

Beyond the Bar: GPT-4 as a Review Assistant in E-Discovery

*Roshanak Omrani, Eugene Yang, Nathan Reff, Evan Curtin,
Jeremy Pickens, Dave Lewis & Aron Ahmadia*

Copyright 2023 Relativity ODA LLC

Reprinted with permission.



Beyond the Bar: GPT-4 as a Review Assistant in E-Discovery

Roshanak Omrani¹, Eugene Yang¹, Nathan Reff¹, Evan Curtin¹,
Jeremy Pickens², Dave Lewis², Aron Ahmadi¹

This draft paper describes Relativity's proprietary experiments using large language models (LLMs) for first-pass responsive, issue, and key document review in electronic discovery (e-discovery). As such, it and its content are Relativity's proprietary information. We are sharing this paper in draft form solely to facilitate discussion at The Sedona Conference Working Group 1 Annual Meeting. The experiments discussed in the document do not reflect evaluation or analysis of products offered by Relativity. Some of the experiments we describe are ongoing and accordingly, we expect that this paper will continue to undergo additional revisions and considerations before publication. This paper is not authorized for redistribution or use of any kind outside of Working Group 1.

Abstract

In the past year, assessing and applying large language models (LLMs) to real-world use cases have become a popular research topic. The legal industry has also started investigating how to incorporate LLMs into their workflows. Prior work has demonstrated that GPT-4 performs well on the Uniform Bar Exam, evidencing the applicability of these LLMs in this highly sensitive industry. In this work, we explore the possibility of replacing a portion of human workload in electronic discovery (e-discovery), a litigation process in the common law for retrieving relevant information from electronically-stored documents, with an LLM. Specifically, a review protocol written by lawyers is transformed into a prompt for GPT-4 to perform document review. In multiple legal matters, we compared the LLM-enhanced pipeline in e-discovery to that of human evaluation by licensed attorneys. We found that using GPT-4 in this fashion reaches effectiveness as high as 95% recall and 85% precision. Effectiveness is highest when the review protocol is self-contained, explicit, and the responsiveness criteria are dependent only on the single contents of the individual document ("four corners" review). Our experiment shows the potential of an LLM-enhanced review process and how it can reduce labor-intensive document review in e-discovery.

1. Introduction

Over the past few decades, legal systems worldwide have needed to adapt as vast amounts of data transitioned into the digital domain. The ubiquity of emails, text messages, documents, and messaging software in personal, business, and criminal activities has led to a massive increase in digital evidence. This increasing volume of data poses unique challenges for the e-discovery process, which involves identifying, collecting, preserving, reviewing, and producing electronically stored information (ESI) for opposing parties during litigation or investigation (Berman, Barton, and Grimm 2011).

¹ Relativity ODA LLC

² Redgrave Data

Throughout a legal matter, parties must navigate vast quantities of collected ESI (often 1M-100M+ documents) to find items “responsive” to the “matter” (i.e., litigation or investigation) at hand. While perfection is not required, and cost considerations play a role (often through the legal notion of “proportionality”), the expectation is that a large proportion of responsive items will be produced to requesting parties. In other words, high recall, as defined in information retrieval, is expected to be achieved. Traditionally, review by large teams (ranging from dozens to over 1000 attorneys) is used to meet these e-discovery obligations. But as ESI volumes have grown, this has led to high costs, difficulties in meeting legal deadlines, challenges in management and coordination, and inconsistencies in responsiveness decisions that are inevitable when using multiple reviewers (Voorhees 2000). To address cost, throughput, and timing challenges, the legal industry has been an avid adopter of natural language processing, information retrieval, and machine learning technology. Boolean keyword queries have been used to pare down ESI collections since the 1980s (Blair and Maron 1985). Supervised machine learning first came into use in e-discovery around 2005, with courts in the United States, England, Ireland, Australia, and other jurisdictions explicitly encouraging its use starting in 2012. An academic research community focused on e-discovery and other high recall tasks emerged over a similar period, starting with the first TREC Legal Track in 2006 (Baron, Lewis, and Oard 2006).

Despite these technological aids, which can significantly reduce the original volume of electronically stored content, volumes ranging from tens to hundreds of thousands of documents (and at times even millions) are still often manually reviewed by large teams of attorneys. A key tool in achieving consistent coding decisions in this context is a review protocol. This document, which can consist of up to dozens of pages of prose, is usually prepared by senior attorneys working on the matter. It typically provides both background information on the legal matter and a detailed rubric for determining responsiveness of documents. It may also provide rubrics for how to assign subcategories (“issue tags”) to reviewed documents, assess legal privilege, flag personally identifiable information (PII), and make other designations. Reviewers are expected to read the review protocol carefully and label (i.e., “code”) documents accordingly.

