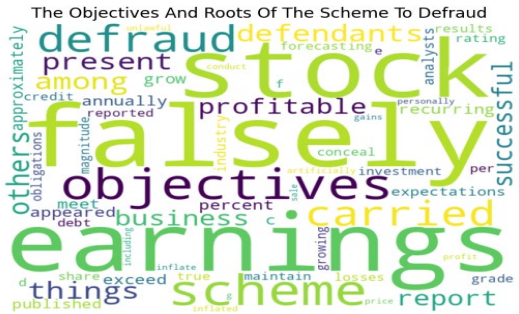
Enron Case Chatbot: Data Prep for an LLM Fine-Tuning

Executive Summary

Milestone 1: Winter 2024
Gary Schaumburg
Venkat Panyam
Samarth Somani
01-gschaumb-vrpanyam-ssomani



Background

LLM RAG vs. Fine-Tuning

Retrieval Augmented Generation (RAG) and Supervised Fine-Tuning (SFT) on Large Language Models (LLM) are hot topics right now. Across industries companies are investigating how to train LLMs on their data with a goal of creating chatbots that can guide their employees or customers on topics related to their business.

There is an ongoing discussion regarding whether RAG or SFT (or other forms) work best and in what situations. Some experts advise that both of these approaches are unreliable, short-term solutions and that it would be wise to wait before deep investment in programs.

Approach

The Project

As a necessary first step in both RAG and SFT, appropriate data sources must be selected that embody the knowledge you want the bot to be an expert in. Then this data must be preprocessed in specific ways to suit the RAG or SFT.

This project focused on the preprocessing of two public datasets from the Enron case: the Enron email corpus and a set of PDF documents.

The objective was to prepare these datasets for use in a Supervised Fine Tuning of a Large Language Model. The project was dedicated solely to the data preparation steps, with associated analysis and visualizations.

Data Sources

Four public Securities and Exchange Commission case complaint PDFs.

Enron public email corpus comprising over 500,000 emails.

Goals

Mission Statement

Our objective was to explore the challenges in creating a training dataset for a moderately difficult SFT scenario – a legal case expert chatbot.

Our hoped for outcome was an efficient method of converting the email corpus and the PDFs into the SFT training dataset.

Questions

We started the project with the following questions:

1. How can these data sources be standardized and processed for integration into an SFT model?
2. How do we maximize information retrieval from the data sources?
3. How do we apply big data concepts to the email corpus?
4. What are the attributes of the source data and how do these guide the preprocessing steps?
5. What types of analyses and visualizations support the above?
6. Where is human intervention necessary?

Results

Selection of SFT over RAG as Goal

SFT was chosen as the end goal due to use-case suitability and Milestone 1 scope. RAG preprocessing relies on concepts which are out of scope for Milestone 1, such as semantic similarity.

Analyses and Visualizations

These were most useful with emails. Keyword searches proved helpful in identifying emails for conversion to Prompt / Response pairs. Visualizations provided big picture context for more detailed search and extraction.

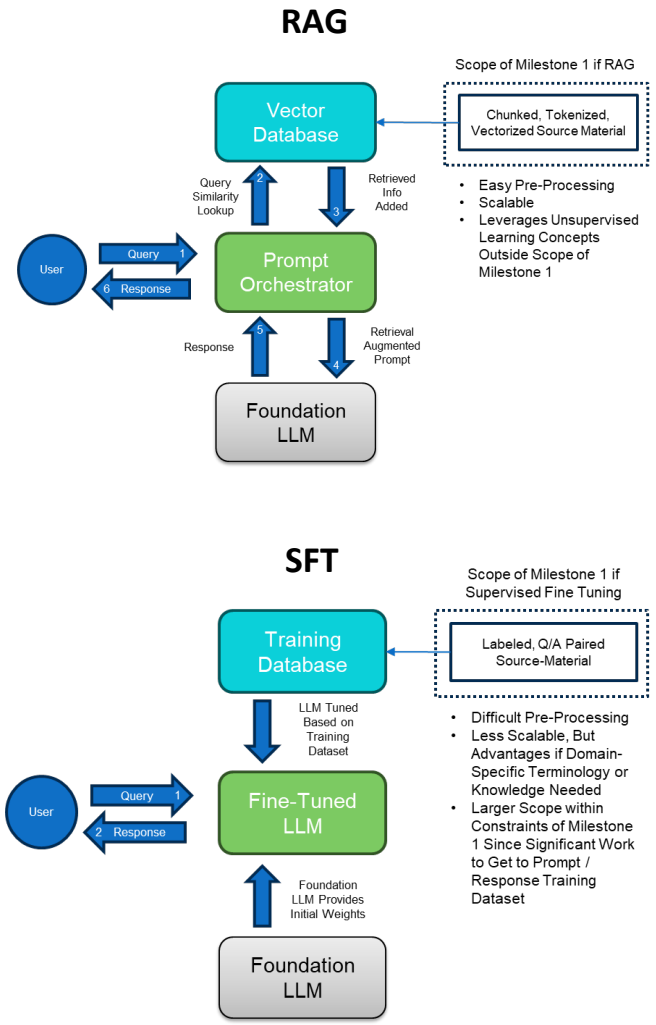
Conclusions

The PDFs proved easiest to preprocess as their structure easily guided topic Prompt / Response creation.

