Seminar
NLP for Finance
Herbstsemester 2024

Automating Financial Text Summarization with LLMs

Verfasserin/Verfasser: Xiaojing Zhang
Matrikel-Nr: 22-741-763

Dozenten: Martin Volk
Institut fur Computerlinguistik

Abgabedatum: 20.01.2025

Automating Financial Text Summarization with LLMs

Xiaojing Zhang (22-741-763)

January 20, 2025

1. Introduction

Financial markets are characterized by complexity and vast amounts of data that are generated daily. Such data, consisting of both textual and numerical information created by companies at different times during their financial year, comes in various forms such as annual financial reports, quarterly reports, preliminary earnings announcements, conference calls and press releases as well as conference calls transcripts, media articles and online social media (El-Haj et al., 2018). These documents are crucial for decision-makers, investors and analysts to gain an insight into a company's performance, market trends, and economic health and to make informed decisions.

However, financial documents tend to be very long, averaging at about 180 pages (Leidner, 2019). Moreover, most of such documents are filled with domain-specific jargons, and contain figures, tables, numbers, and complex textual narratives. Given this, professionals find it challenging to manually read them and efficiently process these documents in order to extract useful information. Even experienced analysts are prone to errors in the face of such length and complexity, which can be costly in an environment where accuracy and time are paramount.

Manual summarization of financial texts, though effective, is time-consuming, expensive, and prone to human errors. As a result, there is a growing need for finding efficient, automated ways using technology to condense these texts into concise and coherent summaries without losing important information. This is where automatic summarization comes into play, as individuals in the field would be able to go through summaries of various reports and other financial documents, and to make informed decisions (Khant and Singh Mehta, 2018). Automatic summarization not only saves time but also ensure faster and more accurate decision-making, risk management, and market analysis, by providing quick access to key insights.

Recently, the advent of Large Language Models (LLMs) (Brown et al., 2020) has offered new possibilities for automating financial text summarization. LLMs, such as GPT and LLaMA, have showed remarkable capabilities in specialized domains across mathematics, coding, medicine, law, and finance (Bubeck et al., 2023). Recent studies in the financial domain (Li et al., 2023; Xie et al., 2023; Lopez-Lira and Tang, 2023) have demonstrated the significant potentials of LLMs in financial text analysis and generation tasks. Summaries generated by state-of-the-art LLMs are comparable to those created by humans (Zhang et al., 2024).

Nevertheless, their application in domain-specific tasks like financial text summarization remains underexplored. Compared to other texts, financial documents pose unique challenges for LLMs, due to their domain-specific terminology, numerical data, and the need for precise, performance-related summaries. Applying

LLMs for these tasks requires not only a lot of annotated domain-specific data but also innovative techniques to achieve optimal performance.

This project is inspired by Task 2 of the FinLLM Challenge 2024, which focuses on financial text summarization using LLMs. The shared task aims to evaluate the ability of LLMs to generate coherent summaries of financial news articles. Although the shared task dataset offers a strong foundation, this project shifts to use earnings call transcripts, as they are more directly relevant to the finance industry. Earnings calls offer valuable insights into a company's financial performance and strategic decisions.

Currently, studies in this field focus on summarizing financial news articles or Environmental, Social and Governance (ESG) reports. Smaller LLMs, especially those with less than 7 billion parameters, have been less explored. This project attempts to fill this gap by exploring the use of smaller and more efficien LLMs to summarize earnings call transcripts.

To explore the potential and challenges of LLMs, this project focuses on the following research questions:
1. How do LLMs perform on summarizing earnings call transcripts?
2. Can smaller LLMs achieve similar performance in financial text summarization compared to larger LLMs?
3. How do parameter-efficient fine-tuning (PEFT) methods like QLoRA perform compared to other approaches?

This project uses two smaller, pre-trained LLMs, microsoft/Phi-3.5-mini-instruct and meta-llama/Llama-3.2-1B-Instruct, which support 4-bit quantization and a 128,000 context length. These models are fine-tuned using Quantized Low-Rank Adaptation (QLoRA), a parameter-efficient method for effective adaptation with minimal computation resources. The dataset consists of 2,424 earnings call transcripts with gold summaries, but due to memory constraints, the dataset is reduced to 125 and 227 samples for the two models, respectively. The fine-tuned models are evaluated using ROUGE and BERTScore metrics, which measure lexical overlap and semantic similarity between generated and gold summaries.

The findings of this project highlight the potential of LLMs in automating financial text summarization while addressing challenges related to dataset size, text length, and resource constraints. By using parameter-efficient fine-tuning techniques, this project shows that smaller models can achieve comparable performance to larger models yet more efficient in terms of parameters and data.

The remainder of this paper is structured as follows: Section 2 provides background information on financial text summarization, LLMs, and related work. Section 3 describes the experiments, including the dataset, models, methodology, and evaluation. Section 4 discusses the results, challenges, and how they compare to prior work. Section 5 concludes with a summary of findings and open research issues.

2. Background

This section provides an overview of financial text summarization, the role of LLMs in finance, and related work. It consists of three subsections: (1) Financial Text Summarization, (2) LLMs in Finance, and (3) Related Work.

2.1 Financial Text Summarization

This subsection explores text summarization techniques, their application to financial documents, and the challenges in this domain.

2.1.1 Text Summarization

Text summarization can be broadly classified into extractive (Saini et al., 2018) and abstractive (Gui et al., 2019) summarization methods.

Extractive summarization involves extracting key sentences or segments directly from the source documents and putting them together to form a summary. Traditional techniques include graph-based methods (Erkan and Radev, 2004; Litvak et al., 2010) and optimization methods such as integer linear programming (Galanis et al., 2012). Recently, transformer-based models (Liu and Lapata, 2019) and methods using summary-level representations (Zhong et al., 2020) have emerged.

Abstractive summarization generates concise text that preserves the essence of the given document (Dong et al., 2021). The dominating techniques are based on sequence-to-sequence (seq2seq) architectures (Fu et al., 2020; Jangra et al., 2020; Krizhevsky et al., 2017). Advances in transformer-based models, including those fine-tuned for specific tasks, have demonstrated great success.

2.1.2 Financial Text Summarization

Summarizing financial documents plays an important role in helping decision-makers, investors and analysts quickly understand a company's financial conditions. Documents such as annual reports, earnings calls, and regulatory filings contain important information about a company's performance, risks and future outlook. Some studies have showed the importance of financial reports, for example, Stanton and Stanton (2002) discuss the critical role of financial reports in company operations and Ghazali and Annum (2010) highlight the value of annual reports for investors.

Earnings calls, which are teleconferences or webcasts hosted by publicly listed companies to discuss their quarterly or annual earnings reports, are valuable (Bowen et al., 2002; Keith and Stent, 2019). These calls provide key insights into a company's financial performance, thus summaries of such calls are essential for financial analysts and investors.

However, earnings calls are usually long and unstructured documents (Mukherjee et al., 2022). Thus, it requires a lot of time and efforts to quickly summarize relevant information, even for trained analysts. Automating the summarization of these documents can significantly reduce the time and efforts for processing and analysis.

