated narrative that mimics a social media responder, which leaves genuine room for disagreement over whether the question was actually answered. Third, because the outputs were rated for whether they answered the question rather than for factual accuracy, the judgment is inherently subjective when the output rambles. The implication is that when domain experts cannot reliably agree on the relevance of outputs derived from this data, the difficulty is located largely in the question and answer pairs themselves, not only in the raters. This observation is what motivated the post-hoc qualitative analysis in Section 4.3.

Although lower than desired, the scores are not surprising to the raters. Several of the question prompts in the FinGPT FiQA_QA dataset seem truncated, are missing context, are grammatically incorrect, and/or make little sense. Further, many of the fine-tuned models suffered from extreme hallucinations in the form of narratives that mimicked a social media responder. These narratives, with answers or non-answers scattered throughout the text, made it difficult to assess whether the output answered the question or not. The model with the highest rating was the Phi-1.5 ZS model with an average rating of 2.60/3.00 as depicted in Table 2. Table 3 shows the quantitative assessments for all models run using Phi-1.5. Although far from ideal, the results show promise for future work and the development of higher quality datasets.

4.3 Post-hoc Qualitative Analysis

This study sought to answer the question: are existing open-source financial question answering datasets and low-cost, efficient, and privacy preserving LMs capable of generating reasonable responses to answer individuals' financial questions? Based on the results of this study, the answer appears to be, not yet with existing datasets. In some

Table 2: Evaluation Metrics for all models, including an instruction prompt. ZS = Zero Shot, FT = Fine Tuned ChatGPT-4o is included for comparison purposes.

Model Name	Hum. Eval	ChatGPT-4o	R-1	R-2	R-L	BertScore	STS
TinyLlama-1.1B FT	2.37	1.20	0.2229	0.0330	0.1246	0.8331	0.4267
TinyLlama-1.1B ZS	2.07	1.20	0.2243	0.0334	0.1255	**0.8334**	0.4280
Phi-1 FT	1.40	1.00	0.2356	0.0348	0.1316	0.8262	0.3710
Phi-1 ZS	1.90	1.00	0.2384	0.0358	0.1320	0.8265	0.3752
Phi-1.5 FT	2.27	**1.60**	0.2395	0.0368	0.1316	0.8324	0.4384
Phi-1.5 ZS	**2.60**	**1.60**	0.2209	0.0209	0.1132	0.8057	0.2460
Phi-2 FT	2.13	1.40	0.2236	0.0339	0.1249	0.8317	0.4307
Phi-2 ZS	2.23	1.40	0.2366	0.0371	0.1296	0.8324	**0.4454**
Gemma-2B FT	2.43	**1.60**	0.1941	0.0286	0.1147	0.8303	0.4073
Gemma-2B ZS	-	-	0.1977	0.0294	0.1157	0.8304	0.4132
OpenELM-1.1B FT	2.27	1.20	0.2348	0.0351	0.1281	0.8319	0.4285
OpenELM-1.1B ZS	2.03	1.00	0.2345	0.0356	0.1283	0.8323	0.4298
OpenELM-3B FT	2.30	1.10	**0.2458**	**0.0373**	**0.1340**	0.8285	0.4061
OpenELM-3B ZS	1.67	1.10	0.2433	0.0351	0.1327	0.8227	0.3472
ChatGPT-4o	-	-	***0.2723***	***0.0420***	0.1310	0.8205	***0.5591***

Table 3: Ablation study for Best Performing Model (Phi-1.5) ZS = Zero Shot, FT = Fine Tuned, I = Instruction Prompt Included

	R-1	R-2	R-L	BS	STS
FT+I	0.240	**0.037**	**0.132**	**0.832**	0.438
ZS+I	0.221	0.021	0.113	0.806	0.246
FT-I	**0.242**	**0.037**	**0.132**	**0.832**	**0.440**
ZS-I	0.239	0.028	0.120	0.811	0.342

Table 4: Readability scores for the top three models and human written ground truth answers

Model	**Readability**
OpenELM-450M FT+I	80.98
OpenELM-270M FT	79.10
OpenELM-270M FT+I	78.50
Ground Truth Answers	63.68

Table 5: Diversity scores for the top three models and human written ground truth answers

Model	**Diversity**
Phi-2B FT+I	72.58
Phi-2B FT	72.02
Gemma-2B FT+I	71.79
Ground Truth Answers	78.38

cases, fine-tuning harmed model performance. In other cases it helped performance. The poor quality of the only openly available dataset made both the qualitative assessment of the results difficult and the quantitative assessment inconclusive. We treat this difficulty as an informative result in its own right: the inability to draw firm conclusions from this dataset localizes the current bottleneck for SLM-based financial question answering in the available open data rather than in model capacity, and it directly motivates the call for better datasets that follows.