The email corpus was more difficult, requiring “needle in a haystack” search techniques and more human support.

An SFT model on Hugging Face was fine-tuned with our resulting training dataset - a topic for future exploration.

Project Motivation



The Issues

RAG and SFT were the two techniques we found discussed most often in the literature related to creating specialized chatbots. Given our project goals it was not immediately obvious that one would be better than the other. RAG is in some ways more scalable, while SFT is said to be better for domain-specific training. So we looked at the two options based on whether the preprocessing could be done within expected Milestone 1 scope.

The RAG technique relies on a prompt orchestrator that supplements the user prompt with a chunk of similar data returned from the training data. The preprocessing of the training data is fairly straightforward, but relies on understanding of vectorization and semantic similarity concepts covered in Unsupervised Learning.

SFT involves tuning the weights of an existing LLM with training data in the form of Prompt / Response pairs. In simple terms, you are actually changing the parameters of the model to suit your body of knowledge through the fine-tuning process. The user interacts with this changed model in prompting.

The Decision - SFT

Milestone 1 Scope: The training data preparation for an SFT better fit within the scope of Milestone 1. Creation of the Prompt / Response training pairs could be done with the basic preprocessing, analysis, and visualization techniques of our prerequisite courses.

Preprocessing Implications

Unlike RAG, which dynamically retrieves information in response to queries, SFT requires a preprocessing step to create specific prompts and corresponding responses that accurately reflect the details and nuances of the Enron case. This process can require a significant amount of human intervention to ensure the training data is representative of the legal and financial terminologies and details contained within the documents.

The possible need for expert knowledge to annotate and create these pairs means that SFT can be more resource-intensive upfront, potentially slowing the initial training phase and requiring a greater investment in domain expertise.

Current Industry Wisdom

SFT Allows for a More Controlled Integration of Domain-Specific Knowledge: SFT directly adjusts the model's parameters based on the information contained in our documents, ensuring the model learns to replicate the patterns, terminology, and nuances of the case. In contrast, RAG combines retrieval mechanisms with generative capabilities, pulling in external information on the fly. While RAG can access a broad range of data, it may struggle with the depth and specificity required to accurately reflect the complexities of the Enron case.

SFT Enables the Creation of a More Consistent and Reliable Chatbot for Legal Case Details: By fine-tuning a model with a curated dataset focused on the Enron case, the resulting chatbot is likely to exhibit higher performance in generating responses that are directly relevant and accurate to the case. RAG's dependence on retrieving and then generating responses can introduce variability in the quality and relevance of the responses.

Data Sources

Name	Description	Size	Links
PDF Documents	Four SEC case documents detailing the charges against defendants that include Enron executives, attorneys, and audit firm lead. The documents are divided into section and paragraph groups that indicate the lower level paragraph topics. “comp18776”: SEC Complaint against Lay, Skilling, and Causey “comp18435”: SEC Complaint against Delainey “comp20441”: SEC Complaint against Duncan “comp20058”: SEC Complaint against Mintz and Rogers	comp18776: 954KB comp18435: 102KB comp20441: 5.98MB comp20058: 603KB	https://www.sec.gov/files/litigation/complaints/comp18776.pdf https://www.sec.gov/files/litigation/complaints/comp18435.pdf https://www.sec.gov/files/litigation/complaints/2008/comp20441.pdf https://www.sec.gov/files/litigation/complaints/2007/comp20058.pdf
Email Corpus	This dataset was collected and prepared by the CALO Project. It contains data from about 150 users, mostly senior management of Enron, organized into folders. The corpus contains a total of about 0.5M messages. This data was originally made public, and posted to the web, by the Federal Energy Regulatory Commission during its investigation. The dataset does not include attachments, and some messages have been deleted as part of a redaction effort due to requests from affected employees.	Full tarred gzipped file: 1.7GB	https://www.cs.cmu.edu/~enron/enron_mail_20150507.tar.gz

Data Sources – Structure of PDF Documents

Section Name

The section names were generally the same for each PDF, with some small differences depending on the number of charges being filed against the defendant and the number of "CLAIM" sections at the end.

Group Name

We've used the term and column "Group Name" to describe the paragraph group topics that sat below the section names and above the paragraphs. Unique to each document, they provided an overview description of the associated paragraphs.

