University of Glasgow at the FinLLM Challenge Task: Adapting Llama for Financial News Abstractive Summarization

Lubingzhi Guo and **Javier Sanz-Cruzado** and **Richard McCreadie**
University of Glasgow
l.guo.1@research.gla.ac.uk and javier.sanz-cruzadopuig@glasgow.ac.uk
and richard.mccreadie@glasgow.ac.uk

Abstract

In this paper, we explore different approaches for aligning Large Language Models (LLMs) with the objectives of the financial abstractive summarization shared task. This shared task focuses on using LLM to abstract news into concise summaries. We investigate three common strategies: few-shot learning, fine-tuning, and reinforcement learning, to adapt LLMs for this purpose, with the fine-tuned model ranked first on the leaderboard.

1 Introduction

Text summarization aims to create coherent and concise summaries from input documents using either extractive and abstractive methods. The extractive approach identifies the most important sentences from the source text(s) and concatenates them into a summary, while the abstractive method focuses on generating novel sentences and words (Widyassari et al., 2022). With the advent of large language models (LLM) in text generation, summaries generated by state-of-the-art LLMs, specifically with instruction tuning, perform comparably to those written by human annotators (Zhang et al., 2024).

The financial text summarization shared task is designed to explore the capabilities of LLMs in the finance domain; the task is focused on generating abstractive news summaries using LLMs. We approach the task by exploring the current tuning strategies for LLMs with the goal of generating concise financial summaries.

2 Dataset

The provided training dataset consists of 8,000 news articles from the EDT corpus (Zhou et al., 2021), which is designed for news event detection and financial domain adaption. For the purpose of financial abstractive summarization, the gold summaries are constructed through distant supervision, using the corresponding news headlines and sub-headlines.

To gain a clear understanding of the summary requirements, we further analyze the provided gold summaries. Figure 1 presents a histogram of the word counts for the gold summaries in the training dataset. The x-axis represents the word count for each summary, calculated using the nltk library (Bird et al., 2009), while the y-axis shows the frequency of summaries at the corresponding word count. From the histogram, it is evident that the majority of summaries have between 10-25 words, with very few extending beyond 100 words. This indicates that the reference summaries are generally brief. Despite the gold summaries comprising both headlines and sub-headlines, the evaluation of this task primarily focuses on headline generation.

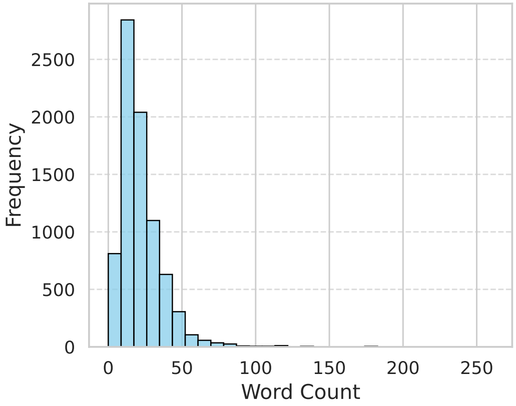

Figure 1: Frequency Distribution of Word Counts in Gold Summaries

3 Methods

We investigate the three techniques described below to adapt the pre-trained LLM from the general domain for this specific summarization task. Figure 2 shows the overall procedure for the three methods described below.