While supervised machine learning has had a large impact on e-discovery efficiency, there are reasons why the technology does not perfectly align with the task. Every legal matter has a separate definition of responsiveness, making it difficult to leverage public data sets or amortize the cost of labeling training examples over multiple matters. Technical approaches to reducing training data requirements include active learning (Cormack and Grossman 2014) and fine-tuning of pre-trained language models whose richer semantic and syntactic representations may reduce training data needs (Yang et al. 2022). However, none of these approaches eliminate the need for substantial matter-specific labeling of training data by attorneys.

An alternative approach is to attempt to guide technological screening of documents for responsiveness not by labeling training data, but instead through the same way that human reviewers are guided: with a review protocol. The emergent capability of large language models (LLMs) to achieve in-context learning (Lampinen et al. 2022) and align with user’s intent (Ouyang et al. 2022) through prompting provides an unprecedented opportunity to leverage this approach. LLMs lend themselves in a generative task-agnostic manner to an array of zero-shot downstream tasks, including text classification (Sanh et al. 2022; Radford et al. 2018). In zero-shot adoption of a language model, no adaptation or tuning of the semantic and

syntactic representations that the model has learned through its pretraining phase takes place. This mode of learning is quite desirable, particularly in the absence of publicly available training data for a specific task. This capability, together with the ability to capture long-range dependencies (Vaswani et al. 2017) make LLMs ideal candidate tools for using a review protocol as context for zero-shot classification of documents.

2. Methodology

In this section, we introduce how to employ an LLM to perform document reviews for e-discovery via a workflow leveraging a review protocol.

2.1 Review Workflow with LLM

The workflow is illustrated in Figure 1. First, we modify the review protocol into a natural language prompt (Radford et al. 2018; Brown et al. 2020) for the LLM. By ingesting the protocol into a pre-defined prompt structure that conveys the general structure of the reviewing task to the LLM, the LLM is transformed into a model that can review documents for a specific legal matter without any human supervision through document labels. The pre-defined prompt structure instructs the LLM to infer and output an ordinal (graded) responsiveness prediction (discussed in detail in the next section) for the provided document based on the protocol. The prompt also instructs the LLM to select informative snippets from the document and provide explanations for its inferences along with potential limitations and issues related to those inferences. These requirements guide the LLM to provide evidence that encourages a user to trust its inferences (Kaur et al. 2022) by grounding the inference on a series of step-by-step decisions, similar to the step-wise decision making examples provided in a chain-of-thought process (Wei et al. 2022).

The ingestion process may require modifying the original protocol with a small development set, where the review team iteratively modifies the prompt based on its observed performance on documents in the development set. The process can be viewed as a form of relevance feedback (Rocchio 1971, Salton 1989) in information retrieval, where the query is modified based on human feedback on the retrieved documents.

At the time of inference, the textual content of the document is concatenated with the final prompt to obtain the ordinal prediction surrounding the document. Some or all predicted responsive documents are then passed to humans for further investigation based on the needs of the legal case.

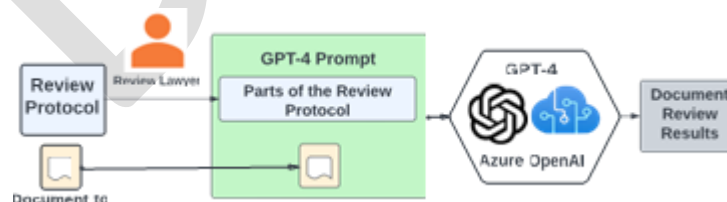


Figure 1: System pipeline with review protocol-prompted GPT-4 mode

2.2 Prompted Ordinal Prediction

Since the purpose of the model is to prioritize the review, providing a sortable output is essential. However, prior work on applying LLMs on pointwise ranking, such as monoT5 (Nogueira, Jiang, and Lin

2020), requires access to the token generation distribution, which restricts the models from generating additional content and prevents the usage of proprietary models that do not expose the generation distributions. In this work, we used an ordinal prediction with a rubric expressed in natural text in the prompt to obtain a coarse sortable score. The instruction of the ordinal rubric is similar to the ones for multiple choice questions (OpenAI 2023), where a textual description follows the grade or the character options.

Our approach defines -1 as the fallback level for any error during inference. Documents with level 0 contain unreadable or empty content, which are classified as non-responsive. Level 1 captures documents that are nonresponsive to the case based on the protocol. Levels 2 through 4 represent responsive documents with different degrees of relevance based on the generated explanation of the protocol expressed in the prompt.

3. Experiment Design

3.1 Large Language Model

In this paper, we describe the use of GPT-4 through Azure OpenAI ChatCompletion endpoint (version 0314) (OpenAI 2023). Since GPT-4 is the most accessible and curated LLM for commercial use, we have experimented with it to gain insights for developing production systems for legal applications.

