LLMs Cannot (Yet) Match the Specificity and Simplicity of Online Finance Communities in Long Form Question Answering

Anonymous ACL submission

Abstract

Retail investing is on the rise, and a growing number of users is relying on online finance communities to educate themselves. However, recent years have positioned Large Language Models (LLMs) as powerful question answering (QA) tools, shifting users away from interacting in communities towards discourse with AI-driven conversational interfaces. These AI tools are currently limited by the availability of labelled data containing domain-specific financial knowledge. Therefore, in this work, we curate a QA preference dataset SOCIALFINANCEQA for fine-tuning and aligning LLMs, extracted from more than 7.4 million submissions and 82 million comments from 2008 to 2022 in Reddit's 15 largest finance communities. Additionally, we propose a novel framework called SOCIALQA-EVAL as a generally-applicable method to evaluate generated QA responses. We evaluate various LLMs fine-tuned on this dataset, using traditional metrics, LLM-based evaluation, and human annotation. Our results demonstrate the value of high-quality Reddit data, with even state-of-the-art LLMs improving on producing simpler and more specific responses.

1 Introduction

Recent years have brought remarkable growth of retail investment activity (Gurrola-Perez et al., 2022), which has been accompanied and potentially fuelled by the rise of trading apps such as Robinhood (Curry, 2024). Retail investors, i.e., non-professional traders, often lack formal financial training, and tend to educate themselves via the internet (Hsieh et al., 2020). The Reddit platform contains the largest domain-specific communities on the internet, has over 73 million daily active unique users[1] in its subreddits (i.e., topic-specific communities), and almost 74 million users in total subscribed to subreddits focused on financial topics. These communities showcase the combined financial knowledge of the (Western) internet through millions of posts and comments, with responses rated by the respective community using Reddit's upvote score mechanism.

However, the information acquisition process has started shifting away from online sources due to the emergence of an accessible and powerful alternative with LLMs such as OpenAI's ChatGPT in 2022 (Wu et al., 2023b) – e.g., studies show that web traffic on sites like Stack Overflow is declining, while usage of AI tools increases, which may be caused by the migration of users (Carr, 2023). Furthermore, early studies suggest that AI tools are already having an impact on academic learning (Xing, 2024) as well as school education (Sidoti and Gottfried, 2023). Smaller, instruction-tuned LLMs like Llama-2-7B-Chat (Touvron et al., 2023) and Zephyr (Tunstall et al., 2023) already provide a remarkable quality of texts while being small enough to run on consumer-grade hardware. Despite their capabilities, LLMs face challenges in finance due to domain-specific jargon, acronyms, and context-dependent nuances. In addition, the dynamic nature of financial markets requires adaptability and domain expertise for accurate analysis.

Our goal is to bridge the gap between the financial knowledge available on Reddit and capabilities of LLMs to provide accurate, personalized insights into financial markets. This raises the following research questions: *How can the domain knowledge of Reddit users be used to improve the performance of the LLMs in this field? And which degree of improvement can we measure?* To this end, we introduce a high-quality question answering (QA) dataset extracted from 7.4 million posts and 82 million comments from the 15 largest finance-related subreddits. We use this dataset to fine-tune various LLMs for financial QA tasks using Supervised Fine-Tuning (SFT; Ouyang et al., 2022) and Direct Preference Optimization (DPO; Rafailov

[1] Current value available via www.redditinc.com

et al., 2024). The evaluation process involves a comprehensive evaluation framework that assesses the quality aspects of question answering, utilizing lexical metrics, human annotation, and LLM-based evaluation.[2] The main contributions of this paper are:

1. We curate a high-quality Reddit Finance dataset called SOCIALFINANCEQA, the first social-media preference dataset for fine-tuning and alignment of LLMs for the finance domain. Additionally we provide insights into the nuanced differences between the source subreddits in our analysis.
2. We demonstrate that fine-tuning LLMs with SOCIALFINANCEQA has positive impacts on question answering capabilities and can help to modify even state-of-the-art instruction-tuned LLMs to generate more specific and simpler responses.
3. We propose SOCIALQA-EVAL, which is inspired by G-Eval (Liu et al., 2023) but re-defined for the domain-independent evaluation of texts in the context of question answering and social media.

2 Background

2.1 Reddit and Internet Datasets

The Reddit platform hosts numerous "subreddits", i.e., topic-specific online forums, where users engage and vote on each other's contributions (Anderson, 2015). These communities serve as valuable research subjects, e.g., with studies investigating user expertise (Strukova et al., 2024; Lim et al., 2017), social roles (Buntain and Golbeck, 2014), and social support mechanisms (De Santo et al., 2023). Reddit's financial subreddits, particularly *wallstreetbets*, have seen rapid growth and widespread attention since the GameStop hype in 2021 and related "meme stock" rallies (Chacon et al., 2023; Agrawal et al., 2022).

Research on these subreddits has analyzed topics and sentiment (Zhu, 2022; Karpenko et al., 2021), predicted stock price movements (Deng et al., 2023; Karnik, 2022; Wang and Luo, 2021; Reichenbach and Walther, 2023) based on post sentiment, and investigated their value as sources for investment advice (Buz and de Melo, 2023, 2024). A Reddit dataset for long-form question answering was proposed (Fan et al., 2019), but is no longer accessible since Reddit's API changes in 2023. Large corpora such as *C4* (Raffel et al., 2019) and *The Pile* (Gao et al., 2020) include Reddit data and have been created to provide suitable datasets for training LLMs, but there is only limited research of the effects of financial datasets on LLMs: FinBERT (Yang et al., 2020) was an early investigation but is already outdated, BloombergGPT (Wu et al., 2023a) is an LLM built with proprietary financial data for financial downstream tasks but inaccessible, and FinGPT (Yang et al., 2023) is a valuable open-source project to help researchers and practitioners in the field, but only includes limited Reddit data. Li et al. (2023) list a handful of additional examples in their survey.

2.2 LLMs, Fine-Tuning, and Benchmarking

Today's state-of-the-art LLMs are designed to generate long texts and are fine-tuned to solve a wide range of tasks and problems. Various LLMs follow the three-step approach of InstructGPT (Ouyang et al., 2022), which involves supervised fine-tuning (SFT) and further alignment with reinforcement learning techniques such as PPO (Schulman et al., 2017) or DPO (Rafailov et al., 2024). Examples include OpenAI's closed-source ChatGPT and smaller models such as Llama-2-7B-Chat (Touvron et al., 2023) and Zephyr (Tunstall et al., 2023), which are instruction-tuned versions of Llama-2-7B and Mistral-7B, respectively.

Benchmarks such as MTBench (Zheng et al., 2024) and G-Eval (Liu et al., 2023) have been created for the challenging task of comparing and evaluating LLM performance, promoting the usage of powerful LLMs to scale and accelerate the evaluation process. However, LLM-based evaluation techniques are still an open research problem (Zheng et al., 2024; Liu et al., 2023) and may not use optimal criteria. This paper proposes an evaluation framework that combines simpler lexical metrics, human annotation, and a novel LLM-based evaluation method for QA data.

3 Dataset

The raw dataset has been collected from the Pushshift API[3] and is a collection of submissions (i.e., posts) and comments made to the top 15 largest finance-related subreddits ranging back from June 2005 until the end of 2022. Unfortunately, we cannot work with more recent data, as Reddit has changed its terms mid-2023 to restrict

[2] Code and data have been provided in the supplementary material. We will release source code via www.github.com/<anonymized> and the dataset via www.hugginface.com/datasets/<anonymized>.

[3] https://api.pushshift.io

access. Table 1 provides an overview of the selected subreddits together with the counts of submissions, comments, and subscribers per subreddit.

Subreddit	Submission Count	Comment Count	Subscribers[4] (Dec 31, 2022)
wallstreetbets	2,218,243	26,012,515	13,368,770
explainlikeimfive	1,803,202	13,194,897	21,809,070
personalfinance	1,334,756	13,614,959	17,113,530
investing	324,784	5,225,983	2,120,613
Economics	314,981	4,716,282	3,049,238
stocks	311,625	5,254,498	5,019,633
RealEstate	262,635	2,743,137	401,797
pennystocks	193,995	2,154,021	1,893,934
StockMarket	140,243	1,322,163	2,553,498
options	105,195	1,297,655	970,365
Wallstreetbetsnew	103,856	1,601,308	819,585
financialindependence	83,885	3,934,876	1,725,319
FinancialPlanning	79,863	408,015	498,798
realestateinvesting	71,025	813,211	1,620,509
AskEconomics	52,285	304,840	982,409
Combined	7,400,573	82,598,360	73,947,068

Table 1: Submission count, comment count, and number of subscribers per subreddit in our dataset

All 15 subreddits combined make up over seven million submissions and more than 82 million comments ranging from January 2008 until December 2022. The size of the considered subreddits varies drastically, with *wallstreetbets*, *explainlikeimfive*, and *personalfinance* contributing a considerable portion of the overall dataset.

Characteristic Words & Topic Modeling: Understanding the topics discussed in financial subreddits is crucial to identify the interests and focus of each community and to provide an overview of the contents of our dataset. For a nuanced analysis, we extract important keywords as well as topics. TF-IDF scores are calculated for all submission titles, treating each subreddit as a separate "document" (Ramos et al., 2003). Subsequently, BERTopic is leveraged for topic modelling on a sample of 50,000 submission titles from each subreddit separately. We identify the following three major areas of knowledge.

Personal finances, budgeting, and real estate: Includes topics around credit cards, debt, retirement, real estate investments. Key subreddits are *personalfinance*, *financialindependence*, *FinancialPlanning*, *RealEstate*, and *realestateinvesting*.

[4]Sourced from https://subredditstats.com

Stocks, trading, and market movements: Consists of topics on the analysis and trading of specific stocks and investment funds, options, share prices, often on a daily or weekly basis. The main subreddits are *investing*, *stocks*, *StockMarket*, *options*, *pennystocks*, *wallstreetbets*, *wallstreetbetsnew*.

Theoretical economic concepts and trends: Includes topics on the economy, inflation, prices, and their effects on people. Salient subreddits include *Economics*, *AskEconomics*, *explainlikeimfive*.

3.1 Data Preprocessing

Our goal is to curate a valuable dataset for question answering from the Reddit data, which requires filtering to enhance data quality by focusing on high-quality, relevant content, and reducing noise. For this purpose, we combine qualitative and quantitative filtering criteria, which leverage Reddit metadata and toxicity metrics, and involve a three-step process: filtering data based on different selection criteria (as shown in Table 2), matching comments to submissions (i.e., posts) into a single dataset, and applying heuristics to curate a preference dataset for improving the performance of LLMs. Our filtering heuristics cover three main aspects: (A) engagement, (B) community approval, and (C) relevance.