Prior studies have focused on summarizing financial reports (Suarez et al., 2020; Abdaljalil and Bouamor, 2021; Orzhenovskii, 2021) and financial news summarization (Passali et al., 2021).

Automatic extractive summarization has been widely explored in the finance domain for summarizing annual reports and news articles. Leidner (2019) exploses different NLP techniques used in text summarization. He discusses the different methods used in summarizing financial texts by reviewing heuristic, statistical, and neural models.

Xu et al. (2013) presents a graph-based model for multi-tweet extractive summarization, making use of the Named Entities and frequency of topics talked about in tweets. Fillippova et al. (2009) propose an extractive summarization system for financial news articles. Their model takes company names as input and retrieves financial news articles about the companies on Yahoo News. They then rank sentences by company relevance.

Statistical features with heuristics have also been used to summarize financial texts (Cardinaels et al., 2019), which produce summaries with reduced positive bias. The financial narrative summarization task (El-Haj et al., 2019) of the Multiling 2019 Workshop (Giannakopoulos, 2019) deals with generating structured summaries for financial narrative disclosures. Zheng et al. (2020) introduce a system that splits annual reports into sections corresponding to the Table of Contents and then applies a BERT-based classifier to include relevant sections in the summary.

Singh (2020) proposes a hybid approach using pointer networks for extractive summarization, followed by T5-based abstractive summarization to paraphrase extracted sentences into a concise summary. This approach combines the strengths of abstractive and extractive summarization methods to general concise and informative summaries.

2.1.3 Challenges in Financial Text Summarization

Summarizing financial documents, especially earnings call transcripts, poses some challenges:

First, financial documents are mostly long and complex, and contain a lot of domain-specific jargons. This requires models to efficiently process and extract key information from these long documents.

Second, the scarcity of large-scale annotated datasets for financial documents (Mukherjee et al., 2022) limits model training and evaluation.

Third, financial documents often include both textual and numerical data, posing a challenge for summarization models to convey both types of data effectively.

2.1.4 Relevance to This Project

This project focuses on summarizing earnings call transcripts, which is an underexplored area in financial text summarization. This work aims to explore abstractive summarization techniques, using state-of-the-art LLMs, so as to address the challenges posed by long, complex and domain-specific texts.

2.2 LLMs in Finance

Recent advancements in LLMs, such as GPT, BERT and LLaMA, have reshaped natural language processing (NLP), achieving cutting-edge performance on a variety of tasks, including text classification, text generation/prediction, sentiment analysis, and summarization. LLMs have also demonstrated impressive capabilities in the finance domain (Bubeck et al., 2023; Li et al., 2023).

LLMs have been applied to tasks such as financial sentiment analysis, argument relation classification, ESG issue identification, news headline classification, text summarization, and financial report generation, etc. (Xie et al., 2024a).

Despite remarkable performance in finance, LLMs face challenges in financial contexts. Financial documents often contain domain-specific terminology, numerical data, and complex narratives that are not common in general-purpose training corpora. To address these issues, domain adaptation through fine-tuning on financial-specific data is necessary.

2.2.1 Parameter-Efficient Fine-Tuning Methods

Though effective, full fine-tuning of LLMs, which retrains all model parameters, on domain-specific data can be computationally expensive, especially when using very large language models. Moreover, it also restricts people without GPU resources to use LLMs for this purpose. To remedy this, parameter-efficient fine-tuning methods have been developed to reduce the computational costs and memory requirements without sacrificing performance. Such methods allow for training only a small set of parameters while freezing the rest.

LoRA (Low-Rank Adaptation) (Hu et al., 2021) reduces the number of parameters that need to be fine-tuned for domain-specific tasks, by freezing the pre-trained model parameters and adds trainable rank decomposition matrices to each layer of the model architecture. LoRA has been shown to achieve similar or better performance compared to full fine-tuning, while significantly reducing the number of trainable parameters.

QLoRA (Quantized LoRA) (Dettmers et al., 2023) builds on LoRA by quantization to further reduce the memory usage and computational costs. For example, QLoRA uses 4-bit precision to represent the model's parameters, significantly reducing computational costs and lowering memory requirements, especially when working with large datasets and limited computational resources. This makes it suitable for large-scale financial text summarization tasks, where LLMs are often needed, but resource constraints may limit people to work with them.

These methods enable the fine-tuning of smaller models for domain-specific tasks, such as financial summarization, without sacrificing accuracy.

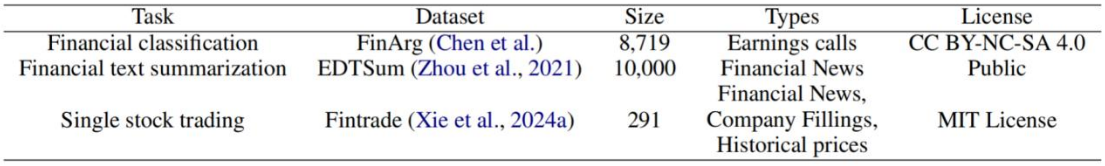
2.2.2 Relevance to This Project

This project uses QLoRA to fine-tune two smaller LLMs, with 1B and 3.8B parameters, respectively, for earnings call summarization. By leveraging parameter-efficient fine-tuning, this project demonstrates the feasibility of using smaller models to achieve competitive results in domain-specific tasks. In this way, the computational costs are reduced significantly and LLMs become more accessible to individuals with limited resources.

2.3 Related Work

Recent studies (Xie et al., 2024b; Lopez Lira and Tang, 2023) have highlighted the remarkable success of LLMs in finance, particularly for financial text analysis and prediction tasks. These models can enhance efficiency and accuracy of predictive models. Yet, their capabilities of comprehensive analysis and decision-making in the finance domain still remain unexplored.

To evaluate the capabilities of LLMs in finance, Xie et al., (2024c) introduce an LLMs-based financial shared task featured at IJCAI-2024 Initiative, FinLLMs Challenges. The Challenges include three subtasks: financial classification, financial text summarization, and single stock trading. Table 1 below shows the datasets, dataset sizes, types, and licenses of these three tasks.

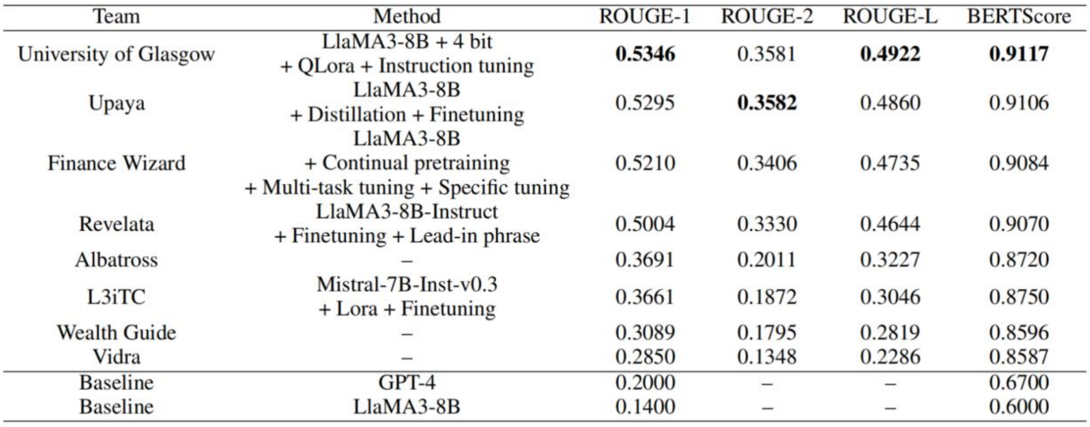


Task	Dataset	Size	Types	License
Financial classification	FinArg (Chen et al.)	8,719	Earnings calls	CC BY-NC-SA 4.0
Financial text summarization	EDTSum (Zhou et al., 2021)	10,000	Financial News	Public
Single stock trading	Fintrade (Xie et al., 2024a)	291	Financial News, Company Fillings, Historical prices	MIT License