3.2 Experiment Legal Matters

We collaborated with multiple law firms, legal service providers, and corporations (collectively referred to as “clients”) to procure a total of four data sets sampled from past legal matters and one additional live matter, for which all documents in the matter were available for experimentation. The procured datasets and the live matter were each accompanied by a review protocol with well-established guidelines for coding documents strictly based on their textual content and the “four corners” of each individual document, (i.e., without reference to external context or other documents in the review population). The protocol was prepared by normal procedures used by the client for all matters, with no influence from the fact that the protocol was also to be used in an LLM experiment. The smaller data sets contained at least 500 responsive and 1000 nonresponsive documents labeled by licensed attorneys prior to the experiment, according to the guidelines specified in the review protocol. The live matter consisted of more than 130K documents, with 20-30% prevalence of responsive documents.

We observed throughout some preliminary experiments that a second pass review may resolve disagreements between GPT-4 and the first pass, human-conducted review, with the outcome in favor of the language model. The majority of such conflicts were due to coding decisions that had been based on external information, even though the client initially assumed that the review team had adhered to the four corners of the document. We asked the legal team involved in the live matter to run a second pass review on a stratified sample of 500 documents. This assessment is ongoing and will allow us to estimate the rate of overturns between first and second pass reviews and its effect on GPT-4’s metrics.

3.3 Baseline and Evaluation

The experiments are designed to assess GPT-4’s document review potential according to two sets of key performance indicators (KPIs); one that is relevant to classification metrics and the other that is relevant to human effort. Regarding the classification metric KPI, GPT-4’s classification ability is being benchmarked

against the learning curve of a text classification model. We used the four data sets, with a limited number of samples, for this purpose and investigated how much training data was needed before the text classifier could catch up or surpass GPT-4's performance on a held-out sample of the data (i.e., a sample that is neither seen by the prompt engineer nor included in the training data of the text classifier). Regarding the KPI relevant to human effort, we are leveraging the live matter to assess GPT-4's performance measured on a random sample and extrapolated that performance to the entire matter to determine the amount of document reduction should a human-led, linear review take place only on the collection deemed responsive by GPT-4.

We found that using GPT-4 can demonstrate effectiveness as high as 95% recall and 85% precision. Effectiveness is highest when the review protocol is self-contained, explicit, and the responsive criteria are dependent only on the single contents of the individual document ("four corners" review). Full discussion of findings and the evaluation process are deferred to the full release of the manuscript in the future.

4. Qualitative Observations

The following observations are based on the outcome of datasets and the live matter that were studied in support of this paper:

1. The approach outlined above is most sensitive to the overall structure and arrangement of the review protocol. It appears that the language model more consistently identifies content that is relevant to various elements of responsiveness if it is presented with the responsiveness criteria arranged as an itemized list instead of a long paragraph.
2. Combining a long review protocol with a short document can confuse the language model as to where the prompt instruction ends and the actual document begins; with such a combination, the language model may point to an excerpt of the review protocol, instead of the document.
3. The language model is capable of extracting instances of exclusion criteria that appear in the document, but often fails to act upon them in the same manner as a human reviewer. For example, GPT-4 recognizes if an email is sent outside the responsive date range but does not apply this finding to overrule responsiveness when the email content relates to other responsiveness criteria. In other words, it can be said that the language model needs very specific instructions to perform logical operations.

5. Limitations and Future Directions

1. The scope of the current study is limited to matters with review protocols that require coding solely on the content of individual documents, excluding external considerations. In reality, legal teams accumulate context relevant to the matter and, inevitably, their understanding of responsiveness drifts as they proceed in the review task. In other words, legal teams leverage the accumulated context when coding documents that are reviewed later in the process. The proposed methodology treats each document in an isolated manner. As such, context is not carried from one document to another.

One solution to this limitation is to iteratively expand and refine the context provided in the prompt by running it against small random sets of human-labeled documents and reviewing

the disagreements between GPT-4's coding and human-generated labels. A workflow similar to fine tuning a machine learning model can be adopted to avoid overfitting the prompt to the documents used for refinement: after refining the prompt according to the residuals in the refinement set, one can test the performance of the prompt on a set of documents unseen by the prompt engineer (who would likely be a member of the legal review team) to ensure that refinements do not overfit to the refinement set and generalize well beyond it. However, three shortcomings are associated with this approach:

- The prompt can become too long as more context is accumulated and consume a significant portion of GPT-4's token limit.
- A long prompt can potentially cause effects similar to the observed sensitivity of results to the example ordering in in-context learning (Lu et al. 2022).
- The review team should terminate the prompt engineering phase at some point; otherwise, it contradicts the purpose of leveraging GPT-4 to reduce human effort.

Another potential solution would be to pair in-context learning with Retrieval Augmented Generation (RAG) (Lewis et al. 2020) and a knowledge base that is built from the review population.