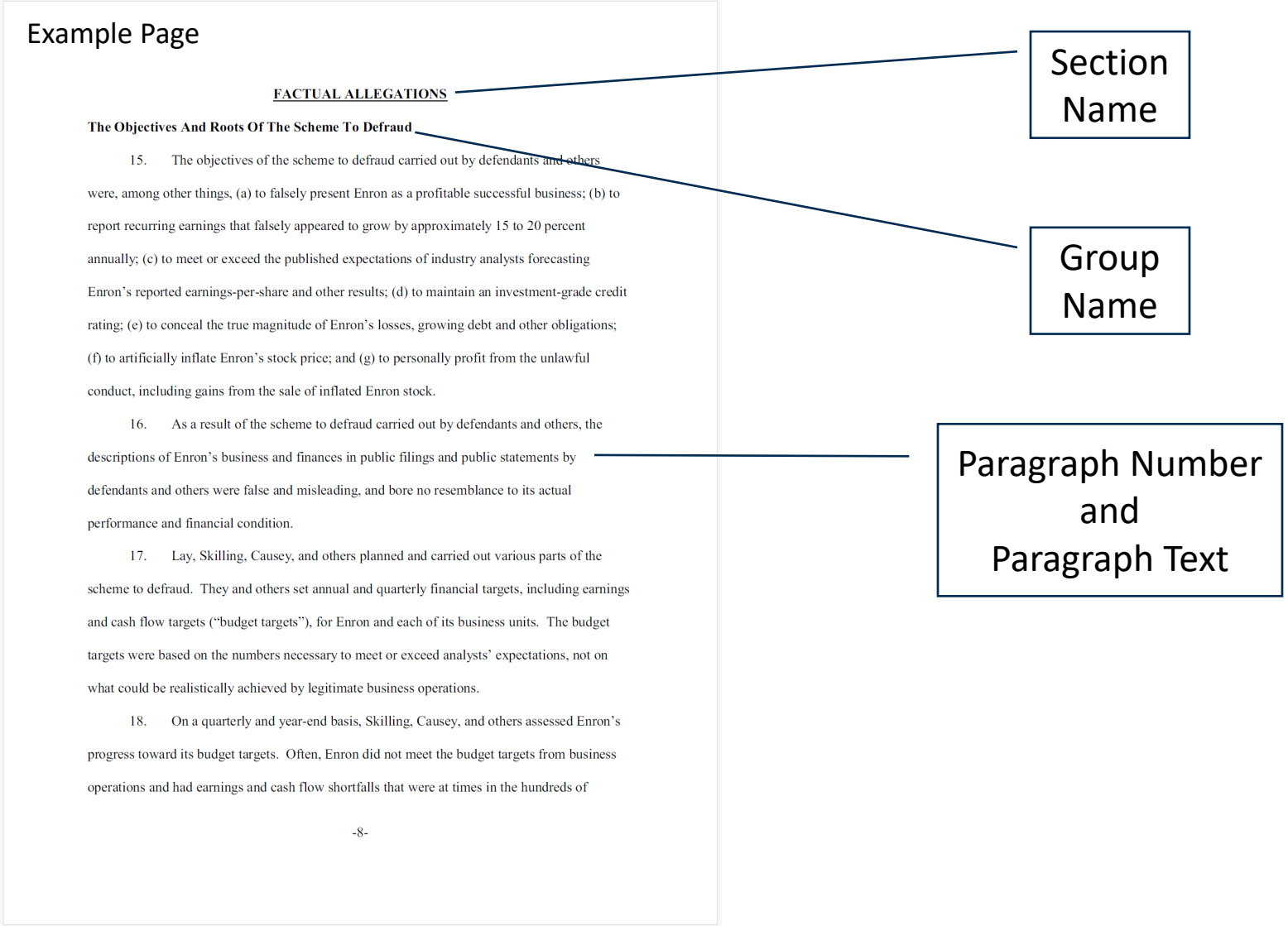
Paragraphs

Each Group Name had one or more paragraphs of description, each of them numbered.

Appendices and Tables

Extractions were modified to exclude appendices and tables.

Example Page



FACTUAL ALLEGATIONS

The Objectives And Roots Of The Scheme To Defraud

15. The objectives of the scheme to defraud carried out by defendants and others were, among other things, (a) to falsely present Enron as a profitable successful business; (b) to report recurring earnings that falsely appeared to grow by approximately 15 to 20 percent annually; (c) to meet or exceed the published expectations of industry analysts forecasting Enron's reported earnings-per-share and other results; (d) to maintain an investment-grade credit rating; (e) to conceal the true magnitude of Enron's losses, growing debt and other obligations; (f) to artificially inflate Enron's stock price; and (g) to personally profit from the unlawful conduct, including gains from the sale of inflated Enron stock.

