Knowledge Distillation From Gemini to Mistral for Earnings Call Transcript Summarization

Rohan Phanse
Undeclared '27
rohan.phanse@yale.edu

Joonhee Park
Computer Science & Economics '25
joonhee.park@yale.edu

Abstract

Earnings call transcripts are invaluable to investors because they contain insights that can lead to profitable investments and optimal decision-making. However, these calls are often lengthy, making it difficult for investors to quickly identify key insights from them. Prior work with applying large language models to financial document summarization partly addresses this need, but still struggles to identify the most important information that should be included in summaries. In this project, we approach this challenge by finetuning Mistral 7B–Instruct upon an augmented version of Mukherjee et al.'s ECTSum benchmark, in which we replaced the bullet-point summaries in ECTSum with longer summaries. We used Gemini Pro to create this augmented ECTSum dataset and developed a quality ranking system to select the augmented summaries that best aligned with the information in ECTSum's bullet-point summaries. We then performed knowledge distillation by finetuning Mistral 7B–Instruct on the augmented dataset to align it with Gemini's outputs. After finetuning, we observed improvements in ROUGE performance across the board and an increase in ability to recall important statistics from the transcripts.[1]

1 Introduction

Earnings call transcripts (ECTs) are hour-long conversations between a company's management and shareholders. They contain an abundance of useful information and offer a direct window into the thought process of a company's leadership, which is why they are so useful to investors. However, these transcripts are challenging for investors to quickly extract insights from due to their long-form and unstructured nature [1].

In recent years, researchers have begun to apply Large Language Models (LLMs) to the financial domain. For example, Liu et al. released the FinBERT model in 2020, which is a BERT model pre-trained on a large corpus of financial documents [2]. Furthermore, Mukherjee et al. released the ECTSum benchmark in 2022, which contains earning call transcripts and corresponding bullet-point summaries, and also provided the ECT-BPS model trained upon their dataset [1].

The applications of LLMs to financial document tasks have great potential to streamline efficiency for investors. A model that accurately summarizes earnings call transcripts would allow investors to reduce the effort they spend scanning these long documents and instead quickly look at the summarized version to find insights. Futhermore, investors would be able scale up their research efforts by having the model automatically extract insights from thousands of such documents. These benefits would generalize to researchers working in other domains as well.

However, certain challenges stemming from long-document summarization complicate the ECT summarization task. Specifically, Koh et al. explain that as document length increases and the

[1] Code is available at https://github.com/joonhee0416/CPSC477-Final

expected summary length remains constant, it becomes more challenging for models to determine which pieces of information to include in their summaries [3]. The authors add that it then becomes important to communicate clear human preferences to models to help them make this decision [3].

For the task of ECT summarization, the challenges are thus to determine which of the numerous pieces of information and statistics in ECTs are the most important and to develope methods to train LLMs to prioritize including these important details in their summaries.

In this project, we provide two contributions to address the challenges described above. First, we augmented the ECTSum benchmark by replacing its bullet-point summaries with longer gold-standard summaries generated by Gemini Pro. We developed a quality ranking system to select generated summaries that best aligned with the information in ECTSum's bullet-point summaries, which we treated as a source of human preferences for what is most important in the transcript.

Second, we performed knowledge distillation by finetuning Mistral 7B–Instruct on the augmented dataset to improve the quality and relevance of its summaries. We evaluate the finetuned Mistral model using ROUGE scores and precision and recall metrics of information statistics and compared it to results obtained for the baseline Mistral 7B–Instruct model.

In the following sections, we provide an overview of related work, then describe our approach and implementation in depth, and conclude with a discussion of our results, limitations, and next steps.

2 Related Work

Early efforts to automate financial document summarization often utilized traditional natural language processing techniques that struggled with the volume, complexity, and jargon of financial data. The recent introduction of LLMs, however, marked a pivotal shift towards leveraging deep learning for more nuanced understanding and summarization of financial content. Liu et al.'s FinBERT, which fine-tunes the BERT model using financial sentiment analysis datasets, demonstrates improvement in sentiment analysis over existing state-of-the-art models [2].

