Beyond the Bar: Generative AI as a Transformative Component in Legal Document Review (2004)

Roshanak Omrani, et al.

Beyond the Bar: Generative AI as a Transformative Component in Legal Document Review

Roshanak Omrani
Relativity
USA
romrani@g.ucla.edu

Eugene Yang
Johns Hopkins University and
Relativity, USA
eugene.yang@jhu.edu

Evan Curtin
Relativity
USA
evan.curtin@relativity.com

Tara Emory
Redgrave Data
USA
tara.emory@redgravedata.com

Lenora Gray
Redgrave Data
USA
lenora.gray@redgravedata.com

Jeremy Pickens
Redgrave Data
USA
jeremy.pickens@redgravedata.com

Nathan Reff
Relativity
USA
nathan.reff@relativity.com

Cristin Traylor
Relativity
USA
cristin.traylor@relativity.com

Sean Underwood
Relativity
USA
sean.underwood@relativity.com

David D. Lewis
Redgrave Data
USA
dave.lewis@redgravedata.com

Aron J. Ahmadia
Relativity
USA
aron.ahmadia@relativity.com

ABSTRACT

Review for responsiveness is a recall-oriented document classification task central to civil litigation. In large legal matters, it may involve the coding of millions of documents by teams of dozens to hundreds of contract attorneys. We describe a prototype document review system based on a large language model (LLM) for replacing the first level of attorney review. Our system accepts the same guidance—a written review protocol—that would be provided to a human review team. We tested our prototype in the context of a live legal matter, evaluating both human review and our LLM-based system against a gold standard coded by expert senior attorneys. Our prototype achieved an estimated 96% recall and 60% precision without matter-specific tuning, and has numerous avenues for further improvement.

1 INTRODUCTION

The ubiquity of email, electronic document creation, and messaging software in personal, business, and criminal activities has led to a massive increase in digital evidence in legal matters. This has posed a challenge for *electronic discovery* (or, as it is commonly referred to in the law, *eDiscovery*): the process of identifying, collecting, preserving, reviewing, and producing electronically stored information (ESI) to opposing parties in litigation, investigations, and other legal matters [10].

In a large litigation, a party may need to search vast quantities of ESI (up to 1M-100M+ items) for material that is *responsive* to requests for production from opposing parties, as well as needed by the party's own attorneys. This is particularly true in countries, such as the United States, with extensive discovery obligations. While perfection is not required, and cost considerations are allowed to play a role through the legal notion of *proportionality* [14], the expectation is that a large proportion of responsive items will be produced to requesting parties. In other words, high recall is paramount [2].

Traditionally, review by teams of attorneys has been used to meet eDiscovery obligations. As ESI volumes have grown, this has led to high costs, difficulties in meeting legal deadlines, and management and coordination challenges [49]. To address cost, throughput, and timing challenges, the legal industry has been an avid adopter of technology, in particular supervised machine learning (Section 3).

However, despite these technologies, manual review of hundreds of thousands of documents or more is still common in large legal matters. A key tool for fighting the inconsistencies that arise with multiple reviewers [57] is a *review protocol*. This is a written set of instructions that defines what does and does not comprise a responsive document, and includes contextual background on the legal matter. The protocol may range from a few to dozens of pages, and its creation requires substantial effort by senior attorneys working on the matter. The reviewing attorneys are expected to read the review protocol carefully and label ("code") documents accordingly.

The role that a review protocol plays in human review is reminiscent of the role a prompt plays in generative AI [39]. This raises an intriguing question: can we guide technological assessment of documents for responsiveness not by labeling of training data for supervised learning, but by using a review protocol of exactly the sort that attorneys already routinely produce?