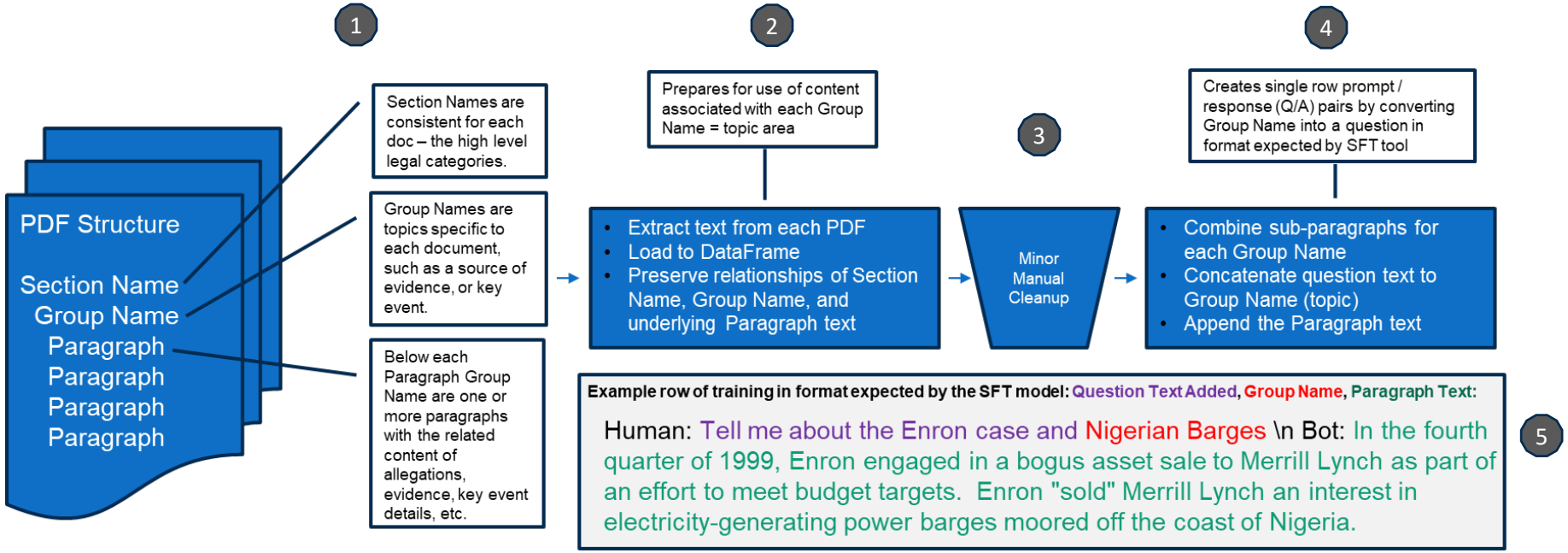
16. As a result of the scheme to defraud carried out by defendants and others, the descriptions of Enron's business and finances in public filings and public statements by defendants and others were false and misleading, and bore no resemblance to its actual performance and financial condition.

17. Lay, Skilling, Causey, and others planned and carried out various parts of the scheme to defraud. They and others set annual and quarterly financial targets, including earnings and cash flow targets ("budget targets"), for Enron and each of its business units. The budget targets were based on the numbers necessary to meet or exceed analysts' expectations, not on what could be realistically achieved by legitimate business operations.

18. On a quarterly and year-end basis, Skilling, Causey, and others assessed Enron's progress toward its budget targets. Often, Enron did not meet the budget targets from business operations and had earnings and cash flow shortfalls that were at times in the hundreds of

-8-

Data Manipulation - PDFs



Preprocessing

1: Somewhat by luck (not known when we scoped the project), the PDFs are well structured for matching content with topics given their Section / Paragraph Group / Paragraphs format.

2: Using PyMuPDF to get a text file from each PDF, we then parsed Section Name, Group Name (paragraph group topic name), and the related Paragraph Numbers and Paragraph Text into DataFrames.

3: The PDF files have some idiosyncrasies with respect to spacing, marks, formats, etc. These documents were likely formatted and scanned with early to mid 2000's techniques. For instance, one of the files seems to have had footers added at point of scanning. This necessitated some workaround coding and some human cleanup of each DataFrame.