Table 6: Average inter-rater reliability metrics between three expert human raters - Rater 1, Rater 2, Rater 3 across all models

Metric	**1 vs 2**	**1 vs 3**	**2 vs 3**
Avg. Pearson Corr.	0.3175	0.3927	0.4321
Avg. Cohen's Kappa	0.1213	0.2236	0.1506

After reviewing the quantitative scores and the qualitative rating results, the authors wanted to better understand why the fine-tuned models didn't perform. One of the business researchers on the team reviewed a small, random subset of the training data and the model results to identify reoccurring patterns. The following are key themes that emerged after analyzing the data further.

4.3.1 Unclear and Cryptic "Questions" in FiQA_QA Dataset

Because the data was scraped from social media sites, it contains a variety of odd, unclear, and grammatically incorrect questions. This created difficulties for the models and human raters. Appendix B.0.1 provides some examples from the dataset.

These types of question prompts provided room for too many possible answers, or the inability to meaningfully provide an answer. The human responses in the data set to question prompts with these issues represent either one of many possible interpretations of the question or unrelated answers to an uninterpretable question. Training data points with this issue provide too much ambiguity to produce a quality model.

4.3.2 Missing Context in FiQA_QA Questions

Second, the FiQA_QA dataset has many questions that seem to be truncated or missing important context. This became evident after reviewing several question/human answer pairs. The human answers seemed to contain important context related to the original question that was missing in the question prompt. This could have occurred if the actual question was truncated during scraping or data cleaning, or if the question was clarified in the discussion thread and therefore not scraped fully. Either way, this issue raises a data quality concern. Appendix B.0.2 provides examples of this data problem.

The missing context in some questions poses a problem when pairing the vague, de-contextualized question prompts to the contextually rich answers. Models, particularly SLMs, trained on too much of this low quality data cannot be expected to perform well.

4.3.3 Unanswered Questions with Hallucinated Context

In some cases, a model did not answer a question, but mimicked the narrative nature of the social media fine-tuning data. It is possible that the hallucinated context in these outputs caused overestimated similarity scores by mimicking the context, but not the answer. These narrative hallucinations also caused issues with qualitative ratings. Appendix B.0.3 provides examples of results with these types of hallucinations.

4.3.4 Hallucinated Context in Relevant Answers

Some models produced relevant answers to questions, but did so through narratives with hallucinated contextual details. Appendix B.0.4 provides examples of some of these hallucinations.

5 Conclusions

Based on the qualitative findings of this study, quality financial question answering datasets will need: 1) clearer questions with less ambiguity that contain the full context of the question, 2) more professional, yet readable answers that remove unnecessary personal information to avoid contextual hallucinations, 3) good instruction prompts to guide the behavior of the model toward answering in a professional manner and to further avoid hallucinations, and 4) models and dataset responses that re-direct uniformed questions toward better questions, as many of the questions asked in the social media dataset exhibited the difficulty many retail investors have in forming appropriate questions.

These findings demonstrate the need for better, open financial question answering datasets. We encourage more faculty, students, and researchers in financial and computing disciplines to engage in a collaborative effort to create these high quality datasets. The democratic benefits of better open datasets could help LMs support retail investors as they develop financial literacy. Other domains are having success with SLMs. It is time to make that a possibility for the finance domain as well. These findings also provide room for other opportunities to improve the quality of SLM output in financial literacy applications, such as through retrieval augmented generation. However, methods like these will also require high-quality open-source data, not opinion-based social media data.

Limitations

This study examined a number of SLMs. Due to the rapid pace at which new models emerge, we did not include more recent SLMs, such as those from Meta (though we did include Tiny-Llama from the Llama community), and Deepseek. We also did not consider other models with more than three billion parameters, such as the common 7B models. However, these models require more computing power. Future studies could compare LORA and QLORA 7B models compared to FP32 or BF16 models with three billion parameters or less.

Although the inter-rater reliability was lower than we expected, we believe it to be less of a limitation of the study and more of a limitation of the existing data. The existing data has many issues that make consistent subjective rating, even among experts, difficult. The difficulty in achieving higher inter-rater reliability helped point the research team toward the deeper qualitative analysis that ultimately answered the research question and identified the larger issue, namely the lack of quality datasets.

