Exploration of Open Large Language Models for eDiscovery

Sumit Pai[1]*, Sounak Lahiri[1]*, Ujjwal Kumar[1]*, Krishanu Das Baksi[1], Elijah Soba[2], Michael Suesserman[2], Nirmala Pudota[1], Jonathan Foster[2], Edward Bowen[2], Sanmitra Bhattacharya[2]

[1]Deloitte & Touche Assurance & Enterprise Risk Services India Private Limited, India
[2]Deloitte & Touche LLP, United States
{sumpai, sanmbhattacharya}@deloitte.com

Abstract

The rapid advancement of Generative Artificial Intelligence (AI), particularly Large Language Models (LLMs), has led to their widespread adoption for various natural language processing (NLP) tasks. One crucial domain ripe for innovation is the Technology-Assisted Review (TAR) process in Electronic discovery (eDiscovery). Traditionally, TAR involves manual review and classification of documents for relevance over large document collections for litigations and investigations. This process is aided by machine learning and NLP tools which require extensive training and fine-tuning. In this paper, we explore the application of LLMs to TAR, specifically for predictive coding. We experiment with out-of-the-box prompting and fine-tuning of LLMs using parameter-efficient techniques. We conduct experiments using open LLMs and compare them to commercially-licensed ones. Our experiments demonstrate that open LLMs lag behind commercially-licensed models in relevance classification using out-of-the-box prompting. However, topic-specific instruction tuning of open LLMs not only improve their effectiveness but can often outperform their commercially-licensed counterparts in performance evaluations. Additionally, we conduct a user study to gauge the preferences of our eDiscovery Subject Matter Specialists (SMS) regarding human-authored versus model-generated reasoning. We demonstrate that instruction-tuned open LLMs can generate high quality reasonings that are comparable to commercial LLMs.

1 Introduction

Electronic discovery (eDiscovery) (Oard et al., 2013) refers to the process of identifying, collecting, and preserving electronic documents and data for the purpose of legal proceedings, investigations, or regulatory compliance. During legal procedures, including litigation proceedings and corporate mergers and acquisitions, the court often issues a *production request*. The request mandates the parties involved to produce documents pertinent to the case. A large team of legal practitioners meticulously examines millions of digital records to identify the ones that are relevant to the production request. This step is a crucial phase of the eDiscovery process, commonly referred to as responsiveness determination. During this early step, legal practitioners often employ Artificial Intelligence (AI)-based tools to help them identify and prioritize the documents for review. This essential component of the eDiscovery process is referred to as Technology-Assisted Review (TAR) (Cormack and Grossman, 2014).

Another related domain focused on management and extraction of relevant information from electronic documents for legal purposes is Legal Information Retrieval (LIR). However, the important distinctions between eDiscovery and the general domain of LIR (Ganguly et al., 2023) are in their purpose, and nature of documents under consideration. eDiscovery is related to request for production of documents and data that are often in the form of email correspondences, text messages, articles, and financial declarations. In contrast, LIR is related to finding legal information in response to specific queries from legal databases, case law repositories, legislative archives, and other legal resources. The language used in the documents for these two tasks differ with the latter being richer in legal terminology. In this paper, our primary focus is on documents pertinent to eDiscovery.

Traditional approaches to TAR have focused on Boolean querying (Blair and Maron, 1985; Baron et al., 2007), information retrieval (Oard et al., 2013) and active learning methodologies (Cormack and Grossman, 2016; McDonald et al., 2018). In recent years, the use of supervised learning on labeled documents, referred to as *predictive cod-*

[*]These authors contributed equally to this work

ing (Brown, 2015; Yang et al., 2021), has gained more popularity within TAR workflows. Predictive coding often employs binary text classification based on textual, syntactic, semantic, and other data-driven features. Motivated by the application of LLMs in zero-shot and few-shot learning in various domains, including recommender systems and document annotation (Hou et al., 2023; Törnberg, 2023; Dai et al., 2022; Ahmed and Devanbu, 2022), we propose a novel approach to predictive coding using LLMs. Furthermore, we go beyond the predictive capabilities of LLMs, and leverage them to provide reasonings for the predictions. This capability can assist human reviewers in making more informed decisions. The use of LLMs for predictive coding and reasoning represents a paradigm shift in a domain burdened by the ever-increasing volume of documents and escalating review costs (Yang et al., 2021).

In this paper, we utilize open LLMs, notably Large Language Model Meta AI v2 (LLaMA2 - 13B and 70B versions) (Touvron et al., 2023) and Falcon-7B[1], as well as the commercially licensed Generative Pre-trained Transformer (GPT)-3.5-turbo model[2], for both predictive coding and reasoning. We explore various methods to use these LLMs effectively, including zero-shot prompting, fine-tuning for instruction following (i.e., instruction tuning), and reasoning generation. Additionally, we conduct a user study involving eDiscovery Subject Matter Specialists (SMSs) to evaluate the quality of reasoning generated by these models, and those authored by a legal SMS. The publicly available Enron email dataset (Grossman et al., 2011) is used in our experiments. This dataset comprises of a large corpus of emails and attachments that had to be reviewed for their responsiveness to production requests. We present experimental results of the various LLM-based predictive coding approaches on this dataset.

The paper presents related research in Section 2, describes the datasets used in Section 3, outlines the experimental approaches in Section 4, and reports the results of classification models as well as user preferences for generated reasonings in Section 5.

2 Related Work

The application of AI in eDiscovery for the purposes of document prioritization, reasoning, and decision-making are not a recent development (Araszkiewicz et al., 2022; Ashley and Bridewell, 2010; Conrad, 2010). Notably, the Text Retrieval Conference (TREC) had consistently featured a dedicated Legal track from 2006 to 2011 (Baron et al., 2006; Tomlinson et al., 2007; Oard et al., 2008; Hedin et al., 2009; Cormack et al., 2010; Grossman et al., 2011), where eDiscovery was given prime importance, with emphasis on retrieval and ranking-based approaches. Several shared tasks related to eDiscovery, including TAR and privilege review, were organized to stimulate and advance research in this specialized domain.