Engagement is measured by the number of comments, and community approval is indicated by the score and upvote ratio. The relevant submissions are selected by identifying posts that (1) contain a question (either as indicated by a community-enforced post "flair" or our question detection heuristic), (2) were not removed or deleted, (3) are an actual post on Reddit (as opposed to a link to a different website), (4) were not written by an administrator or moderator, (5) were not a message made "sticky" by the moderators, (6) are not written by one of the community's bots.

Our study defines thresholds for score and upvote ratio in subreddits using the upper 20^{th} percentile of submissions for each community, with a minimum score of 3 and a minimum upvote ratio of 0.75 (i.e., 75% of users voted positively for the post). This helps to filter for quality while avoiding selection biases due to different community sizes and voting behaviours. While a few subreddits use "flairs" (i.e., category tags) to indicate whether a submission is a question, we use a question detection heuristic to cover all subreddits. The heuristic matches phrases indicating a question, such as "please help" or "should I", and checks whether the last two sentences of a post end with a question

Attribute	Selection Criteria
Score	>= min(3, percentile[80])
Upvote Ratio	>= min(0.75, percentile[80])
# Comments	>= 3
Question	Flair contains "Question" or satisfies our heuristic
Content	selftext and title not empty, removed or deleted
Domain	Reddit (contain .self)
Author Flair	not ["Admin", "Moderator"]
Stickied	False
Author	not ["IndexBot", "AutoModerator", "Moderation Bot"]
Link Flair	**AskEconomics:** ["Approved Answers", "Good Question", "Simple Questions/Career"]
	financialindependence: not ["Mod Post", "Case Study", "Moderator Meta", "Personal Journey"]
	explainlikeimfive: ["Economics"]
Distinguished	not ["moderator", "admin"]

Table 2: Filtering criteria for submissions using Reddit metadata

mark. This method is robust and well-aligned with human judgment, reaching an accuracy of over 90% on a random test sample after human evaluation.

There are also question-thread posts specifically for answering community questions – we treat their first-level comments as submissions. These comments are filtered using the same mechanisms as submissions, using filters that are applicable to comment attributes. Comments that are empty, removed, deleted, short, collapsed, or created by moderators or bots are filtered out.

3.2 SocialFinanceQA Preference Dataset

In the next step, we aim to create a preference dataset SocialFinanceQA that can be used for LLM alignment, based on the curated Reddit data we filtered as described above. For this purpose, we use question–answer pairs corresponding to each eligible submission with its top-level comments, identifying "better" and "worse" answers by leveraging Reddit's upvote scores. A score difference rule is introduced to ensure better comments have at least a 10-point higher score, while only comments with a maximum score of 3 qualify as the "worse" option. This results in a preference dataset with 61,610 question–answer–answer tuples.

The study aims to improve the quality of social media data by using the HateBERT OffenseEval (Caselli et al., 2020) model to detect and remove "better" comments with toxicity in terms of offensive, abusive, or hate speech. The model removes 5.4% of entries, which are unequally distributed across subreddits with *wallstreetbets* having the highest ratio of offensive texts (see Table 3). To further reduce toxicity, we apply a fine-tuned RoBERTa model[5] to remove further remaining offensive texts (0.9% of the dataset).

Subreddits	QA Pairs	# off.	% off.
wallstreetbets	3,155	1,141	36%
Wallstreetbetsnew	6	1	17%
Economics	85	9	11%
pennystocks	156	16	10%
StockMarket	528	37	7%
RealEstate	5,800	317	6%
stocks	2,466	150	6%
options	767	35	5%
investing	6,082	306	5%
financialindependence	2,323	98	4%
realestateinvesting	1,167	41	4%
personalfinance	37,818	1,146	3%
explainlikeimfive	508	10	2%
FinancialPlanning	732	10	1%
AskEconomics	17	0	0%
Total	61,610	3317	5%

Table 3: Number of comments (before filtering) and absolute amount (#) and percentage (%) of comments identified as offensive (off.) with HateBERT OffenseEval toxicity detection per subreddit

When creating the dataset for fine-tuning the LLMs, it's crucial to consider the maximum length of question–answer pairs to prevent truncation and model degradation, as training with cut-off mid-sentence inputs can be detrimental. Due to limitations in computational resources, we set a maximum length of 1,024 tokens and removed all longer entries (reducing the dataset by 7.3%).

The final preference dataset, after filtering, contains 53,561 entries. The subreddit *personalfinance* dominates, accounting for over 62% of the dataset, as it is one of the largest subreddits in our dataset. Despite their large subscriber counts, *explainlikeimfive* and *wallstreetbets* have fewer entries: *explainlikeimfive* only partially covers Economics, and *wallstreetbets* focuses on discussions and stocks analyses rather than questions, while a large portion of it is excluded due toxicity.

[5] https://huggingface.co/badmatr11x/distilroberta-base-offensive-hateful-speech-text-multiclassification

4 Methodology

4.1 Models

We use the pre-trained LLMs Llama-2-7B (Touvron et al., 2023) and Mistral-7B-v0.1 (Jiang et al., 2023), as well as instruction-tuned models Llama-2-7B-Chat (Touvron et al., 2023) and Zephyr β (Tunstall et al., 2023), to evaluate the effect of fine-tuning our dataset on variants of different LLM families. We leverage different prompt templates based on the prior training of the considered models (details in Section C.1 in the Appendix). Mistral-7B enhances efficiency by introducing changes like Sliding Window Attention, Rolling Buffer Cache as well as Pre-fill and Chunking to their training compared to Llama (Jiang et al., 2023). Zephyr-7B was created from Mistral-7B using SFT with the UltraChat dataset (Ding et al., 2023) and DPO with the UltraFeedback dataset (Cui et al., 2023). We refer to these models as Llama-2, Mistral, Llama-2-Chat, and Zephyr for simplicity.

4.2 Fine-tuning and Alignment

We expect fine-tuning LLMs on finance-related subreddits to significantly enhance the quality in providing insightful responses to finance-related queries, reflecting the models' exposure to diverse perspectives from Reddit's discussions.

Similarly to the training of Zephyr, we employ SFT on the four LLMs followed by DPO. Additionally, we perform DPO directly on the base versions of Llama-2-Chat and Zephyr to measure the impact our dataset can have with DPO only. We apply QLoRa (Dettmers et al., 2023) adapters for training with 4-Bit quantization, which has been shown to only have a marginal impact on the model performance compared to fine-tuning in full-precision. Furthermore, we split our dataset into a test dataset of 500 entries used for evaluation, a validation dataset of 1,000 entries used within the training, and a training dataset used for SFT and DPO with 52,061 entries. We use approximately 200 GPU-hours on an RTX 3090 Ti with 24 GB for training and inference.

Supervised Fine-tuning Configuration We expect SFT on Reddit finance data to enable domain adaptation, improved task performance, and adaptability to user preferences. We leverage the SFTTrainer of huggingface's `trl` library, which enables SFT on custom datasets (von Werra et al., 2020). We align our hyperparameter choices with research best practices (Dettmers et al., 2023), such as the LoRa alpha (32) being twice as large as the rank (16), addressing all linear layers of the model as target modules within the fine-tuning, setting the Dropout rate to 0.05 and bias to "none". Additionally, we use a limited learning rate (1×10^{-4}) with cosine-decay and a warm-up ratio (0.05) throughout the training to support the convergence of the model during the training.[6]

The default setting for SFTTrainer uses the packing algorithm for higher training efficiency. This is useful for pre-training but leads to undesirable splits and truncations during SFT. As roughly 40% of our dataset was affected by truncation, we disabled the packing algorithm for SFT and instead modified SFTTrainer to use a custom padding token.

Direct Preference Optimization Configuration We leverage DPOTrainer from `trl` for DPO training on our dataset. We align our hyperparameter choices with settings that have been found to work in the DPO training of Zephyr-7B, with the difference that we set LoRa rank and alpha to 128 and the learning rate to a lower 5×10^{-6}.[7]

Text Generation Hyperparameters For text generation, we use standard hyperparameters of the huggingface transformers library, with the exception of a token limit of 1,000 (to avoid excessively long texts) and a repetition penalty of 1.2, which obtained the best results in our preliminary testing and is in line with prior research (Keskar et al., 2019).

4.3 Evaluation Methods

As explained above, we use three different approaches for a robust evaluation: traditional metrics, LLM-based evaluation, and human evaluation.

4.3.1 Traditional Metrics

We evaluate the test data on the following set of more traditional metrics:

Readability metrics: We compare the generations of our models on a variety of metrics provided by textstat, specifically text length, reading time, and complexity as measured by the Flesch-Kincaid score, which indicates the required school-grade level to understand a text (Kincaid et al., 1975).

Perplexity: We compare the perplexities of our models given the question–answer pairs of our test dataset to measure how well each model learns the writing style of Reddit. A lower perplexity score

[6] See Table 13 in the appendix for all hyperparameters.

[7] See Table 14 in the appendix for all hyperparameters

reflects a better alignment with Reddit users' word choices, indicating a better understanding of their writing style. **Toxicity:** We compare across models how many generations of the test dataset are categorized as offensive or hateful by the RoBERTa-based model used for generating the preference dataset.

4.3.2 LLM-based & Human Evaluation

This work proposes a novel evaluation framework called SOCIALQA-EVAL, which we use to evaluate all model variants. SOCIALQA-EVAL is inspired by the principles of G-Eval (Liu et al., 2023), but has been modified significantly to (1) apply to the linguistic characteristics that are relevant for the QA context (particularly in social media), and (2) to receive a wider range of evaluation scores from the judge model (our preliminary tests yielded homogeneous scores with the original G-Eval).

SOCIALQA-EVAL aims to provide five evaluation criteria (detailed prompts can be found in the appendix): Relevance, Specificity, Simplicity, Helpfulness, and Objectivity. These criteria reflect the qualities of good responses for online questions: They are expected to be relevant (i.e., provide an answer and not a joke or anecdote), specific to the question (i.e., match the specificity of the question instead of providing a broad response from a textbook), written in simple language (i.e., suitable for the linguistic context of casual subreddits), helpful (i.e., friendly and polite), and objective (i.e., provide an unbiased perspective that is not influenced by the response author's personal views).

