FINALE : Finance Domain Instruction-Tuning Dataset with High-Quality Rationales via Chain-of-Thought Prompting

Sangmin Lee
Korea University
sangmin_lee@korea.ac.kr

Suzie Oh
KT
suzie.oh@kt.com

Saeran Park
Korea University
saeran_park@korea.ac.kr

Gyujin Son
Yonsei University
spthsrbwls123@yonsei.ac.kr

Pilsung Kang *
Korea University
pilsung_kang@korea.ac.kr

Abstract

Recent research on financial domain large language models (LLMs) progress by applying instruction tuning to general-domain LLMs, which are known for their powerful reasoning and generation capabilities. However, specialized domains such as finance and legal are replete with arcane terminology and require specialized knowledge, resulting in a diminished user understanding of the outputs generated by LLMs. Therefore, it is crucial to augment user comprehension by accompanying the model-generated responses with detailed rationales. Nevertheless, previous works focus primarily on training to generate the answer, failing to generate appropriate rationales in the financial context. Therefore, we propose FINALE, a financial instruction tuning dataset that includes high-quality rationales generated through the use of a Chain-of-Thought (CoT) prompting and quality filtering. A model trained on FINALE shows an average improvement of 9% across nine sub-tasks compared to models trained on other instruction tuning datasets. Additionally, human evaluation results show that the comprehensibility of outputs from models trained on FINALE is rated four times higher. Through various analytical experiments, we demonstrate the effectiveness of FINALE and emphasize the importance of training models to generate high-quality rationales.

1 Introduction

Recent advancements in Large Language Models (LLMs), such as LLaMA (Touvron et al., 2023), GPT-4 (OpenAI et al., 2024), and Alpaca (Taori et al., 2023), have generated significant interest in their application across diverse domains. Researchers are actively exploring how these models' powerful generative capabilities can be leveraged to tackle various of tasks. Notably, using instruction tuning methodologies (Wei et al., 2022), LLMs are fine-tuned to perform various tasks through domain adaptation (Bao et al., 2023; Yue et al., 2023a). In the financial domain, downstream tasks are defined, training data is compiled, and general-domain LLMs are fine-tuned to enhance performance across multiple tasks through by applying instruction tuning (Wu et al., 2023a; Wang et al., 2023; Xie et al., 2023).

Most tasks evaluated within the finance domain generate short answers. However, finance documents are characterized by specialized knowledge and rare words (Mik, 2017), making it difficult for users to comprehend the outputs generated by LLMs in financial tasks (Misheva and Osterrieder, 2023; Hicham Sadok and Maknouzi, 2022). Therefore, it is essential to verify 1) whether the financial knowledge is understood and 2) whether the answers are derived based on correct reasoning. However, previous studies overlook these considerations.

We think adding reasoning steps that serve as rationales to short answers is a feasible alternative. In the general domain, enhancing the quality and length of reasoning used in training data improves model performance (Wang et al., 2022; Sanh et al., 2022; Mukherjee et al., 2023). Similarly, InvestLM (Yang et al., 2023b) enhances performance by utilizing long answers in its training data. However, it does not apply the concept of rationales that consider the constructed questions and context. Additionally, this method requires substantial human resources because it relies on manually curated selections.

In this study, we introduce FINALE, a dataset that provides rationales for texts generated in the finance domain. Additionally, we present a construction pipeline that aims to develop high-quality rationales with minimal human effort.

When trained with the developed FINALE, it exhibits an average performance enhancement of 9% over other instruction-tuned models. This aligns

* Corresponding author.

Proceedings of the Joint Workshop of the 8th Financial Technology and Natural Language Processing (FinNLP) and the 1st Agent AI for Scenario Planning (AgentScen), Jeju, South Korea, August 3, 2024

with findings in the general domain that training models to generate reasoning steps improves performance (Wei et al., 2023; Nye et al., 2021; Zhou et al., 2023b; Gao et al., 2023), and marks the first study to demonstrate its applicability in the financial domain. Figure 3 shows that models trained on FINALE exhibit an average win rate four times higher than the baseline model in human evaluation comparisons of rationale quality. This demonstrates that training with high-quality rationales enhances user comprehension.

Additionally, Son et al. (2024); Bi et al. (2024) report a degradation in general performance when domain-specific LLMs are trained with instruction tuning. Yet, models trained with FINALE exhibit less degradation in performance compared to the baseline model. The contributions of this research can be summarized as follows:

- We propose FINALE, a dataset that enriches short answers in the finance domain with high-quality rationales.
- We provide a pipeline for constructing rationales with minimal human effort.
- Our evaluation results indicate that models trained with FINALE demonstrate improved performance and enhanced comprehensibility of generated text compared to those trained with other instruction-tuning datasets.

2 Background

2.1 Finance-Specific LLMs

Traditional financial language models have been studied in various scenarios based on BERT (Devlin et al., 2019), a representative encoder-based pre-trained model. These studies have focused on Named Entity Recognition (Nakayama and Wan, 2017), News Sentiment Analysis (Araci, 2019), and Text Summarization (La Quatra and Cagliero, 2020).

The advancement of LLMs such as ChatGPT (OpenAI, 2022), GPT-4, Alpaca, and LLaMA has led to research applying their reasoning and generation capabilities to the financial domain. One of the primary methods for domain adaptation of these LLMs is instruction tuning (Wei et al., 2022), which involves fine-tuning LLMs using various instructional data to achieve desired behavioral patterns (Bao et al., 2023; Yue et al., 2023a). A notable example related to this is BloombergGPT (Wu et al., 2023b), which proposes a 50 billion parameter language model trained on a finance-specific corpus. However, the non-disclosure of the dataset poses challenges for further development of financial LLMs. To address this issue, research has focused on training on small LLMs (Son et al., 2023) and multi-task training (Wang et al., 2023; Xie et al., 2023) for specific financial tasks using publicly available financial data. Specifically, FinGPT adopts a data-centric approach and trains using LoRA (Hu et al., 2021), providing useful resources for researchers to develop their financial LLMs. InvestLM constructs its training data by manually collecting long answers based on results that extended reasoning steps significantly improve performance (Wang et al., 2022; Sanh et al., 2022; Mukherjee et al., 2023).