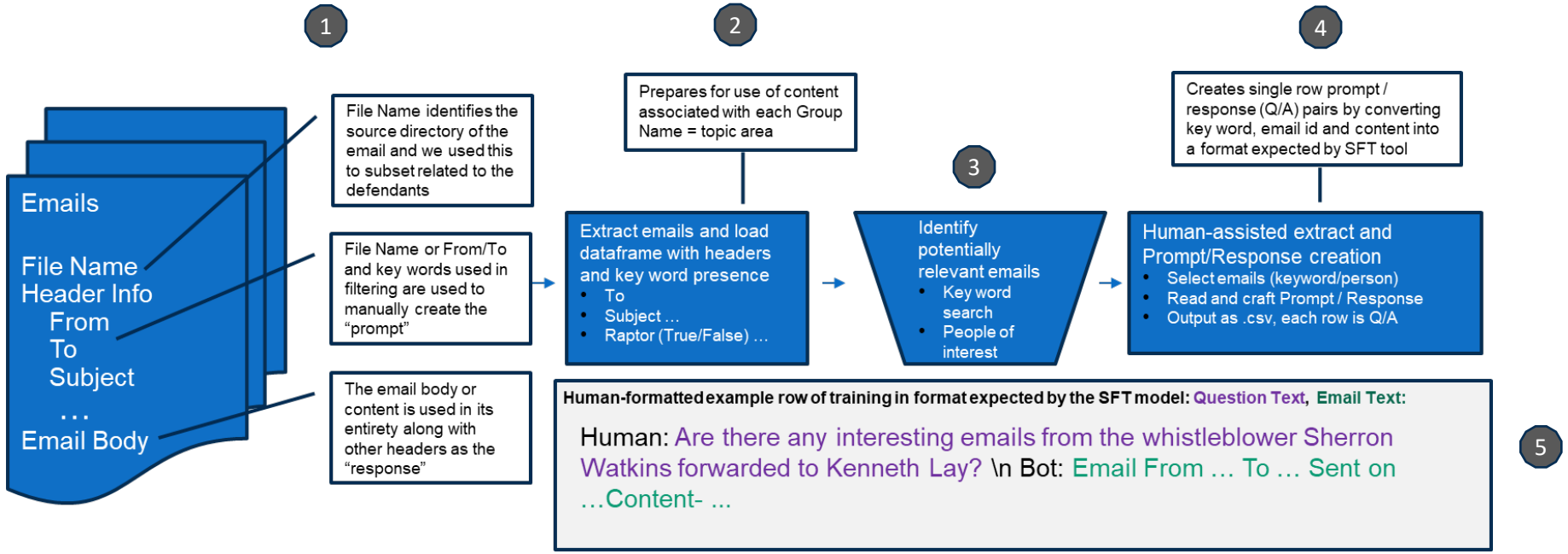
Prompt / Response Extraction

4: Paragraph data for each unique Group Name was combined – resulting in an association between Group Name topic and its sub-paragraphs in a DataFrame.

5: Because each set of paragraphs associated with a Group Name is essentially the answer to a question such as "What is [Group Name]", it was easy to create code to parse the DataFrame into our Prompt / Response pairs. We appended "Tell me about the Enron case and" to the front of each Group Name, as the Prompt. And then used the content of the associated paragraphs as the Response.

We did not dig too deeply into best practices of SFT training dataset creation since it is out of scope for this Milestone. It is therefore possible the above format could be improved.

Data Manipulation – Emails



Preprocessing

1: The source for email data is a gzip file containing a directory structure with emails in sub-directories for each employee. Using the requests library and the tarfile module, we downloaded the .gz file and extracted the individual email files.

2: With the package, we extracted various components/headers of each email like the From, To, message id, Subject, Content, etc. to create a dataframe containing details of all 517,401 emails.

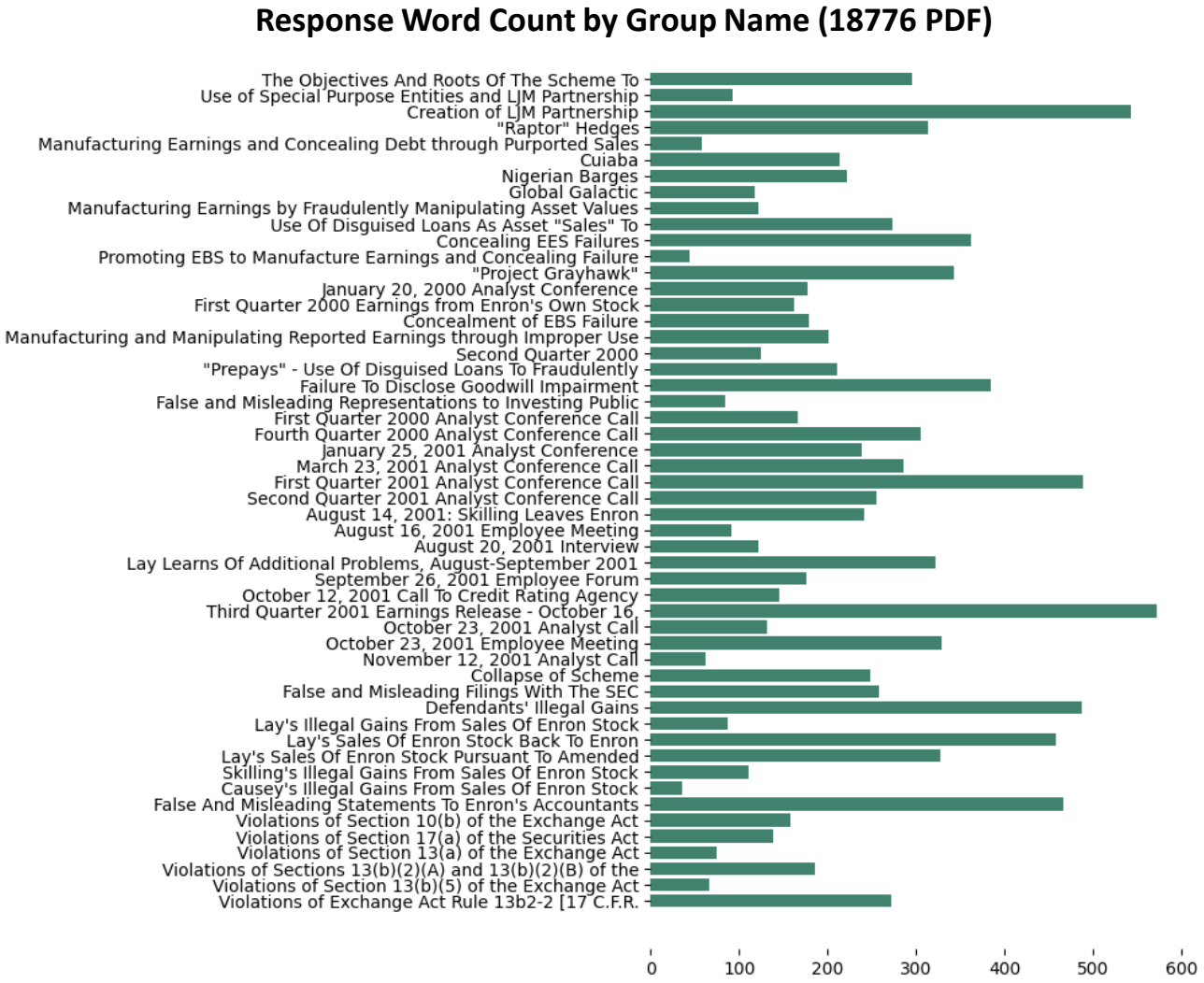
3: We found that the dataset had no email directories for some defendants, the CFO Andrew Fastow and the Enron whistleblower, Sherron Watkins. To identify a subset of relevant emails, we analyzed for occurrence of key words from the case such as "Cuiaba" and "Raptor". We decided to focus on emails that contained keywords or were from the case defendants.