Table 1: Summary of the tasks and datasets in FinNLP-AgentScen-2024

Table 1: Summary of the tasks and datasets in FinLLMs Challenges (Xie et al., 2024c)

Task 2 of the FinLLM Challenges focuses on smmarization of financial news articles, evaluating models using the EDT corpus (Zhou et al., 2021) and metrics like ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020). Table 2 shows the evaluation results of Task 2. The baseline models are GPT-4 and LLaMA3-8B using zero-shot prompting from (Xie et al., 2024a).



Team	Method	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
University of Glasgow	LlaMA3-8B + 4 bit + QLora + Instruction tuning	**0.5346**	0.3581	**0.4922**	**0.9117**
Upaya	LlaMA3-8B + Distillation + Finetuning	0.5295	**0.3582**	0.4860	0.9106
Finance Wizard	LlaMA3-8B + Continual pretraining + Multi-task tuning + Specific tuning	0.5210	0.3406	0.4735	0.9084
Revelata	LlaMA3-8B-Instruct + Finetuning + Lead-in phrase	0.5004	0.3330	0.4644	0.9070
Albatross	–	0.3691	0.2011	0.3227	0.8720
L3iTC	Mistral-7B-Inst-v0.3 + Lora + Finetuning	0.3661	0.1872	0.3046	0.8750
Wealth Guide	–	0.3089	0.1795	0.2819	0.8596
Vidra	–	0.2850	0.1348	0.2286	0.8587
Baseline	GPT-4	0.2000	–	–	0.6700
Baseline	LlaMA3-8B	0.1400	–	–	0.6000

Table 3: Evaluation results of Task 2 - Financial Text Summarization.

Table 2: Results of Task 2 - Financial Text Summarization (Xie et al., 2024c)

University of Glasgow (Guo et al., 2024) experimented with few-shot learning with Chain-of-Thought prompting, fine-tuning, and reinforcement learning to adapt their models to abstract news into concise and coherent summaries. Their fine-tuned model ranked first among the eight participating teams in this shared task. They fine-tuned a 4-bit quantized Llama3-8b model (AI@Meta, 2024), employing instruction tuning, parameter-efficient fine-tuning, and QLoRA techniques.

Upaya (Jindal et al., 2024) used distillation-based fine-tuning for financial tasks, using a smaller model to mimic the behavior of a larger, more complex model. Their model ranked the second in the leaderboard. This approach demonstrates the trade-offs between model size and performance, highlighting that smaller models can achieve competitive results with efficient training methodologies.

Finance Wizard (Lee and Lay-Ki, 2024) focused on continual pretraining of financial data to improve the model's performance in the finance domain. They first used continual pre-training on Llama3-8B foundation model, then multi-task tuning to the finance domain, and finally fine-tuning for specific tasks. By exposing the model to continuous financial texts, they were able to enhance the model's understanding of domain-specific terminology and context.

Revelata (Kawamura et al., 2024) first used a set of prompts and then fine-tuning Llama3-8B-Instruct on each prompt separately. L3iTC (Pontes et al., 2024) chose to fine-tune Mistral-7B-Inst-v0.3 model with 4-bit quantization and LoRA.

Despite of these advancements, research on summarizing earnings call transcripts remains limited. Most studies focus on finnacial news or ESG reports, and the use of smaller LLMs with fewer than 7 billion parameters for financial summarization is underexplored. This project addresses the these gaps by applying QLoRA to fine-tune smaller, efficient models for summarizing earnings call transcripts.

3. Experiments

This section describes the dataset, methodology, and evaluation metrics used in this project, as well as the results of fine-tuning the microsoft/Phi-3.5-mini-instruct and meta-llama/Llama-3.2-1B-Instruct models for summarization of earnings call transcripts. This project adopts the same fine-tuning approach as the winning team of FinLLM Challenges (Guo et al., 2024).

3.1 **Dataset**

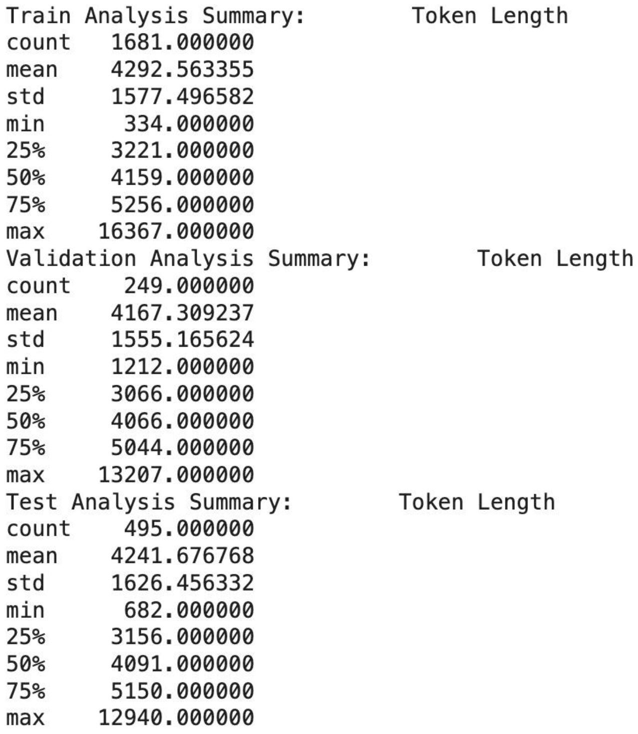
The dataset consists of 2,424 earnings call transcripts, each with a telegraph-style bullet-point gold summary. Earnings calls provide insights into a company's performance, strategic direction, and market outlook. These transcripts are typically lengthy with numerical and textual data, posing challenges for manual summarization. The gold summaries are performance-focused, highlighting key financial metrics and strategic developments.

The training set has 1,681 documents, with an average token length of 4,292 tokens. The length distribution ranges from 334 tokens (minimum) to 16,367 tokens (maximum). The 25th percentile is 3,221 tokens, the median is 4,159 tokens, and the 75th percentile is 5,256 tokens. This indicates that most training documents are relatively long and require extensive summarization.

The validation set contains 249 documents with an average length of 4,167 tokens, slightly shorter than the training set. Token lengths vary between 1,212 tokens (minimum) and 13,207 tokens (maximum). The median token length is 4,066, with the 25th and 75th percentiles at 3,066 and 5,044 tokens, respectively.