6 Ethics Statement

To avoid copyright violations, we chose to fine-tune the model on a widely used open-source dataset. Our focus on democratizing LMs and LM training data also helped to keep our focus grounded in ethical pursuits. Our call for more accurate open-source financial question answering datasets is also in line with our ethical pursuit to produce low-cost, privacy preserving models for financial literacy applications for those who lack an appropriate financial education to benefit from societal financial structures.

Acknowledgments

This study was made possible with funding from Michigan Technological University's College of Computing and College of Business.

References

AbciNLP. Investmentlm on github (abacinlp/investlm). https://github.com/AbaciNLP/InvestLM. [Accessed 7-11-2024].

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

ai4finance.org. GitHub - AI4Finance-Foundation/FinGPT: FinGPT: Open-Source Financial Large Language Models! Revolutionize We release the trained model on HuggingFace. — github.com. https://github.com/AI4Finance-Foundation/FinGPT. [Accessed 11-06-2024].

Mousumi Akter, Naman Bansal, and Shubhra Kanti Karmaker. 2022. Revisiting automatic evaluation of extractive summarization task: Can we do better than rouge? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1547–1560.

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

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. *Advances in neural information processing systems*, 13.

Neil Burgess, Jelena Milanovic, Nigel Stephens, Konstantinos Monachopoulos, and David Mansell. 2019. Bfloat16 processing for neural networks. In *2019 IEEE 26th Symposium on Computer Arithmetic (ARITH)*, pages 88–91. IEEE.

Jian Chen, Peilin Zhou, Yining Hua, Yingxin Loh, Kehui Chen, Ziyuan Li, Bing Zhu, and Junwei Liang. 2024. Fintextqa: A dataset for long-form financial question answering. *arXiv preprint arXiv:2405.09980*.

Zuliani Dalimunthe, Ananda Chairunnisa, and Rachmadi Agus Triono. 2023. Are social media users blindly following influencers' recommendations on investing? *The Indonesian Capital Market Review*, 15(1):1.

Jill E Fisch. 2022. Gamestop and the reemergence of the retail investor. *BUL Rev.*, 102:1799.

Rudolf Flesch. 2007. Flesch-kincaid readability test. *Retrieved October*, 26(3):2007.

Xue-Yong Fu, Md Tahmid Rahman Laskar, Elena Khasanova, Cheng Chen, and Shashi Bhushan TN. 2024. Tiny titans: Can smaller large language models punch above their weight in the real world for meeting summarization? *arXiv preprint arXiv:2402.00841*.

Xiaohui Gao and Tse-Chun Lin. 2015. Do individual investors treat trading as a fun and exciting gambling activity? evidence from repeated natural experiments. *The Review of Financial Studies*, 28(7):2128–2166.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Michael Hanna and Ondřej Bojar. 2021. A fine-grained analysis of bertscore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517.

Jimin Huang, Ruodyu Xiang, Chen Zhengyu, and Clémentine Fourrier. Thefinai/flare-finqa. https://huggingface.co/datasets/TheFinAI/flare-finqa. [Accessed 7-11-2024].

Tagore Rao Kosireddy, Jeffrey D. Wall, and Evan Lucas. 2024. Exploring the readiness of prominent small language models for the democratization of financial literacy. *Preprint*, arXiv:2410.07118.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023a. Large language models in finance: A survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 374–382.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023b. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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

Annamaria Lusardi and Olivia S. Mitchell. 2014. The economic importance of financial literacy: Theory and evidence. *Journal of Economic Literature*, 52(1):5–44.

Macedo Maia, André Freitas, Alexandra Balahur, Siegfried Handschuh, Manel Zarrouk, Ross McDermott, and Brian Davis. Financial opinion mining and question answering. https://sites.google.com/view/fiqa/home. [Accessed 7-11-2024].

Ggaliwango Marvin, Nakayiza Hellen, Daudi Jjingo, and Joyce Nakatumba-Nabende. 2023. Prompt engineering in large language models. In *International conference on data intelligence and cognitive informatics*, pages 387–402. Springer.

Sachin Mehta, Mohammad Hossein Sekhavat, Qingqing Cao, Maxwell Horton, Yanzi Jin, Chenfan Sun, Iman Mirzadeh, Mahyar Najibi, Dmitry Belenko, Peter Zatloukal, et al. 2024. Openelm: An efficient language model family with open-source training and inference framework. *arXiv preprint arXiv:2404.14619*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. *Preprint*, arXiv:2104.08773.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Lasse Heje Pedersen. 2022. Game on: Social networks and markets. *Journal of Financial Economics*, 146(3):1097–1119.