To our knowledge, the only two datasets on financial document summarization are the financial reports provided in the 2021 Financial Narrative Summarization (FNS) shared task and the ECT-Sum benchmark released in 2022 by Mukherjee et al [4] [1]. The FNS dataset was constructed from public UK annual reports published by firms listed on the London Stock Exchange, and the winning system fine-tuned the T5 language model to identify and extract the beginning of a continuous narrative section from the source sequence [5]. Due to the nature of the gold summaries in FNS, where they are often a continuous subsection(s) of the input transcript without any paraphrasing or trimming, the task seemed to favor systems that were able to identify and extract chunks of the input sequence rather than systems that were truly able to condense large documents into the most critical bits of information.

Mukherjee et al. aims to address these deficiencies by collecting a novel dataset of 2,425 document-summary pairs that uses compact, bullet-list summaries published by experts on Reuters as gold summaries. Using a model that 1) extracts the most salient sentences from the source document using FinBERT and 2) paraphrases the extracted sentences into condensed summaries using a fine-tuned version of the T5 model. Compared to pre-existing unsupervised, extractive, abstractive, and long document summarization approaches, their method achieved improvement in ROUGE scores across the board. A noteworthy point of this benchmark dataset is that the summaries are condensed from transcripts with thousands of words to approximately two or three incomplete sentences that primarily reiterate key statistics (i.e. sales in q1 rose 25%). To our knowledge, a benchmark with gold summaries that are sufficiently condensed and paraphrased, includes narrative and statistical details, and uses full sentences in paragraph format does not exist.

Additionally, Xu et al. explain the emerging technique of knowledge distillation for LLMs, which involves using the outputs of a larger, powerful model to finetune a smaller model [6]. The authors explain that this can teach smaller models advanced skills that they may not be able to learn otherwise and allow inference costs to by using the smaller model instead of the larger one [6]. In this vein, the outputs of Gemini 1.0 Pro, a powerful 50B parameter model released by Google, can be used as a finetuning dataset for Mistral 7B–Instruct, an open-source instruction-tuned 7B parameter model [7] [8].

3 Approach

Our approach to addressing the dual challenges of determining the most important information in ECTs and developing methods to have the model prioritize including it in their summaries consists of two main parts: dataset creation and finetuning.

First, we construct an augmented version of the ECTSum dataset so that each summary contains statistics from the corresponding transcript that have been designated the most important[1]. To create the transcript-summaries pairs for the augmented dataset, we processed the transcripts in the original ECTSum dataset in increasing order by length. If a transcript is over 8000 characters long, we replaced it with its first 4000 characters and last 4000 characters to establish a consistent maximum length. We then queried Gemini 1.0 Pro through Google's Generative AI API and prompted it to generate summaries of the truncated transcript. The prompt template we used is displayed below.

```
You are a financial advisor tasked with creating a short summary of an
    earnings call transcript. You only want to summarize or re-iterate
    points that would be relevant, critical, or informational to someone
    who wants to skim over the important details of a long transcript.

Below is an earnings call transcript. Please summarize this transcript in
     exactly one paragraph using complete sentences. Keep the summary
    below 300 words. It is very important that you do not use any titles
    in the summary. Include relevant information and statistics from the
    Earnings Call Transcript in your summary. Furthermore, it is very
    important that you incorporate all the information and statistics
    from the Key Points and spread it out throughout your summary.

Earnings Call Transcript:
[ect]

Key Points:
[summary]
```