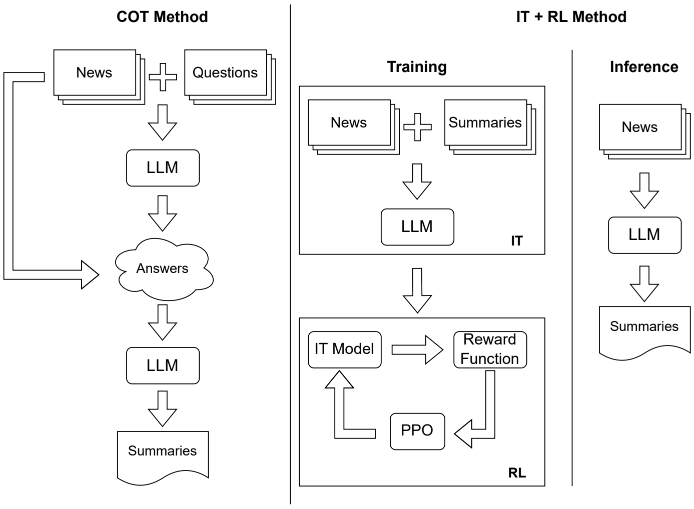

Figure 2: A diagram of the three applied methods

3.1 Chain of Thoughts

LLMs exhibit strong few-shot learning capabilities, effectively using a few demonstrations to perform a wide range of downstream tasks through in-context learning (Brown et al., 2020). Chain of thought (COT) prompting further augments the approach via step-by-step reasoning examples instead of standard question-answer pairs (Wei et al., 2022b; Nye et al., 2022). Wang et al. (2023) improved the summarization ability of LLMs by employing guiding questions as prompts to generate step-by-step, by adapting 5W1H (who, what, when, where, why and how) framework to represent semantic elements of news events, the answers to these key questions are considered to result in summaries with more fine-grained elements. Therefore, following the SumCOT (Wang et al., 2023) approach, we first employ the same set of questions to prompt LLM to generate answers that identify elements such as entities, dates, and events. Moreover, given that the summaries for this task need to be concise, we adjust the questions to align more closely with 5W rules, as shown in Table 1. These answers, along with the corresponding news articles and questions, are then used as input to generate the final summary.

SumCOT
What are the important entities in this document? What are the important dates in this document? What events are happening in this document? What is the result of these events?
5WCOT
Who is involved? (Identify all key entities.) What happened? (Describe the main event or action.) Where did it occur? (Provide the location or setting.) When did it take place? (Specify the date and time, if applicable.) Why did it happen? (Explain the causes, reasons, or purposes behind the event.)

Table 1: Guiding Questions for COT Method

3.2 Instruction Tuning

The above method enables task adaption for LLMs without updating any parameters. However, fine-tuning can be a more effective method to align with desired downstream tasks when the examples from the target domain are available (Ouyang et al., 2022; Taori et al., 2023). Specifically, instruction tuning(IT) is the process of fine-tuning LLMs with instruction-response pairs that use labeled data to improve performance (Wei et al., 2022a). Since full-model fine-tuning requires significant computational resources, parameter-efficient fine-tuning (PEFT) has been introduced, which allows for training only on a small set of additional parameters (Houlsby et al., 2019). Therefore, in this work, we use the QLoRA (Dettmers et al., 2023) method for supervised instruction tuning on the given query-answer pairs using the labeled dataset, which allows for the fine-tuning of a quantized 4-bit model with low-rank adapter weights (Hu et al., 2022). To construct the input prompt for training, we follow the provided instruction in the dataset (Xie et al., 2024), as detailed in Table 2, where the '{text}' and '{answer}' denote the corresponding fields in the dataset. During the inference phase, we exclude the content after 'Answer:[/INST]' to prompt the model to generate summaries.

Training Prompt Template
<s>[INST]You are given a text that consists of multiple sentences. Your task is to perform abstractive summarization on this text. Use your understanding of the content to express the main ideas and crucial details in a shorter, coherent, and natural sounding text. Text:{text} Answer:[/INST] {answer}</s>

Table 2: Template for IT Method

3.3 Reinforcement Learning

The IT method can improve performance (Wei et al., 2022a), however, there's still room for improvement using the reinforcement learning. Lambert et al. (2022) proposed training a language model using proximal policy optimaization (PPO) to further align the model human feedback. Recently, approaches that adapt the final result as the reward signal as outcome supervision has been to solve math problems (Lightman et al., 2024). Inspired by these works, we further investigate the outcome supervision for this task by using a combination of final performance metrics as the reward function to provide reward signals in PPO training. We construct the summary-level reward S by averaging the ROUGE-1/2/L scores and BERTScore, as detailed in Equation 1. Moreover, we incorporate a length penalty L, derived from the BLEU score method (Papineni et al., 2002), to constrain the length of the generated summary relative to the reference summary. As shown in Equation 2, c and r represent the word counts of the generated candidate summary and the reference summary, respectively, with words separated by blank spaces.

$$\begin{aligned} S = L \times \Bigg[& \frac{\text{ROUGE-1} + \text{ROUGE-2}}{4} + \\ & + \frac{\text{ROUGE-L} + \text{BERTScore}}{4} \Bigg] \end{aligned} \tag{1}$$

$$L = \begin{cases} e^{(1-c/r)} & \text{if } c > r, \\ e^{(1-r/c)} & \text{if } c \leq r. \\ 0 & \text{if } c = 0 \end{cases} \tag{2}$$

4 Experimental Setup

Data Preprocessing We split the dataset into two subsets for model training and validation, allocating 80% training and 20% for validation via random seed.