Typical TAR solutions can be decomposed into retrieval-based (Oard et al., 2013) and classification-based approaches (Barnett et al., 2009) (Lewis, 2010). Both of these approaches are instrumental in facilitating the prioritization of documents for subsequent review processes. Alongside these methods, active learning strategies (Cormack and Grossman, 2016; McDonald et al., 2018) have been extensively explored to boost the operational efficiency of these models. These strategies incorporate relevance feedback provided by human reviewers to enhance model performance. However, it's imperative to acknowledge that certain feedback-based processes can be inherently slow, expensive and inconsistent. In this paper, we primarily focus on the classification-based approach, also referred to as predictive coding, which has gained prominence in recent years.

Conventional TAR approaches are often perceived as black box systems, making it challenging to trust their decisions. The generation of reasoning and explanation for TAR models has received limited attention. Previous approaches (Chhatwal et al., 2018; Villata et al., 2020) focused on training models to identify snippets (at sentence level) within documents that are classified as either responsive or non-responsive. However, these approaches necessitate training on annotated data, which is both time-consuming and resource-intensive.

In this work, we focus on leveraging the capabilities of generative LLMs for both classification and reasoning. These architectures, rooted in the Transformer framework (Vaswani et al., 2017), are characterized by their auto-regressive decoders. Typi-

[1] https://falconllm.tii.ae/

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

cally, these models are trained on vast and diverse corpora of textual data, demonstrating remarkable proficiency in comprehending natural language instructions across diverse domains. These LLMs are able to generate coherent natural language responses when provided with appropriate prompts (Liu et al., 2023). They can interpret the topics specified in the prompts and execute them effectively. In this study, we experiment with some open LLMs, such as LLaMA2 (13B and 70B versions) and Falcon-7B, as well as commercially licensed models like GPT-3.5-turbo.

3 Dataset

In this section, we describe the dataset used for our experiments. We used the Electronic Discovery Reference Model (EDRM) Enron Email Data Set Version 2, which is the post-processed version employed in the TREC 2011 Legal eDiscovery track (Grossman et al., 2011). This dataset serves as a rich resource for exploring various facets of eDiscovery, particularly in the context of email communication.

3.1 Source and Composition

The EDRM Enron Email Data Set Version 2 originates from the Enron Corporation, once a leading energy company in the United States that collapsed in 2001 due to accounting malpractices. The dataset was used in the TREC Legal eDiscovery track from 2009 to 2011. The post-processed dataset is meticulously organized into 159 directories, each capturing the email communication records of individuals associated with the company, offering a distinct insight into corporate communication.

3.2 Formats

The dataset is divided into two primary categories:

- **Emails**: This subset comprises a substantial portion of the dataset, with a total of 455,449 email messages. Each email has a distinct document ID and is stored as plain text with the naming convention `doc_id.txt`. These emails are central to our analyses and experiments.

- **Attachments**: In addition to the emails, the dataset contains 230,143 attachment files. Each attachments is linked to a specific email message and follows the naming format `doc_id.number.txt`, where `number` indicates its sequential order.

Attachments are treated as separate documents.

3.3 Experimental Focus

Following the specifications of the Learning Task of the TREC 2011 Legal Track, we focused on three available topics, numbered 401, 402, and 403. Detailed description of each topic is provided in Appendix A.1. The task organizers offered a choice to use either emails, attachments, or both. Evaluations of the submitted runs for this shared task encompassed these three options. In our experiments, we concentrate on analyzing emails from each of the aforementioned topics. For simplicity and consistency, attachments were intentionally excluded due to their potentially large size and varied formats, such as spreadsheets, presentations, and webpages.

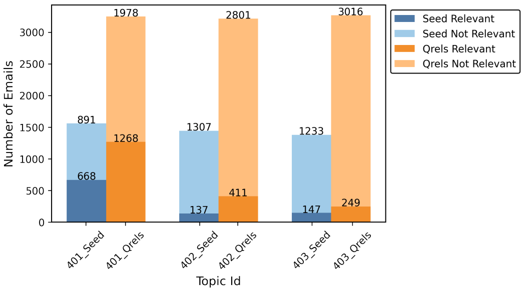


Figure 1: Distribution of emails across the seed dataset and qrels. The color-coded stacked bars show the number of relevant and non-relevant emails for each topic.

3.4 Training and Testing Data

For our experiments, we used two primary components from the Learning Task of the TREC 2011 Legal Track:

- **Seeds**: The seed data was used to fine-tune the LLMs to adapt to the eDiscovery domain, especially in the context of Enron's email communications.

- **Qrels**: The query relevance judgments (qrels) served as our test set. This dataset allowed us to evaluate the effectiveness of the LLMs in identifying relevant documents within the Enron Dataset for the given topics.

Figure 1 shows the distribution of emails across the three topics used in our study. The seed data and qrels for each topic are highly imbalanced with a majority of documents being non-relevant.

We focused on topics from the 2011 TREC Legal track, rather than prior years (i.e. 2009 and 2010 TREC Legal tracks), for two main reasons. Firstly, each topic from this year featured a relatively larger number of labeled documents in both the seed and qrel sets, and showcased a diverse range of label distributions. For instance, Topic 401 displayed a fairly balanced distribution between relevant and non-relevant documents, whereas other topics exhibited significant imbalances. Secondly, the dataset from this year provided a sufficient number of topic statements to effectively evaluate the capability and generalizability of our approach when applied to predictive coding.