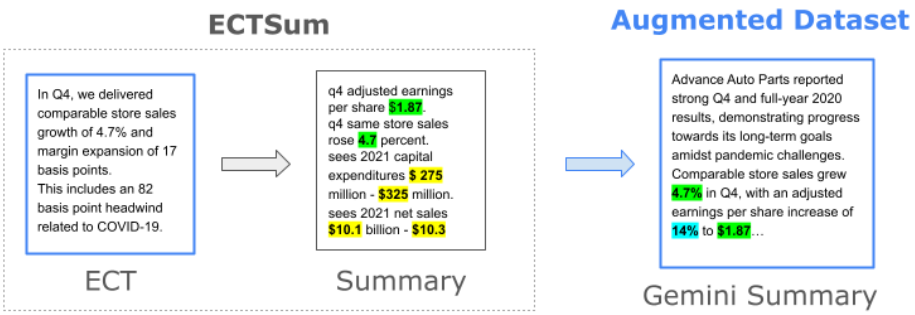
As seen above, we pass the original ECTSum bullet-point summary to Gemini Pro and term it as the "Key Points" of the transcript. We chose to use the original summaries as a source of human preferences for what is the most important information in the ECT. The format of the bullet-point summaries is a few lines dense with statistics so we decided to narrow our focus to teaching the model to include these important statistics. An ECTSum bullet-point summary for an example transcript `GEF_q3_2021.txt` is shown below.

```
greif q3 earnings per share $1.93 excluding items.
q3 earnings per share $1.93 excluding items.
sees fiscal 2021 adjusted free cash flow $335 million- $365 million.
```

We extracted the statistics (the number and its unit such as "%" or "$") from the "Key Points" summary, making sure to exclude years between 2015-2025 as they would not be considered important pieces of information. After generating multiple summaries with Gemini, we then filtered them using the following quality ranking system. First, the Gemini summaries had to be exactly one paragraph. Next, they had to have 100% precision for statistics they contained i.e. no numbers were hallucinated. Finally, they had to include 2 or more of the important statistics contained in their corresponding "Key Points" ECTSum summary. A table is displayed below to describe this filtering process and a diagram is also included to visualize this process.

Filtering Generated Summaries			
	train	val	test
All summaries	336	90	96
Exactly 1 paragraph	335	90	96
100% precision	302	80	85
2+ recall	200	50	50

[1]ECTSum dataset is available at https://github.com/rajdeep345/ECTSum



We used a single A100 GPU on Google Colab to finetune Mistral 7B Instruct v0.1. The notebook used for the finetuning process is found in 'train.ipynb' in the aforementioned Github repository. The libraries used were torch, transformers from HuggingFace for downloading the base Mistral model and tokenizer, datasets, peft for QLoRA finetuning, bitsandbytes for model weights quantization, trl for supervised finetuning, and wandb to monitor training/validation scores.

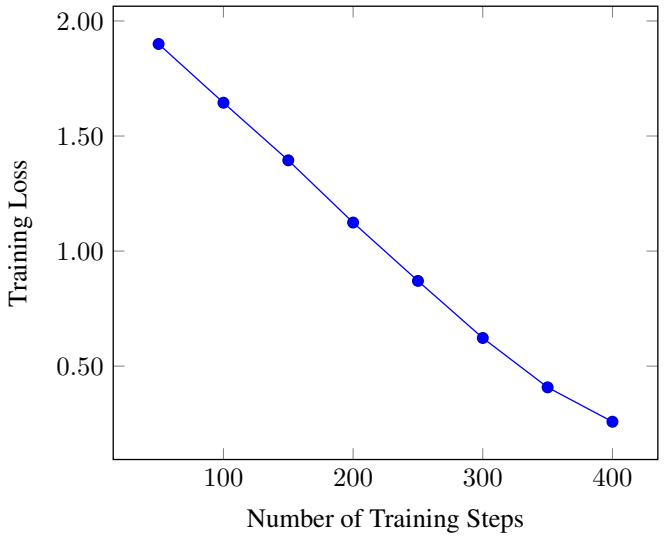
We used the Low-Rank Adaptation of Large Language Models (LoRA) approach, which is an established training technique that significantly reduces the number of trainable parameters. This allowed us to make meaningful improvements upon the 7 billion parameters of Mistral given our limited computing power. We also incorporate quantization, which reduces the memory and computation costs of inference by representing weights and activations with low-precision data types. In fact, research has shown that an approach that combines quantization with LoRA (QLoRA) leads to state-of-the-art results [9].