The test set includes 495 documents with an average token length of 4,242 tokens, similar to the training set. Token lengths range from 682 tokens (minimum) to 12,940 tokens (maximum). The median length is 4,091 tokens, while the 25th and 75th percentiles are 3,156 and 5,150 tokens, respectively.

Table 3 and Figure 1 show the token length distribution of the dataset. We can see the dataset is complex, withsignificantly long and dense documents, particularly in the training and test sets. This reinforces the need for efficient summarization techniques to condense detailed financial information into concise, performance-focused summaries.



```
Train Analysis Summary:        Token Length
count    1681.000000
mean     4292.563355
std      1577.496582
min       334.000000
25%      3221.000000
50%      4159.000000
75%      5256.000000
max     16367.000000
Validation Analysis Summary:        Token Length
count     249.000000
mean     4167.309237
std      1555.165624
min      1212.000000
25%      3066.000000
50%      4066.000000
75%      5044.000000
max     13207.000000
Test Analysis Summary:        Token Length
count     495.000000
mean     4241.676768
std      1626.456332
min       682.000000
25%      3156.000000
50%      4091.000000
75%      5150.000000
max     12940.000000
```

Table 3: Token Length Statistics for Training, Validation, and Test Sets

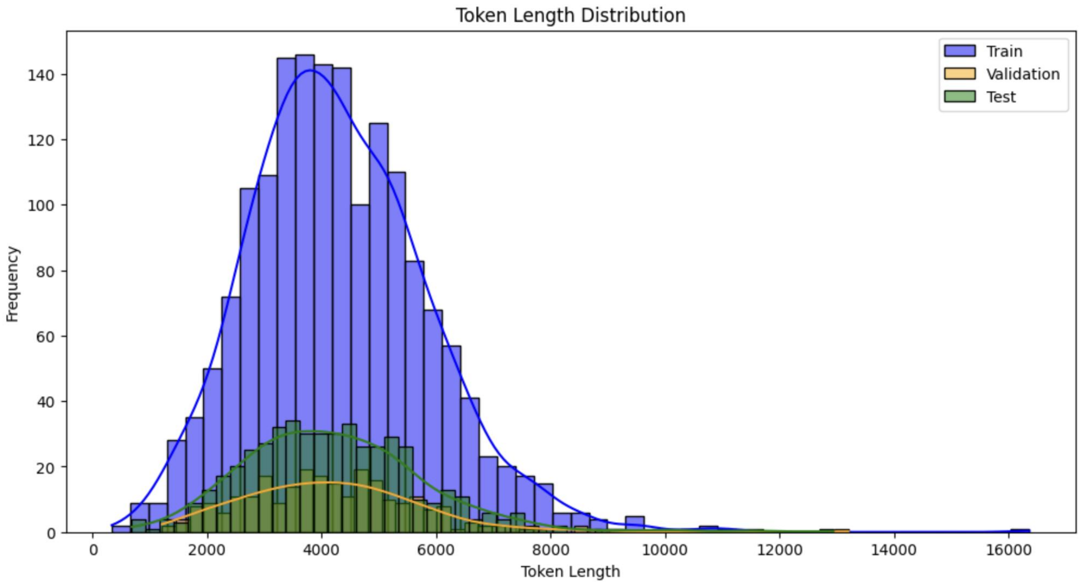
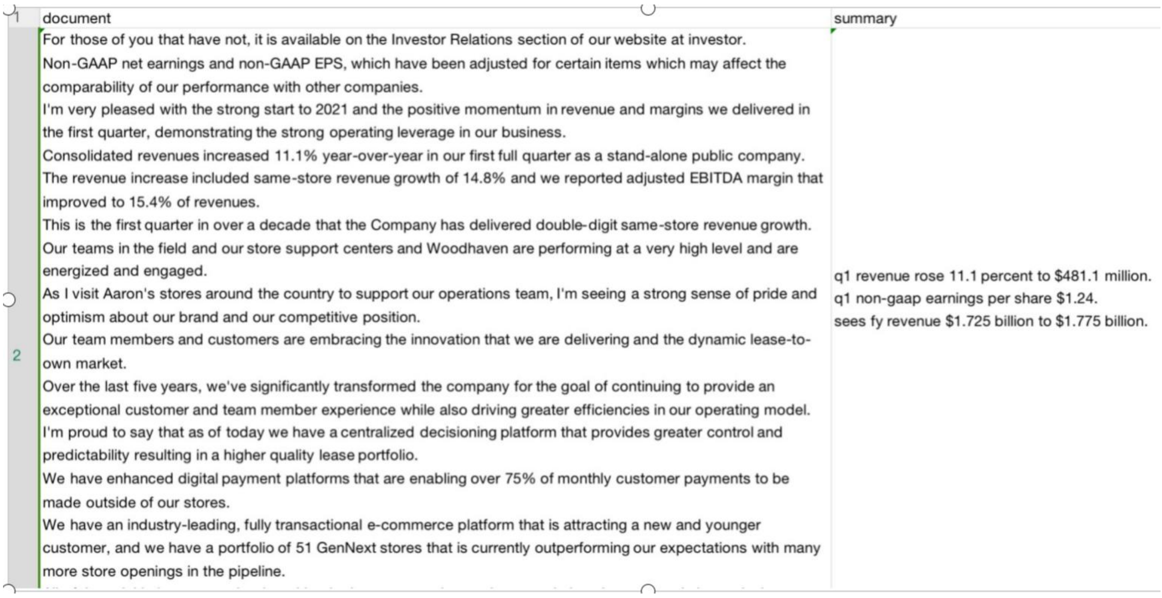


Figure 1: Token Length Distribution

Table 4 shows an example from the dataset, showing the structure of a document and its corresponding gold summary. The document contains 1775 words, with detailed financial and operational updates, such as revenue growth (11.1% year-over-year to $481.1 million), non-GAAP earnings per share ($1.24), and future revenue projections ($1.725 billion to $1.775 billion). The gold summary, only 22 words, extracts these key insights into a concise, performance-focused format, highlighting crucial metrics and company performance. This example illustrates that the dataset emphasizes summarizing dense financial information into very telegraph-style bulletin-point insights, which is essential for decision-makers. The summarization ratio of this example is 1.2%, calculated by comparing the length of the document to its summary. This reflects the challenge of condensing complex and long financial texts into concise and precise summaries while keeping the core information.