2. The ordinal rubric used in the current study does not generate calibrated outputs. Furthermore, its coarse scoring schema does not leave much room for threshold tuning, and as such, it is more suitable for classification than ranking. Aggregating the scores generated according to this rubric with scores generated by a rank ordering function such as BM25 allows for prioritization within the set of documents that are categorized as responsive by GPT-4.
3. As discussed earlier, the main objective of the current study was to assess GPT-4's performance in a standalone zero-shot setting. Devising an efficient workflow to couple conventional active learning protocols with GPT-4 demands further investigation.

6. Responsible AI Considerations

Novel applications of technology such as LLMs require careful attention to issues in responsible AI. In relation to the approach adopted in the current study, two issues particularly raise concerns:

- **Hallucinations:** A hallucinated passage that does not exist in the document, or a reasoning that is backed by hallucinated "facts", could potentially mislead an attorney, and would significantly diminish a legal team's trust in the model and its reliability.
- **Bias:** It is well-known that language models trained on massive collections of text are prone to generate inferences that are biased with respect to race, gender, and other protected classes. A LLM's ability to frame socially biased content as part of the "reasoning" associated with the model's inferences can increase the persuasive power of the content to influence decisions of a person using the output (e.g., an attorney making a document review decision). This is of particular concern in situations such as employment litigation and internal investigations.

Our research and development team has focused on responsible AI issues from the start of this effort and is developing a range of approaches for addressing these issues.

Acknowledgements

We would like to acknowledge Sarah Green, Cristin Traylor and Tara Emory for their thoughtful commentary on an early draft of this manuscript.

References

Baron, J. R.; Lewis, D. D.; and Oard, D. W. 2006. TREC 2006 Legal Track Overview. In *TREC*.

Berman, M. D.; Barton, C. I.; and Grimm, P. W. 2011. *Managing E-Discovery and ESI: From Pre-Litigation Through Trial*. ABA Section of Litigation.

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

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Cormack, G. V.; and Grossman, M. R. 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*, 153–162.

Kaur, D.; Uslu, S.; Rittichier, K. J.; and Durresi, A. 2022. Trustworthy artificial intelligence: a review. *ACM Computing Surveys (CSUR)*, 55(2): 1–38.

Lampinen, A. K.; Dasgupta, I.; Chan, S. C. Y.; Matthewson, K.; Tessler, M. H.; Creswell, A.; McClelland, J. L.; Wang, J. X.; and Hill, F. 2022. Can language models learn from explanations in context? ArXiv:2204.02329 [cs].

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W. t.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 33: 9459–9474.

Lu, Y.; Bartolo, M.; Moore, A.; Riedel, S.; and Stenetorp, P. 2022. Fantastically ordered prompts and where to find them: overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th annual meeting of the association for computational linguistics*, 1: 8086–8098.

Nogueira, R.; Jiang, Z.; and Lin, J. 2020. Document ranking with a pretrained sequence-to-sequence model. ArXiv:2003.06713.

OpenAI. 2023. GPT-4 Technical Report. ArXiv:2303.08774 [cs].

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P. F.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. In *Advances in neural information processing systems*, 35: 27730–27744.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2018. language models are unsupervised multitask learners.

Rocchio, J. 1971. *Relevance feedback in information retrieval*. Prentice Hall.

Salton, G. 1989. *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Reading: Addison-Wesley.

Sanh, V.; Webson, A.; Raffel, C.; Bach, S. H.; Sutawika, L.; Alyafeai, Z.; Chaffin, A.; Stiegler, A.; Le Scao, T.; Raja, A.; Dey, M.; Bari, M. S.; Xu, C.; Thakker, U.; Sharma, S.; Szczechla, E.; Kim, T.; Chhablani, G.; V. Nayak, N.; Datta, D.; Chang, J.; Jiang, M. T. J.; Wang, H.; Manica, M.; Shen, S.; Yong, Z. X.; Pandey, H.; Mckenna, M.; Bawden, R.; Wang, T.; Neeraj, T.; Rozen, J.; Sharma, A.; Santilli, A.; Fevry, T.; Fries, J. A.; Teehan, R.; Bers, T.; Biderman, S.; Gao, L.; Wolf, T.; and Rush, A. M. 2022. Multitask prompted training enables zero-shot task generalization. In *Tenth international conference on learning representations*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *Proceedings of the 31st international conference on neural information processing Systems*, 6000–6010.

Voorhees, E. M. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information processing and management*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q. V.; and Zhou, D. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in neural information processing systems*, 35: 24824–24837.

Yang, E.; MacAvaney, S.; Lewis, D. D.; and Frieder, O. 2022. Goldilocks: Just-right tuning of BERT for technology-assisted review. In *European Conference on Information Retrieval*, 502–517.