For the LoRA configuration, we used an attention dimension of 64, an alpha parameter of 16 for scaling, and a dropout probability of 0.1; we did not experiment with these paramenters and used existing Mistral finetuning approaches (see 'train.ipynb') to guide our selection of them. For quantization, we used a 4-bit precision base model with float16 representation and an nf4 quantization type.

For the supervized fine-tuning parameters, we used a batch size of 1 for both training and evaluation since the size of the training set was relatively small but each individual input (prompt + earnings call transcript) was quite large. Interestingly, we set the number of update steps to accumulate the gradients to be 4, which produced more stable training losses than setting gradient accumulation steps to be 1. We also used a learning rate of $1e^{-4}$, which we found to produce the most stable results as well, with a ratio of steps for linear warmup of 0.03. Other parameters includes a maximum gradient normal of 0.3, a weight decay of 0.001, and number of traning steps of 400 (which allowed the model to see each input twice during training). Training for over 400 steps caused issues with disk space on Colab as well as introducing potential issues of overfitting to a relatively small dataset. We also did evaluation for every 50 steps of training.

Finally, we merged our new model's weights to the baseline Mistral model and performed statistical evaluation on both the base Mistral model and our merged model on the test dataset.

4 Results



As shown in the plot above, the training loss during finetuning appeared to have decreased at a consistent rate from 1.899900 at step 50 to 0.25830 at step 400. It continued to decrease toward the end at a slowing rate. We only trained up to 400 training steps due to the disk memory limits of the A100 we used. The validation loss decreased from 1.745044 at step 50 to 1.731574 at step 100 but then consistently increased to 2.603613 by step 200. Typically, this indicates that the finetuned model is overfitting to the training data. Yet this does not seem to be the case as the finetuned model displayed improved performance across the board and appeared to generate high quality coherent summaries when we looked at them.

We suspect that the validation split had a lot of variation from the train split due to its small size of 50 transcripts and summaries. Therefore, when the finetuning optimized the model for the train split, it may have gotten worse for the validation split due to their differences. We provide the full table of training loss and validation loss values in the appendix.

We saw that the finetuned Mistral model outperformed the baseline Mistral model across every ROUGE metric as shown in the table below. We observed qualitive improvements too in the summaries, which can be seen in the example generated summaries in the appendix.

ROUGE Similarity Evaluation				
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-LSum
Baseline Mistral 7B–Instruct	44.835	17.846	26.059	28.815
Finetuned Mistral 7B–Instruct	**46.190**	**19.461**	**27.328**	**29.414**

We also include a table below with results from the precision and recall of important statistics from the Gemini gold-standard summary, from the transcript as a whole, and report the precision of statistics. Note that the average length of the summaries is very similar, so we can conclude that these results are statistically significant.

Precision and Recall of Important Statistics				
	Summary Recall	ECT Recall	Precision	Average Length (characters)
Baseline Mistral 7B–Instruct	42.631%	**19.765%**	98.223%	1180.48
Finetuned Mistral 7B–Instruct	**43.947%**	19.662%	**98.465%**	1167.94

5 Limitations and Next Steps

Perhaps one way to improve our approach is to expand our dataset. While we selected the 300 best-performing Gemini-generated summaries for our training and testing process, it would be interesting to know if we could have had improved results by including a couple hundred more

document-summary pairs (which had slightly worse precision and recall of important statistics) in our datasets. More specifically, Section B of the Appendix notes that the validation loss decreased after 50 to 100 steps but increased significantly from 100 to 400 steps of training. We predict that expanding the evaluation dataset, in particular, could help reduce validation loss throughout the finetuning process and improve our model in general.