We use Gemini Pro (et al., 2023) as the evaluator or "judge LLM" due to its comparable performance to GPT-3.5 and availability through the Vertex AI API of Google Cloud. For our evaluation shown in this paper, we used approximately 10 hours of the Gemini Pro API. To be able to assess how well our LLM-based evaluation works, we also manually annotated 800 question–answer pairs (100 human-written and 700 AI-generated) with all five criteria of the SOCIALQA-EVAL framework, resulting in a total of 4,000 annotated scores, which are provided by four different annotators (the paper's authors). This is conducted via blind evaluation (without knowledge of the source model nor the scores of the LLM-based evaluation) of randomly sampled and randomly ordered responses from the test data. We measure the alignment of the human and LLM-based scores with Cohen's κ as well as the correlation.

5 Results

5.1 Traditional Metrics

5.1.1 Readability Metrics

The `textstat` library measures the length and complexity of Reddit texts. Table 4 reveals that base models (especially instruction-tuned variants) have higher sentence counts than fine-tuned models and our dataset. SFT significantly affects the sentence count, while DPO-only alignment reduces text lengths (especially for Zephyr). DPO applied after SFT has a smaller impact and reverses SFT's effect by increasing the sentence count to some extent. The Flesch-Kincaid score reflects sentence length and word length, and correlates with reading time. Mistral produces the easiest-to-read texts, as reflected in the fine-tuned variants.

Source	Sentence Length	Sentence Count	Reading Time	Flesch-Kincaid score
Better Answer	17.4	6.1	6.6	6.5
Worse Answer	17.0	5.3	5.6	6.1
Llama-2-Chat	20.6	23.7	38.2	11.0
-SFT	20.2	2.8	3.6	7.7
-SFT-DPO	19.8	5.4	7.0	8.2
-DPO	20.6	22.1	36.1	11.2
Zephyr	20.6	14.4	22.4	10.7
-SFT	16.0	4.3	4.3	5.5
-SFT-DPO	13.4	3.4	2.9	4.7
-DPO	18.1	7.9	10.1	8.7
Llama-2	18.4	11.1	18.8	7.9
-SFT	21.8	2.6	3.5	8.2
-SFT-DPO	19.7	4.8	6.5	8.1
Mistral	14.0	9.3	9.1	3.6
-SFT	16.4	2.9	2.9	5.6
-SFT-DPO	14.5	3.6	3.4	5.4

Table 4: Textstat readability-related metrics (mean values) of original and generated answers for the test dataset

5.1.2 Perplexity

Our study measures the perplexity of each model as a metric for their ability to learn to emulate Reddit's writing style. The results reported in Table 5 show that after applying SFT, the perplexity decreases as expected, indicating effective fine-tuning. All model variants have lower perplexity values for the "better" answers compared to the "worse" answers within our preference dataset. Applying DPO after SFT slightly increases perplexity again, but widens the gap between better and worse answers, indicating a stronger alignment towards better answers.

Source	Better A. (mean)	Worse A. (mean)	Better A. (median)	Worse A. (median)
Llama-2-Chat	20.24	22.47	16.60	17.73
-SFT	8.60	9.20	8.11	8.68
-SFT-DPO	10.29	11.77	9.46	10.65
-DPO	24.37	29.72	18.37	20.85
Zephyr	13.16	14.25	11.80	12.37
-SFT	10.57	11.40	9.68	10.34
-SFT-DPO	11.09	12.01	10.16	10.83
-DPO	13.01	14.10	11.74	12.49
Llama-2	10.94	11.74	9.97	10.57
-SFT	8.27	8.84	7.82	8.34
-SFT-DPO	9.52	10.74	8.77	9.77
Mistral	10.81	11.61	9.87	10.40
-SFT	8.35	8.96	7.90	8.47
-SFT-DPO	9.18	9.93	8.59	9.27

Table 5: Mean and median perplexity values for each model's generated answers on test dataset, compared to the baselines of the better and worse answers (A.) from the preference dataset

However, when DPO is used without SFT, the alignment is less effective, with LlamaChat-DPO (tuned with DPO only) having higher perplexity than the baseline and Zephyr-DPO only showing minor changes. We conclude that the model variants learn to reproduce Reddit's writing style differently, depending on the applied fine-tuning process, with instruction tuning showing no improvement in perplexity when only applying DPO.

5.1.3 Toxicity

As described above, we use a pre-trained RoBERTa model to detect hate speech and offensive language in the LLM-generated answers. While the model does not identify any hate speech in our model outputs, there are rare cases of detected offensive language. Out of 500 answers, only one or two answers are labeled offensive per model. SFT with Reddit data appears to introduce a slight amount of toxicity to the models, which is then usually reduced by DPO[8]. It is noteworthy that Mistral's baseline model seems to be more toxic with three texts classified as offensive, so that it improves after SFT.

5.2 LLM-based & Human Evaluation

To enrich the insights gained from the analysis above, we finally assessed the quality of generated answers using the SOCIALQA-EVAL framework. The results summarized in Table 6 show that fine-tuning with the SOCIALFINANCEQA dataset consistently improves the performance of baseline models Llama-2 and Mistral with respect to all evaluation criteria of SOCIALQA-EVAL (except for Mistral's helpfulness). Llama-2's overall score doubles from the baseline to the SFT-DPO variant in the human evaluation. This is observed in both LLM-based and human evaluation. Our results indicate that human evaluation scores show larger differences, with annotators using a wider range of scores, while LLM judges tend to rate more "mildly", providing more homogeneous results.

For instruction-tuned models, the results are more nuanced, with baseline models winning on relevance, helpfulness, and objectivity in both LLM-based and human evaluation. The DPO-only variants achieve similar scores, only slightly lower. SFT has significant impact on these criteria, with lower scores for both Llama-2-Chat and Zephyr. Contrasting these results, the fine-tuning with our preference dataset seems to improve both models' abilities to generate specific and simple answers, as both the LLM-based and human scores show.

Regarding the overall scores, both LLM-based and human evaluations agree that Llama-2's and Mistral's SFT-DPO variants perform best, while Llama-2-Chat and Zephyr have the LLM judge preferring the baseline versions and human evaluators rating Zephyr-DPO as the best model overall. Our results indicate that the LLM judge rates human-written texts worse compared to human evaluators, possibly due to a bias towards LLM-generated texts, as discussed by Zheng et al. (2024).

6 Discussion

Our results demonstrate that fine-tuning LLMs with the SOCIALFINANCEQA dataset leads to responses with improved conciseness and simpler language, with reduced sentence count and length, reduced reading time, and a lower Flesch-Kincaid readability score (ranging between 4.7 – 8.2, i.e., the text is comprehensible for 5^{th} to 8^{th} graders). Our evaluation using SOCIALQA-EVAL shows that the fine-tuned models are perceived as more specific and simple by both LLM and human evaluators, benefitting the Llama-2 and Mistral models significantly. Llama-2-Chat and Zephyr improve on specificity and simplicity, but the scores of the other three evaluation criteria decrease slightly. This is likely due to Reddit responses being more subjective and less polite and helpful compared to

[8]More details can be found in the appendix Section E.3.

Source	Relevance		Specificity		Simplicity		Helpfulness		Objectivity		Overall Score	
	LLM	HU	LLM	HU	LLM	HU	LLM	HU	LLM	HU	LLM	HU
Better Answer	3.35	4.72	3.11	4.10	3.49	4.34	3.34	4.32	3.48	3.72	16.77	21.20
Worse Answer	2.84	4.06	2.70	3.58	3.15	4.28	3.07	3.82	3.17	3.42	14.93	19.16
Llama-2-Chat	**4.76**	**4.56**	3.30	2.98	**3.88**	2.92	**4.83**	**4.52**	**4.65**	4.70	**21.42**	**19.68**
-SFT	2.84	3.74	2.52	3.70	3.26	**4.46**	2.88	3.64	3.27	3.36	14.77	18.9
-SFT-DPO	3.50	3.98	3.26	**3.80**	3.57	3.80	3.22	3.74	3.70	3.06	17.25	18.38
-DPO	4.73	4.46	**3.36**	2.88	**3.88**	2.92	4.78	**4.52**	4.61	**4.74**	21.36	19.52
Zephyr	**4.62**	**4.60**	3.68	3.42	3.98	3.34	**4.63**	**4.66**	**4.51**	**4.72**	**21.42**	20.74
-SFT	3.10	3.88	2.83	3.88	3.53	4.54	3.12	3.74	3.48	3.12	16.06	19.16
-SFT-DPO	3.37	4.22	3.08	4.10	3.61	**4.68**	3.10	3.76	3.58	3.56	16.74	20.32
-DPO	4.40	4.46	**3.82**	**4.16**	**3.99**	4.02	4.21	4.30	4.44	4.18	20.86	**21.12**
Llama-2	2.43	2.08	2.22	1.60	2.83	2.18	2.85	2.12	3.04	1.78	13.37	9.76
-SFT	2.74	3.72	2.49	3.68	3.21	**4.40**	2.82	3.66	3.16	3.26	14.42	18.72
-SFT-DPO	**3.55**	**4.08**	**3.24**	**4.00**	**3.61**	4.16	**3.26**	**3.86**	**3.74**	**3.36**	**17.41**	**19.46**
Mistral	3.23	2.96	2.91	2.68	3.56	2.98	**3.35**	2.98	3.57	2.70	16.62	14.3
-SFT	2.99	3.96	2.71	4.02	3.48	**4.76**	2.93	3.76	3.38	3.38	15.48	19.88
-SFT-DPO	**3.48**	**4.26**	**3.11**	**4.18**	**3.69**	4.62	3.19	**3.94**	**3.65**	**3.52**	**17.12**	**20.52**

Table 6: SOCIALFINANCEQA scores from LLM-based (LLM) and human (HU) evaluation for all text sources (best model variant per column highlighted in bold; overall best score per column underlined)

instruction-tuned LLMs. The judge LLM, Gemini Pro, prefers the baseline Llama-2-Chat and Zephyr models, while our human evaluators prefer Zephyr-DPO and Mistral-SFT-DPO overall.

The quantitative evaluation of long-form responses remains challenging and requires careful development and iteration to minimize randomness and subjectivity. Despite the significant overlap between LLM-based and human evaluation, LLM-based evaluation still has limitations and is affected by the following biases.

Preference of LLM-generated texts: Our LLM-based evaluation rates Llama-2 and Mistral much higher than human evaluators do, despite the texts often containing long, illogical lists of unrelated words. In addition, the LLM judge rates the original Reddit answers consistently lower than the human evaluators. This issue can only be mitigated by continuing to conduct additional human evaluation but may improve with larger, more powerful judge LLMs in the future.

More homogeneous scores: The LLM judge penalizes poor-quality texts less and gives the highest score more rarely. This leads to more homogeneous results that restrict the score range, making it more difficult to identify differences in text quality when comparing scores. Besides conducting human evaluation, a potential mitigation is to normalize or recalibrate the LLM scores across the full score range after the experiments.