Prompt / Response Extraction

4: We found that emails do not lend themselves to be easily tagged to a topic, as word references are mostly tangential or completely unrelated to the case. Manual effort is required to ensure any data that is tagged as related to the Cuiaba or Raptor evidence (as an example) does in fact have material information related to these topics.

5: Because of our findings, we used just a few results from the analysis to create a small sample of Prompt / Response pairs as an example of how it could be done with a lot of manual human effort to validate email content relevance.

Analysis & Visualizations: PDFs



Useful Analyses & Visualizations

Although the actual fine-tuning of the LLM is out of scope for this project, our research indicated that understanding the length of training responses in words or tokens is useful. (Tokens are typically 1X to 1.5X word count and depend on model chosen for SFT.)

Word count of responses can aid in identifying imbalances in the training data. It helps understand how to apply techniques such as truncation, padding, or augmentation to normalize response lengths, which may help with consistent model performance.

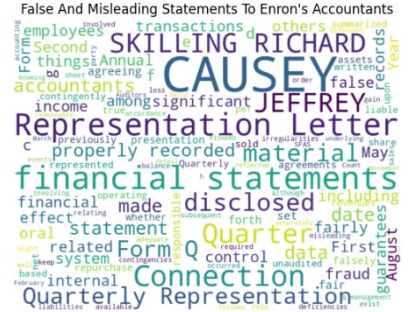
Models pre-trained on datasets with similar response length may require less adjustment and achieve better performance more quickly.

Also, insights from word count may guide customization of loss functions or the implementation of specific training objectives that emphasize brevity, detail, or style in responses.

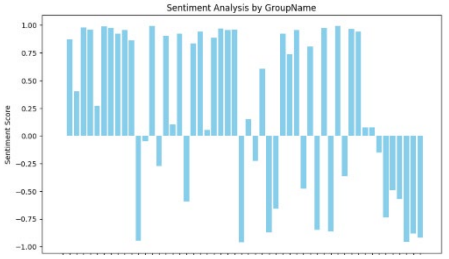
Since the structure of the PDFs easily guided our Prompt / Response extraction, we struggled to find useful analyses beyond word count. Some 'Dead Ends' follow.

Dead Ends

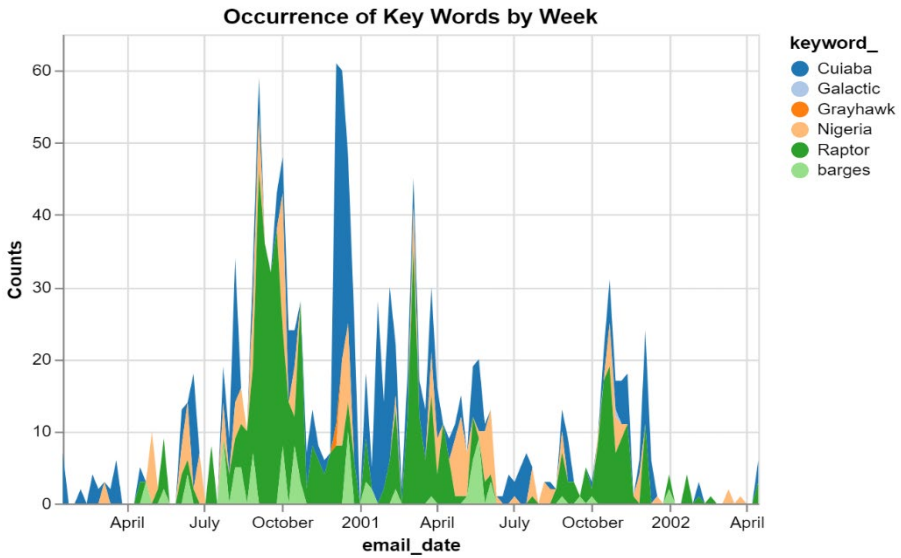
Word Clouds: Word clouds for each Group Name were explored, but found to not reveal anything useful. Each group's content is more easily understood by quick human reading.