1	document	summary
2	For those of you that have not, it is available on the Investor Relations section of our website at investor. Non-GAAP net earnings and non-GAAP EPS, which have been adjusted for certain items which may affect the comparability of our performance with other companies. I'm very pleased with the strong start to 2021 and the positive momentum in revenue and margins we delivered in the first quarter, demonstrating the strong operating leverage in our business. Consolidated revenues increased 11.1% year-over-year in our first full quarter as a stand-alone public company. The revenue increase included same-store revenue growth of 14.8% and we reported adjusted EBITDA margin that improved to 15.4% of revenues. This is the first quarter in over a decade that the Company has delivered double-digit same-store revenue growth. Our teams in the field and our store support centers and Woodhaven are performing at a very high level and are energized and engaged. As I visit Aaron's stores around the country to support our operations team, I'm seeing a strong sense of pride and optimism about our brand and our competitive position. Our team members and customers are embracing the innovation that we are delivering and the dynamic lease-to-own market. Over the last five years, we've significantly transformed the company for the goal of continuing to provide an exceptional customer and team member experience while also driving greater efficiencies in our operating model. I'm proud to say that as of today we have a centralized decisioning platform that provides greater control and predictability resulting in a higher quality lease portfolio. We have enhanced digital payment platforms that are enabling over 75% of monthly customer payments to be made outside of our stores. We have an industry-leading, fully transactional e-commerce platform that is attracting a new and younger customer, and we have a portfolio of 51 GenNext stores that is currently outperforming our expectations with many more store openings in the pipeline.	q1 revenue rose 11.1 percent to $481.1 million. q1 non-gaap earnings per share $1.24. sees fy revenue $1.725 billion to $1.775 billion.

Table 4: An Example from the Dataset

3.2 Preprocessing

Due to memory constraints, transcripts exceeding 2,048 tokens were excluded. Truncation was not used to keep more data because some important information is at the end of the document. Filtering documents longer thann 2,048 can retain the whole text so that the model will get full context for summarization. This preprocessing step reduced the dataset to 125 and 227 samples for the two models, respectively, possibly due to different tokenizers used by the two models.

For microsoft/Phi-3.5-mini-instruct model, 125 samples are kept after the preprocessing, distributed as follows:

- Training set: 85 samples
- Validation set: 16 samples
- Test set: 24 samples

For meta-llama/Llama-3.2-1B-Instruct model, there are 227 samples after filtering, distributed as follows:

- Training set: 149 samples
- Validation set: 29 samples
- Test set: 49 samples

Although the reduced dataset may limit the diversity of training data, it ensures that fine-tuning can be conducted efficiently within the available computational resources.

3.2 Models

The microsoft/Phi-3.5-mini-instruct and meta-llama/Llama-3.2-1B-Instruct models were selected for their small size (with 3.8B and 1B parameters, respectively), efficient memory usage, and extended context length (128k tokens). Both models are available for 4-bit quantization, which significantly reduces memory requirements while maintaining performance. These features make the two models particularly suitable for summarizing long financial documents like earnings call transcripts within resource constraints.

3.3 QLoRA Fine-Tuning

Fine-tuning was applied using QLoRA technique, which combines 4-bit quantization with low-rank adaptation. This approach significantly reduces the number of parameters for training, thus making it very efficient. The following setup was used:

- Rank and Alpha was set to 8 to balance efficiency and model capacity.
- Learning Rate was set to 2e-4, with a weight decay rate of 0.01, for stable training.
- Epochs were set to 10 to ensure sufficient learning without overfitting.
- Early Stopping was set to 3, i.e., the training will be terminated when the validation loss does not decrease for 3 consective epochs, to prevent overfitting.

After applying QLoRA, only about 11 million parameters (28.73% of the total parameters) were trainable for microsoft/Phi-3.5-mini-instruct and 1.3 million parameters (10.59% of the total parameters) meta-llama/Llama-3.2-1B-Instruct model. This parameter-efficient setup allows us to fine-tune with minimal computational resources while maintaining competitive performance.

3.3 Prompt Construction

The prompt was structured as an instruction-response pair to guide the model's output. The system instruction defines the model's role as a helpful assistant for financial text summarization, specify task requirements, and guide the model's behavior, and the user provides the document for summarization. The assistant's response begins with the hardcoded phrase "Sure! Here is the summary:".

Here is the prompt used for fine-tuning, which provides the gold summary to the model for learning:

```
messages = [
  {"role": "system", "content": "You are a helpful assistant that summarizes earnings call documents. Your task is to perform abstractive summarization on the given text. Pay special attention to performance-related numbers and key metrics. Summarize the text in a few short and concise sentences, focusing on the most important information. Always start your summary with 'Sure! Here is the summary:\n'."},

  {"role": "user", "content": "Summarise the following text: {input_text}"},

  {"role": "assistant", "content": "Sure! Here is the summary:\n {example['summary']}"}
]
```

The prompt used for inference to evaluate model performance does not provide the gold summary to the model, with the rest part remaining the same as the prompt for fine-tuning to maintain consistency.

3.4 Evaluation Metrics

ROUGE and BERTScore were used to evaluate model performance.

ROUGE-1, ROUGE-2, and ROUGE-L measure the overlap of unigrams, bigrams, and the longest common subsequence between generated and Gold summaries, emphasizing lexical overlap and content retention.

BERTScore calculates cosine similarity between the embeddings of generated and Gold summaries, assessing semantic similarity and overall quality.

ROUGE offers a straightforward measure of how well the generated summaries match Gold summaries in terms of word overlap. BERTScore provides a more nuanced evaluation, focusing on semantic similarity, especially valuable for abstractive summarization.

3.5 Results

3.5.1 Training Results

This subsection discusses the training losses, validation losses, and perplexity scores of the two models.

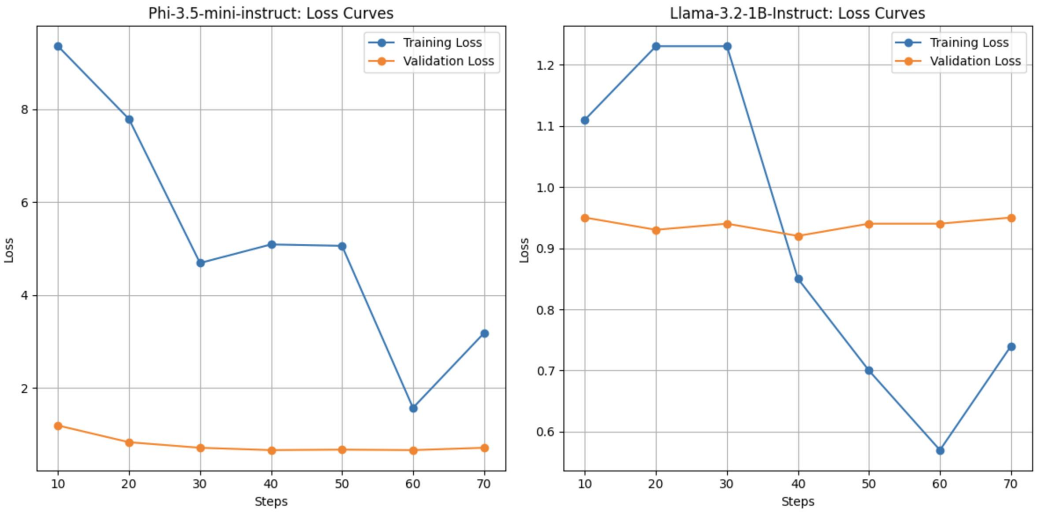


Figure 2: Training and Validation Loss Curves

The training process for both models shows obvious trends in training and validation loss decrease over the steps.

(1) Meta’s Llama-3.2-1B-Instruct

The plot above shows that the training and validation losses decrease steadily during training. At step 70, the training loss is 0.74 and the validation loss is 0.95, indicating smooth convergence without overfitting. Moreover, the validation loss remains consistently close to the training loss, highlighting the model’s stability throughout training. This trend indicates that the model generalizes well to unseen data.