4 Methodology

In this study, we primarily focus on examining several generative models from the decoder-only auto-regressive family, a subset of the broader category of LLMs. These models have gained significant attention and adoption following the introduction of ChatGPT[3]. Legal departments and law firms have expressed interest in utilizing these models to enhance their document review processes, achieving both scalability and cost-efficiency. We explore four models: Falcon (7B version), LLaMA2 (both 13B and 70B versions), and GPT (3.5 turbo version). Our experimentation involves two distinct methodologies: out-of-the-box (OOB) prompting approach and fine-tuning of open LLMs, with a special emphasis on Falcon 7B and LLaMA2 13B, which have been shown to outperform other open LLMs of similar size across various benchmarks[4]. We detail these methodologies in the subsequent sections.

4.1 Prompt Engineering

A common method to interact with a decoder-based auto-regressive LLMs is through *prompts*. A prompt is a natural language instruction that combines topic specifications, contextual information, and input parameters for executing the specified tasks. The LLM interprets the task described in the prompt and generates corresponding responses. Prompting greatly improves the usability of decoder-only models, allowing users without technical knowledge to effectively interact with them using just natural language. Without the knowledge of the technical intricacies, such as coding or statistical methodologies, one can easily interact with these models using this zero-shot approach. The art of crafting effective prompts has paved the way for a new area of research known as *prompt engineering*. Prompt engineering (Liu et al., 2023) is an iterative process which entails the creation, evaluation, and refinement of prompts, all aimed at enhancing performance outcomes for the targeted topic.

In this paper, we present a systematic approach towards prompt engineering grounded on performance evaluation on a sample set of qrels. We started with an initial prompt for each topic, and iteratively refined it to derive a final optimized prompt. To ensure that the prompt template, topic statement, and email content all fit within the context length limitations of the LLMs used in this study, we truncated each emails to its first 300 tokens (emails from our dataset have a median word count of < 50). To evaluate the effectiveness of incremental updates to the prompts, we computed macro F1 score on a randomly chosen subset of 50 emails from the qrels for each topic, and then compared the results through each iteration. The prompt which gave us optimal performance on this smaller prompt-evaluation subset was then applied on the entire set of qrels (test set) for a comprehensive evaluation. In the early stages of prompt engineering, our primary emphasis was on classification metrics. However, after achieving optimal performance, we shifted our focus to producing high-quality reasoning, achieved by providing the model with explicit guidance through the prompt.

4.2 Instruction Tuning

Instruction tuning is the process of fine-tuning LLMs to follow instructions for targeted tasks (Wei et al., 2021). The process necessitates the use of an instruction dataset, comprised of instruction-output pairs that function as the training data. Within the scope of the predictive coding application, our primary objective is to evaluate the relevance of emails in relation to topic statements (typically characterized as production requests). To facilitate this, instructions or prompts are crafted in simple English by describing the topic of relevance determination, and providing the email and topic statement, and finally providing the expected output in plain text

[3] https://chat.openai.com/

[4] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

– either ‘Yes’ or ‘No’ – based on the email’s relevance to the given topic. To fine-tune the models, we create the prompt and its corresponding output using the seed set from our dataset. The evaluation is conducted on the qrels.

LLMs can be instruction tuned through various methods. While a typical approach involves full fine-tuning, it often comes with significant infrastructure requirements. As an alternate, parameter-efficient techniques like Low Rank Adaptation (LoRA) (Hu et al., 2021) and Quantized LoRA (qLoRA) (Dettmers et al., 2023) offer a way to fine-tune on relatively modest infrastructure, while maintaining performance that’s comparable to full fine-tuning (Liu et al., 2022). This can be further augmented by aligning the model with human preferences using Reinforcement learning with Human Feedback (RLHF). However, RLHF demands a vast collection of manually curated preference data. In this paper, our primary delve into the exploration of LoRA and qLoRA. Their minimal infrastructure requirements and cost-effectiveness make these techniques especially fitting for our research objectives.

LoRA is a fine-tuning methodology that decomposes the model’s weight matrix into a low-rank approximation. This approximation is subsequently trained on topic-specific data, capturing nuanced, topic-specific intricacies while the original weights are kept frozen. As a result, this approach introduces fewer trainable parameters, requiring significantly less compute power. It also requires a lower volume of training data compared to more extensive fine-tuning approaches.

Upon the completion of the learning process, the resultant LoRA adapter weights can be merged with the original LLM weights. This produces an updated set of weights for subsequent inference. LoRA adapters can either be integrated across the model weight layers or be selectively applied to specific layers, offering flexibility in adaptation. This choice dictates the number of trainable parameters utilized during the learning process. We conducted a hyperparameter search on various LoRA parameters, to identify the optimal combination of these values for each topic statement.

qLoRA is a fine-tuning technique designed to reduce the Graphics Processing Unit (GPU) memory requirements of a model by employing weight quantization. In essence, this method involves converting the base model’s wide range of weight values into a more compact range through min-max scaling. As a result, the transformed weights require less memory for storage. To illustrate this, consider the storage of integers ranging from -128 to 127, which requires 8 bits of storage. However, by mapping these values to a narrower range, like -16 to 15, only 4 bits are needed. This reduction in bit width effectively decreases the memory footprint by a factor of 2. When applied across the model’s weight matrix, this quantization leads to significant memory savings. As an example, if the initial model occupies 50GB of memory, the quantized model only needs 25GB of GPU memory. Importantly, the original weights can be restored by implementing an inverse scaling operation, albeit with a marginal loss of information.

4.3 Explanation Generation

One notable advantage of using LLMs over traditional text classification-based models is their ability to generate explanations. By providing appropriate prompts, these models can generate coherent reasoning to explain model predictions. Smaller LLMs, such as LLaMA2-13B, typically exhibit limited proficiency in reasoning through OOB prompting, especially when compared to their larger counterparts like LLaMA2-70B (Wei et al., 2022). These smaller models are often less verbose and tend to repetitively reiterate the topic statement.