This paper presents the first study of the use of review protocol-prompted LLMs (GPT-4 and GPT-4 Turbo) to conduct responsiveness review in a live legal matter. We begin by discussing the legal

and technical background on large-scale reviews in litigation (Sections 2 and 3). We then present our approach for eliciting prompt-based responsiveness decisions and explanations from an LLM (Section 4). Section 5 presents our experimental design and effectiveness estimation approach. Our results (Section 6) show that our prototype achieves competitive effectiveness with a team of human contract attorney reviewers. We conclude by discussing the limitations of this work, future directions, and societal implications of our results (Sections 7 and 8).

2 THE LEGAL CONTEXT

Review for responsiveness is a small but important part of the legal system. Its larger context is important to understand our task constraints. Our focus here is on the legal system in the United States, but some of the same issues arise in other jurisdictions.

Deciding whether a document is responsive to a request for production is just one aspect of the document review. Attorney reviewers are also often called upon to flag key (particularly important) documents, to assign subcategories ("issue tags") to responsive documents, to identify which documents are covered by any of several legal privileges, and to flag occurrences of personally identifiable information (PII), protected health information (PHI), and trade secrets. As with responsiveness, reviewers are provided guidance on these topics through the written review protocol.

Large scale litigations involve multiple phases of review. The initial first level (1L) review of all documents that make it through technological filters (Section 3) is typically carried out either by junior law firm attorneys or professional contract document review attorneys provided by a legal service provider. These attorneys follow the review protocol produced by senior attorneys, may direct questions to those senior attorneys, and operate under their legal direction. This process may result in modifications to the review protocol to clarify existing issues or address newly discovered ones. The size and composition of the 1L review team may fluctuate during the review, introducing additional management challenges.

Review of random or targeted samples of 1L review decisions for quality control is typical. Beyond that, second level (2L) review of some or all 1L coding decisions may be carried out by more senior contract attorneys or law firm attorneys. Second level review tends to be focused on documents that were found to be responsive and/or privileged in 1L review, or are otherwise particularly sensitive. In some cases, further quality control or even third level review may be used. Documents that are found to be responsive and not subject to a legal privilege are produced to other parties in the litigation.

Document review in turn is just one aspect of the fact discovery process in litigation, which also includes, for instance, deposing (formally interviewing) people with knowledge of the matter. These processes inform each other, with reviewed documents perhaps triggering identification of new deponents or custodians, and facts uncovered in interviews leading to additional collection and review. The result of the fact discovery process then informs negotiations among parties and, if a case goes to trial, provides the evidence that is evaluated by a court.

3 RELATED WORK

Technology has a long history of adoption in eDiscovery. Boolean keyword queries have been used since the 1980s [11] to pare down collected ESI for review. Since the early 2000s, a much wider range of technologies have been deployed and are now common in eDiscovery software. These include exact duplicate detection, near-duplicate detection, document clustering, term clustering and relationship finding, entity detection, and statistically ranked retrieval combined with various forms of query reformulation (known as "concept search" in eDiscovery) [45].

However, it is supervised machine learning for producing text classifiers (known as *technology-assisted review* or *TAR* in eDiscovery) which has led to the greatest change in eDiscovery workflows, and which has received the greatest judicial scrutiny. The combination of a trained text classifier and a coded random sample allowing estimation of effectiveness makes it possible to review only a small fraction of collected data while achieving review of a high proportion of responsive material (high recall) with a specified statistical level of confidence.

Supervised machine learning first came into use in eDiscovery around 2005, with courts in the United States[1], England[2], Ireland[3], Australia[4], and other jurisdictions explicitly encouraging its use starting in 2012. Supervised learning is now a routine capability in eDiscovery software, though the original hopes that it would lead to major reductions in document review costs [49] have dimmed a bit.

While there has been some academic research on document retrieval [3, 63], clustering [27, 29, 45], and other technologies for eDiscovery, the bulk of eDiscovery academic research (leaving aside scholarship on purely legal questions) has been on supervised learning and its evaluation. In recent years, a broader *technology-assisted review* (TAR) research community has emerged [16, 17] that encompasses uses of supervised learning to support review not only in eDiscovery, but for systematic review in medicine [21, 30, 31] and content moderation [65], among other areas.