Baseline Journalists commonly address the 5W questions within the first few sentences of an article to highlight the core event (Hamborg et al., 2018). Hence, we use the first sentence of each news article as a summary to serve as our baseline for analysis (First Sentence).

Implementation For the few-shot method, following the SumCOT approach, we use the GPT-3.5 model through the OpenAI API[1]. For instruction fine-tuning, we employ 4bit quantized Llama3-8b model (AI@Meta, 2024)[2], using the unsloth library[3]. This model training is conducted using a rank and alpha of 16 across all applicable modules, with a learning rate of 2e-4. As for the final submission, we choose the fine-tuned checkpoint with the highest performance on the validation set to generate the results, which is trained for 600 steps. For the reinforcement learning approach, we employ the PPOTrainer from the trl (von Werra et al., 2020) library based on the best fine-tuned checkpoint with the learning rate of 5e-6. Since the reward continues to decrease throughout the training process, we only report the model performance after 200 steps.

Evaluation In this shared task, We evaluate the quality of generated summaries through unigram (ROUGE-1) and bigram (ROUGE-2) overlap as well as the longest common subsequence (ROUGE-L) comparison to reference summaries (Lin, 2004). Besides using the n-gram based metrics, BERTScore (Zhang et al., 2020) is also employed, which computes the cosine similarity between their textual embeddings from a BERT-based model. Specifically, we use the evaluate library[4] to calculate the performance scores, and use the multilingual BERT model[5] for BERTScore F1 measurement.

5 Results

In this section, we compare the results of the three different approaches for generating financial abstractive summaries. In particular, we investigate the following research question:

- RQ: How effective are the three different methods in adapting LLMs for abstractive summarization in the financial domain?

To answer this question, we evaluate the three methods by comparing the generated summaries

[1] https://platform.openai.com/docs/models/gpt-3-5-turbo

[2] https://huggingface.co/unsloth/llama-3-8b-bnb-4bit

[3] https://github.com/unslothai/unsloth

[4] https://github.com/huggingface/evaluate

[5] https://huggingface.co/google-bert/bert-base-multilingual-cased

Table 3: Overall Performance

Model	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
First Sentence	0.3443	0.1808	0.2872	0.6992
SumCOT	0.3002	0.1453	0.2397	0.69
5WCOT	0.3082	0.1511	0.2439	0.6923
IT	0.5348	0.358	0.4924	0.8074
RL	0.4944	0.3294	0.4577	0.7906

Table 4: Our Submission on Leaderboard

Metrics	Performance
ROUGE-1	0.5346
ROUGE-2	0.3581
ROUGE-L	0.4922
BERTScore	0.9117
BARTScore	-3.4076

against reference summaries using evaluation metrics such as ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore, with higher scores indicating better performance. In addition to LLM-based models, we have included a baseline denoted First Sentence. As we can see from Table 3, this model establishes foundational performance with a ROUGE-1 score of 0.3443, ROUGE-2 score of 0.1808, ROUGE-L score of 0.2872, and a BERTScore of 0.6992.

Chain-of-Thought Techniques: In few-shot scenarios, both SumCOT and 5WCOT show comparable performance, with 5WCOT slightly improving overall performance when refined guiding questions are used. However, in terms of ROUGE scores, COT methods perform worse than the baseline, with the highest ROUGE-1 score achieved by a COT method being only 0.3082, the highest ROUGE-2 score being oly 0.1511, the best ROUGE-L score being 0.2439 and the best BERTScore only reaching 0.6923.

Instruction tuning (IT): When using this approach, we observe a considerable improvement over the baseline (between 15% and 99% improvements, depending on the metric), chain-of-thought and reinforcement learning methods, achieving the highest performance scores. Therefore, we have submitted this result as our submission to the shared task, with the detailed performance on the leaderboard shown in Table 4.

Reinforcement learning (RL): Finally, reinforcement learning achieves a notorious improvement over the First Sentence baseline and the chain-of-thought approaches. However, it lowers the performance of the best fine-tuned checkpoint (IT) in all the studied metrics (around 7-8% in the case of ROUGE metrics, 2.1% for BERTScore). This suggests that employing standard metrics with reference summaries as reward signals may not effectively guide the model toward developing better strategies for generating financial news summaries.

Conclusions: Overall, when assessing abstractive summarization with headlines as the gold standard, it is clear that using the first sentence as a summary forms a strong baseline. Additionally, instruction tuning is essential to ensure that the model's output aligns with the desired summaries.