(2) Microsoft’s Phi-3.5-mini-instruct

The Phi-3.5-mini-instruct model shows a more rapid decrease in losses compared to the Llama model. By step 70, the training loss reaches 3.18, while the validation loss stabilizes at 0.71. Although the training loss starts at a much higher value than the

Llama model, the validation loss at the final step suggests that the Phi-3.5-mini-instruct model adapts efficiently to the task.

The perplexity scores computed at the end of training offer an additional insight into the model's fluency and coherence in summarization. Llama-3.2-1B-Instruct achieves a perplexity score of 11.77. This suggests the model generates coherent text, but the higher perplexity compared to Phi-3.5-mini-instruct indicates slightly less fluency for this dataset. Phi-3.5-mini-instruct outperforms in terms of perplexity, with a score of 6.74. This lower perplexity highlights the model's ability to generate more coherent and fluent summaries.

The lower validation loss and perplexity means that the Phi-3.5-mini-instruct model has a stronger generalization capability. However, the Llama-3.2-1B-Instruct model also demonstrates reliable performance with its consistently low training and validation losses.

3.5.2 Model Performance

In this section, we evaluate the performance of the fine-tuned Phi-3.5-mini-instruct and Llama-3.2-1B-Instruct models on the earnings call dataset using ROUGE and BERTScore metrics.

(1) Quantative Metrics

For context, we compare the results of the two models used in this project with models from Task 2 of FinLLM Challenges that evaluated financial news summarization, including the best-performing fine-tuned Llama3-8B-Instruct and zero-shot baselines like GPT-4 and Llama3-8B.

Table 5 summarizes the results across models and datasets in this project and Task 2.

Model	Techniques	Dataset	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
Phi-3.5-mini-instruct	Fine-tuning	Earnings Call - Val	0.15	0.06	0.10	0.83
	Fine-tuning	Earnings Call - Test	0.11	0.04	0.08	0.82
Llama-3.2-1B-Instruct	Fine-tuning	Earnings Call - Val	0.15	0.06	0.10	0.83
	Fine-tuning	Earnings Call - Test	0.11	0.03	0.07	0.82
Universit of Glasgow	Fine-tuning	Financial News	0.53	0.36	0.49	0.91
GPT-4	zero-shot	Financial News	0.20			0.67
Llama3-8B	zero-shot	Financial News	0.14			0.60

Table 5: Results across Models and Datasets

We can see that the fine-tuned Phi-3.5-mini-instruct and Llama-3.2-1B-Instruct models achieved similar performance across both validation and test sets on the earnings call dataset. For example, both models achieved a ROUGE-1 score of 0.15 on the validation set and 0.11 on the test set. And their BERTScores were also the same, with 0.83 (validation) and 0.82 (test).

The low performance results reflect the unique challenges of the earnings call dataset. Texts are very long and require considerable compression, with gold summaries often containing just a few telegraph-style bullet points. Moreover, the gold summaries tend to use domain-specific acronyms, while the fine-tuned models usually use complete words, in a more formal style.

The Llama3-8B model fine-tuned by the University of Glasgow on the financial news dataset achieved much higher performance, with ROUGE-1 and BERTScore values of 0.53 and 0.91, respectively. This large performance gap highlights the relative simplicity of financial news articles compared to earnings calls.

The baseline models (GPT-4 and Llama3-8B) of Task 2 in FinLLM Challenges were evaluated on financial news dataset, with zero-shot prompting (no other techniques). Both models performed significantly worse compared to fine-tuned models, whether on earnings call or financial news dataset. For instance, GPT-4 achieved a ROUGE-1 score of 0.20 and a BERTScore of 0.67. The BERTScore of both baseline models was lower than that of the two smaller models fine-tuned on the earnings call dataset. This demonstrates the importance of fine-tuning. Smaller models fine-tuned for a more difficult task can even perform better than larger ones.

(2) Qualitative Analysis

A qualitative analysis for the two models is also performed on their generated summaries. Several representative examples are selected to highlight their strengths and limitations.

Example 1:
Gold Summary:
second quarter sales of $404 million, down 28% from the prior year reflecting the impact of COVID-19.
q2 adjusted diluted earnings per share from continuing operations of $0.31, which excludes the impact of CARES Act support and other items.
q2 sales $404 million versus Refinitiv IBES estimate of $409.4 million.

Phi-3.5-mini-instruct Generated Summary:
The company reported a 14% increase in sales from $401 million to $455 million, with adjusted diluted earnings per share rising from $0.17 to $0.52.
Sales to commercial customers increased by 52%, while sales to government and defense customers decreased by 17%.
The company announced new business wins and expects overall Q2 performance to be similar to Q1.

Llama-3.2-1B-Instruct Generated Summary:
We had a strong second quarter, with sales decreasing 28% year-over-year to $403.6 million. Our Aviation Services segment grew 6% sequentially, driven by increased volume in our MRO business.
Adjusted diluted earnings per share from continuing operations was $0.31.
The company remains well-positioned to grow its business and drive further margin improvements.

Llama-3.2-1B performs better in capturing the core financial metrics and aligns more closely with the Gold summary, while Phi-3.5-mini-instruct struggles with factual accuracy and relevance.

Example 2:
Gold Summary:
qtrly adjusted earnings per share from continuing operations $2.79.
qtrly revenue $5.90 billion, up 27%.
qtrly same-store used vehicle unit sales increased 28%.
targets selling 1 million combined new and pre-owned vehicles annually.

Phi-3.5-mini-instruct Generated Summary:
The company reported Q1 adjusted earnings per share of $2.79 and revenue of $5.9 billion, up 27% year-over-year.
Same-store vehicle unit sales increased 28%.
The company announced plans to sell 1 million vehicles annually.

Llama-3.2-1B-Instruct Generated Summary:
We reported all-time record quarter results with adjusted earnings per share from continuing operations of $2.79, an increase of 207% compared to last year.
Revenue was $5.90 billion, driven by strong demand for vehicles.
The company remains focused on acquiring vehicles and expects to sell 1 million combined new and pre-owned vehicles annually.

Phi-3.5-mini-instruct aligns more closely with the telegraph style of the gold summary, while Llama-3.2-1B generates a more verbose and explanatory output with inaccuracies.

Llama-3.2-1B-Instruct is more accurate in capturing key financial metrics compared to Phi-3.5-mini-instruct. It provides detailed context, which may be beneficial for broader financial analysis. Phi-3.5-mini-instruct tends to produce more concise summaries that sometimes match the style of the gold summaries.

Both models often include irrelevant or speculative details. Moreover, neither model consistently generates the concise, telegraphic format of the gold summaries. Both models tend to omit important financial metrics.

In conclusion, Llama-3.2-1B-Instruct performs better in factual accuracy and capturing key metrics, while Phi-3.5-mini-instruct generates more concise summaries. However, both models face challenges with style mismatches, over-generation, and occsional errors, which might due to small dataset size (125 and 227 samples, respectively) and model capacity (3.8B and 1B parameters, respectively).