A primary focus of research on supervised learning in eDiscovery has been on reducing the need for expensive senior attorneys to spend time on coding documents for use in training classifiers. Training data reduction has been sought through active learning [15], workflow design [64], training from larger amounts of lower quality review decisions [52], and fine-tuning of pre-trained language models [66].

Even with these techniques, however, substantial training sets must be labeled to use supervised learning. Further, there are ways in which supervised learning has always been an awkward fit to the document review task. Every legal matter has its own definition of responsiveness, which reduces opportunities to leverage public training data or amortize the cost of labeling training examples over multiple matters. Collection of documents typically goes on in parallel with review, with collected documents arriving over a

[1] Da Silva Moore v. Publicis Groupe (Da Silva Moore 17), No. 11 Civ. 1279(ALC)(AJP), 2012 WL 607412. (S.D.N.Y. Feb. 24, 2012)
[2] Pyrrho Inv. Ltd. v. MWB Bus. Exch., Ltd., [2016] EWHC 256 (Ch) [1] (Eng.)
[3] Irish Bank Resolution Corp. v. Quinn [2015] IEHC 175 (H. Ct.) (Ir).
[4] McConnell Dowell Constructors (Aust) Pty Ltd v Santam Ltd & Ors (No.1) [2016] VSC 734

period of months or years. This makes the labeling of a training set not a one-and-done task, but an ongoing and distracting chore.

From a legal standpoint, the fact that a training set typically contains both positive and negative examples means that it contains material which is not responsive, and thus for which the producing party has no obligation to show a requesting party. This, combined with traditional supervised learning systems' inability to explain their decisions, has created legal debate over what should or must be revealed about training sets to other parties [34]. The risk of expensive conflicts with opposing counsel over how supervised learning is to be used, and what must be disclosed about it, has limited the use of supervised learning in practice [60].

These challenges to the use of supervised learning have set the stage for recent interest in applying LLMs to eDiscovery and the law more broadly. The emergent capability of LLMs for in-context learning [36] and alignment with user intent through prompting [48] is highly suggestive given the availability of written review protocols. LLMs have been applied in a task-independent manner to a range of zero-shot downstream tasks, including text classification [51, 53], information extraction [20, 42], and summarization [67]. This capability, together with the ability to capture long-range dependencies [8] makes it plausible to use a multi-page review protocol as context for zero-shot document classification [13].

The work most closely related to ours is a recent study by Pai et al. [50] of prompted LLMs applied to three TREC Legal Track topics [22]. This study compared several commercial and open-source LLMs, and explored a variety of prompt engineering and refinement approaches for improving effectiveness. The most interesting result was that human subject matter experts in a blind study preferred LLM-based explanations of responsive decisions over those of a separate human subject matter expert. However, in contrast to our study, the tasks used were simulated, the techniques investigated required the use of coded training examples, and there was no evaluation of human versus LLM-based review against a common standard.

Baron et al. [4] investigated the application of a prompt-based LLM (GPT 3.5) to determine whether government records could be withheld under a deliberative process privilege exception to the US Freedom of Information Act. This task is recall-oriented and is similar to privilege review in eDiscovery. The authors found LLM-based predictions to have similar effectiveness to those produced by classifiers trained by supervised learning, when both were evaluated against a common expert standard.

A group of researchers from eleven institutions examined the prospects for using LLMs to replace manual review in producing labeled data sets for evaluating information retrieval systems [18]. Both an extensive literature review and a pilot study were conducted. The pilot study found only fair agreement (Cohen's κ=0.26) between human and LLM-based labels. However, the focus was on high precision information retrieval tasks, rather than high recall ones. Prompting also used the short topic descriptions typical of information retrieval test collections, rather than the extensive rubrics used in document review. Arguments were presented both for and against the use of LLMs for this purpose, as well discussions of a number of ways in which human and LLM-based coding decisions can be combined.