6 Conclusions

This work mainly explores the application of COT, IT and PPO method to adapting the LLM for financial abstractive summarization task. Surprisingly, the IT method surpasses both COT and PPO methods, achieved the highest performance and the 1st rank in this shared task. Although the other two approaches failed at this task, their results also indicate that relying only on the standard performance metrics based on a single reference summary to evaluate the quality of the LLM-generated summary may be insufficient and may not provide a useful signal for the LLM to learn more effective summarization strategies. Additionally, the brevity of headline contents may limit the evaluation in terms of informativeness and user interest. The effectiveness of automatic metrics is closely dependent on the quality of reference summaries and the preferences of different annotators can vary when evaluating the same summary (Zhang et al., 2024). Therefore, particularly in the domain of finance, it is beneficial to identify the target consumer and their preferences. For example, previous tracks on temporal and crisis summarization (Aslam et al., 2014; McCreadie and Buntain, 2023) evaluated performance by assessing the coverage of information and the similarity to the user query. Similarly, Böhm et al. (2019); Lambert et al. (2022) suggest constructing the reward function directly from human ratings instead of the existing metrics. Overall, for the future direction, we would employ human-in-loop approaches that evaluate summaries based on the aspects that are important to the target user. By integrating human feedback into the evaluation process, it becomes more possible that the summaries capture essential topics while addressing the specific interests of the financial domain.

References

AI@Meta. 2024. Llama 3 model card.

Javed A. Aslam, Fernando Diaz, Matthew Ekstrand-Abueg, Richard McCreadie, Virgil Pavlu, and Tetsuya Sakai. 2014. TREC 2014 Temporal Summarization Track Overview. In *23rd Text REtrieval Conference (TREC 2014)*, Gaithersburg, Maryland, USA.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Florian Böhm, Yang Gao, Christian M. Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. 2019. Better rewards yield better summaries: Learning to summarise without references. In *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*, pages 3110–3120, Hong Kong, China. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, pages 1877–1901, Virtual Event. Curran Associates, Inc.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. In *37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, New Orleans, LA, USA.

Felix Hamborg, Soeren Lachnit, Moritz Schubotz, Thomas Hepp, and Bela Gipp. 2018. Giveme5W: Main Event Retrieval from News Articles by Extraction of the Five Journalistic W Questions. In *iConference 2018*, pages 356–366, Sheffield, United Kingdom.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *36th International Conference on Machine Learning (ICML 2019)*, pages 2790–2799. PMLR.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *10th International Conference on Learning Representations (ICLR 2022)*, Virtual Event.

Nathan Lambert, Louis Castricato, Leandro von Werra, and Alex Havrilla. 2022. Illustrating Reinforcement Learning from Human Feedback (RLHF). *Hugging Face Blog*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's Verify Step by Step. In *12th International Conference on Learning Representations (ICLR 2024)*, Vienna, Austria.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Richard McCreadie and Cody Buntain. 2023. Crisisfacts: Buidling and evaluating crisis timelines. In *20th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2023)*, pages 320–339, Omaha, NE, USA.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. 2022. Show your work: Scratchpads for intermediate computation with language models. In *Deep Learning for Code Workshop (DL4C 2022) at the 10th International Conference on Learning Representations (ICLR 2022)*, Virtual Event.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, New Orleans, LA, USA.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *40th Annual Meeting on Association for Computational Linguistics (ACL 2002)*, page 311–318, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

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`.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. In *61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8640–8665, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *10th International Conference on Learning Representations (ICLR 2022)*, Virtual Event.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *36th Annual Conference on Neural Information Processing Systems (NeurIPS 2022)*, pages 24824–24837, New Orleans, LA, USA. Curran Associates, Inc.

Adhika Pramita Widyassari, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, Affandy Affandy, and De Rosal Ignatius Moses Setiadi. 2022. Review of automatic text summarization techniques & methods. *Journal of King Saud University - Computer and Information Sciences*, 34(4):1029–1046.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2024. Pixiu: a large language model, instruction data and evaluation benchmark for finance. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations (ICLR 2020)*, Virtual Event.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Zhihan Zhou, Liqian Ma, and Han Liu. 2021. Trade the Event: Corporate Events Detection for News-Based Event-Driven Trading. In *Findings of the Association for Computational Linguistics (ACL-IJCNLP 2021)*, pages 2114–2124, Virtual Event. Association for Computational Linguistics.