4. Discussion

4.1 Challenges with Evaluation Metrics

The gold summaries in the earnings call dataset are written in a highly compressed telegraph style, often using acronyms (e.g., “qtrly” for “quarterly” and “fy” for “financial year”) and short bullet points. In contrast, the generated summaries are verbose and written in full sentences. This style mismatch leads to low ROUGE scores, as ROUGE relies on word overlap and does not account for differences in phrasing or structure.

However, the BERTScore, which measures semantic similarity using contextual embeddings, shows that the generated summaries effectively capture the meaning of the gold summaries. Both models achieve high BERTScores (0.82–0.83), reflecting their ability to produce semantically accurate outputs despite structural differences.

These results demonstrate the limitations of ROUGE for evaluating tasks with reference and generated summaries differing in style and form. BERTScore becomes a more appropriate metric for this dataset.

4.2 Model Comparison

The two models used in this project perform similarly across both ROUGE and BERTScore metrics. For instance, both models achieve a ROUGE-1 score of 0.11 on the test set and a BERTScore of 0.82.

This finding is noteworthy given the size difference between the two models. Phi-3.5-mini-instruct has 3.8 billion parameters, while Llama-3.2-1B-Instruct has just 1 billion parameters. The comparable performance suggests that smaller models can be highly efficient when fine-tuned effectively, making them a cost-effective choice for domain-specific tasks, especially with limited computational resources.

5. Conclusion

This project has showed that smaller LLMs, i.e., Phi-3.5-mini-instruct and Llama-3.2-1B-Instruct models, can effectively summarize complex financial texts. The findings indicate that these models provide a resource-efficient alternative to larger models without sacrificing much performance.

However, there are several challenges. Due to resource and memory constraints, the dataset was filtered by excluding samples with more than 2,048 tokens, as a result, there were only 125 and 227 samples for the two models, respectively. This limitation affects the model's ability to generalize across all earnings calls. It likely limited the learning of the models and the representativeness of the training data.

Balancing model size with performance is another challenge. This project reflected that while smaller models can perform comparatively, there might be performance trade-offs compared to larger models.

The low ROUGE scores indicate the challenges posed by the telegraph style of the gold summaries, while the high BERTScores (0.82–0.83) demonstrate that the generated summaries capture the semantic meaning of the references effectively.

To enhance the performance of summarization models, future research will focus on expanding the training dataset to include more data. This will help with addressing current limitations related to dataset size and token length. Furthermore, experimenting with larger models may provide better performance in handling complex financial texts such as earnings calls.

References:

[Abdaljalil and Bouamor, 2021] Abdaljalil, S. and Bouamor, H. (2021). An exploration of automatic text summarization of financial reports. In *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing*, pages 1–7.

[Bowen et al., 2002] Robert M. Bowen, Angela K. Davis, and Dawn A. Matsumoto. (2002). Do conference calls affect analysts' forecasts? *The Accounting Review*, 77(2):285–316.

[Brown et al., 2020] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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.

[Bubeck et al., 2023] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4.

[Cardinaels et al., 2019] Eddy Cardinaels, Stephan Hollander, and Brian J. White, (2019). "Automatic summarization of earnings releases: attributes and effects on investors' judgments," *Review of Accounting Studies*, Springer, vol. 24(3), pages 860-890, September.

[Dettmers et al., 2023] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. (2023). Qlora: Efficient finetuning of quantized llms. arXiv preprint arXiv:2305.14314.

[Dong et al., 2021] Luobing Dong, Meghana N Satpute, Weili Wu, and Ding-Zhu Du. Two-phase multidocument summarization through content-attention-based subtopic detection. *IEEE Transactions on Computational Social Systems*, 8(6):1379–1392, 2021.

[El-Haj et al., 2018] El-Haj, M., Dr., Rayson, P., and Moore, A. (2018). Towards a Multilingual Financial Narrative Processing System. *The First Financial Narrative Processing Workshop: Proceedings of the 11th Edition of the Language Resources and Evaluation Conference*, 52–58.

[El-Haj et al., 2019] El-Haj, M., Rayson, P., Young, S., Bouamor, H., and Ferradans, S. (2019). Proceedings of the Second Financial Narrative Processing Workshop (FNP 2019). *Proceedings of the Second Financial Narrative Processing Workshop (FNP 2019).*

[Filippova et al., 2009] Filippova, K., Surdeanu, M., Ciaramita, M., and Zaragoza, H. (2009). Company-oriented extractive summarization of financial news. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics EACL '09.*

[Fu et al., 2020] Xiyan Fu, Jun Wang, Jinghan Zhang, Jinmao Wei, and Zhenglu Yang. Document summarization with vhtm: Variational hierarchical topic-aware mechanism. *Proceedings of the AAAI Conference on Artifcial Intelligence*, 34(05):7740–7747, Apr. 2020.

[Galanis et al., 2012] Dimitrios Galanis, Gerasimos Lampouras, and Ion Androutsopoulos. Extractive multidocument summarization with integer linear programming and support vector regression. In *Proceedings of COLING 2012*, pages 911–926, 2012.

[Ghazali and Anum, 2010] Ghazali, Mohd. and Anum, N. (2010). The importance and usefulness of corporate annual reports in malaysia. *Gadjah Mada International Journal of Business*, 12(1), 31.

[Giannakopoulos, 2019] George Giannakopoulos. (2019). Proceedings of the Workshop MultiLing 2019: Summarization Across Languages, Genres and Sources. INCOMA Ltd., Varna, Bulgaria.

[Erkan and Radev, 2004] Gunes Erkan and Dragomir R. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artifcial intelligence research*, 22:457–479, 2004.

[Gui et al., 2019] Min Gui, Junfeng Tian, Rui Wang, and Zhenglu Yang. Attention optimization for abstractive document summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1222– 1228, Hong Kong, China, November 2019. ACL.

[Guo et al., 2024] Lubingzhi Guo, Javier Sanz-Cruzado, and Richard McCreadie. (2024). University of glasgow at the finllm challenge task: Adapting llama for financial news abstractive summarization. In *proceedings of Joint Workshop of the 8th Financial Technology and Natural Language Processing (FinNLP) and the 1st Agent AI for Scenario Planning (AgentScen): FinNLP AgentScen @ IJCAI 2024*. International Joint Conference on Artificial Intelligence.

[Hu et al., 2021] 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. arXiv preprint arXiv:2106.09685.

[Jangra et al., 2020] Anubhav Jangra, Raghav Jain, Vaibhav Mavi, Sriparna Saha, and Pushpak Bhattacharyya. Semantic extractor-paraphraser based abstractive summarization. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 191–199, Indian Institute of Technology Patna, Patna, India, December 2020. NLP Association of India (NLPAI).

[Jindal et al., 2024] Ashvini Kumar Jindal, Pawan Kumar Rajpoot, and Ankur Parikh. (2024). Upaya at the finllm challenge task 1 and 2: Distfin: Distillation based fine-tuning for financial tasks. In *proceedings of Joint Workshop of the 8th Financial Technology and Natural Language Processing (FinNLP) and the 1st Agent AI for Scenario Planning (AgentScen): FinNLP-AgentScen @ IJCAI 2024*. International Joint Conference on Artificial Intelligence.