Preference of longer answers: Our analysis confirms previous observations that the LLM judge favours longer, verbose responses (Zheng et al., 2024). We mitigated this by introducing evaluation criteria that address specificity (coverage of only those topics that are relevant to answering the question) and simplicity (usage of short and easy-to-understand explanations).

7 Conclusion

This paper proposes a novel QA preference dataset, SOCIALFINANCEQA, based on finance discussions on social media, and investigates the impact of using it to fine-tune and align LLMs. We conduct a multifaceted evaluation including our novel evaluation framework SOCIALQA-EVAL and reveal Reddit's strengths in specificity and simplicity of responses, benefitting both baseline and advanced, instruction-tuned LLMs.

Our findings can be beneficial for researchers and practitioners who (1) aim to create domain-specific datasets of high quality, (2) modify existing state-of-the-art LLMs for their specific context, or (3) evaluate the QA performance of their LLMs for social media text or similar. The SOCIALFINANCEQA preference dataset offers benefits for QA in social media contexts, but the evaluation criteria of SOCIALQA-EVAL can be generalized to text generation capabilities of LLMs in other domains, providing a better approximation of human author responses to questions.

Ethical Considerations and Risks

Social media should never be the only source for investors planning to make financial decisions. To provide a broad foundation, we have chosen a large number of 15 different communities that cover various topics, as described in Section 3. Nonetheless, we recommend retail investors to additionally consult other sources such as relevant books, scientific literature, and trained professionals before making significant investment decisions. Investing very often bears the risk of losing money, which is why investors should be careful with money that they cannot afford to lose.

Inappropriate, toxic, or offensive language is always a risk when working with social media data. We mitigate this by filtering our dataset carefully leveraging Reddit's internal mechanisms as well as additional techniques such as toxicity detection, as explained in Section 3. Nonetheless, there is a possibility that our dataset still harbors remaining instances of text that may be considered offensive or may cause fine-tuned LLMs to generate such.

The dataset contains the user names of the texts' authors. These user names are pseudonyms that do not contain information about the identity of the author, unless the author actively decides to share their personal information in their Reddit profile. We have not obtained a dedicated permission from the authors of the texts to include them in our dataset. However, this is mitigated by the fact that the users have agreed to Reddit's terms that allow data extraction services such as Pushshift, which is the source for our dataset and was accessed legitimately. Additionally, the texts may contain the names of individuals, usually public figures relevant for the stock market, which the authors included as part of their questions or responses. We have decided to leave these names in the dataset, as they are usually critical for understanding the text (e.g., when a company's CEO is mentioned in a text about to how the same company performs).

Current LLMs are able to generate high-quality texts that are near-impossible to distinguish from human-written ones. This brings various beneficial opportunities but also paves the path for bad actors and unethical behavior. While we do not endorse unethical practices that exploit these technologies, we acknowledge that the techniques presented in our work may assist malicious actors in harmful endeavors. For instance, LLMs could be trained to deliberately provide harmful advice. We believe that the best way to counter such practices is to pursue open science and share datasets and techniques for others to reproduce and validate – this creates accessible repositories of verifiable resources and builds trust.

Limitations

With the vast landscape of available LLMs and fine-tuning techniques, it is impossible to cover all possible options and techniques in a study and we expect that there will always be additional topics left for future work. We have chosen highly popular LLMs and training techniques, and conducted a large number of experiments in the context of our restricted resources with the aim of providing an empirically sound study. We have perceived DPO training to be much more costly and slower than initially anticipated. With budget and time restrictions, we stopped DPO training after what we deemed a sufficient training duration. However, there is the possibility that continuing DPO even further could yield better results (in terms of more significant differences) for those models.

It should be noted that our SOCIALFINANCEQA dataset may not provide as significant overall LLM quality improvements as high-quality fine-tuning datasets such as UltraChat and UltraFeedback. However, we argue that achieving overall highest ratings may not always be a researcher's or practitioner's goal – they may want to focus on specific linguistic characteristics as we have done in our study. For this purpose, our dataset is valuable.

Our SOCIALQA-EVAL framework has been developed to measure the qualities of question answering datasets we have identified as most critical. We argue that the application of this framework is not limited to the financial domain, which we have tested it on. Still, there may be additional linguistic characteristics relevant for specific domains, which our framework does not include. We look forward to seeing future research advance our work.

References

Pratik Agrawal, Tolga Buz, and Gerard de Melo. 2022. WallStreetBets beyond GameStop, YOLOs, and the moon: The unique traits of Reddit's finance communities. In *AMCIS 2022 Proceedings*, volume 8.

Katie Elson Anderson. 2015. Ask me anything: What is Reddit? *Library Hi Tech News*, 32(5):8–11.

Cody Buntain and Jennifer Golbeck. 2014. Identifying social roles in Reddit using network structure. In

Proceedings of the 23rd International Conference on World Wide Web, WWW '14 Companion, page 615–620, New York, NY, USA. Association for Computing Machinery.

Tolga Buz and Gerard de Melo. 2023. WallStreetBets: An analysis of investment advice democratization. In *Proceedings of the 56th Hawaii International Conference on System Sciences*, volume 56, page 2150.

Tolga Buz and Gerard de Melo. 2024. Democratisation of retail trading: a data-driven comparison of Reddit's wallstreetbets to investment bank analysts. *Journal of Business Analytics*, 0(0):1–17.

David F. Carr. 2023. Stack Overflow is ChatGPT Casualty: Traffic Down 14% in March.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. HateBERT: Retraining BERT for abusive language detection in English. *arXiv preprint arXiv:2010.12472*, abs/2010.12472.

Ryan G Chacon, Thibaut G Morillon, and Ruixiang Wang. 2023. Will the Reddit rebellion take you to the moon? Evidence from WallStreetBets. *Financial Markets and Portfolio Management*, 37(1):1–25.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *ArXiv*, abs/2310.01377.

David Curry. 2024. Stock trading & investing app revenue and usage statistics (2024). *Business of Apps*.

Alessio De Santo, Arielle Moro, Bruno Kocher, and Adrian Holzer. 2023. Helping each other quit online: Understanding user engagement and real-life outcomes of the r/stopsmoking digital smoking cessation community. *Trans. Soc. Comput.*, 6(1–2).

Xiang Deng, Vasilisa Bashlovkina, Feng Han, Simon Baumgartner, and Michael Bendersky. 2023. What do LLMs know about financial markets? a case study on Reddit market sentiment analysis. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 107–110, New York, NY, USA. Association for Computing Machinery.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, 36.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*, abs/2305.14233.

Gemini Team Google et al. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, abs/2312.11805.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, volume abs/1907.09190, pages 3558–3567.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling. *ArXiv*, abs/2101.00027.

Pedro Gurrola-Perez, Kaitao Lin, and Bill Speth. 2022. Retail trading: An analysis of current trends and drivers. *Available at SSRN 4562259*.

Shu-Fan Hsieh, Chia-Ying Chan, and Ming-Chun Wang. 2020. Retail investor attention and herding behavior. *Journal of Empirical Finance*, 59:109–132.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*, abs/2310.06825.

Gauri Karnik. 2022. BERT vs. VADER: Stock price prediction with analysis of financial sentiment from Reddit.

Valeriya Karpenko, Kirill Mukhina, Daria Rybakova, Irina Busurkina, and Denis Bulygin. 2021. A study of personal finance practices. the case of online discussions on Reddit. In *Proceedings of the International Conference "Internet and Modern Society" (IMS-2021), St. Petersburg, Russia 23-26 June 2021*, volume 3090 of *CEUR Workshop Proceedings*, pages 206–211. CEUR-WS.org.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*, abs/1909.05858.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023. Large language models in finance: A survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, ICAIF '23, page 374–382, New York, NY, USA. Association for Computing Machinery.

Wern Han Lim, Mark James Carman, and Sze-Meng Jojo Wong. 2017. Estimating relative user

expertise for content quality prediction on Reddit. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, HT '17, page 55–64, New York, NY, USA. Association for Computing Machinery.

Yang Liu, Dan Iter, Yichong Xu, Shuo Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Conference on Empirical Methods in Natural Language Processing*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer.

Felix Reichenbach and Martin Walther. 2023. Financial recommendations on Reddit, stock returns and cumulative prospect theory. *Digital Finance*, 5(2):421–448.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, abs/1707.06347.

Olivia Sidoti and Jeffrey Gottfried. 2023. About 1 in 5 us teens who've heard of ChatGPT have used it for schoolwork.

Sofia Strukova, José A Ruipérez-Valiente, and Félix Gómez Mármol. 2024. Computational approaches to detect experts in distributed online communities: a case study on Reddit. *Cluster Computing*, 27(2):2181–2201.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, abs/2307.09288.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of LM alignment. *arXiv preprint arXiv:2310.16944*, abs/2310.16944.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Charlie Wang and Ben Luo. 2021. Predicting $gme stock price movement using sentiment from Reddit r/wallstreetbets. In *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing*, pages 22–30.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023a. Bloomberggpt: A large language model for finance. *ArXiv*, abs/2303.17564.

Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023b. A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136.

Yixun Xing. 2024. Exploring the use of ChatGPT in learning and instructing statistics and data analytics. *Teaching Statistics*, 46:95–104.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. FinGPT: Open-Source Financial Large Language Models. *Preprint*, arXiv:2306.06031.

Yi Yang, Mark Christopher Siy UY, and Allen Huang. 2020. FinBERT: A pretrained language model for financial communications. *Preprint*, arXiv:2006.08097.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging LLM-as-a-judge with MT-Bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Jinfei Zhu. 2022. What people worry about: Top personal finance concerns with Reddit data.

A Dataset Details

A.1 Growth of Finance-Related Subreddits

Figure 1 provides an overview of the growth of the reviewed subreddits for the last 10 years. Due to the strong exponential growth of the communities, we have chosen a logarithmic scale for representation.

A.2 Description of Subreddits

In Table 7, we provide short descriptions of the reviewed subreddits. Each subreddit is accessible via www.reddit.com/r/[subreddit name].

A.3 Additional Details on Subreddit Characteristics

Table 8 provides an overview of the top five words identified per subreddit via TF-IDF as well as the main topics identified using BERTopic.