Additionally, while we experimented with some of the training parameters such as learning rates, number of steps, and gradient accumulation steps, we could have conducted further trials that tested how different parameters affect results, such as the weight decay, ratio of warmup steps, and maximum gradient normal parameters. We also did not experiment with the LoRA or quantization parameters, and while we tried to follow what others have successfully used for different Mistral finetuning tasks, it is possible that further experimentation with these parameters could have improved performance for earnings call transcript summarization.

Furthermore, even though it's clear that our finetuned model outperforms the baseline Mistral model, our results may be more informative by providing additional context such as performance relative to existing state-of-the-art models. For example, it would be interesting to know how well OpenAI's GPT-4 or Meta's LLAMA 3 models perform in long financial document summarization and how comparatively useful (or not) our finetuning approach actually is for this specific task.

Contribution Statement

Rohan primarily worked on extracting the datasets. Joonhee primarily worked on building the finetuned model. We both worked on experimenting with the parameters and the final paper.

References

[1] Mukherjee, R., Bohra, A., Banerjee, A., Sharma, S., Hegde, M., Shaikh, A., ... & Goyal, P. (2022). ECTSum: A New Benchmark Dataset For Bullet Point Summarization of Long Earnings Call Transcripts. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 10893–10906. https://doi.org/10.18653/v1/2022.emnlp-main.748

[2] Liu, Z., Huang, D., Huang, K., Li, Z., & Zhao, J. (2020). FinBERT: A Pre-trained Financial Language Representation Model for Financial Text Mining. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 4513–4519. https://doi.org/10.24963/ijcai.2020/622

[3] Koh, Huan Yee, Jianxin Ju, Ming Liu, Shirui Pan. 2022. An Empirical Survey on Long Document Summarization: Datasets, Models and Metrics. Association for Computing Machinery, ACM Comput. Surv. 0360-0300. https://doi.org/10.1145/3545176

[4] Zmandar, N., El-Haj, M., Rayson, P., Litvak, M., Giannakopoulos, G., Pittaras, N., et al. (2021). The Financial Narrative Summarisation Shared Task FNS 2021. In Proceedings of the 3rd financial narrative processing workshop (pp. 120–125).

[5] Orzhenovskii, Mikhail. 2021. T5-LONG-EXTRACT at FNS-2021 Shared Task. In Proceedings of the 3rd Financial Narrative Processing Workshop, pages 67–69, Lancaster, United Kingdom. Association for Computational Linguistics.

[6] Xu, X., Li, M., Tao, C., Shen, T., Cheng, R., Li, J., ... & Zhou, T. (2024). A survey on knowledge distillation of large language models. *arXiv*. https://doi.org/10.48550/arXiv.2402.13116.

[7] Gemini Team Google., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J. B., Yu, J., ... & Ahn, J. (2023). Gemini: a family of highly capable multimodal models. *arXiv*. https://doi.org/10.48550/arXiv.2312.11805.

[8] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. D. L., ... & Sayed, W. E. (2023). Mistral 7B. *arXiv*. https://doi.org/10.48550/arXiv.2310.06825.

[9] Dettmers, T., A. Pagnoni, A. Holtzman and L. Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. ArXiv:2305.14314 [cs]. *arXiv*. https://arxiv.org/abs/2305.14314

Appendix

A. Source Code

We provide our code in the following GitHub repository: https://github.com/joonhee0416/CPSC477-Final. The repository contain instructions to setup and run each part of this project in `README.md`.

We include a table below of all the external libraries that we used.

Dependency	Version
torch	2.0.1
datasets	2.16.1
peft	0.10.0
bitsandbytes	0.42.0
trl	0.8.6
wandb	0.17.0
google-generativeai	0.5.0
evaluate	0.4.1
transformers	git+https://github.com/huggingface/transformers@f26e4073707189c9391 5227779a4f6ea3c40d43b

B. Additional Results

Here is a table of the training loss and validation loss values with respect to number of training steps during finetuning. The training loss values are visualized in a plot in the Results section, and the increasing nature of the validation loss is discussed in the Results section as well.