A number of studies have investigated the promise [24, 38, 62] or threat [56, 58] of LLMs replacing, impersonating, or augmenting human crowdsourced workers. He, et al. found that using GPT 3.5 to provide ground truth labels training downstream NLP systems compared favorably with crowdsourced annotation [24]. This is somewhat analogous to our comparison of LLMs with 1L reviewers, but the coding decisions in this case affect real-world tasks only indirectly, through their role as training data. The stakes are higher for 1L review, in that coding decisions themselves are considered legal judgments, and 1L reviewers are acting as attorneys with responsibilities to the legal system.

4 AN LLM-BASED REVIEW PROTOTYPE

We developed an LLM-based prototype review architecture based on accessing GPT-4 via the Azure OpenAI ChatCompletion endpoint (version 0314) [46]. Our main criteria for selecting GPT-4 (versus other choices of LLMs) was its availability for commercial use. This was important both because our experiment was in the context of a commercial litigation, and because our study was part of the development process for a commercial product.

During the course of our study, OpenAI released GPT-4 Turbo [47]. GPT-4 Turbo has more recent training data than GPT-4, is claimed to have a stronger ability to follow instructions, is substantially less expensive per API call, and has other improvements. Our experiments (Section 5) therefore evaluated the use of GPT-4 Turbo as well, using the same architecture and prompting strategy used with GPT-4.

4.1 Workflow

Figure 1 compares the workflow for using our prototype with that of typical first level review. In both cases, senior attorneys produce a review protocol which is used to guide the review of each document. In our LLM-based prototype, information from the review protocol is used to create a prompt (see next section) which describes how to review a document. The same prompt is used for all documents. This prompt is combined with each individual document and sent to the LLM one at a time for inference. The resulting completions are saved alongside the document in a document review platform, which enables senior attorneys to review the results.

4.2 Prompt Generation

The prompt is generated by extracting selected sections from the review protocol and injecting them into a prompt template that includes additional instructions to the LLM [12, 51]. The pertinent sections of the review protocol include the description of responsiveness, identification of relevant entities and terms, as well as descriptions of subcategories ("issues") that comprise the legal matter. By contrast, sections on attorney-client privilege and redaction of PII are not included. While human attorneys can easily ignore such information in making their responsiveness assessments, superfluous information is likely to harm LLM assessments.

Our template first instructs the model to identify citations from the document that might suggest it is responsive, and to explain its reasoning both for and against a prediction of responsiveness. These requirements instruct the LLM to ground the inference based on chain-of-thought [59] and provide evidence to establish trust

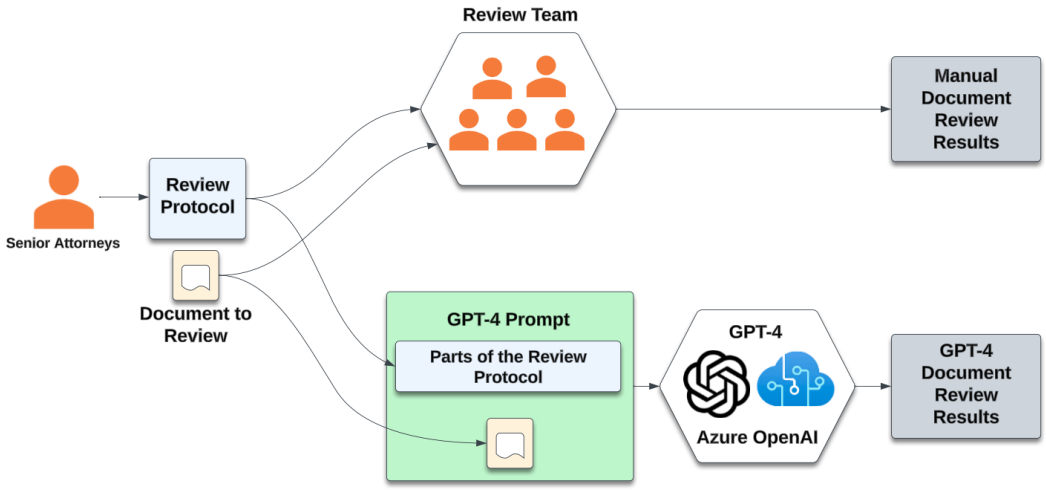

Figure 1: Parallel workflows of a standard manual document review compared to the system pipeline of the GPT-4 document review.