A.4 Submission Data Format

The raw dataset is split across individual subreddit files, which have a uniform format for submissions. Within the submission data format, we find 124 different attributes with varying levels of meaningfulness. We only report the attributes we find to be relevant for the creation of our datasets. In order to match it with parent_id's or link_id's of the respective comments one has to prepend "t3_" with the ID, which works as an identifier for submissions. The score denotes the difference between upvotes and downvotes of a submission. It cannot be negative, as submissions with a majority of downvotes are masked with a score of zero. The upvote_ratio is a ratio of upvotes over total votes (e.g., a submission with three upvotes and one downvote has an upvote_ratio of 0.75). The attribute num_comments denotes the amount of comments a submission has received. The title of a submission is a descriptive headline for the submission. The selftext of a submission contains all the content of a submission. The domain attribute denotes whether a submission is contained within the Reddit domain or rather a link to external resources. The author_flair_text attribute denotes the role of an author, which can be a moderated role like "admin" or a custom role chosen by the author depending on the subreddit. The stickied attribute denotes whether a submission is showcased on the subreddit's front page, which usually tends to denote administrative messages. The author attribute is simply the user name of the author of the submission. The link_flair attribute classifies a submission with a topic, which depending on the subreddit can either be a moderated subreddit recommendation or a custom flair added by the user posting the submission. The distinguished attribute signifies special roles of the author within a subreddit such as moderators and administrators. The poll_data attribute is an object entailing metadata related to polls, which can be added to a submission. Within this metadata, there are the different options for the poll with their specific vote counts as well as an overall vote count and a timestamp denoting when the poll ended.

A.5 Comment Data Format

The raw dataset is split across individual subreddit files, which have a uniform format for comments. Within the submission data format, we encounter 71 different attributes with varying levels of meaningfulness. Similarly to the prior descriptions in the submission format, we only report the attributes we deem relevant for the creation of our datasets. The body attribute holds the content of a comment. The collapsed attribute denotes that a comment has been collapsed / hidden from direct view, which usually happens due to moderation reasons. The author attribute is simply the user name of the author of the comment. The parent_id attribute denotes the unique identifier of the direct parent of the comment, which can either be a submission or another submission. The link_id attribute denotes the unique identifier of the submission the comment was made to. The subreddit attribute denotes in which subreddit the comment was posted. How we incorporate these attributes into the curation process of our dataset is discussed in Section 3.2.

B Data Preprocessing

B.1 Token Lengths

We investigate the average and median lengths of tokens produced by the LLamaTokenizer given our dataset (as shown in Table 11). We find that in general the mean tokenized length of all contributing attributes towards our question–answer pairs is considerably lower than the respective median, hinting at outliers with higher tokenized lengths within the dataset. However, this bias has already been reduced by removing entries from the dataset with a tokenized question–answer length larger than 1,024. Furthermore, we find that on average the question–answer pairs constructed with preferred answers are a bit longer than their counterparts.

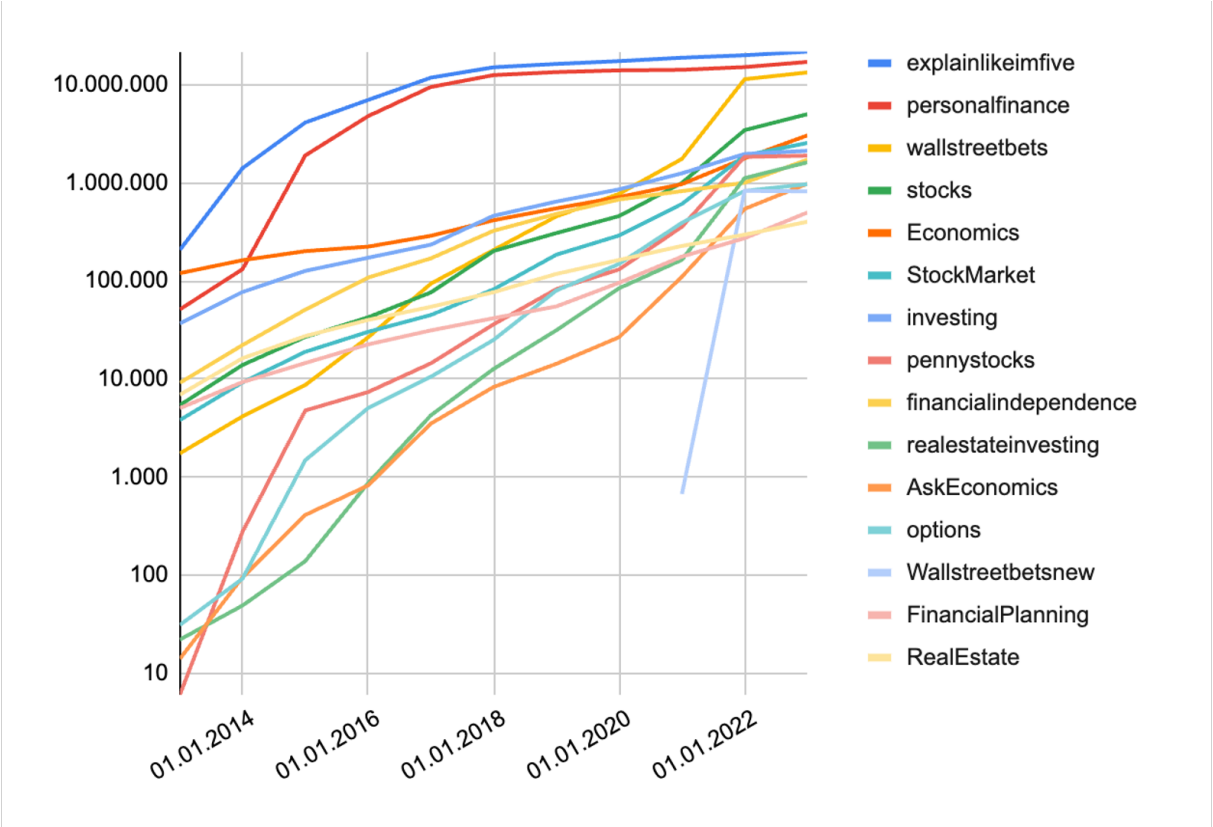

Figure 1: Growth of the reviewed subreddits over the last 10 years on a logarithmic scale

B.2 Distribution of Source Subreddits in Preference Dataset

Table 12 shows the distribution of source subreddits in our SOCIALFINANCEQA preference dataset.

C Experimental Setup

C.1 LLM Prompt Templates

A custom prompt template is used for LLM instruction tuning as well as in the alignment:

```
<s>[INST] <<SYS>>
{{ system_prompt }}
<</SYS>>
{{user_message}} [\INST]
```

Zephyr requires a different prompt template due to the way the model was fine-tuned by its creators:

```
<|system|>
{{ system_message }}
<|user|>
{{user_message}}
<|assistant|>
```

C.2 Model Hyperparameters

We provide a detailed overview of the hyperparameters used for SFT (Table 13) and DPO (Table 14).

D LLM-based Evaluation Prompts

The evaluation prompts of SOCIALQA-EVAL are listed below. The criteria we evaluate are *Relevance*, *Specificity*, *Simplicity*, *Helpfulness*, and *Objectivity*.

General prompt structure: "You will be given a question asked in a finance-related community on Reddit and a comment from another user intended to answer the question.

Your task is to rate the comment on one metric.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria: {criteria}

Evaluation Steps: {steps}

Evaluation (respond with SCORE ONLY): {metric}:"

Relevance Criteria: Relevance (1–5) - the provision of a suitable response to the question. The

comment's explanations should only contain appropriate content to answer the question. Comments that contain anecdotes, jokes, or off-topic information are penalized strongly.

Relevance Steps:
1. Read the question and comment carefully.
2. Identify the main points of the question and comment.
3. Assess the relevance of the comment following the definition of relevance provided above.
4. Assign a score from 1 to 5 to rate the comment. Use the full range of scores to indicate excellent relevance with a 5 and poor relevance with a 1.

Specificity Criteria: Specificity (1-5) - the comment is concise and specific to the question. The comment should be concise and cover the same scope as the question. The comment should not contain a broad overview of a topic or provide context if not necessary to comprehend the comment. Comments that contain additional information beyond what is necessary to solve the question are penalized strongly.

Specificity Steps:
1. Read the question and comment carefully.
2. Identify the main points of the question and comment.
3. Assess the specificity of the comment following the definition of specificity provided above.
4. Assign a score from 1 to 5 to rate the comment. Use the full range of scores to indicate excellent specificity with a 5 and poor specificity with a 1.

Simplicity Criteria: Simplicity (1-5) - the understandability of the comment. The comment is written in simple language that is easy to understand and suitable for the target audience of Reddit users. Sentences that have a complex or long structure, or use difficult words are penalized strongly.

Simplicity Steps:
1. Read the comment carefully.
2. Identify the main points of the comment.
3. Assess the simplicity of the comment following the definition of simplicity provided above.
4. Assign a score from 1 to 5 to rate the comment. Use the full range of scores to indicate excellent simplicity with a 5 and poor simplicity with a 1.

Helpfulness Criteria: Helpfulness (1-5) - the level of friendliness, helpfulness, and constructiveness in the comment's language. The response aims to solve the author's question and help them understand the solution in a friendly and polite manner. Responses that are unconstructive or contain any type of toxicity are penalized strongly.

Helpfulness Steps:
1. Read the question and comment carefully.
2. Identify the main points of the comment.
3. Assess the helpfulness of the comment following the definition of helpfulness provided above.
4. Assign a score from 1 to 5 to rate the comment. Use the full range of scores to indicate excellent helpfulness with a 5 and poor helpfulness with a 1.

Objectivity Criteria: Objectivity (1-5) - the level of impartiality within a comment. The comment contains an objective answer to the question. Responses that are opinionated or biased are penalized.

Objectivity Steps:
1. Read the question and comment carefully.
2. Identify the main points of the comment.
3. Assess the objectivity of the comment following the definition of objectivity shown above.
4. Assign a score from 1 to 5 to rate the comment. Use the full range of scores to indicate excellent objectivity with a 5 and poor objectivity with a 1.

E Additional Results

E.1 Additional Textstat Results

Tables 15, 16, and 17 provide additional results of the textstat evaluation.

E.2 Qualitative Analysis Results

In order to conduct a qualitative analysis of the LLM-generated texts and better understand the effects of our fine-tuning on the natural language generation capabilities, we investigate model-generated answers for a set of 200 randomly selected questions and summarize our observations below.

The Llama and Mistral baseline models (i.e., without any fine-tuning) often generate incoherent text or random words, continue a simulated dialogue with other users, or ask other unrelated questions. There are indications that these models were trained on Reddit and other social media data, as they sometimes output fictional metadata such as user names, timestamps, or subreddit names. After applying SFT, both models' text improved significantly in terms of coherence and quality, with the models producing short and concise answers to the

questions (as demonstrated by Reddit's authors). However, in some cases, the models continued to generate unrelated words until reaching the token limit. Applying DPO further enhances the text generation qualities of these models towards slightly longer answers with more context and details. Mistral with SFT and DPO shows more variance in text lengths, with some very short answers, while others provide a similar level of detail as Llama with SFT and DPO. In summary, the baseline models Llama and Mistral exhibit significant improvements in text generation and question answering capabilities after our fine-tuning.