Mercy Ranjit, Gopinath Ganapathy, Shaury Srivastav, Tanuja Ganu, and Srujana Oruganti. 2024. Rad-phi2: Instruction tuning phi-2 for radiology. *arXiv preprint arXiv:2403.09725*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Julian Risch, Timo Möller, Julian Gutsch, and Malte Pietsch. 2021. Semantic answer similarity for evaluating question answering models. *arXiv preprint arXiv:2108.06130*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. *Preprint*, arXiv:2110.08207.

David N Sattler, Patrick E McKnight, Linda Naney, and Randy Mathis. 2015. Grant peer review: improving inter-rater reliability with training. *PloS one*, 10(6):e0130450.

Natalie Schluter. 2017. The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 41–45, Valencia, Spain. Association for Computational Linguistics.

Tanmay Sen, Ansuman Das, and Mrinmay Sen. 2024. Hatetinyllm: Hate speech detection using tiny large language models. *arXiv preprint arXiv:2405.01577*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

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

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *Preprint*, arXiv:2109.01652.

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 preprint arXiv:2303.17564*.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *Preprint*, arXiv:2306.05443.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023a. Fingpt: Open-source financial large language models. *FinLLM Symposium at IJCAI 2023*.

Yi Yang, Yixuan Tang, and Kar Yan Tam. 2023b. Investlm: A large language model for investment using financial domain instruction tuning. *arXiv preprint arXiv:2309.13064*.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.

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

A Appendix: Dataset Instruction Prompts

The FinGPT FiQA_QA dataset assigns one of five fixed instruction prompts to each example. We did not author or modify these instructions; they are used verbatim as provided by the dataset. The five instructions are:

1. *Utilize your financial knowledge, give your answer or opinion to the input question or subject. Answer format is not limited.*
2. *Offer your insights or judgment on the input financial query or topic using your financial expertise. Reply as normal question answering.*
3. *Based on your financial expertise, provide your response or viewpoint on the given financial question or topic. The response format is open.*
4. *Share your insights or perspective on the financial matter presented in the input.*
5. *Offer your thoughts or opinion on the input financial query or topic using your financial background.*

These five instructions are near-synonymous paraphrases of the same request. Because they were applied per example rather than chosen or tuned by us, the small effect of the instruction prompt observed in our ablation (Section 4.1) most plausibly reflects that generic, semantically overlapping instructions add little signal for these models on this data, rather than indicating that a poorly chosen prompt or rigid instruction inherently harms performance.

B Appendix: Examples of Poor Quality Data

The following sections provide examples of issues that exist within the dataset. Where quantitative analysis was feasible, quantitative data is also provided.

B.0.1 Examples of Unclear and Cryptic "Questions" in FiQA_QA Dataset

The following question is a statement and not a well formed question:

> *Question: Living in my own rental property*

The next question prompt is also a statement, but there is at least an implicit question:

> *Question: Short Term Capital Gains tax vs. IRA Withdrawal Tax w/o Quarterly Est. Taxes*

Further, cryptic question prompts exist in the dataset. The intention of the questioner is open to vast interpretation in the following question prompt:

> *Question: Buying an option in the money, at the money, or out of the money*

As a simple quantitative test of the number of potential questions that were unclear and poorly formed, we examined whether all of the questions ended with a question mark. Of the 17,110 questions in the dataset, 4,614 questions (26.967%) lacked an ending question mark.

B.0.2 Examples of Questions with Missing Context in FiQA_QA Dataset

The following question was missing context brought up in the human response. Even the question itself makes reference to "advice" that is not mentioned in the question prompt. Clearly, something is missing from the question prompt.

> *Question: Is this mortgage advice good, or is it hooey?*
> *Answer: That makes no sense at all. They try to compare and that's exactly the same as comparing apples versus oranges. Mortgage is long-term loan, so for the first many years the huge part of the payment will go to repaying interest, so that ratio 1 will indeed be something like 20% or more despite the fact that the interest rate on the mortgage is much lower - something around 6%. HELOC will have the interest rate of 6%, but it will have the same structure so that you have equal payments, so if you compute that ratio 1 it will be very close to that of the mortgage. The bottom line is - if HELOCs were that great noone would apply for mortgages. You should stick to making*

extra payments towards the principal on the mortgage.