Nonetheless, there’s potential to improve the reasoning capabilities of these models through instruction tuning. Such improvement demands access to annotated reasoning data, which is not readily available for the Enron emails dataset. In this dataset, our seed set is restricted to binary labels (yes/no) of relevance. To address this limitation, we utilize the LLaMa2-70B model with OOB prompting to generate reasoning on the seed set. Subsequently, the generated reasoning (on correctly predicted labels) serves as the foundation for instruction tuning of the LLaMA2-13B model for improved explanation generation.

4.4 User Study

We conducted a user study to evaluate the quality of reasoning that different models generated concerning the relevance of emails to given topic statements. For this assessment, we curated a set of 20 distinct predictions by randomly sampling from the qrels. We focused solely on cases where every model made accurate predictions. We created a questionnaire that was distributed to five eDiscovery SMSs who has knowledge in the domain of legal document review.

The primary objective of this study was to gauge human preferences for reasoning generated by different models. Each annotator was presented with a question comprising the email text, the topic statement, the actual label, and five separate reasoning outputs generated by five different techniques. A snapshot of this survey is shown in Appendix A.2.

Of the five techniques used, one reasoning instance was crafted by a SMS from our eDiscovery team (not involved in the subsequent survey). The other four were generated using the following methods: LLaMA2-13B OOB, LLaMA2-70B OOB, LLaMA2-13B fine-tuned, and GPT-3.5 OOB. To reduce biases, these reasoning outputs were presented to annotators in a random sequence. Annotators were then asked to rate each reasoning on a scale of 1 to 5, where 5 denoted the top preference and 1 the lowest. Annotators were urged to avoid assigning tied scores whenever possible, aiming to derive a clear ranking indicative of preference.

Interestingly, the 13B fine-tuned model occasionally generated outputs similar to those of the 70B OOB model. This overlap was expected since the 70B OOB model was used to create the ground truth data for the 13B model. Consequently, occasional ties in the reasoning scores were anticipated. After gathering the preference data, we analysed the results and ranked the models based on the preferences of the SMSs.

5 Results

In this section, we present the results of applying LLMs to the predictive coding process. The training and evaluation were conducted on the seed data and qrels for topics 401, 402, and 403 from the TREC 2011 Legal track. Model performance was evaluated on the qrels using standard classification metrics, such as precision, recall, and macro F1 scores.

5.1 Prompt Engineering

Table 1 shows the initial prompt, which through several iterations of performance improvements on a sample set of qrels, was refined to derive the final prompt. For the sake of brevity, we focus exclusively on results for the LLaMA2-13B model. In the final prompt, several elements collectively contribute to achieving the improved outcome: the strategic emphasis on distinct terms such as “email”, “topic”, and “relevant”, the consistent use of these terms throughout the prompt, a clear demarcation between inputs and their surrounding context, precise specification of the desired output structure, and intentional guidance provided to the model for task execution.

The results of OOB prompting for Falcon-7B, LLaMA2-13B, LLaMA2-70B, and GPT-3.5 are presented in the top third of Table 2. The results clearly show that GPT-3.5 outperforms other open LLMs when used through OOB prompting (even with the more effective model-specific prompt engineering).

We make an interesting observation that a chosen prompt for one model might not produce favorable results when used with other models. As a result, we crafted prompts tailored to each model. The underlying rationale behind these varied outcomes are still a subject of ongoing research and fall outside the scope of this paper.

In terms of GPU memory requirements, we were able to conduct OOB inference on two 40GB A100 GPUs for both Falcon-7B and LLaMA2-13B. However, for the LLaMA2-70B model, a more substantial hardware configuration with eight 40GB A100 GPUs was required.

5.2 Instruction Tuning

We limited our model selection for instruction tuning to those compatible with two 40GB A100 GPUs. Consequently, the instruction tuning process was solely performed on the Falcon-7B and LLaMa2-13B models. Due to the significant cost of instruction tuning for GPT-3.5, we opted not to include it in this study. The instruction tuning process followed the methodologies outlined in Section 4.2, and the improved results achieved after extensive hyperparameter tuning (described in Appendix A.3) are detailed in the last four rows of Table 2.

In order to instruction tune our models, we used the improved prompt described in the preceding subsection. This fine-tuning process utilized the seed set and was subsequently evaluated against the qrels. As shown in Table 2, for both LLaMA2-13B and Falcon-7b models, the LoRA instruction tuned models outperform their OOB counterparts in determining relevancy. This is quite significant for topic 401 as the seed set was sufficiently balanced. For Falcon, we saw an improvement from an F1 score of 0.39 to 0.79, while for LLaMA2-13B this

You are a subject matter expert reviewing a document to evaluate if it is related to a topic. Respond with Yes if the document is directly or indirectly related to the topic or No if not related. On a new line, give a reason. *Topic: topic_statement* *Email: email_text*	*As a subject matter expert, youQr task is to read the following email and determine from its contents whether it is related to the provided topic.* *Email: """email_text"""* *Topic: """topic_statement"""* *Task: Decide whether the email is related to the topic or not. Provide a simple 'yes' or 'no' answer. Additionally, give a reason to support your answer, by summarizing parts of email that helped you make this decision.* *Answer and Reason:*
(a)	(b)

Table 1: The initial prompt (a) and the final improved prompt (b) were iteratively developed for the LLaMA2-13B model. An evaluation was conducted using a random sample of 50 questions spanning the three topics. The initial prompt yielded a macro average F1 score of 0.29, whereas the improved prompt achieved a significantly improved F1 score of 0.51. Improved prompt (b) exhibits a deliberate emphasis on specific terms, such as "email", "topic", and "related", which are consistently employed throughout the prompt. Notably, the email and topic statements are distinctly delineated by the use of triple quotes. Additionally, a structured output format is outlined. The model is directed to provide its reasoning by "summarizing parts of the emails". This iterative refinement process represents a systematic and methodical approach to enhance the prompt, resulting in improved performance for classification and reasoning.