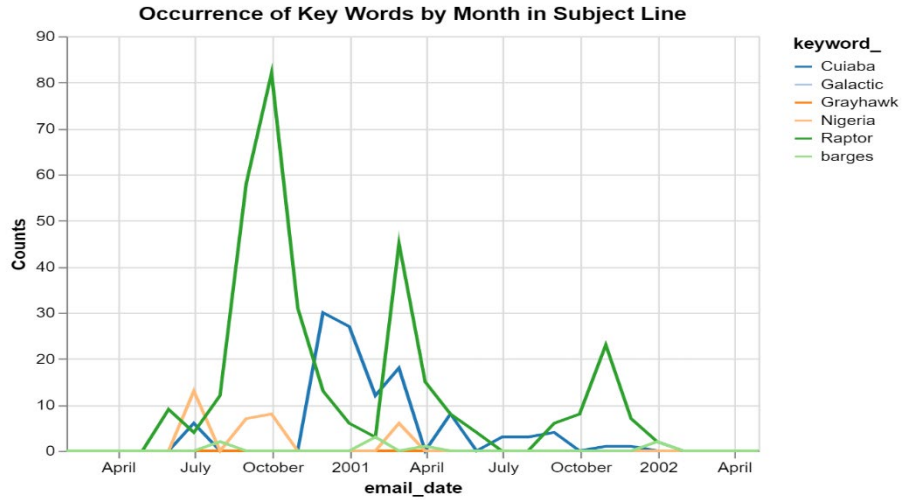


Sentiment Analysis: Running sentiment analysis on each paragraph topic group was also experimented with. Lower sentiment resulted from the presence of words such as "misled", "fraudulent", "misrepresented", etc. But we didn't find this helpful in distinguishing the importance of content.



Analysis & Visualizations: Emails





Useful Analysis & Visualization

Key Word Appearances: In order to analyze the Enron email corpus for creating the prompt-response training data, we looked at ways to filter the 500,000 plus emails to a more manageable number for further analysis.

First, we looked at emails that contained words related to the fraudulent practices and projects mentioned in the PDF court documents. The Occurrence of Key Words by Week chart (top left) gave us confidence that we might find some useful data if we looked at emails that had these words like Raptor and Cuiaba in the email body content. A distribution of the same keywords in the Subject line of emails aggregated at the week level was sparser.

The chart (bottom left) shows the counts by month. Comparing it with the top left visualization, the same two top key words (Raptor and Cuiaba) were present in the Subject line as with the email body content.

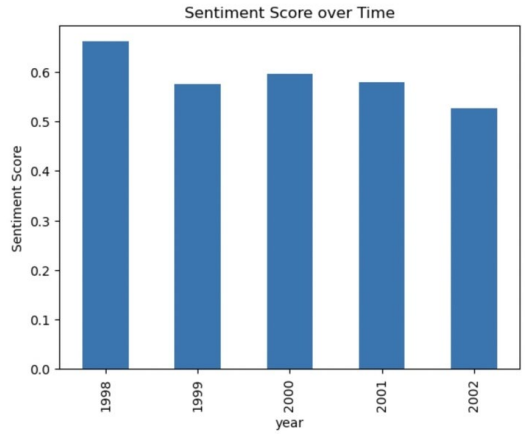
Since emails do not have a similar structure as the PDF court documents, it was not easy to clearly identify the section/topic for use in formulating the prompt-response data and some manual effort was required. Using these word count visualizations helped us focus the manual effort on emails that are likely to be relevant to the case.

Dead Ends

Sentiment Analysis: We also attempted to do sentiment analysis on the email data, including general sentiment for a period of 5 years that included years before and after the court case. (Shown below.)

The sentiment results did not help us in our goal to find a subset of potentially useful emails for generating the SFT training data.

The yearly analysis did not show any significant changes (other than a slight drop between 2000 and 2002) and this was not explored further.



Discussion

Choice of Supervised Fine Tuning

Our most influential driver for selecting SFT was that the training data preprocessing fit within the scope of Milestone 1. After getting our hands dirty with the identification and extraction of relevant content for our use case, we still believe SFT is the overall correct choice. SFT allows nuanced training, which seems to be a requirement for the Enron Chatbot.

We still have a lot to learn regarding SFT and RAG, but the project gave us a great opportunity to struggle with some hands-on questions regarding the training preparation. We suspect the amount of data we had for training was not sufficient.

There are easy ways to test an SFT using cloud services. We briefly experimented with Hugging Face AutoTrain – the model was trained successfully, but our first-pass test output was inconsistent.

Preprocessing

The PDF complaint files appear to be optical scans of word processor documents from the mid-2000s. In one case, the scanning process introduced footer information that wasn't present in the other documents. In general, extracting from the PDFs was experimental in nature, with a lot of human review and adjustments.