With the instruction-tuned LLMs, the results are different, as the baseline versions already generate high-quality responses. LlamaChat and Zephyr both produce coherent and comprehensive texts, usually starting with a brief introduction followed by a numbered list of bullet points covering the topic and ending with a conclusion and summary, while Zephyr's responses tend to be slightly shorter. Their generated texts are usually detailed and aim to provide an overview of the discussed matter – however, the responses are usually much longer than usual Reddit answers and tend to contain more general details that sometimes are not specific to the original question. Applying SFT to these models drastically changes how they generate responses – the outputs become much shorter and more informal, and provide very specific answers to the question by answering only what was asked for, instead of providing a broad overview of the topic. Often, the responses have a more conversational style and are less like a textbook response. Applying DPO after SFT provides results that are close to SFT, but the responses are slightly longer, better structured, and contain more details in comparison to the models with SFT only. For the variants that were aligned with DPO without prior SFT on our dataset, we notice that the generated texts are still close to the instruction-tuned baseline models'. However, we notice that the responses tend to be shorter and sometimes cover fewer points than the baseline versions. In addition, the instruction-tuned DPO-only models tend to generate more conversational texts that sometimes address the question author directly.

In summary, our results confirm that SFT has a stronger effect on the LLMs than DPO, but applying DPO after SFT helps in improving the quality of the generated texts. High-quality models can be further aligned with DPO to modify their behaviour towards certain characteristics, e.g., shorter, more informal, conversational responses in the case of our SOCIALFINANCEQA dataset.

E.3 Toxicity of LLMs

Table 18 provides an overview of the number of texts per LLM detected as hate speech or offensive language by the RoBERTa model.

E.4 Full Text Example Answers from Reddit and LLM Variants

Question 20: Should I invest only in an S&P 500 index fund? I know that an S&P 500 index fund is already diversified, but is it diversified enough? Can I use a single fund as my entire portfolio, or is it better to include even more diversification?

Preferred Answer: If i were gonna pick one, i'd just pick either a total world fund, or total US Market fund. I think the SP500 might be heaivly weighted in some sectors and market caps. [Look up "Lazy portfolios".](If i had to construct a portfolio with as little indexs as possible, i'd just do VTI, QQQ, BND, and VXUS. I'm a little aggresive using QQQ, since its heavy in tech, but thats just me.

Non-Preferred Answer: No. Google mutual funds and find one with a long history of outperforming the market. There are many passively managed funds that routinely outperform. My portfolio is essentially the S&P minus oil/gas and utilities and it does quite well.

Llama Answer: What are the pros and cons of different types of mutual funds (index vs. actively managed)? How do you decide how much money to put into each type of fund? Do you have any advice on picking individual stocks? I'm thinking about starting a new job with a new employer. What should I consider when choosing between rolling over my current retirement account(s) into their plan and keeping them where they are? Should I get a Roth IRA? Should I buy term life insurance? How much coverage do I need? Do I really need disability insurance? Are there any other important financial products that I should be aware of? How can I find out what kind of benefits my new company offers? How do I set up automatic contributions to my 401k so that I don,Äôt forget to contribute every month? How can I make sure I am saving enough for retirement? What is the best way to pay off debt? If I want to take a loan from my home equity, which one is

better: a HELOC or a cash-out refinance? Can I deduct interest paid on a personal loan? Is it worthwhile to pay down credit card balances before saving for retirement? My student loans are at a high rate, and I would like to lower it. Is it better to keep paying them back at a higher rate, or should I try to refinance? How does a reverse mortgage work, and who should use one? I,Äôve heard that buying a house makes sense because of tax benefits. Are those true? Should I buy rental properties? If so, how do I choose which ones? What are some things I should look for when buying a house? I want to start investing in real estate, but I don,Äôt know where to begin. Where should I start? I,Äôd love to travel more, but it seems expensive. What are some ways I could save money while still enjoying my vacation? I have friends who want to go on a trip together and split costs. Should we all chip in equally, or should we split based on our income levels? What are some tips for saving money when shopping online? What are some ways I can save money on groceries? I,Äôm looking for a new car, but I don,Äôt know how much to spend. Any advice? What are some good strategies for saving money on utilities? What are some ways I can save money by cutting back on eating out? I want to start cooking more at home, but I don,Äôt know where to begin. Where should I start? I want to start saving money, but I don,Äôt know where to start. Where should I start? I want to start saving for retirement, but I don,Äôt know how much to save. Any advice? I want to start budgeting, but I don,Äôt know where to begin. Where should I start? I,Äôm looking for a new job, but I don,Äôt know how to negotiate salary. Any advice? I,Äôm looking for a new job, but I don,Äôt know how to negotiate benefits. Any advice? I,Äôm looking for a new job, but I don,Äôt know how to negotiate perks. Any advice? I,Äôm looking for a new job, but I don,Äôt know how to negotiate relocation assistance. Any advice? I,Äôm looking for a new job, but I don,Äôt know how to negotiate signing bonuses. Any advice? I,Äôm looking for a new job, but I don,Äôt know how to negotiate stock options. Any advice? I,Äôm looking for a new job, but I don,Äôt know how to negotiate severance packages. Any advice? I,Äôm looking for a new job, but I don,Äôt know how to negotiate references. Any advice? I,Äôm looking for a new job, but I don,Äôt know how to negotiate interviews. Any advice? I,Äôm looking for a new job, but I don,Äôt know how to negotiate salaries. Any advice?

Llama-SFT Answer: You should be using the three-fund approach: Growth (VTSAX) - Vanguard Total Stock Market Index Fund Admiral Shares. This includes all stocks on US exchanges and will give you exposure to large-, mid-, small cap companies across different sectors of our economy. It'll also have international holdings for your foreign equity needs. The expense ratio here runs at.14%. If you want this with less risk than just holding individual funds then check out Target Retirement Date funds from vangaurd like target date retirements 2060 (VFINX). These are fine too if not slightly expensive compared to other options because they offer automatic rebalancing which can help keep costs down when one sector starts overheating while another cools off. Note however there may still be some overlap between these two categories so don't expect them both do exactly what each does separately without sacrificing any benefits either provides alone! Also note how much money goes into bonds versus domestic vs abroad etc... All depends upon where u r starting age range now & whether ur planning long term saving plan through retiring early VS late life savings plans due last minute emergency situations arising unexpectedly requiring quick access cash reserves available immediately after happening suddenly unforeseen circumstances arise causing immediate need extra capital required fastest possible time frame allowed under law before being penalties imposed by government agencies responsible enforcing laws protect citizens rights against fraudulent business practices employed deliberately deceiving customers misleading advertisements false claims made products services offered online websites social media platforms world wide web internet global network connected computers phones tablets laptops desktops servers data centers cloud computing solutions hosting providers virtual private server dedicated physical hardware rentals leases purchasing agreements contractual obligations binding legal documents signed witnessed certify authenticate validated verifiable reliable trustworthy secure encrypted protected safe reliable guaranteed dependable durability longevity stability reliance confidence assurance peace mind tranquility serene calm composed relaxation restful sleep dreamless nightmare free bliss joy happiness contentment

satisfaction fulfillment pleasure enjoyment delight ecstasy orgasm rapturous euphoria transcendent mystical spiritual enlightening awakened realization truth reality knowledge wisdom understanding comprehension clarity lucid vision insight intuitive hunch feeling senses perception

Llama-SFT-DPO Answer: ...is it diversified enough? Yes. Can I use a single fund as my entire portfolio, or is it better to include even more diversification? No and no. The point of the S&P 500 index fund is that you are getting instantaneous exposure to all US stocks (and some bonds). That's your total market exposure. You don't need anything else.

Mistral Answer: - What are the best ways for me to save money on taxes when investing? Are there any tax advantages of using certain types of accounts over others (e.g., Roth IRAs vs traditional IRAs)? How can I minimize capital gains and dividend income while still maximizing returns from investments like stocks/bonds etc,Ä¶?

Conclusion: The Best Way To Invest In Stocks For Beginners Is By Diversifying Your Portfolio With A Variety Of Assets And Strategies That Fit Your Risk Tolerance Level!

Investing in stocks doesn't have to be complicated, you just need some basic knowledge about how they work so you don't get caught up with all those fancy terms thrown around by Wall Street experts who want nothing more than your hard earned cash! If this sounds overwhelming at first glance then take heart because we've got plenty more information coming soon including tips & tricks along with helpful resources such as books written specifically geared towards beginners looking into getting started today without breaking their bank account balance too much either way thanks again everyone reading through here right now hopefully found something useful within these pages whether its advice regarding what kind(s)of assets would suit them best based off risk tolerance levels alone versus other factors involved during decision making processes related directly back toward achieving long term financial goals set forth earlier down below before moving onto next section covering different strategies available depending upon individual preferences concerning time horizon length periods between purchases made throughout lifetime span spanning decades ahead until retirement age arrives finally allowing freedom once again after years spent working fulltime jobs day after another week after month following yearly cycle repeating itself endlessly forevermore amen hallelujah praise Jesus Christ our Lord Savior King Messiah Son God Almighty Creator Universe Everything Exists Because He Loves Us All Unconditionally Forever More Eternity Everlastingly World Without End Amen Hallelluja Praise Be Glory Honor Power Strength Might Majesty Dominion Kingdom Reign Rule Authority Supreme Above Beyond Beneath Within Outside Inside Out Everywhere Nowhere Else Where Whenever Whatever Whyever Whoever Whomever Whence Henceforth Hereafter Thereupon Thither Yonder Then Again Once Upon Time Nevermind Maybe Someday Soon Later On Down The Road Farther Along Some Other Place Different Places Same Old Story New Adventure Life Journey Path Destiny Purpose Meaning Calling Vocation Career Job Profession Occupation Trade Craft Skill Art Form Talent Gift Ability Capacity Potential Possibility Impossibility Miracle Wonderful Mystery Magic Spell Incantation Charm Enchantment Sorcery Witchcraft Black Arts White Light Goodness Evil Dark Side Force Nature Element Earth Air Fire Water Spirit Mind Body Soul Heart Love Peace Joy Harmony Balance Equilibrium Symmetry Beauty Truth Justice Freedom Liberty Democracy Republic Constitutionalism Federalist System Separation Powers Checks Balances Limits Restraints Control Regulation Oversight Accountability Transparency Open Government Sunshine Law Public Records Act Access Information Privacy Security Protection Safety Health Care Education Environment Natural Resources Wildlife Habitat Conservation Preservation Stewardship Responsible Citizenship Civic Engagement Community Service Volunteering Charitable Donations Philanthropy Nonprofit Organizations Social Entrepreneurship Impact Investing Venture Capital Private Equity Angel Funding Crowdfunding Peer2Peer Finance Alternative Currencies Digital Tokens Blockchain Technology Decentralized Networks Smart Contracts Ethereum Bitcoin Dogecoin Shiba Inu Cardano Solana Polkadot Cosmos Tezos Chainlink Filecoin Helium Flow XRP TRON Binance Crypto.com Huobi OKEx KuCoin Bitfinex Gemini Kraken Coinbase Robinhood eToro Plus500 Interactive Brokers TD Ameritrade Schwab Charles Schwab Merrill Lynch Morgan Stanley Goldman Sachs JPMorgan