		401			402			403		
Type	**Model**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**
OOB	Falcon 7B	0.39	0.51	0.50	0.33	0.48	0.46	0.28	0.47	0.42
	LLaMa v2 13B	0.56	0.56	0.56	0.54	0.55	0.59	0.47	0.52	0.56
	LLaMa v2 70B	0.58	0.71	0.66	0.63	0.69	0.61	0.66	0.66	0.65
	GPT 3.5	0.61	0.76	0.62	**0.66**	**0.82**	0.62	0.66	0.64	0.68
LoRA	Falcon 7B	0.79	0.79	0.80	0.60	0.59	0.60	0.60	0.59	0.66
	LLaMa v2 13B	**0.85**	**0.84**	**0.86**	0.59	0.59	**0.66**	**0.68**	**0.68**	0.67
qLoRA	Falcon 7B	0.74	0.79	0.78	0.60	0.76	0.58	0.53	0.58	**0.77**
	LLaMa v2 13B	0.69	0.71	0.69	0.62	0.61	**0.66**	0.63	0.61	0.69

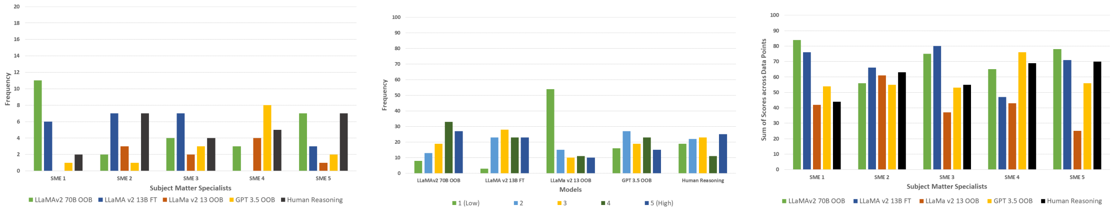
Table 2: Comparison of LLM Performance Metrics on TREC 2011 Topics 401, 402, and 403 for the assessment of emails relevance to topic statements. The uppermost four rows pertain to Out-of-Box (OOB) prompting, the subsequent two rows are dedicated to LoRA-based fine-tuned models, and the final two rows concern qLoRA fine-tuned models.

jumped from 0.56 to 0.85. Both these models, surpassed GPT-3.5-turbo OOB results on this topic statement. These results empirically highlight the advantages of LoRA instruction tuning. Similarly, for qLoRA finetuned models, we see a performance improvement compared to OOB prompting across all the topics. However, this is not as significant as LoRA finetuned models.

5.3 Quality of reasoning

The summary of the user study for evaluating the quality of reasoning can be found in Fig 3. The figure shows the mean score obtained by different techniques that were used for reasoning generation. Notably, of the five reasoning techniques assessed, the rationale generated by the LLaMA2-70B OOB model was the one preferred by the SMSs of our eDiscovery team, achieving a mean preference score of 3.58 with a standard error of 0.12. On the other hand, the reasoning produced by the LLaMA2-13B OOB model was the least favored, as reflected by its mean score of 2.08 and a standard error of 0.14. Furthermore, the reasoning generated by the LLaMA2-70B OOB and LLaMA2-13B fine-tuned models were preferred over SMS generated reasoning, which had a mean score of 3.01 and standard error of 0.15.

It is interesting to observe that the reasoning capability of the LLaMA2-13B model significantly improved after fine-tuning, as supported by the



(a)

(b)

(c)

Figure 2: Frequency distribution of reasoning preference scores ascribed to various techniques by the 5 Subject Matter Specialists (SMSs) engaged in the survey. A total of 20 email samples were presented to each SMSs, along with their relevance to a topic (401). They were provided with five distinct rationales, generated by five distinct techniques, in support of the email's topical relevance. SMSs were instructed to score the reasonings on a scale of 1-5, wherein 5 signified the highest preference and 1 indicated the lowest preference. 2(a) shows the number of times a reasoning generation technique is ranked at the first position by SMSs. 2(b) shows the sum of scores received by each technique on the 20 questions for each SMSs. 2(c) shows the distribution of scores received by each model across the SMSs and questions.

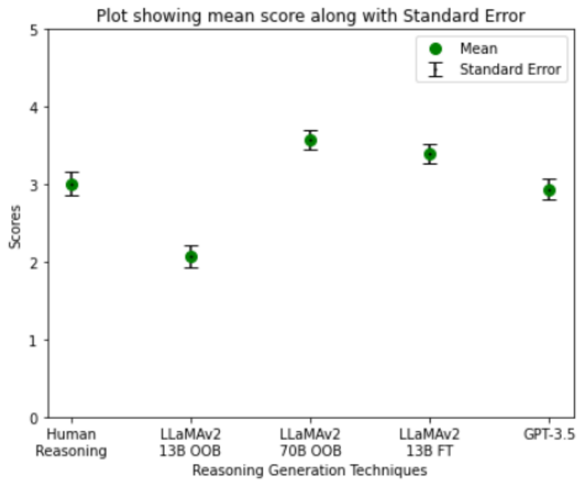


Figure 3: Mean and standard error of reasoning preference scores for each model as rated by SMSs

SMS preferences. This improvement is reflected in a mean score of 3.4 and a standard error of 0.12, further corroborating our initial hypothesis that finetuning contributes to improved performance. Surprisingly, the average preference score for GPT-3.5 was notably lower, coming in at 2.94 with a standard error of 0.13.

Detailed results of the survey are shown in Figure 2. Figure 2(a) provides insights into the frequency with which a technique was ranked highest by a SMS. Figure 2(b) shows the cumulative scores accrued by various techniques during when assessed by different SMSs. Figure 2(c) displays the distribution of scores that individual models received from the entire pool of SMSs and questions. Notably, of all the techniques reviewed, the LLaMA2-13B OOB model consistently registered a low score in a majority of evaluations conducted by the SMSs.