Table 1: GPT-4 scoring rubric

Score	Description
-1	Fallback score for undefined errors
0	Not enough information to score
1	Not Responsive
2	Borderline Responsive
3	Responsive
4	Very Responsive

from users [32]. Finally, the prompt instructs the LLM to provide an ordinal score between -1 and 4 to indicate the degree of responsiveness of the document. This rubric-based classification is similar to other multi-class classification prompts [46].

Our use of an ordinal scoring rubric in the prompt allows the model to provide a prediction of the strength of responsiveness in a way that is easily understood. In order to integrate the model's outputs with other software components, we prompt the LLM to return its predictions using JSON formatting.

5 EXPERIMENTAL PROTOCOL

5.1 Dataset

We report a test of our LLM-based review prototype in support of the review of documents for responsiveness in a lawsuit in the United States. The ability to conduct a test during an active litigation provided the most realistic assessment possible, but also imposed a number of constraints on us, most notably limitations on the amount of ground truth data available for evaluation. The population of documents to review was hosted in a workspace within a commercial review platform. For our study, we omitted those documents that were not amenable to language-based analysis (e.g., image files), resulting in a test collection of 133,638 documents. The constraints of the litigation do not allow us to share this data publicly.

5.2 1L Review as a Classifier

First level review in this legal matter involved contract attorneys coding documents for responsiveness using a three-level ordinal scheme: *Key* (highly responsive), *Responsive*, or *Not Responsive*. The legal service provider followed their usual quality control procedures for litigation coding. These procedures involve double-checking of some fraction of contract attorney decisions by more experienced contract attorneys, with guidance from senior attorneys as necessary. In addition to the first level review of interest in this study, reviewers also carried out other tasks not studied here, such as coding for issues and attorney-client privilege. We do not study this additional coding.

This manual first level review served as the baseline "classifier" in our study. To support evaluation using binary classification effectiveness measures, ordinal level *Not Responsive* was treated as a prediction that a document was not responsive, and ordinal levels *Key* and *Responsive* were both treated as predictions that a document was responsive.

5.3 LLM-Based Classifiers

We used GPT-4 and GPT-4 Turbo to produce an ordinal prediction for each document in our Phase 1 sample (Section 5.6). For evaluation using binary classification effectiveness measures, we converted the ordinal predictions from both the GPT-based systems and manual 1L review to binary classifications for responsiveness. GPT-based ordinal prediction levels {0,1} were treated as a prediction that the document was not responsive, and levels {2,3,4} as a prediction that the document was responsive.

Our LLM-based prototype was guided by a prompt produced from the review protocol as discussed in Section 4. As is typical in legal matters, both the 1L review attorneys and the senior attorneys

learned a great deal about the collected documents and the facts of the case during the review process. However, in order to provide a conservative evaluation of the capabilities of our prototype versus 1L review, we based our prompt only on the initial review protocol produced by the senior attorneys. Therefore, unlike the 1L review team, our system did not benefit from learning about the case as the review progressed.

5.4 Gold Standard

Both 1L review and our LLM-based classifiers were evaluated against a gold standard (set of assumed correct coding decisions) based on labeling of a random sample (Section 5.6) of documents by the senior attorneys on the case. We refer to this as second level (2L) review in this study, though it was not the formal 2L review for the actual litigation. In producing this gold standard, the senior attorneys had access to 1L review decisions, GPT-4 output, the document itself, the rubric, and their additional knowledge of the case. This means the gold standard was not produced blindly, but was produced in the mode that senior attorneys who must sign off on the case legally operate: as the final arbiters of any disagreements about responsiveness.