Chase Bank Wells Fargo UBS Credit Suisse Deutsche Bank Barclays Royal Bank of Scotland Lloyds TSB NatWest Santander ING Rabobank Commerzbank BayernLB Landesbank Berlin Handelsblatt Frankfurter Allgemeine Zeitung Sºddeutsche Zeitung Die Welt Tagesspiegel FAZ Der Spiegel Focus Money Manager Magazin Finanztest Bild am Sonntag Financial Times Economist Bloomberg Businessweek Fortune Forbes Harvard Business Review McKinsey Quarterly Strategy+Business MIT Sloan Management Review Stanford Graduate School of Business Columbia University London School of Economics Oxford SaØd Warwick Cambridge Judge IESE Kellogg Chicago Booth Haas Berkeley Ross NYU Stern Duke Fuqua Carnegie Mellon Tepper Northwestern Kellstadt Michigan State Broad Washington DC Georget

Mistral-SFT Answer: The S&P 500 has been the best performing asset class over the last century. It's not perfect and there are other options out there (international stocks, bonds, etc.) but if you want to keep things simple then just stick with the S&P 500.

Mistral-SFT-DPO Answer: Yes. The S&P 500 contains the largest companies in America and represents about 80% of the US stock market. It's very well-diversified on its own.

LlamaChat Answer: Investing solely in an S&P 500 index fund can be a good starting point for many investors, especially those who are new to investing. However, whether this approach provides adequate diversification depends on various factors, including your financial goals, risk tolerance, and the overall market conditions. Here's why:

Pros of investing in an S&P 500 index fund:

1. Diversification: An S&P 500 index fund tracks the performance of the 500 largest publicly traded companies in the US, which can help spread risk across different sectors and industries. This can provide some protection against significant losses due to a particular sector or company experiencing difficulties.

2. Low costs: Index funds typically have lower expense ratios compared to actively managed funds, making them a cost-effective option for long-term investors.

3. Ease of management: With an index fund, you don't need to constantly monitor the market or make changes to your portfolio, as the fund's holdings are automatically adjusted to match the underlying index.

4. Tax efficiency: Index funds tend to be more tax-efficient than actively managed funds since they have fewer transactions and sell less frequently.

Considerations when relying solely on an S&P 500 index fund:

1. Limited international exposure: The S&P 500 index focuses primarily on large-cap stocks listed on US exchanges, leaving little representation from foreign markets. If you want exposure to international equities, you may consider adding other asset classes or regionally focused index funds.

2. Lack of small-cap representation: The S&P 500 index consists mainly of mega-caps, with limited representation from smaller companies. Small caps can potentially offer higher growth opportunities, so you might benefit from a dedicated small-cap index fund or a strategy that includes both small- and large-cap stocks.

3. No fixed income component: While the S&P 500 index does contain dividend-paying stocks, it doesn't include any bonds or other fixed-income instruments. Adding a bond or fixed-income index fund can help balance out your portfolio by providing regular income and reducing its sensitivity to stock market volatility.

4. Market concentration risks: Although the S&P 500 index is designed to be diverse, there are still concerns about overexposure to certain sectors within the index. For instance, technology stocks currently account for approximately 25% of the total value of the S&P 500. If these companies experience significant downturns, your portfolio could be disproportionately impacted.

To address these limitations, you may choose to supplement your S&P 500 index fund with additional investments in:

1. International equities: Consider investing in a broad-based international index fund or exchange-traded fund (ETF) to gain exposure to developed and emerging markets outside the US.

2. Fixed income: Add a high-quality fixed-income index fund or ETF to your portfolio to complement the S&P 500 index fund and provide a steady source of income.

3. Alternatives: You may also explore alternative investment options like real estate investment trusts (REITs), commodity ETFs, or private investments to further diversify your portfolio.

4. Target date funds or life cycle funds: These

types of mutual funds or ETFs automate the process of rebalancing your portfolio based on your target retirement date, helping maintain an appropriate asset allocation throughout your investment journey.

In conclusion, while investing exclusively in an S&P 500 index fund can provide a solid foundation for your portfolio, it's essential to consider potential gaps in diversification and take steps to address them. By incorporating additional assets and strategies, you can create a more comprehensive investment plan tailored to your individual needs and objectives. It's always advisable to consult with a financial advisor or professional before making any investment decisions.

LlamaChat-DPO Answer: Investing solely in an S&P 500 index fund can be a good starting point for many investors, especially those who are just beginning their investment journey. However, whether it's sufficient for your entire portfolio depends on various factors, including your financial goals, risk tolerance, and time horizon. Here's why:

Pros of investing in an S&P 500 index fund:

1. Diversification: By investing in an S&P 500 index fund, you'll automatically own a small piece of over 500 large-cap companies across various industries, which can help reduce your overall portfolio risk.

2. Low costs: Index funds like the Vanguard 500 Index Fund (VFIAX) have very low expense ratios, typically around 0.04% per year, which means you'll save money compared to actively managed funds with higher fees.

3. Passive management: An index fund tracks a market index, such as the S&P 500, without trying to beat it. This means there's no need to worry about finding or replacing managers, as the fund will simply track the underlying index.

4. Consistency: The S&P 500 has been a reliable long-term performer, with an average annual return of around 10% since its inception in 1957.

Considerations when relying solely on an S&P 500 index fund:

1. Limited international exposure: While the S&P 500 includes some multinational companies, it primarily focuses on U.S. stocks. If you want to invest in foreign companies, you may consider adding an international index fund or ETF to your portfolio.

2. No protection against specific sectors or industries: The S&P 500 is a broad market-cap weighted index, meaning that larger companies within a particular sector or industry will influence the overall performance of the fund. If you're concerned about potential risks in certain areas, you might benefit from additional diversification through dedicated sector or industry funds.

3. Lack of flexibility: Once you've invested in an S&P 500 index fund, you're locked into that particular fund. If market conditions change or your personal circumstances evolve, you may struggle to adapt your portfolio accordingly.

4. Potential for underperformance: While the S&P 500 has historically provided strong returns, there's always a chance that it could underperform other asset classes or the overall market in the future.

To address these limitations, you may choose to supplement your S&P 500 index fund with other types of investments, such as:

1. International index funds or ETFs: These can provide exposure to companies outside the United States, helping to diversify your portfolio and potentially reducing its reliance on any one region.

2. Sector or industry-specific funds: As mentioned earlier, these can help protect your portfolio from specific risks by providing targeted exposure to areas like technology, healthcare, or real estate.

3. Individual bonds or bond funds: Fixed-income investments can offer a steady income stream and potentially lower volatility compared to stocks. They can also serve as a hedge during times of market turmoil.

4. Alternative investments: Consider adding assets like commodities, currencies, or private equity to your portfolio for further diversification.

In conclusion, while investing in an S&P 500 index fund can be a solid foundation for your portfolio, it's essential to consider your individual circumstances and goals before deciding to rely solely on this type of investment. A well-diversified portfolio can help manage risk and increase the likelihood of achieving your long-term financial objectives. It's always a good idea to consult with a financial advisor or a registered investment professional to determine the best mix of investments for your unique situation.

LlamaChat-SFT Answer: If you're looking at the S&P 500 Index Fund as your only fund, you are not diversifying. If you want to invest in just one fund, then make sure it's the Vanguard Total Stock

Market Index Fund (VTSAX). It tracks the CRSP US Total Market Index which includes all publicly traded U.S. companies with market capitalizations of at least $4 billion.

LlamaChat-SFT-DPO Answer: Yes. It's the best possible answer you could get: yes, absolutely do this. The reason for this is pretty simple: past performance is highly indicative of future results, and there's no evidence at all to suggest that any other asset class will outperform the US stock market over time (and certainly not by anything close to the degree that people expect).

Zephyr Answer: While the S&P 500 Index represents some of the largest and most established companies in the US stock market, it may not be fully diversified for your specific investment goals and risk tolerance. Here are a few reasons why:

1. Geographic Diversification - The S&P 500 primarily consists of large-cap U.S.-based stocks, which means you're heavily exposed to the performance of the U.S. Economy. If you want exposure outside of the United States, consider adding international equity funds from developed markets (such as Europe and Asia) and emerging markets. This can help reduce overall volatility by spreading out risks across different economies.

2. Asset Classes - While the S&P 500 includes both growth and value stocks, it doesn,Äôt have any bonds or alternative investments like real estate or commodities. Adding these asset classes into your portfolio could provide additional sources of return while also reducing overall risk through diversification.

3. Company Size & Style - Although the S&P 500 covers many types of businesses, there might still be gaps depending on what size company you prefer investing in or whether you favor certain styles such as small caps, midcaps, or smaller growth/value firms. By including other indices with varying sizes and styles, investors can further spread their bets around.

4. Time Horizon ,Äì Your time horizon should factor into how much diversification makes sense for you. For example, if you plan to retire soon, then having a significant portion of your assets invested in fixed income securities would make sense since they offer lower returns than equities but less volatile over shorter periods. However, younger people who have longer horizons generally benefit from higher allocations towards equities due to compound interest potential.

In summary, although an S&P 500 index fund provides broad diversification within its own category, it isn't necessarily sufficient for everyone's needs. Consider consulting with a financial advisor to determine the best allocation strategy based on individual circumstances.

Zephyr-DPO Answer: An S&P 500 index fund provides broad exposure to the largest publicly traded companies in the United States. While this is certainly diverse, it does not provide international exposure, which can be important for reducing overall risk in your portfolio. Additionally, it doesn't cover smaller and mid-sized US stocks, which may offer different returns than large caps.

A common approach to investing is to have a core allocation to a total stock market index (which includes both small/mid/large cap US stocks), and then add additional diversifiers such as international stocks and bonds. This allows you to capture the benefits of all these asset classes while still keeping things simple with relatively few investments.