6 Conclusion

This study is among of the first works to assess the efficacy of generative LLMs in the eDiscovery document review process. We compare the classification performance of various open and commercially-licensed LLMs, and demonstrate that although OOB performance of open LLMs are worse compared to GPT-3.5, these models can be finetuned to achieve comparable results to GPT-3.5. Moreover, we conduct a user study and show that the SMSs favored AI-generated reasoning over human reasoning, underscoring the viability of these approaches in eDiscovery.

In terms of scalability, we suggest that there is no inherent need to deploy large models such as GPT-3.5 or LLaMa2-70B for production purposes. The LLaMA2-13B model, which fits on just two 40GB A100 GPUs, can be fine-tuned to yield similar classification performance and reasoning that rivals human reasoning – as supported by our user study. This approach offers significant savings in infrastructure costs associated with model deployment.

7 Future Work

In future, we aim to extend the scope of this work by including all the TREC topics from 2009 to 2011. Additionally, we intend to enhance our framework by including attachments, allowing for a deeper analysis. We also plan to undertake the prioritization task, wherein we focus on computing ranking metrics such as F1@k, precision@k, recall@k, where k represents the number of documents reviewed. This will allow us to conduct a thorough examination of model performance across the Enron emails corpus and in turn help us identify the specific 'k' at which recall surpasses certain court-mandated thresholds.

Furthermore, we plan to conduct a large-scale user study, involving a broader cohort of annotators and an expanded array of survey questions. We also plan to assess the degree of inter-annotator agreement within this extended study, employing established metrics such as Cohens Kappa (Smeeton, 1985) or Krippendorf Alpha (Gwet, 2011) to quantify the level of consensus among annotators.

References

Toufique Ahmed and Premkumar Devanbu. 2022. Few-shot training llms for project-specific code-summarization. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, pages 1–5.

Michał Araszkiewicz, Trevor Bench-Capon, Enrico Francesconi, Marc Lauritsen, and Antonino Rotolo. 2022. Thirty years of artificial intelligence and law: overviews. *Artificial Intelligence and Law*, 30(4):593–610.

Kevin D Ashley and Will Bridewell. 2010. Emerging ai & law approaches to automating analysis and retrieval of electronically stored information in discovery proceedings. *Artificial Intelligence and Law*, 18:311–320.

Thomas Barnett, Svetlana Godjevac, Jean-Michel Renders, Caroline Privault, John Schneider, and Robert Wickstrom. 2009. Machine learning classification for document review. In *DESI III: The ICAIL Workshop on Globaal E-Discovery/E-Disclosure*. Citeseer Princeton, NJ, USA.

Jason R Baron, R Braman, K Withers, T Allman, M Daley, and G Paul. 2007. The sedona conference® best practices commentary on the use of search and information retrieval methods in e-discovery. In *The Sedona conference journal*, volume 8.

Jason R Baron, David D Lewis, and Douglas W Oard. 2006. Trec 2006 legal track overview. In *TREC*. Citeseer.

David C Blair and Melvin E Maron. 1985. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM*, 28(3):289–299.

Shannon Brown. 2015. Peeking inside the black box: A preliminary survey of technology assisted review (tar) and predictive coding algorithms for ediscovery. *Suffolk J. Trial & App. Advoc.*, 21:221.

Rishi Chhatwal, Peter Gronvall, Nathaniel Huber-Fliflet, Robert Keeling, Jianping Zhang, and Haozhen Zhao. 2018. Explainable text classification in legal document review a case study of explainable predictive coding. In *2018 IEEE international conference on big data (Big Data)*, pages 1905–1911. IEEE.

Jack G Conrad. 2010. E-discovery revisited: the need for artificial intelligence beyond information retrieval. *Artificial Intelligence and Law*, 18(4):321–345.

Gordon V Cormack and Maura R Grossman. 2014. Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 153–162.

Gordon V Cormack and Maura R Grossman. 2016. Scalability of continuous active learning for reliable high-recall text classification. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 1039–1048.

Gordon V Cormack, Maura R Grossman, Bruce Hedin, and Douglas W Oard. 2010. Overview of the trec 2010 legal track. In *TREC*.

Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B Hall, and Ming-Wei Chang. 2022. Promptagator: Few-shot dense retrieval from 8 examples. *arXiv preprint arXiv:2209.11755*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Debasis Ganguly, Jack G Conrad, Kripabandhu Ghosh, Saptarshi Ghosh, Pawan Goyal, Paheli Bhattacharya, Shubham Kumar Nigam, and Shounak Paul. 2023. Legal ir and nlp: the history, challenges, and state-of-the-art. In *European Conference on Information Retrieval*, pages 331–340. Springer.

Maura R. Grossman, Gordon V. Cormack, Bruce Hedin, and Douglas W. Oard. 2011. Overview of the TREC 2011 legal track. In *Proceedings of The Twentieth Text REtrieval Conference, TREC 2011, Gaithersburg, Maryland, USA, November 15-18, 2011*, volume 500-296 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Kilem L Gwet. 2011. On the krippendorff's alpha coefficient. *Manuscript submitted for publication. Retrieved October*, 2(2011):2011.

Bruce Hedin, Stephen Tomlinson, Jason R Baron, and Douglas W Oard. 2009. Overview of the trec 2009 legal track. In *TREC*.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2023. Large language models are zero-shot rankers for recommender systems. *arXiv preprint arXiv:2305.08845*.

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.

David D Lewis. 2010. Afterword: data, knowledge, and e-discovery. *Artificial Intelligence and Law*, 18(4):481–486.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pretrain, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Graham McDonald, Craig Macdonald, and Iadh Ounis. 2018. Active learning strategies for technology assisted sensitivity review. In *European Conference on Information Retrieval*, pages 439–453. Springer.

Douglas W Oard, Björn Hedin, Stephen Tomlinson, and Jason R Baron. 2008. Overview of the trec 2008 legal track. In *TREC*, pages 500–277.