Our study deviates from previous research in the form of the target answers for finance LLMs. Most studies train models to generate short answers, which do not provide the rationales before the final answers. We further construct rationales suitable for the financial context. Additionally, we enhance quality by using filtering methods rather than manually selecting all data, minimizing human resources.

2.2 Chain-of-Thought Fine-tuning

Chain-of-Thought prompting (CoT) (Wei et al., 2023) is a method that encourages the model to generate reasoning steps before providing an answer. This approach effectively enhances the reasoning capabilities of language models (Nye et al., 2021; Zhou et al., 2023b; Gao et al., 2023). Furthermore, CoT-Collection (Kim et al., 2023a) has shown that training models to generate reasoning steps improves both zero-shot and few-shot performance in the general domain. Inspired by this, we aim to construct an instruction tuning dataset that includes high-quality rationales to assist the model's reasoning capabilities in the financial domain. Through this approach, we seek to enhance both the performance of the model and the quality of the generated rationales, improving user comprehension.

3 FINALE

3.1 Task Overview

Our work extensively addresses whether including rationales for diverse forms, such as numbers and tables, in addition to the traditional text-based

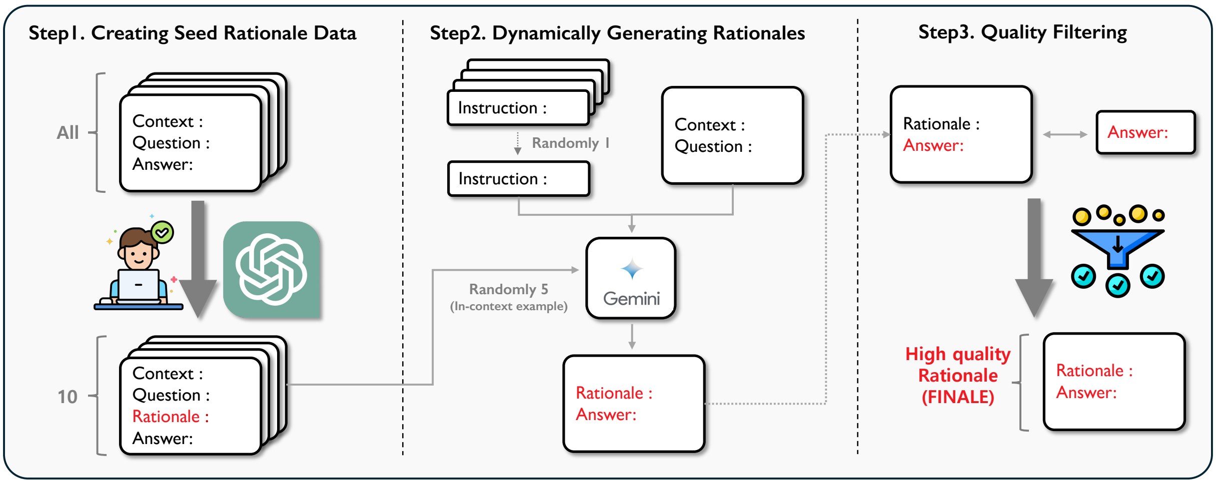

Figure 1: Overview of the FINALE creation pipeline consisting of 1) Creating seed rationale data and 2) Dynamically generating rationales and 3) Quality filtering.

Dataset	Sub-Task	# of Rationale (BF)	# of Rationale (AF)
	Sentiment Analysis		
(Malo et al., 2014)	Sentiment Analysis (FPB)	4836	2888 (-41%)
(Organizers)	Impact Type Prediction (ESG)	790	289 (-63%)
	Numerical Reasoning		
(AiHUB, 2023b)	Arithmetic (Arith.)	23064	5368 (-76%)
	Extraction (Extract.)	21000	11061 (-47%)
	Comparison (Comp.)	23016	14844 (-35%)
	Question Answering		
(AiHUB, 2023a)	Multiple-Choice Question Answering (MCQA)	5265	4715 (-20%)
	Extractive Question Answering (EQA)	8248	652 (-92%)
	Binary Question Answering (BQA)	6368	4730 (-25%)
(AiHUB, 2023c)	Table Question Answering (TQA)	50000	31886 (-36%)
Total		98681	76433 (-23%)

Table 1: An overview of the FINALE dataset. *# of Rationale (BF)* denotes the total number of rationales generated, *# of Rationale (AF)* denotes the total number of rationales after the filtering process.

instruction tuning dataset, enhances generative capabilities in the financial domain. To this end, the task selection criteria adhere to these principles.

Firstly, instead of selecting homogeneous tasks to collect data, we diversify by including different data types, such as numbers and tables. The downstream tasks are divided into four categories, including Numerical Reasoning (AiHUB, 2023b), Question Answering (AiHUB, 2023a), and Table Question Answering (AiHUB, 2023c), which require proficiency within the financial domain.

Secondly, nine sub-tasks are selected, and data is collected to ensure coverage across diverse areas. For example, the Sentiment Analysis task includes: classifying the sentiment of news sentences using a financial phrase bank and classifying risk from an ESG perspective, considering detailed scopes. Each sub-task comprises a dataset ranging from a minimum of 790 to a maximum of 50,000 instances. For more information about FINALE, refer to Table 1.

3.2 Dataset Creation

The objective of FINALE is to construct high-quality rationales in addition to the existing answers corresponding to contexts and questions. Inspired by Chung et al. (2022), we generate rationales for all instances in a CoT prompt format. In this study, the reasoning steps generated by the model are considered as rationales for the answers. Additionally, a three-step filtering process is applied to select high-quality rationales.