[Kawamura et al., 2024] Ken Kawamura, Zeqian Li, Chit-Kwan Lin, and Bradley McDanel. (2024). Revelata at the finllm challenge task: Improving financial text summarization by restricted prompt engineering and fine-tuning. In *proceedings of Joint Workshop of the 8th Financial Technology and Natural Language Processing (FinNLP) and the 1st Agent AI for Scenario Planning (AgentScen): FinNLP-AgentScen @ IJCAI 2024*. International Joint Conference on Artificial Intelligence.

[Keith and Stent, 2019] Katherine Keith and Amanda Stent. (2019). Modeling financial analysts' decision making via the pragmatics and semantics of earnings calls. In *Proceedings* *of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 493–503, Florence, Italy. Association for Computational Linguistics.

[Khant and Singh Mehta, 2018] Khant, A., and Singh Mehta, M. (2018). Analysis of Financial News Using Natural Language Processing and Artificial Intelligence. *Conference: International Conference on Business Innovation 2018*.

[Krizhevsky et al., 2017] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classifcation with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, May 2017.

[Leidner, 2019] Leidner, J. L. (2019). Summarization in the Financial and Regulatory Domain. In *Trends and Applications of Text Summarization Techniques* (pp. 187–215).

[Lee and Lay-Ki, 2024] Meisin Lee and Soon Lay-Ki. (2024). ‘finance wizard’ at the finllm challenge task: Financial text summarization. In *proceedings of Joint Workshop of the 8th Financial Technology and Natural Language Processing (FinNLP) and the 1st Agent AI for Scenario Planning (AgentScen): FinNLP-AgentScen @ IJCAI 2024*. International Joint Conference on Artificial Intelligence.

[Li et al., 2023] Xianzhi Li, Samuel Chan, Xiaodan Zhu, Yulong Pei, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. (2023). Are chatgpt and gpt-4 general-purpose solvers for financial text analytics? a study on several typical tasks.

[Lin, 2004] Chin-Yew Lin. (2004). ROUGE: A package for auto_x0002_matic evaluation of summaries. In *Text Summariza_x0002_tion Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

[Litvak et al., 2010] Marina Litvak, Mark Last, and Menahem Friedman. A new approach to improving multilingual summarization using a genetic algorithm. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 927–936, 2010.

[Liu and Lapata, 2019] Liu, Y. and Lapata, M. (2019). Text summarization with pretrained encoders. arXiv preprint arXiv:1908.08345.

[Lopez-Lira and Tang, 2023] Alejandro Lopez-Lira and Yuehua Tang. (2023). Can chatgpt forecast stock price movements? return predictability and large language models.

[Mukherjee et al., 2022] Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. (2022). ECTSum: A New Benchmark Dataset For Bullet Point Summarization of Long Earnings Call Transcripts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10893–10906, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[Orzhenovskii, 2021] Orzhenovskii, M. (2021). T5-long-extract at fns-2021 shared task. In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 67–69.

[Passali et al., 2021] Passali, T., Gidiotis, A., Chatzikyriakidis, E., and Tsoumakas, G. (2021). Towards human-centered summarization: A case study on financial news. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 21–27.

[Pontes et al., 2024] Elvys Linhares Pontes, Carlos-Emiliano González Gallardo, Mohamed Benjannet, Caryn Qu, and Antoine Doucet. 2024. L3itc at the finllm challenge task: Quantization for financial text classification summarization. In *proceedings of Joint Workshop of the 8th Financial Technology and Natural Language Processing (FinNLP) and the 1st Agent AI for Scenario Planning (AgentScen): FinNLP-AgentScen @ IJCAI 2024*. International Joint Conference on Artificial Intelligence.

[Saini et al., 2018] Naveen Saini, Sriparna Saha, Anubhav Jangra, and Pushpak Bhattacharyya. Extractive single document summarization using multi-objective optimization: Exploring self-organized differential evolution, grey wolf optimizer and water cycle algorithm. *Knowledge-Based Systems*, 164, 11 2018.

[Singh, 2020] Singh, A. (2020). PoinT-5: Pointer Network and T-5 based Financial Narrative Summarisation. *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, 104–110.

[Suarez et al., 2020] Suarez, J. B., Martinez, P., and Martinez, J. L. (2020). Combining financial word embeddings and 58 knowledge-based features for financial text summarization uc3m-mc system at fns-2020. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 112–117.

[Xie et al., 2023] Qianqian Xie, Weiguang Han, Yanzhao Lai, Min Peng, and Jimin Huang. (2023). The wall street neophyte: A zero-shot analysis of chatgpt over multimodal stock movement prediction challenges.

[Xie et al., 2024a] Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang Kang, Ziyan Kuang, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, Guojun Xiong, Zhiyang Deng, Yuechen Jiang, Zhiyuan Yao, Haohang Li, Yangyang Yu, Gang Hu, Jiajia Huang, Xiao-Yang Liu, Alejandro Lopez-Lira, Benyou Wang, Yanzhao Lai, Hao Wang, Min Peng, Sophia Ananiadou, and Jimin Huang. (2024a). The finben: An holistic financial benchmark for large language models.

[Xie et al., 2024b] Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. (2024b). Pixiu: A comprehensive benchmark, instruction dataset and large language model for finance. *Advances in Neural Information Processing Systems*, 36.

[Xie et al., 2024c] Qianqian Xie, Jimin Huang, Dong Li, Zhengyu Chen, Ruoyu Xiang, Mengxi Xiao, Yangyang Yu, Vijayasai Somasundaram, Kailai Yang, Chenhan Yuan, Zheheng Luo, Zhiwei Liu, Yueru He, Yuechen Jiang, Haohang Li, Duanyu Feng, Xiao-Yang Liu, Benyou Wang, Hao Wang, Yanzhao Lai, Jordan Suchow, Alejandro Lopez-Lira, Min Peng, and Sophia Ananiadou. (2024c). FinNLP-AgentScen-2024 Shared Task: Financial Challenges in Large Language Models - FinLLMs. In *Proceedings of the Eighth Financial Technology and Natural Language Processing and the 1st Agent AI for Scenario Planning*, pages 119–126, Jeju, South Korea. -.

[Xu et al., 2013] Xu, W., Grishman, R., Meyers, A., and Ritter, A. (2013). A Preliminary Study of Tweet Summarization using Information Extraction. *Proceedings of the Workshop on Language in Social Media (LASM 2013)*, 20–29.

[Zhang et al., 2024] 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.

[Zhang et al., 2020] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. (2020). BERTScore:Evaluating Text Generation with BERT. In *8th In_x0002_ternational Conference on Learning Representations (ICLR 2020)*, Virtual Event.

[Zhent et al., 2020] Zheng, S., Lu, A., and Cardie, C. (2020). SUMSUM@FNS-2020 Shared Task. *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, 147–151.

[Zhong et al., 2020] Zhong, M., Liu, P., Chen, Y., Wang, D., Qiu, X., and Huang, X. (2020). Extractive summarization as text matching. arXiv preprint arXiv:2004.08795.

[Zhou et al., 2021] 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.