Similarly, the following question seems to be missing important context. The human response makes mention of conversation provided by the questioner that doesn't appear in the question itself.

Question: Is Cost of Living overstated?
Answer: You mention: High rent places are usually also high property value places. Given the tax incentives, it seems like a good long-term idea to grab a house, so if we assume you have the option of working and buying a house in a high CoL or a low CoL city, I think you'd prefer the high cost. Because essentially, after 30 years, you'd have a million dollar house vs a quarter million dollar house. You've captured three quarters of a million dollars in rent, given my napkin math hypothetical. I think you're forgetting about some of the associated costs with "owning" a home, including:

B.0.3 Examples of Unanswered Questions with Hallucinated Context

The following response from the OpenELM-270M model does not directly answer the question, but simply tells a story about starting a business in the UK while asking questions related to the original prompt.

Question: How to safely earn interest on business profits (UK)
Answer: I recently bought my first business. I am a bit nervous as I do not know much about the UK market, so I have some questions. The business is a food shop with a very small store. I'm currently working part time and earning £15/hour. I would like to be able to earn £20/hour. The business operates Monday to Friday. I have to work on my lunch break, and I do not have any time to work after work. My only plan is to buy a business loan. My lender says that I could be eligible to get up to £4000 if I buy the business outright. Is this correct? I have no idea what a business loan is and I am worried about the interest rates. I know that it's not ideal to buy a business if I do not have the cash to pay it off in full. Is there any way I could avoid this? Thank you so much in advance for any advice you could give me! I would be really happy if you could tell me how I can achieve the £2

Even the 3B parameter version of OpenELM showed similar issues. For example, the following response shows the model responding as though it had an investment account with Schwab. Further, it hallucinated multiple index funds and their prices. The answer only tangentially answered the question.

Question: Should I continue to invest in an S&P 500 index fund?
Answer: You're right. I've been using Schwab for a while and haven't noticed the 500-dollar error in a 5000 index fund. But let's say you don't want to use Schwab. You can use any "Index funds" from Vanguard or TI (Trudeau). 10000 Vanguard's "Tax-Free" funds will cost you $10,000, and $5000 will go to buy the same number of shares of 500 Indexes. (If you're using Schwab's website, $10,500 will go for the same thing). 3000 Indexes (500-funds) will cost you $3000 (you can use Schwab, or a bank or a mutual fund company like Fidelity). 500 (5000-fund) will cost you about $500. The Vanguard and TI are all mutual fund companies, so you pay a fund company to manage your money (in a bank, you would go

The OpenELM models were not the only models to suffer from unanswered questions with hallucinations. Examples of this issue can be found across models. For example, the Gemma-2B model hallucinated that is was an insurance agent with a client whose house burned down. Although the response was tangentially related to the question, it made no mention of high-net worth individuals and the necessity of insurance for individuals with substantial wealth.

Question: For very high-net worth individuals, does it make sense to not have insurance?
Answer: You should have insurance to cover your potential liability. For example, I have

a client with a house that burned down. His insurance policy covered the cost to replace the house. If he had not had that coverage, he would have been forced to come up with a lot of cash to pay for a new home.

B.0.4 Examples of Hallucinated Context Surrounding Relevant Answers

The following response from OpenELM-270M hallucinates a wife and child while answering the question (though with some contradictions in the answer).

Question: Is this mortgage advice good, or is it hooey?
Answer: I think the advice is a bit overstated. I am not a realtor, but my gut says that this advice is overstated. This is the advice that I would give my wife and daughter if they are considering buying a house. If I had to buy a house and it was my wife and daughter, I would buy it for a number of reasons. For example: I know that my wife and daughter are young and don't want to spend a lot of money on a house. They are not ready for a house yet. This advice would help them realize that they are ready to buy a house. It could save them from a lot of regret later on.

Although the response answers the question to some extent, it does so through a made up narrative.

Similarly, in the following response, the same model invented a life in Toronto while at least partially answering the question.

Question: Is Cost of Living overstated?
Answer: It's not overstated. I live in Toronto, and we do pay about $2500/yr for utilities. If you were living in a place with a $1500/year utilities bill, I'd still suggest you move. The $300 you mentioned as your rent for the unit is actually a discount on your monthly cost. The cost to you is $4800/yr. If you were $150k/yr out of work, I'd suggest you move. You can probably save money in your own area and you'll be a better tenant. Edit - The $30k is actually the rent. I've had tenants pay $30k for their unit. They have had no problems with it, and it is just an extra $3000/yr to have a place that is only $250/yr.