3.2.1 Creating Seed Rationale Data

Initially, the authors use GPT-4 to craft high-quality rationales, which are later used to guide following generations as in-context examples. Therefore, it is necessary to select a variety of high-quality rationale types. The selection criteria adhere to two main principles. Firstly, we select ten examples per sub-task while ensuring diversity. For instance, in the Arithmetic sub-task, Arithmetic is categorized into addition, subtraction, multiplication, and division. The authors reviews whether instances fall into these types and selects them in equal numbers.

Secondly, to ensure high-quality rationales, rationales are generated using GPT-4. We employ answer-based filtering and manual review to further filter for quality. If the final answer of a generated rationale differs from the gold answer, it is considered low quality. Therefore, only those with matching answers are selected. Furthermore, incorrect arithmetic operations and brief rationales, considered inadequate for explaining answers, are removed.

As discussed in LIMA (Zhou et al., 2023a), manually evaluating the quality and diversity of rationales ensures higher data quality when humans select sentence-form data. For more information about the prompts for generating seed rationales, refer to Appendix A.

3.2.2 Dynamically Generating Rationales

The seed rationale data is used as an in-context example within Gemini-Pro to generate rationales for all instances. We opt to use Gemini-Pro due to cost constraints. Nevertheless, it provides generative capabilities comparable to GPT-4, making it a reasonable alternative (Team et al., 2023). When generating rationales, we adhere to the following protocol:

The diversity in instruction data, as evidenced by WizardLM (Xu et al., 2023), has shown significant performance improvements. Therefore, instead of using fixed in-context examples, we dynamically change the in-context examples and instructions according to the principles of dynamic prompting (Yang et al., 2023a). Five out of ten seed data are randomly selected for the in-context examples, and one out of five instructions is chosen at random. This approach effectively prevent the monotony of rationales within the dataset. For more information about in-context example prompts for generating rationales, refer to Appendix B.

3.2.3 Quality Filtering

We designed a selection process to guarantee the selection of only high-quality rationales for all generated instances. While manual review of each instance is the most effective method to ensure fidelity and relevance, this approach is time-consuming, costly, and inefficient. Therefore, an automatic filtering method is applied to all instances. This filtering technique involves selecting rationales where the final answer included in the generated rationale matches the gold answer. For EQA and TQA, we filtered by ROUGE score because the answers were very long. Filtering by EM significantly reduced the data size. Instances are selected only if their ROUGE scores exceed the threshold. The criterion for choosing the threshold value was established empirically through manual inspection, determining that quality is assured when the value is 0.6 or higher.

3.3 Dataset Analysis

As illustrated in Figure 2, FINALE proposed in this paper consists of data that is, on average, longer and has a more varied distribution compared to the KOR-OpenOrca-Playti-V3 (KyujinHan) and CoT-Collection-Ko (Kim et al., 2023b) datasets. Liu et al. (2023) has demonstrated that the length of instruction data is crucial in building better models. Table 2 compares the datasets' total number of instances and the count of unique words. FINALE contains the highest number of instances and unique words among these datasets. Moreover, unlike the other two datasets that cover various domains, FINALE includes a diverse vocabulary within the confined domain of finance. A diverse vocabulary can help improve performance (Choe et al., 2023).

Dataset	Instances	Unique Words	Rate
FINALE (Ours)	78k	96k	123%
KOR-Platypus-v3[1]	34k	73k	213%
CoT-Collection-Ko	77k	89k	155%

Table 2: Comparison of number of instances and unique words. unique words is the number of unique words after tokenize the Mecab-Tokenizer (Kudo et al., 2004), and rate is the ratio of the total number of unique tokens divided by the total number of instances.

4 Experiments

4.1 Experimental setting

Baseline Models We utilize two open-source models that show high performance on Korean tasks as the foundation and baseline model, selecting models with 7 billion parameters due to resource limitations.

[1]KOR-OpenOrca-Playti-V3

Models	FPB	ESG	Arith.	Ext.	Comp.	MCQA	EQA	BQA	TQA	Avg
Yi-6B-Ko	28.0	9.6	9.2	<u>55.2</u>	58.0	44.0	43.2	39.6	21.6	34.3
Ko-Platyi-6B	54.0	**23.6**	9.2	52.4	60.0	65.2	**56.8**	58.0	22.8	44.7
FINALE (OURS)										
Single (per100)	56.8	15.2	10.0	52.0	60.0	70.8	41.2	85.6	15.6	45.2
Single (per400)	54.4	<u>21.6</u>	<u>14.8</u>	53.6	57.2	83.2	46.8	83.2	15.2	47.8
Single (All)	<u>71.2</u>	15.2	<u>14.8</u>	54.8	**68.8**	**89.2**	<u>50.0</u>	<u>86.8</u>	**24.8**	<u>52.8</u>
Multi (All)	**76.0**	19.6	**15.6**	**55.6**	<u>68.4</u>	<u>83.6</u>	<u>50.0</u>	**87.2**	**24.8**	**53.4**
Δ	+22.0%	-4.0%	+6.4%	+3.2%	+8.4%	+18.4%	-6.8%	+29.2%	+2.0%	+8.7%

Table 3: Performance of models with different samples of FINALE. The highest-scoring model per task is highlighted in **bold**, and the second-highest is <u>underlined</u>. Δ values indicate the percentage change in performance of FINALE trained with Full data, compared to the Ko-Platyi-6B model.