5.5 Effectiveness Measures

We evaluated manual 1L review and the two GPT-based reviews using estimated values of standard information retrieval metrics:

$$
\begin{aligned}
R &= \frac{TP}{TP + FN} \\
P &= \frac{TP}{TP + FP} \\
F_\beta &= \frac{(1+\beta^2)TP}{(1+\beta^2)TP + (\beta^2)FN + FP}
\end{aligned}
$$

where TP is the number of true positive predictions, FN is the number of false negative predictions, FP is the number of false positive predictions, R is recall, P is precision, and F_β is Van Rijsbergen's F-measure [55]. The value β may be chosen between 0 and infinity to adjust the weight the measure puts on recall versus precision.

Note that the true values of the effectiveness measures are unknown in our experiments, since we do not have gold standard coding for the full collection. We instead estimate these values using a gold-coded random sample, as is common in practice when supervised-learning based systems are used in eDiscovery.

At the time our random sample was drawn, 15104 of the 133683 documents in our collection had not received a 1L review decision. For the purpose of comparing GPT-based reviews to 1L review, we use estimates of effectiveness on the 118534 documents that had received 1L review at the time of sampling. For the purpose of comparing GPT-4 and GPT-4 Turbo to each other, we also compute estimates of effectiveness on the complete collection.

5.6 Random Sampling

Cost, in terms of both resource availability and attorney time during an active litigation, were major limiting factors in our evaluation. We had a limited budget for calls to GPT-4 at the time of evaluation, so running it on the entire collection was not possible. In addition, the senior attorneys in the case were willing to code only 500-600 documents for use as a gold standard. A further complexity, common in eDiscovery data sets, was the presence of duplicate documents.

Table 2: File type breakdown of Phase 1 sample

File Type	Count
email	1954
pdf	564
spreadsheet	370
document	315
slides	294
image	26
plaintext	11
html	9
diagram	3
video	2
calendar	2

To reduce uncertainty in our estimates, we adopted a sampling strategy that leveraged the classifiers being evaluated. We split our collection into four strata (K, R, N, and U) based on the category assigned during 1L review (U corresponds to the documents that had not gotten 1L review at the time of sampling). Within each stratum we then used *double sampling for stratification* [40, Section 12.2]. Briefly, this sampling method works as follows:

(1) A relatively large random sample of units (the Phase 1 sample) is taken from a population.
(2) A cheap (though not free) attribute is measured for each unit in the Phase 1 sample.
(3) The Phase 1 sample is stratified on the measured attribute.
(4) A Phase 2 stratified sample is drawn from the Phase 1 sample and coded for the expensive attribute that is actually of interest.

Double sampling has been used, for instance, in evaluating medical tests (e.g., a cheap existing test or patient information is used in stratification for evaluation of a new, more expensive test) [7], in survey research (e.g., a mail survey or demographic information is used in stratifying people for phone interviews) [25], and in environmental studies and natural resource management [5, 23, 61].

For us, the Phase 1 sample was itself a stratified sample [40, Chapter 3], in that we chose a separate Phase 1 subsample from each top-level stratum, and thus repeated the double sampling process four times. Each Phase 1 subsample was a simple random sample without replacement (SRSWOR) cluster sample [40, Chapter 5]. By cluster sample, we mean that the stratum was treated not as a set of documents, but as a set of clusters, which for us were groups of duplicate documents determined by cryptographic file hashes. This follows the eDiscovery practice of reviewing duplicate documents together. The "cheap" attribute computed for each document in the Phase 1 sample and used for second level stratification was the binary output of our GPT-4 based classifier. It imposed a further 2-way stratification within each top level stratum.