The challenge with the email processing was filtering the very large corpus based on metadata.

Analysis & Visualizations

With the PDFs, the analyses and visualizations did not provide additional useful information – their structure was already apparent. The word count would help with downstream training, but that is out of scope for Milestone 1. With a larger set of source material unsupervised techniques such as topic identification might be helpful.

With the email corpus, the analyses and visualizations proved necessary to locate content for extraction. We utilized key word searches to identify emails of interest. The key words were based on the content of the complaint PDFs and the facts of the cases.

Prompt / Response Extraction

We got lucky with the structure of the SEC complaint case PDFs. Their layout with Topic (Group Name) / Content (Paragraph Text) associations made the training data extraction easy. This could be enhanced by legal domain knowledge and more specific / nuanced extractions.

Email extraction was more difficult. The size of the corpus and the lack of structure from a content perspective forced us to rely on subject, keyword search, and human identification of Prompt / Response pairs. Sometimes when specific emails containing key word content were located, however, we were still uncertain how to word the Prompt.

Lessons Learned & Next Steps

It's important to know up front what your unstructured data sources bring in terms of to-be-structured knowledge and how you will get to Prompt / Response pairs.

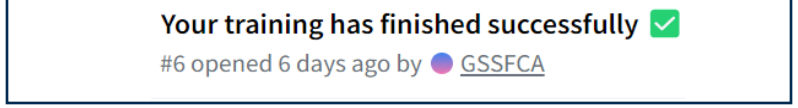
It's important to have the right end solution in mind, since this dictates the training data format. RAG preprocessing is very different from SFT.

A real-life project of this nature may require a much larger training set – necessitating use of many more sources and possibly much more human assistance.

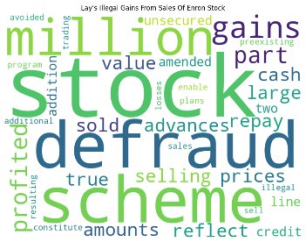
Using a PDF extract approach based on deep learning text recognition may be more effective, by extracting to a consistent intermediate file format before parsing the information.

Fine-Tuning Test Run: To prove to ourselves that we had correctly formatted our training dataset "train.csv", we used it to fine-tune an LLM with AutoTrain on Hugging Face (default parameters).

We still have a lot to learn about LLM fine-tuning. If we were to continue this effort, we would explore SFT model properties, training parameters, and any improvements we could make to the training data.



Work Statement & References



Project Environment:

We initially experimented with Google Colaboratory, Jira and Confluence and setup a kanban sprint/backlog/issue tracking system with integration to BitBucket. We ended up finding this more complex than needed for Milestone 1. We were able to keep the project moving adequately with a combination of Deepnote for coding collaboration and Slack for communication.

If we were to continue this project and take the environment up a level, from a collaboration and DevOps perspective we would like to explore integration of online file storage such as AWS, possible Docker use, Git for collaboration on code, and revisit a more formal project / work assignment and tracking tool such as Jira, incorporating agile principles.

Team and Tasks:

Gary Schaumburg

- PDF extraction
- PDF Analysis & Visualization
- Report draft

Venkat Panyam

- Email file handling
- Email extraction
- Email Analysis & Visualization

Samarth Somani

- Email Analysis & Visualization
- Email Prompt / Response extraction

1. Wikipedia contributors. (2024, February 4). Enron: The Smartest Guys in the Room. In *Wikipedia, The Free Encyclopedia*. Retrieved February 6, 2024, from https://en.wikipedia.org/w/index.php?title=Enron:_The_Smartest_Guys_in_the_Room&oldid=1203293550
2. Hugging Face. (n.d.). *LLM Finetuning*. AutoTrain documentation. Retrieved February 9, 2024, from https://huggingface.co/docs/autotrain/llm_finetuning
3. Alpin. (n.d.). *The Novice's LLM Training Guide*. Rentry.org., Retrieved February 9, 2024, from https://rentry.org/llm-training
4. June, Florian. (2024 February 2). Advanced RAG 02: Unveiling PDF Parsing. In *Towards AI* on Medium. Retrieved February 9, 2024, from https://pub.towardsai.net/advanced-rag-02-unveiling-pdf-parsing-b84ae866344e
5. Hosni, Youssef. (2024 January 20). LLM Researcher and Scientist Roadmap: A Guide to Mastering Large Language Models Research. In *Towards Data Science* on Medium. Retrieved February 9, 2024, from https://pub.towardsai.net/llm-researcher-and-scientist-roadmap-a-guide-to-mastering-large-language-models-research-bd179f873a21?sk=v2%2F9ad799ea-d23d-4996-b3b4-73372e781ec2
6. Catalan, Nati. (n.d.). Unleashing the Power of LLM Fine-Tuning, On *tasq.ai*, from https://www.tasq.ai/blog/unleashing-the-power-of-llm-fine-tuning/#
7. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language Models are Few-Shot Learners. https://arxiv.org/abs/2005.14165
8. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research, 21(140), 1-67. http://jmlr.org/papers/v21/20-074.html