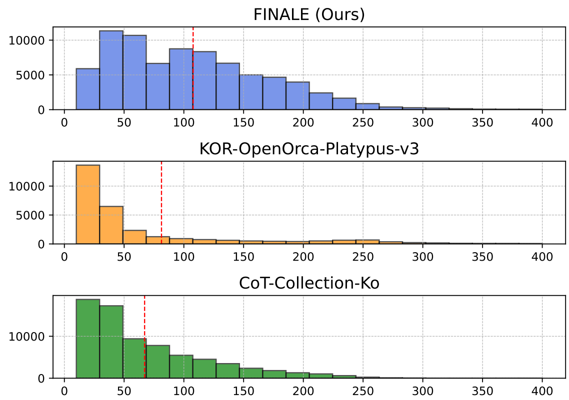

Figure 2: Length distribution of different datasets. The X-axis represents the number of tokens in each instance after tokenization, and the Y-axis represents the number of instances.

Yi-6b-ko is a model that is further pre-trained on Korean and English datasets using the Yi-6B (Yue et al., 2023b). We chose this model because it is recognized for its superior Korean language comprehension among smaller-sized models, as confirmed by the Korean LLM Leaderboard (Park et al., 2023) and KMMLU (Son et al., 2024).

Ko-Platyi-6B is a model that is instruction-tuned using the KOR-OpenOrca-Playti-V3 dataset. The KOR-OpenOrca-Playti-V3 dataset is a Korean translation of the Open-Platypus dataset. The reason for selecting Ko-Platyi-6B as the baseline is that it is an instruction-tuned model, which allows for an equivalent comparison. For more information about the training method and hyperparameters, see Appendix C.

Evaluation Method Given the varying sizes of the evaluation datasets across different datasets, we randomly sample 250 instances from the original evaluation dataset for each sub-task. Consequently, the evaluation dataset comprises 2250 instances across the nine sub-tasks.

The evaluation prompt is assessed in a few-shot setting rather than a zero-shot setting to measure the maximum performance of the model. In the few-shot setting, three in-context examples are used, selected considering for the model's maximum input length. To ensure a fair comparison of model performance, the same prompt is used for all models, although the instructions are different for each task. The details of the evaluation prompt can be found in Appendix D.

To ensure precise performance measurement of the model, sophisticated post-processing is applied to each sub-task. For instance, in the Arithmetic sub-task, the number following the "=" symbol in expressions like "3+5=2" is extracted as the final answer. For generated text that is not addressed by post-processing, we apply the Cover EM method (Rosset et al., 2021), a technique also used in the existing financial LLM, FinGPT.

4.2 Experimental Results

4.2.1 Performance Comparison

The results in Table 3 demonstrate that the model trained using FINALE (Multi) outperforms the Ko-Platyi-6b in most sub-tasks, with an average performance that is 20% higher than Yi-6B-Ko and 9% higher than Ko-Platyi-6B. This indicates that training the model to generate rationales before generating the final answers leads to more accurate answer generation.

4.2.2 Rationale Quality Comparison

Methodology To demonstrate the efficacy of FINALE, which is constructed to ensure high quality, we measure the quality of generated text from models trained using FINALE. We select Ko-Platyi-6b, an instruction-tuned model, as the baseline. Three human annotators evaluate the answer generated by both models according to the assessment guidelines provided in Appendix E. Human annotators are composed of experts in economics. Ten samples for each sub-task are randomly selected for assessment.

Annotators are directed to determine the more comprehensible rationale from different models. If no rationale is deemed superior, evaluators score it as a "Tie" (2). Ultimately, the average score calculated by the three annotators for each sub-task determines the final score. The inter-annotator agreement shows a high consistency rate of 83% on average across sub-tasks.

Results Figure 3 indicates that outputs from the model trained with FINALE average scored four times higher than the baseline, demonstrating that FINALE was significantly helpful. This finding suggests that high-quality rationales improve the model's ability to reason correct answers and significantly enhance comprehension for the users viewing the generated text. Particularly in tasks like ESG, EQA, and TQA, where quantitative performance is low, the comprehensibility of the generated text is higher than that of the baseline. This indicates that even if the model generates incorrect answers, a high-quality rationale increases the likelihood of human understanding. Therefore, we emphasize that in the finance domain, the rationale quality is as crucial as the correctness of the answers generated by the model.

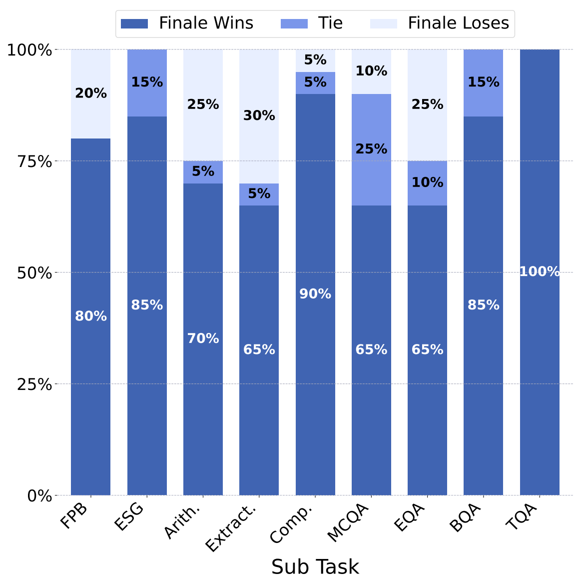

Figure 3: Result of the human evaluation of rationales generated by the model trained on Yi-6B using FINALE and the rationales generated by the `Ko-Playti-6B`.

5 Demonstrate the Effectiveness of Data Construction

To demonstrate the effectiveness of the FINALE construction method, we conduct additional experiments.