A SRSWOR cluster sample was then chosen from each of the resulting 8 substrata and sent for 2L coding by senior attorneys. This resulted in a total of 547 documents getting 2L review. The

counts for strata, Phase 1 samples, and Phase 2 samples are shown in Table 3. The sizes of Phase 1 and Phase 2 samples were chosen based on projections of the relative proportions of 2L Responsive in top level strata, but no formal optimal allocation of the sampling budget to substrata was done.

As in many complex sampling exercises in operational settings, there were a number of small deviations from our formal sampling design:

- We ran our GPT-4 based prototype on batches of roughly 320 documents for each stratum until our target number of positive and negative predictions for that stratum was reached. This termination procedure in theory introduces a small upward bias in the proportion of the minority GPT-4 prediction in a stratum [19].
- For the K stratum, a bug resulted in ending first level sampling after only 74 rather than 80 negative GPT-4 predictions.
- For a small number of documents the GPT-4 call failed, and this was not noticed until later in our analysis. Failures did not appear to be related to document content. In cases where one or more duplicates of the document were included in the sample, we imputed a GPT-4 label by selecting a label at random from the duplicates for which the call succeeded. When no duplicate was available, we dropped the document from our analysis.
- For 3 documents, we did not receive 2L coding decisions from the senior attorneys. In two cases a duplicate that did receive a 2L label was present in the sample, so we imputed the missing value by copying its duplicate's label. The third document was dropped from the analysis.

Given the small proportion of documents involved, we do not believe these significantly affected our results.

We stress that the double sampling scheme used in this paper was driven by the limited budget of GPT4 calls available at the time of our experiment. A full LLM-based review of a document population could be evaluated by using the same industry methods applied to manual and supervised-learning based review.

5.7 Estimation

Point estimates of the population contingency table quantities (e.g., true positives, false positives, etc.) were produced for each classifier by summing stratum level estimates, with each stratum level estimate produced using the double expansion estimator (DEE) [40, Section 12.2]:

$$\hat{N}_{\cdot\cdot u} = \left(\sum_{h \in \{K,R,N,U\}} \hat{N}_{h \cdot u} \right) = \sum_{h \in \{K,R,N,U\}} \left(\sum_{s \in \{R,U\}} N_h \frac{n_{hs}}{n_h} \frac{m_{hsu}}{m_{hs}} \right)$$

Here,

- $\hat{N}_{\cdot\cdot u}$ is our estimate of the number of members of class u in the population (e.g., estimated number of true positives), with $\hat{N}_{h \cdot u}$ the stratum-level estimate
- N_h is the number of documents in stratum h
- n_h is the size of the Phase 1 sample from stratum h, and n_{hs} the number of members of substratum s in that sample
- m_{hs} is the size of the Phase 2 sample from substratum hs, and m_{hsu} the number of documents from class u in that sample

Point estimates of the effectiveness measures were then produced using the combined ratio estimator [40, Section 4.5], which simply plugs the population contingency table estimates into the effectiveness measure definitions.

6 RESULTS AND DISCUSSION

Table 4 shows our sample-based point estimates of recall, precision, and F_1 (F-measure with $\beta = 1$), corresponding to the harmonic mean of recall and precision. Results both with (K, R, N, U) and without (K, R, N) documents lacking first level review are shown; differences are slight. Our core result is that our LLM-based prototypes are capable of achieving high recall and acceptable precision with the review protocol as the only case-specific information used.

The F_1 measure, which gives recall and precision equal weight, shows the LLM-based methods having slightly higher effectiveness than 1L review, with the GPT-4 prototype slightly outperforming GPT-4T. If higher weight was put on recall, the performance difference would be even larger, though there is no exact translation from the desirable eDiscovery recall levels to values of β for F_β.