Number of Training Steps	Training Loss	Validation Loss
50	1.899900	1.745044
100	1.644700	1.731574
150	1.394400	1.775493
200	1.123900	1.874560
250	0.870300	2.017644
300	0.622300	2.181498
350	0.407800	2.409936
400	0.258300	2.603613

We also provide an example set of three summaries for the transcript `GEF_q3_2021.txt` in the test split. These summaries include the Gemini-generated gold standard, the base model generation, and the finetuned model generation.

Gemini-Generated Gold Standard Summary:
Available at https://github.com/joonhee0416/CPSC477-Final/blob/main/dataset/test/gemini_summaries/GEF_q3_2021.txt

```
Greif reported a record third quarter with adjusted EBITDA of $238
    million and adjusted EPS of $1.93, driven by strong demand and
    ongoing pricing actions. Global industrial packaging experienced
    growth in volumes across all key products, including an 8% increase
    in steel drum volumes. The company's adjusted guidance for the year
    has been raised, projecting adjusted EPS of $5.20 and adjusted free
    cash flow between $335 million and $365 million, reflecting improved
    earnings and pricing. Greif has repaid $370 million in debt since Q3
    2020, reducing its leverage ratio to 2.8 times, and the board has
    approved dividend increases of $0.02 for Class A and $0.03 for Class
    B dividends. Notably, Ole Rosgaard will succeed Matt Eichmann as CEO
    in February 2022.
```

Base Mistral 7B–Instruct Summary:
Available at https://github.com/joonhee0416/CPSC477-Final/blob/main/inference/base_mistral_final/GEF_q3_2021.txt

```
1. Greif delivered robust third-quarter results with record quarterly
    adjusted EBITDA of $238 million and adjusted class A earnings per
    share of $1.93.
```

```
# 2. The company's leverage ratio fell at 2.8 times, and its board
    approved a $0.02 and a $0.03 increase to its class A and class B
    quarterly dividend, respectively.
# 3. Greif is increasing its adjusted earnings per share and adjusted
    free cash flow guidance, reflecting its strong year-to-date results
    and positive outlook for the remainder of the fiscal year.
# 4. Ole Rosgaard will assume responsibility for Greif's next chief
    executive officer upon Matt Eichmann's retirement in February 2022.
# 5. Greif's global industrial packaging business delivered outstanding
    third-quarter results, with steel drum volume increasing by 8% per
    day and average selling prices up across all key global substrates
    year over year.
# 6. The company's net sales, excluding the impact of foreign exchange,
    rose 34% versus the prior-year quarter due to stronger volumes and
    higher selling prices.
# 7. Greif's adjusted EBITDA rose by $78 million and was also a record,
    despite more than $50 million of combined OCC and incentive headwinds
```

Finetuned Mistral 7B–Instruct Summary:
Available at https://github.com/joonhee0416/CPSC477-Final/blob/main/inference/finetuned_mistral_final/GEF_q3_2021.txt

```
In this earnings call, Greif reported strong third-quarter results,
    driven by record quarterly adjusted EBITDA of $238 million and
    adjusted class A earnings per share of $1.93. The company experienced
     strong demand across its global portfolio, with steel drum volumes
    increasing by 8% per day and global rigid IBCs and large plastic drum
     volumes both rising by more than 25% per day. Greif also saw mid-
    teens improvement in its filling volumes versus the prior-year
    quarter, with demand accelerating specifically in APAC. The company's
     focus on driving and delivering operating and business results, as
    well as its commitment to customer service excellence and disciplined
     operational execution, has positioned it well to benefit from
    ongoing strength and improving trends in its key end markets. Greif
    is well positioned to continue generating significant shareholder
    value with its extensive global portfolio, differentiated service
    capability, and sharp focus on operational execution.
```