Douglas W Oard, William Webber, et al. 2013. Information retrieval for e-discovery. *Foundations and Trends® in Information Retrieval*, 7(2–3):99–237.

Nigel C. Smeeton. 1985. Early history of the kappa statistic. *Biometrics*, 41(3):795–795.

Stephen Tomlinson, Douglas W Oard, Jason R Baron, and Paul Thompson. 2007. Overview of the trec 2007 legal track. In *TREC*.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

S Villata et al. 2020. Sentence embeddings and high-speed similarity search for fast computer assisted annotation of legal documents. In *Legal Knowledge and Information Systems: JURIX 2020: The Thirty-third Annual Conference, Brno, Czech Republic, December 9-11, 2020*, volume 334, page 164. IOS Press.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Eugene Yang, David D. Lewis, and Ophir Frieder. 2021. On minimizing cost in legal document review workflows. In *Proceedings of the 21st ACM Symposium on Document Engineering*. ACM.

A Appendix

A.1 Topic Statements

The three topics, numbered 401, 402, and 403, used in the TREC 2011 Legal Track learning task are described below:

A.1.1 Topic 401

All documents or communications that describe, discuss, refer to, report on, or relate to the design, development, operation, or marketing of enrononline, or any other online service offered, provided, or used by the Company (or any of its subsidiaries, predecessors, or successors-in-interest), for the purchase, sale, trading, or exchange of financial or other instruments or products, including but not limited to, derivative instruments, commodities, futures, and swaps.

A.1.2 Topic 402

All documents or communications that describe, discuss, refer to, report on, or relate to whether the purchase, sale, trading, or exchange of over-the-counter derivatives, or any other actual or contemplated financial instruments or products, is, was, would be, or will be legal or illegal, or permitted or prohibited, under any existing or proposed rule(s), regulation(s), law(s), standard(s), or other proscription(s), whether domestic or foreign.

A.1.3 Topic 403

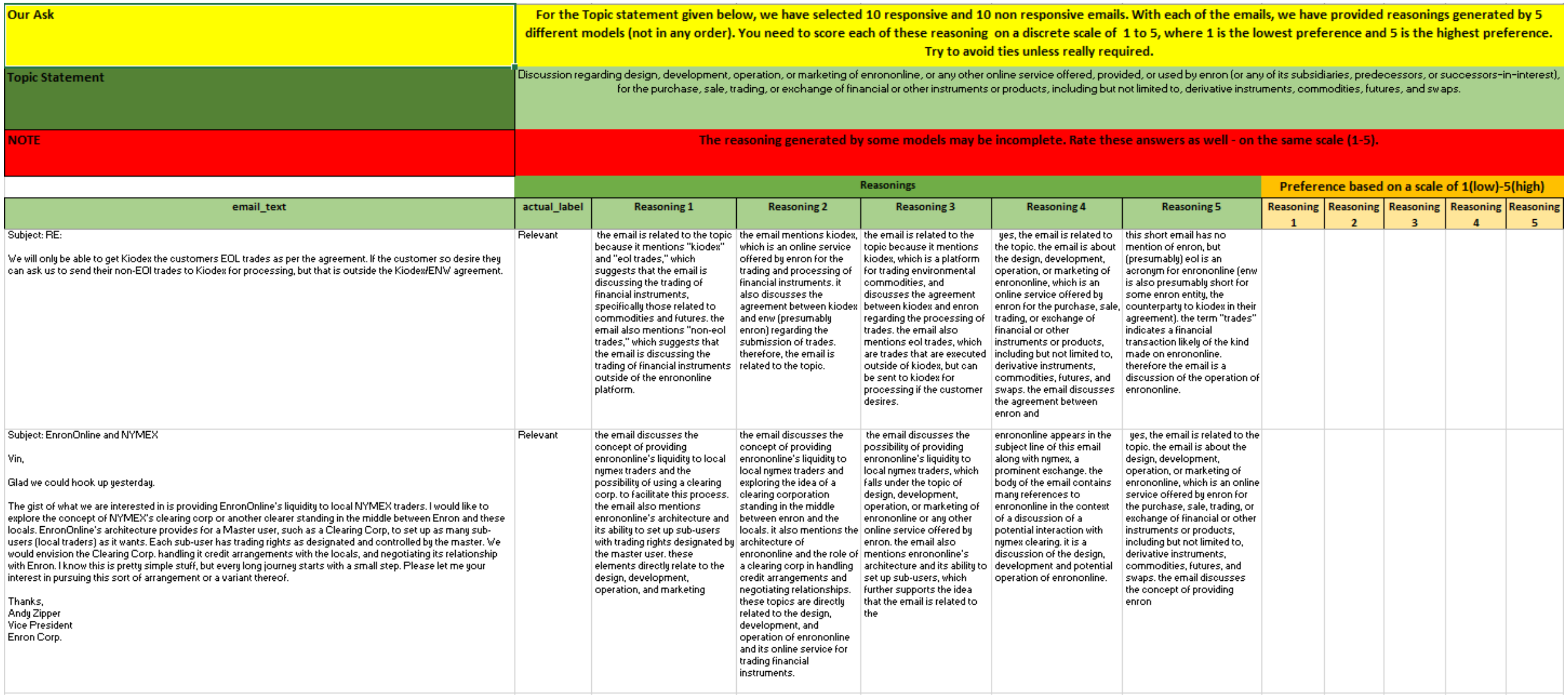
All documents or communications that describe, discuss, refer to, report on, or relate to the environmental impact of any activity or activities undertaken by the Company, including but not limited to, any measures taken to conform to, comply with, avoid, circumvent, or influence any existing or proposed rule(s), regulation(s), law(s), standard(s), or other proscription(s), such as those governing environmental emissions, spills, pollution, noise, and/or animal habitats.