5.1 Compare other learning methods

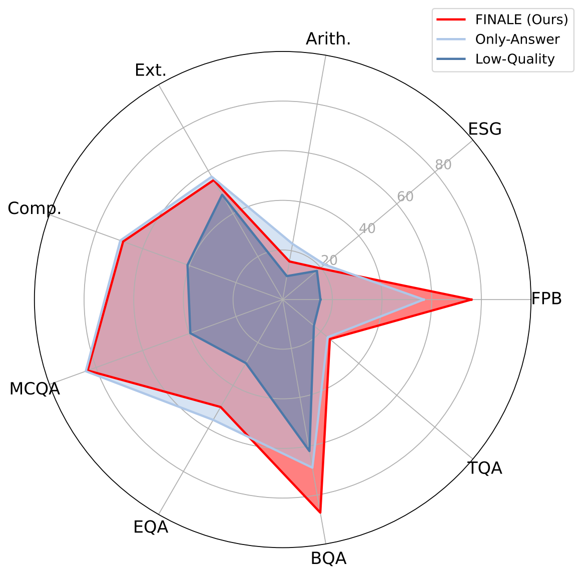

Figure 4: "FINALE" indicates training of the `Yi-6B` model with full data. "Only-Answer" indicates the results of training exclusively on answers, excluding rationales. "Low-Quality" indicates the results of training solely with data that has been removed through a filtering method.

Figure 4 compares the performance when using the same dataset as FINALE but excluding rationales and using only answers (light blue area in Figure 4) and when using data identified as low quality during the filtering process (blue area in Figure 4). The results show that FINALE significantly outperforms models trained solely on answers in the binary classification tasks of FPB and BQA. However, similar or significantly degraded performances are observed in other sub-tasks. This can be interpreted due to two factors.

Firstly, when the rationale length is excessively long, there is a tendency to generate new labels not specified in the task. For example, in the ESG sub-task, the average rationale length is 475, which is considerably longer than the average of around 100 in other tasks, leading to the generation of different answers, such as "economic" and "environmental" instead of the final gold answers like "indistinct," "opportunity" or "risk". The EQA sub-task, with an average rationale length of 486, also indiscriminately generated answers that differ from the gold answer. This suggests that contrary to existing studies showing that reasoning steps improve model generate capabilities (Wei et al., 2023; Nye et al., 2021; Zhou et al., 2023b; Gao et al., 2023), excessively long rationales may hinder the model's ability to generate the final gold answer.

Secondly, the rationale for the second numeric reasoning task emphasizes the need for specialized explanations. The rationales of FINALE tend to be short and concise, which do not sufficiently reflect the challenging nature of the numeric reasoning task. Therefore, numerical reasoning must consider the specialized rationale of arithmetic operations and the characteristics of the finance domain, suggesting that a sufficiently lengthy rationale is needed to explain the final answer adequately.

Moreover, models trained with datasets classified as low-quality (blue area in Figure 4) show significantly reduced performance across all sub-tasks, as evidenced by Figure 4. Despite the minimal difference in data quantity between Low-Quality and High-Quality datasets for most sub-tasks. In particular, in Extractive Question Answering (EQA), it is observed that performance significantly declines despite the notably large quantity of training data from low-quality cases. This emphasizes that rationales containing incorrect answers can substantially hinder model training. Therefore, the filtering process has a significant impact on enhancing the model's performance.

5.2 Compare training data quantity

Table 3 presents a performance comparison based on the quantity of training data. Specifically, Single (per100) and Single (per400) represent results from models trained with 100 and 400 randomly sampled instances in each sub-task, respectively. Despite the very small amount of data, these models surpass the performance of the baseline `Ko-Platyi-6B`. This indicates that the superior performance of FINALE is more than merely due to the large data size.

Furthermore, when comparing the results of training only on a single task (All) versus training on all tasks simultaneously multi-task (All), it is found that training in a multi-task (All) setting yields higher performance in all tasks except MCQA and Comp. This indicates that more extensive training on financial domain data simultaneously enhances performance across various tasks. Especially, FPB and ESG are classification tasks based on financial terminology. By training with other sub-tasks, additional financial knowledge is acquired, which consequently has been observed to enhance performance.

5.3 Compare General Performance

Appendix F presents the effects of using the domain-specific dataset FINALE on general performance. Performance changes are observed using the Korean benchmark dataset KMMLU (Son et al., 2024), leading to the following key findings:

Models trained on `Ko-Platyi-6B` and FINALE (Ours) exhibit lower performance compared to `Yi-6B-Ko`. This aligns with previous studies suggesting that Instruction Tuning can degrade general performance and negatively impact knowledge-based benchmarks (Son et al., 2024; Bi et al., 2024). However, despite being a finance domain instruction tuning dataset, FINALE shows a less performance decline than `Ko-Platyi-6B`.

This can be analyzed for two reasons. Firstly, `Ko-Platyi-6B`, derived from translated English datasets, may suffer from quality degradation due to translation errors (Xia et al., 2019; Riley et al., 2023; Yao et al., 2024). In contrast, FINALE is not a translated dataset and ensures high data quality through a filtering process that removes 54% of the original dataset. Secondly, while `Ko-Platyi-6B` focuses on generating diverse instructions and answers, FINALE emphasizes training models on high-quality rationales before the final answers. These results underscore the importance of data quality management and stringent filtering processes, indicating that methodologies like instruction tuning that consider rationales are essential to minimize declines in general performance.

6 Conclusion

In this paper, we propose FINALE, an instruction-tuning dataset with high-quality rationales for the financial domain. Furthermore demonstrate that training the foundation model on FINALE enhances the generative capabilities of LLMs. Notably, the performance is approximately 9% better than the baseline. Human comprehension of the model-generated text is shown to be four times better. Additionally, despite using only a very small amount of data (100 or 400 instances), the performance exceeded that of the baseline, with minimal degradation in general performance. Through this study, we anticipate an increased recognition of the importance of rationales in the finance domain.

Limitations This research is conducted solely in Korean. However, our data generation method applies to all languages, leaving research in other languages as a future work. Additionally, we utilized Gemini-Pro for rationale generation due to the high API prices. However, by using more powerful models such as GPT-4, the quality of the rationales and performance can be further enhanced. Lastly, due to the lack of specific criteria for rationale quality in the financial domain, this study utilizes an automatic metric for filtering based on the final answers. Consequently, we plan future research to establish criteria for rationale quality suitable for the finance domain.