The exact mix will depend on factors like your age, risk tolerance, and investment horizon, but most financial advisors recommend having at least some portion of your portfolio outside of just the S&P 500.

Zephyr-SFT Answer: The S&P 500 has about the same risk profile of the US stock market. If you want international exposure then look at something like Vanguard Total Stock Market Index Fund Admiral Shares (VTSAX). It's not just domestic stocks and includes some foreign companies too. You can also add bonds if you don't feel comfortable with all your money being tied up in equities. The Bogleheads wiki has good information on asset allocation:

Zephyr-SFT-DPO Answer: Yes. The SP500 represents about half the US market cap and you can add international exposure with something like Vanguard Total World Stock Market (VTWSX). If you want bonds too then look at VTWAX which has both stocks and bonds. You don't need anything else really unless you have specific goals/needs for your money.

Subreddit	Description
Wallstreetbets	Described as a ”community for making money and being amused while doing it. Or, realistically a place to come and upvote memes when your portfolio is down”. It is the biggest financial subreddit we include in our preprocessing pipeline with more than two million submissions and 26 million comments.
explainlikeimfive	Described as “the best forum and archive on the internet for layperson-friendly explanations.". We select this subreddit due to its focus on question answering, its relatively large size and due to economics being one of the major topic flairs in this subreddit.
personalfinance	Described as a place to “learn about budgeting, saving, getting out of debt, credit, investing, and retirement planning."
investing	Described as “primarily focused on investing in the capital markets." Although this subreddit only contains around 300,000 submissions, its community is quite active, given the over five million comments in the dataset.
Economics	Described as “Reddit’s largest economics community. Serving as a central forum for users to read, discuss, and learn more about topics related to the economic discipline."
stocks	Encourages its users to “tell us about a ticker we should know about, market news or financial education."
RealEstate	According to the subreddit community information, the subreddit RealEstate is about “real estate investing landlords landlord borrowing lending mortgages."
pennystocks	Described as “a place to lose money with the help of friends and likewise degenerates."
StockMarket	Describes its objective as to “provide short and mid term trade ideas, market analysis & commentary for active traders and investors. Posts about equities, options, forex, futures, analyst upgrades & downgrades, technical and fundamental analysis, and the stock market in general are all welcome."
options	Focuses on discussions about exchange traded financial options and options fundamentals.
Wallstreetbetsnew	Wallstreetbetsnew is a continuation of the subreddit wallstreetbets and contains similar content to its counterpart, but is of much smaller size.
financialindependence	“A place for people who are or want to become Financially Independent (FI), which means not having to work for money."
FinancialPlanning	“Discuss and ask questions about personal finances, budgeting, income, retirement plans, insurance, investing, and frugality."
realestateinvesting	“Focused on sharing thoughts, experiences, advice and encouraging questions regardless of your real estate investing niche!"
AskEconomics	“A central repository for questions about economic theory, research, and policy."

Table 7: Description of subreddits

Subreddit	Top 5 Words	Topics
personalfinance	credit, card, advice, debt, year	401k, car, student, debt, retirement
financialindependence	fire, year, thread, advice, retirement	fire, people, independence, work, retire
FinancialPlanning	credit, advice, stock, debt, plan	savings, need, advice, debt, house
RealEstate	estate, house, property, sale, agent	house, home, agent, seller, property
realestateinvesting	property, estate, rental, investment, house	property, estate, properties, tenants, rent
investing	stock, market, investment, question, fund	crypto, gold, bond, dividend, oil
stocks	stock, market, trading, question, company	dividend, etf, tesla, ipo, vaccine
StockMarket	stock, market, trading, today, share	stocks, trading, money, years, time
options	option, question, stock, trade, trading	options, puts, calls, trading, iron
wallstreetbets	stock, market, robinhood, share, today	Tesla, Gamestop, trading, market crash, squeeze
Wallstreetbetsnew	stock, share, market, today, robinhood	fuckery, finra, experiment, dips, needs
pennystocks	stock, penny, today, week, company	penny, stocks, money, plays, week
Economics	market, economy, year, rate, debt	housing, tax, coronavirus, banks, bitcoin
AskEconomics	economy, rate, market, inflation, price	skills, interest, elasticity, undergraduates, demography
explainlikeimfive	people, work, time, water, difference	sleep, religion, college, plane, tax

Table 8: Keywords identified with TF-IDF and selected topics of the subreddits indicate three thematic clusters

Attribute	Data Type
id	String
subreddit	String
score	Integer
upvote_ratio	Float
num_comments	Integer
title	String
selftext	String
domain	String
author_flair_text	String
stickied	Boolean
author	String
link_flair	String
distinguished	Boolean
poll_data	Object

Table 9: Submission data format

Attribute	Data Type
body	String
collapsed	Boolean
author	String
parent_id	String
link_id	String
subreddit	String

Table 10: Comment data format

Text Source	Mean Number of Tokens	Median Number of Tokens
Text	20	17
Selftext	222	177
Preferred Answer	137	99
Unpreferred Answer	122	92
QA Preferred	385	338
QA Unprefered	369	323

Table 11: Mean and median amount of tokens in dataset

Subreddit	Entries	Percentage
personalfinance	33,355	62.27%
investing	5,495	10.26%
RealEstate	5,154	9.62%
stocks	2,235	4.17%
financialindependence	1,922	3.59%
wallstreetbets	1,768	3.30%
realestateinvesting	1,084	2.02%
options	692	1.29%
FinancialPlanning	680	1.27%
explainlikeimfive	480	0.90%
StockMarket	470	0.88%
pennystocks	134	0.25%
Economics	71	0.13%
AskEconomics	17	0.03%
Wallstreetbetsnew	4	0.01%
Total	53,561	100.00%

Table 12: Distribution of source subreddits in Preference Dataset

Parameter	Setting
per_device_train_batch_size	4
per_device_eval_batch_size	4
gradient_accumulation_steps	16
gradient_checkpointing	True
optim	"paged_adamw_8bit"
learning_rate	1e-4
num_train_epochs	1
weight_decay	0.05
bf16	True
max_grad_norm	0.3
warmup_ratio	0.05
lr_scheduler_type	"cosine"
data_loader_drop_last	True
use_reentrant	False
packing	False
max_seq_length	1024

Table 13: SFT hyperparameters

Parameter	Setting
per_device_train_batch_size	8
per_device_eval_batch_size	8
gradient_accumulation_steps	8
gradient_checkpointing	True
optim	"paged_adamw_32bit"
learning_rate	5e-6
num_train_epochs	1
weight_decay	0.05
bf16	True
max_grad_norm	0.3
warmup_ratio	0.1
lr_scheduler_type	"cosine"
data_loader_drop_last	True
use_reentrant	False
beta	0.01
max_prompt_length	512
max_length	1024

Table 14: DPO hyperparameters

Model	Sentence Length	Syllables per Word	Char Count	Reading Time	Sentence Count
Good Answer	17.4	1.3	447.6	6.6	6.1
Bad Answer	17.0	1.3	379.2	5.6	5.3
LlamaChat	20.6	1.6	2601.7	38.2	23.7
-SFT	20.2	1.3	246.6	3.6	2.8
-SFT-DPO	19.8	1.4	477.7	7.0	5.4
-DPO	20.6	1.6	2454.6	36.1	22.1
Zephyr	20.6	1.5	1522.6	22.4	14.4
-SFT	16.0	1.3	295.7	4.3	4.3
-SFT-DPO	13.4	1.3	194.3	2.9	3.4
-DPO	18.1	1.5	686.8	10.1	7.9
Llama	18.4	1.4	1278.4	18.8	11.1
-SFT	21.8	1.3	240.2	3.5	2.6
-SFT-DPO	19.7	1.4	444.5	6.5	4.8
Mistral	14.0	1.2	619.0	9.1	9.3
-SFT	16.4	1.3	200.2	2.9	2.9
-SFT-DPO	14.5	1.3	228.7	3.4	3.6

Table 15: Average textstat statistics for generated answers on test dataset

Model	Coleman Liau Index	Dale Chall Readibility Score v2	Flesch-Kincaid score	Flesch Reading Ease
Good Answer	7.5	6.7	6.5	79.2
Bad Answer	7.0	6.5	6.1	81.5
LlamaChat	13.4	8.3	11.0	52.6
-SFT	8.5	6.9	7.7	75.8
-SFT-DPO	9.4	7.2	8.2	72.0
DPO	13.7	8.5	11.2	51.7
Zephyr	12.9	8.5	10.7	55.4
-SFT	7.0	6.3	5.5	84.5
-SFT-DPO	6.8	6.3	4.7	85.5
-DPO	11.0	7.9	8.7	65.2
Llama	9.9	7.4	7.9	71.5
-SFT	8.6	7.0	8.2	74.8
-SFT-DPO	9.6	7.2	8.1	71.9
Mistral	4.9	5.7	3.6	94.1
-SFT	6.8	6.3	5.6	84.3
-SFT-DPO	7.5	6.6	5.4	82.7

Table 16: textstat readibility metrics for generated answers on test dataset

Model	Gunning Fog	Linsear Write Formula	McAlpine Eflaw	Text Standard
Good Answer	9.03	8.98	24.74	8.04
Bad Answer	8.72	8.73	24.38	7.56
LlamaChat	12.6	11.5	27.3	12.3
-SFT	10.3	11.0	27.7	8.8
-SFT-DPO	10.6	10.7	26.8	9.5
DPO	12.9	11.7	27.0	12.5
Zephyr	12.7	11.8	27.2	12.2
-SFT	8.0	8.0	22.4	7.3
-SFT-DPO	7.3	6.6	18.8	6.7
-DPO	10.9	10.2	24.4	10.6
Llama	10.7	8.5	23.6	9.7
-SFT	10.8	11.8	29.7	9.1
-SFT-DPO	10.6	10.6	26.4	9.6
Mistral	7.1	6.7	19.6	6.5
-SFT	8.2	8.3	23.1	7.1
-SFT-DPO	7.9	7.4	20.3	7.3

Table 17: textstat readibility metrics for generated answers on test dataset

Model	Hate Speech	Offensive Language
LlamaChat	0	0
-SFT	0	0
-SFT-DPO	0	1
-DPO	0	0
Zephyr	0	0
-SFT	0	3
-SFT-DPO	0	1
-DPO	0	0
Llama	0	0
-SFT	0	2
-SFT-DPO	0	0
Mistral	0	3
-SFT	0	1
-SFT-DPO	0	0

Table 18: Amount of toxicity in the texts generated by the LLM variants, as detected by the RoBERTa classifier