A.2 User Survey

Figure 4 shows the snapshot of the survey spreadsheet provided to the SMSs. The first part contains the instructions and the topic statement. These panes were frozen while the annotator could scroll through the rest of the data. Each row consisted of an email content, followed by the relevancy to the topic statement, and the five reasonings which the annotators had to rate. The reasonings were randomly shuffled to prevent annotator bias.

A.3 Hyperparameter Ranges

We conducted an extensive hyperparameter search to tune model parameters for instruction tuning. We primarily tuned LoRA parameters such as rank (r), scaling factor (alpha), target modules (the layers where the adapters are inserted), learning rate, and number of epochs. For the rank parameter, we considered values of 4, 8, 16, and 32. The scaling factor was checked for values of 16 and 32. For adapter placement, we examined insertion into one or more of the following layers: query-key-value, attention output, upsampling, or downsampling. The learning rate was varied across three orders of magnitude: 1e-3, 1e-4, and 1e-5. Similarly, the number of epochs was chosen from the range of 1 to 4. To ensure a detailed exploration of the hyperparameter space, we randomly sampled 16 distinct combinations from these ranges for each model. All the qLoRA experiments were done in 4 bit precision with same hyperparameter ranges as described above.



Our Ask	For the Topic statement given below, we have selected 10 responsive and 10 non responsive emails. With each of the emails, we have provided reasonings generated by 5 different models (not in any order). You need to score each of these reasoning on a discrete scale of 1 to 5, where 1 is the lowest preference and 5 is the highest preference. Try to avoid ties unless really required.										
Topic Statement	Discussion regarding design, development, operation, or marketing of enrononline, or any other online service offered, provided, or used by enron (or any of its subsidiaries, predecessors, or successors-in-interest), for the purchase, sale, trading, or exchange of financial or other instruments or products, including but not limited to, derivative instruments, commodities, futures, and swaps.										
NOTE	The reasoning generated by some models may be incomplete. Rate these answers as well - on the same scale (1-5).										
	Reasonings						Preference based on a scale of 1(low)-5(high)				
email_text	actual_label	Reasoning 1	Reasoning 2	Reasoning 3	Reasoning 4	Reasoning 5	Reasoning 1	Reasoning 2	Reasoning 3	Reasoning 4	Reasoning 5
Subject: RE: We will only be able to get Kiodex the customers EOL trades as per the agreement. If the customer so desire they can ask us to send their non-EOl trades to Kiodex for processing, but that is outside the Kiodex/ENW agreement.	Relevant	the email is related to the topic because it mentions "kiodex" and "eol trades," which suggests that the email is discussing the trading of financial instruments, specifically those related to commodities and futures. the email also mentions "non-eol trades," which suggests that the email is discussing the trading of financial instruments outside of the enrononline platform.	the email mentions kiodex, which is an online service offered by enron for the trading and processing of financial instruments. it also discusses the agreement between kiodex and enw (presumably enron) regarding the submission of trades. therefore, the email is related to the topic.	the email is related to the topic because it mentions kiodex, which is a platform for trading environmental commodities, and discusses the agreement between kiodex and enron regarding the processing of trades. the email also mentions eol trades, which are trades that are executed outside of kiodex, but can be sent to kiodex for processing if the customer desires.	yes, the email is related to the topic. the email is about the design, development, operation, or marketing of enrononline, which is an online service offered by enron for the purchase, sale, trading, or exchange of financial or other instruments or products, including but not limited to, derivative instruments, commodities, futures, and swaps. the email discusses the agreement between enron and	this short email has no mention of enron, but (presumably) eol is an acronym for enrononline (enw is also presumably short for some enron entity, the counterparty to kiodex in their agreement). the term "trades" indicates a financial transaction likely of the kind made on enrononline. therefore the email is a discussion of the operation of enrononline.					
Subject: EnronOnline and NYMEX Vin, Glad we could hook up yesterday. The gist of what we are interested in is providing EnronOnline's liquidity to local NYMEX traders. I would like to explore the concept of NYMEX's clearing corp or another clearer standing in the middle between Enron and these locals. EnronOnline's architecture provides for a Master user, such as a Clearing Corp, to set up as many sub-users (local traders) as it wants. Each sub-user has trading rights as designated and controlled by the master. We would envision the Clearing Corp. handling it credit arrangements with the locals, and negotiating its relationship with Enron. I know this is pretty simple stuff, but every long journey starts with a small step. Please let me your interest in pursuing this sort of arrangement or a variant thereof. Thanks, Andy Zipper Vice President Enron Corp.	Relevant	the email discusses the concept of providing enrononline's liquidity to local nymex traders and the possibility of using a clearing corp. to facilitate this process. the email also mentions enrononline's architecture and its ability to set up sub-users with trading rights designated by the master user. these elements directly relate to the design, development, operation, and marketing	the email discusses the concept of providing enrononline's liquidity to local nymex traders and exploring the idea of a clearing corporation standing in the middle between enron and the locals. it also mentions the architecture of enrononline and the role of a clearing corp in handling credit arrangements and negotiating relationships. these topics are directly related to the design, development, and operation of enrononline and its online service for trading financial instruments.	the email discusses the possibility of providing enrononline's liquidity to local nymex traders, which falls under the topic of design, development, operation, or marketing of enrononline or any other online service offered by enron. the email also mentions enrononline's architecture and its ability to set up sub-users, which further supports the idea that the email is related to the	enrononline appears in the subject line of this email along with nymex, a prominent exchange. the body of the email contains many references to enrononline in the context of a discussion of a potential interaction with nymex clearing. it is a discussion of the design, development and potential operation of enrononline.	yes, the email is related to the topic. the email is about the design, development, operation, or marketing of enrononline, which is an online service offered by enron for the purchase, sale, trading, or exchange of financial or other instruments or products, including but not limited to, derivative instruments, commodities, futures, and swaps. the email discusses the concept of providing enron					

Figure 4: Snapshot of the survey spreadsheet.