7 Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2022R1A2C2005455). This work was also supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2021-0-00471, Development of Autonomous Control Technology for Error-Free Information Infrastructure Based on Modeling Optimization). We express our sincere gratitude to Yukyung Lee for her thoughtful advice. We also appreciate Jinwoo Park, Hanwool Lee, and Inyeol Choi for their invaluable feedback and evaluation. Additionally, we are thankful to the researchers at the DSBA lab for their comments on the paper.

References

AiHUB. 2023a. Aihub; financial & legal machine reading comprehension dataset. https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=&topMenu=&aihubDataSe=data&dataSetSn=71610. Accessed: 2024-02-21.

AiHUB. 2023b. Aihub; numerical machine reading comprehension dataset. https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=71568. Accessed: 2024-02-21.

AiHUB. 2023c. Aihub; tabular machine reading comprehension dataset. https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=71565. Accessed: 2024-02-21.

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

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-medllm: Bridging general large language models and real-world medical consultation. *Preprint*, arXiv:2308.14346.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Jaeyoung Choe, Keonwoong Noh, Nayeon Kim, Seyun Ahn, and Woohwan Jung. 2023. Exploring the impact of corpus diversity on financial pretrained language models. *Preprint*, arXiv:2310.13312.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. *Preprint*, arXiv:2211.10435.

Fadi Sakka Hicham Sadok and Mohammed El Hadi El Maknouzi. 2022. Artificial intelligence and bank credit analysis: A review. *Cogent Economics & Finance*, 10(1):2023262.

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

Seungone Kim, Se June Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. 2023a. The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. *Preprint*, arXiv:2305.14045.

Seungone Kim, Se June Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. 2023b. The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. *arXiv preprint arXiv:2305.14045*.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Barcelona, Spain. Association for Computational Linguistics.

KyujinHan. Kor-openorca-platypus-v3. https://huggingface.co/datasets/kyujinpy/KOR-OpenOrca-Platypus-v3. Accessed: 2024-03-01.

Moreno La Quatra and Luca Cagliero. 2020. End-to-end training for financial report summarization.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2023. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. *arXiv preprint arXiv:2312.15685*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65.

Eliza Mik. 2017. Smart contracts: terminology, technical limitations and real world complexity. *Law, Innovation and Technology*, 9:1–32.

Branka Hadji Misheva and Joerg Osterrieder. 2023. A hypothesis on good practices for ai-based systems for financial time series forecasting: Towards domain-driven xai methods. *Preprint*, arXiv:2311.07513.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.

Makoto Nakayama and Yun Wan. 2017. Exploratory study on anchoring: Fake vote counts in consumer reviews affect judgments of information quality. *Journal of theoretical and applied electronic commerce research*, 12:1–20.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. 2021. Show your work: Scratchpads for intermediate computation with language models. *Preprint*, arXiv:2112.00114.

OpenAI. 2022. Openai: Introducing chatgpt. `https://openai.com/blog/chatgpt`.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

FinNLP Shared Task Organizers. Multi-lingual esg impact duration inference (ml-esg-3). https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp-kdf-2024/shared-task-ml-esg-3. Accessed: 2024-02-21.

Chanjun Park, Hwalsuk Lee, Hyunbyung Park, Hyeonwoo Kim, Sanghoon Kim, Seonghwan Cho, Sunghun Kim, and Sukyung Lee. 2023. Open ko-llm leaderboard. https://huggingface.co/spaces/upstage/open-ko-llm-leaderboard.

Parker Riley, Timothy Dozat, Jan A. Botha, Xavier Garcia, Dan Garrette, Jason Riesa, Orhan Firat, and Noah Constant. 2023. Frmt: A benchmark for few-shot region-aware machine translation. *Preprint*, arXiv:2210.00193.

Corby Rosset, Chenyan Xiong, Minh Phan, Xia Song, Paul Bennett, and Saurabh Tiwary. 2021. Knowledge-aware language model pretraining. *Preprint*, arXiv:2007.00655.

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

Guijin Son, Hanearl Jung, Moonjeong Hahm, Keonju Na, and Sol Jin. 2023. Beyond classification: Financial reasoning in state-of-the-art language models. *arXiv preprint arXiv:2305.01505*.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2024. Kmmlu: Measuring massive multitask language understanding in korean. *arXiv preprint arXiv:2402.11548*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Martin Chadwick, Gaurav Singh Tomar, Xavier Garcia, Evan Senter, Emanuel Taropa, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Yujing Zhang, Ravi Addanki, Antoine Miech, Annie Louis, Laurent El Shafey, Denis Teplyashin, Geoff Brown, Elliot Catt, Nithya Attaluri, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaly Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan,

Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, Hanzhao Lin, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yong Cheng, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, YaGuang Li, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Gamaleldin Elsayed, Ed Chi, Mahdis Mahdieh, Ian Tenney, Nan Hua, Ivan Petrychenko, Patrick Kane, Dylan Scandinaro, Rishub Jain, Jonathan Uesato, Romina Datta, Adam Sadovsky, Oskar Bunyan, Dominik Rabiej, Shimu Wu, John Zhang, Gautam Vasudevan, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Betty Chan, Pam G Rabinovitch, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Sahitya Potluri, Jane Park, Elnaz Davoodi, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Chris Gorgolewski, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Paul Suganthan, Evan Palmer, Geoffrey Irving, Edward Loper, Manaal Faruqui, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Michael Fink, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marin Georgiev, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Minnie Lui, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Lam Nguyen Thiet, Daniel Andor, Pedro Valenzuela, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Sarmishta Velury, Sebastian Krause, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Tejasi Latkar, Mingyang Zhang, Quoc Le, Elena Allica Abellan, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi

Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Sid Lall, Ken Franko, Egor Filonov, Anna Bulanova, Rémi Leblond, Vikas Yadav, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Hao Zhou, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Jeremiah Liu, Mark Omernick, Colton Bishop, Chintu Kumar, Rachel Sterneck, Ryan Foley, Rohan Jain, Swaroop Mishra, Jiawei Xia, Taylor Bos, Geoffrey Cideron, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Petru Gurita, Hila Noga, Premal Shah, Daniel J. Mankowitz, Alex Polozov, Nate Kushman, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Anhad Mohananey, Matthieu Geist, Sidharth Mudgal, Sertan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Quan Yuan, Sumit Bagri, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Aliaksei Severyn, Jonathan Lai, Kathy Wu, Heng-Tze Cheng, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Mark Geller, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Andrei Sozanschi, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Abhimanyu Goyal, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Sabaer Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Tao Zhu, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Dustin Tran, Yeqing Li, Nir Levine, Ariel Stolovich, Norbert Kalb, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Balaji Lakshminarayanan, Charlie Deck, Shyam Upadhyay, Hyo Lee, Mike Dusenberry, Zonglin Li, Xuezhi Wang, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Summer Yue, Sho Arora, Eric Malmi, Daniil Mirylenka, Qijun Tan, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Steven Zheng, Francesco Pongetti, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Ragha Kotikalapudi, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Chenkai Kuang, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Pei Sun, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Ishita Dasgupta, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Yuan Liu, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Ivo Penchev, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Adam Kurzrok, Lynette Webb, Sahil Dua, Dong Li, Preethi Lahoti, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Taylan Bilal, Evgenii Eltyshev, Daniel Balle, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Adams Yu, Christof Angermueller, Xiaowei Li, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Kevin Brooks, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Komal Jalan, Dinghua Li, Ginger Perng, Blake Hechtman, Parker Schuh, Milad Nasr, Mia Chen, Kieran Milan, Vladimir Mikulik, Trevor Strohman, Juliana Franco, Tim Green, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2023. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Neng Wang, Hongyang Yang, and Christina Dan Wang. 2023. Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets. *Preprint*, arXiv:2310.04793.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *Preprint*, arXiv:2204.07705.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

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. *Preprint*, arXiv:2303.17564.

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

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized data augmentation for low-resource translation. *Preprint*, arXiv:1906.03785.

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

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *Preprint*, arXiv:2304.12244.

Xianjun Yang, Wei Cheng, Xujiang Zhao, Wenchao Yu, Linda Petzold, and Haifeng Chen. 2023a. Dynamic prompting: A unified framework for prompt tuning. *Preprint*, arXiv:2303.02909.

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

Binwei Yao, Ming Jiang, Diyi Yang, and Junjie Hu. 2024. Benchmarking llm-based machine translation on cultural awareness. *Preprint*, arXiv:2305.14328.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, and Zhongyu Wei. 2023a. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *Preprint*, arXiv:2309.11325.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023b. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *Preprint*, arXiv:2311.16502.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023a. Lima: Less is more for alignment. *Preprint*, arXiv:2305.11206.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023b. Least-to-most prompting enables complex reasoning in large language models. *Preprint*, arXiv:2205.10625.

A Rationale Generation Prompt

Below are examples of the prompts we used to generate a seed rationale for two sub-tasks: the sentiment analysis task, which receives only one sentence as input, and the multiple-choice question answering task, which receives context and a question. These prompts aim to instruct the model to generate appropriate rationales for each sub-task.

Instruction: Please classify the sentiment of the sentence as positive, negative, or neutral, and explain the reasons step by step in Korean. Finish your answer in the following format. "Therefore, the answer is X."

Input: During the past decade, POSCO has gradually divested noncore assets and bought several sports equipment makers, including California-based Fitness Products International and Sparks, Nevada-based ATEC, a leading maker of baseball and softball pitching machines.

Response:

Figure 5: Sentiment Analysis Prompt used in Rationale Generation.

Instruction: Please answer the question based on the context, and explain the reasons step by step in Korean. Finish your answer in the following format. "Therefore, the answer is X."

Context: During the first quarter, capital investment was surveyed to have increased slightly compared to the fourth quarter of the previous year. The semiconductor sector saw an increase due to the import of manufacturing equipment such as EUV (Extreme Ultraviolet) lithography equipment. Additionally, steel saw a slight increase due to facility maintenance during the Lunar New Year holidays, and machinery and equipment due to the anticipation of economic recovery. On the other hand, automobiles, displays, petrochemicals, and refining showed levels similar to the fourth quarter of the previous year. The monitoring results indicate that capital investment is expected to increase slightly in the future. Semiconductors are expected to see a slight increase due to the introduction of new production equipment, and petrochemicals and refining due to the gradual recovery in demand for petroleum products. Meanwhile, automobiles, displays, steel, and machinery and equipment are expected to maintain levels similar to the first quarter.

Question: What is expected to slightly increase in capital investment due to the anticipated gradual recovery in demand for petroleum products according to monitoring results?

Options: A. Steel B. Petrochemicals/Refining C. Automobiles D. Displays

Response:

Figure 6: Multi-Choice Question Answering Prompt used in Rationale Generation.

B Rationale Generation In-Context Example Prompt

Below is the prompt for a sentiment analysis task for rationale generation. The purpose of the prompt is to generate a high-quality rationale by providing five in-context examples.

Instruction: Please classify the sentiment of the sentence as positive, negative, or neutral, and explain the reasons step by step in Korean. Finish your answer in the following format. "Therefore, the answer is X."

Input: During the past decade, POSCO has gradually divested noncore assets and bought several sports equipment makers, including California-based Fitness Products International and Sparks, Nevada-based ATEC, a leading maker of baseball and softball pitching machines.

Response: This sentence describes how POSCO has disposed of non-core assets over the past decade and acquired a major sports equipment manufacturer. These strategic actions reflect the strengthening of the company's portfolio and focus on its business, receiving positive evaluations. Therefore, the answer is positive.

Input: Net sales of Finnish Sanoma Learning & Literature, of Finnish media group Sanoma, decreased by 3.6% in January-June 2009 totalling EUR 162.8 mn, down from EUR 168.8 mn in the corresponding period in 2008.

Response: The decrease in sales can be considered negative because it is generally interpreted as a negative indicator for a company. A decline in sales can suggest a decrease in profitability, market share, and economic soundness, which can cause concern among relevant stakeholders. This sentence is classified as having a negative sentiment because the decrease in sales is perceived as an event that can negatively affect the performance of a company. Therefore, the answer is negative.

Input: Vaisala Oyj Stock exchange release 26.03.2010 klo 09.00 1 (1) The Annual General Meeting of Vaisala Oyj decided on March 25, 2010 to approve the Company's annual accounts for 2009.

Response: This sentence reports on the approval of the annual accounts for 2009 at the annual general meeting of Vaisala Oyj. This reporting of facts is intended to convey information rather than evoke emotions. It does not explicitly contain elements that are emotionally positive or negative, but simply provides information about the event of the company's annual accounts approval. Therefore, classifying this sentence as neutral is due to the content not being biased towards any specific emotion and being faithful to the purpose of providing information. Therefore, the answer is neutral.

Input: 26 October 2010 - Finnish environmental management company Lassila & Tikanoja Oyj (HEL: LAT1V), or L&T, said today its net profit declined to EUR 20.9 m for the first nine months of 2010 from EUR 27.6 m for the same period a year earlier.

Response: A decrease in net profit generally indicates a deterioration in corporate performance, so such news can lead to a negative evaluation of the company. Therefore, the reason for the negative classification is that the decrease in net profit represents concerns about the financial soundness of the company. Therefore, the answer is negative.

Input: Finnish sports equipment maker Amer Sports Oyj (HEL: AMEAS) said today that its net loss narrowed to EUR 16.9 million (USD 22.3m) in the second quarter of 2010 from EUR 23.2 million in the corresponding period a year earlier.

Response: The reason this statement is classified as having a positive sentiment is because the decrease in net loss signifies an improvement in the company's financial performance, which can send a positive signal to investors and the market. Therefore, the answer is positive.

Input: Over the course of 6 months, operating profit and net sales increased from 0.4 million euros to 3.2 million euros compared to the same period in 2005.

Response:

Figure 7: Sentiment Analysis Prompt used in Rationale Generation.

C Training Method and Hyper parameters

We conducted fine-tuning on the Yi-Ko-6B model using a single A100 80GB GPU. For training, we selected the AdamW (Loshchilov and Hutter, 2019) optimizer along with a linear learning rate scheduler, setting the learning rate at 3e-4 and a batch size of 16 with a gradient accumulation of 4. Additionally, to optimize the training process, LLaMA-Factory (Zheng et al., 2024) was utilized as the code base during fine-tuning.

Hyperparameter	**Value**
Model	Yi-6B-Ko
Number of epochs	3
Devices	1 Tesla A100 GPU (80GB)
Learning rate	3e-4
Scheduler	Linear
Batch size	16
Input max length	2,048
Output max length	512
Temperature	0.0

Table 4: Hyperparameters used for training and inference.

D Evaluation In-Context Example Prompt

Instruction: Please classify the sentiment of the sentence as positive, negative, or neutral.

Input: Over a three-month period, both operating profit and net sales increased to 160,000 euros and 1.39 million euros respectively, compared to the same quarter in 2006.

Response: Positive tone: The use of the word "increased" suggests that the company's performance is improving, which conveys a positive meaning. Such increases imply the company's profitability and potential for growth. Positive outcome: Increases in operating profit and net sales generally indicate good financial performance and overall health of a company. This can elicit positive responses from investors and market participants. Taking these factors into consideration, this sentence contains positive content that reflects the improvement and financial growth of the company. Therefore, the answer is positive.

Input: In Finland, sales decreased by 10.5% in January, while international sales fell by 17%.

Response: Mention of sales decrease: The sentence states, "In Finland, sales decreased by 10.5% in January, while international sales fell by 17%." A decrease in sales can negatively impact a company's revenue and prospects for growth. Deterioration in market conditions: The decrease in sales can occur due to a worsening of market conditions, increased competition, or internal problems within the company. These factors can negatively affect the future growth and profitability of the company. Therefore, the answer is negative.

Input: According to Gran, while the company is growing, there are no plans to move all production to Russia.

Response: The sentence conveys information about whether the company will relocate its production to Russia but does not express a clear opinion or emotion. This sentence is interpreted as neutral since it lacks a positive or negative tone. Therefore, the answer is neutral.

Input: {sentence}

Response:

Figure 8: Sentiment Analysis Prompt used in evaluation.

E Human annotation guideline

Read the question and the context, and choose which of the two rationales given is more helpful in reasoning the correct answer. Your choice should not be based on whether you got the answer right but on whether the rationale helps you make an inference even if you got the answer wrong.

Rationale 1 -> 1 in the selection box
Rationale 2 -> 2 in the selection box
Neither rationale is helpful -> 0 in the selection box

Context: {context}
Question: {question}
Answer: {answer}

Rationale 1: {$model_A$ rationale}
Rationale 2: {$model_B$ rationale}

Selection:

Figure 9: Human Annotation Guideline.

F Compare Generation Performance

Models	HUMSS	STEM	Applied Science	Other	Total
Yi-6B-Ko	39.76	40.49	39.51	41.62	40.33
Ko-PlatYi-6B	39.13	36.94	37.21	39.00	38.05
FINALE (Ours)	39.71	40.14	38.96	40.36	39.77

Table 5: The result of the General Performance comparison using the KMMLU dataset