When 1L review and the GPT predictions agreed, the gold standard almost always agreed as well (compare R vs. N values in Table 3). When the two disagreed, the gold standard split roughly evenly between the two classes. While in this study 1L review had high precision and comparatively low recall, this is not the case with all 1L review. We also stress that this result in no way indicates that the production of responsive documents in this particular legal matter was inadequate, since a variety of quality control procedures were applied subsequent to 1L review.

The very high recall achieved by the GPT-4 prototype was arguably overkill. In general, 70-85% recall is considered to meet requirements of reasonableness and proportionality, at least in civil litigation in the United States.[5] The GPT-4 Turbo prototype achieved defensible recall of 89% while also slightly outperforming the GPT-4 prototype's precision.

7 LIMITATIONS AND FUTURE DIRECTIONS

As discussed above, our prototypes achieved recall higher than necessary. We expect that with tuning of our prompting strategy we can achieve substantial improvements in precision (and thus lower costs for QC and second level review) while maintaining acceptable recall levels.

The approach used in our prototypes require that determination of responsiveness can be made for each document based on

[5] *See* Fed. R. Civ. P. 26(b)(1) (scope of discovery includes relevant material with consideration for "whether the burden or expense of the proposed discovery outweighs its likely benefit"); Fed. R. Civ. P. 26(g) (attorneys must make reasonable inquiry to completeness and correctness of disclosures); Lawson v. Spirit AeroSystems, Inc., No. 18-1100-EFM-ADM, 2020 WL 1813395 (D. Kan. April 9, 2020) (85% recall was reasonable and typical for TAR, and where defendant agreed to 80% recall after initial results under 70%); In re Diisocyanates Antitrust Litig., No. MC 18-1001, MDL No. 2862, 2021 WL 4295729 (W.D. Pa. Aug. 23, 2021), adopted by In re Diisocyanates, 2021 WL 4295719 (W.D. Pa. Sept. 21, 2021) (where parties agreed 70-80% recall was generally acceptable, Special Master held results ranging from 74-89% were reasonable, though required additional review and production based on other considerations).

Table 3: Sizes of first level strata (based on 1L review), Phase 1 samples from first level strata, and Phase 2 samples from Phase 1 samples. We show counts of GPT-4 predictions bucketed into Responsive (R) and Not Responsive (N), and similarly show bucketing of 2L review coding decisions into those two buckets.

		GPT-4				GPT-4 Turbo			
1L Coding	Stratum Size	Phase 1 Sample	GPT-4 Prediction	Phase 2 Sample	2L Label	Phase 1 Sample	GPT-4 Turbo Prediction	Phase 2 Sample	2L Label
Key	9204	1599	R: 1524	103	R: 84 N: 19	1595	R: 1337	109	R: 90 N: 19
			N: 75	74	R: 33 N: 41		N: 258	68	R: 27 N: 41
Responsive	22851	1003	R: 953	59	R: 57 N: 2	982	R: 871	70	R: 67 N: 3
			N: 50	48	R: 29 N: 19		N: 111	37	R: 19 N: 18
Not Responsive	86479	432	R: 274	90	R: 38 N: 52	311	R: 160	80	R: 33 N: 47
			N: 158	90	R: 3 N: 87		N: 151	97	R: 8 N: 89
Uncoded	15104	311	R: 266	43	R: 23 N: 20	310	R: 225	42	R: 22 N: 20
			N: 45	40	R: 1 N: 39		N: 85	41	R: 2 N: 39
Total	133638	3345	3345	547	547	3198	3198	544	544

Table 4: Effectiveness of 1L review and our prototype review. Distribution of phase I and II sample in 2-way and 3-way partitions; K: key and responsive, R: responsive, N: not responsive, U: not coded or needs further review.

Population	Classifier	Recall	Precision	F_1
K,R,N	1L review	0.55	0.91	0.68
	GPT-4	0.96	0.60	0.74
	GPT-4 Turbo	0.89	0.61	0.72
K,R,N,U	GPT-4	0.97	0.59	0.73
	GPT-4 Turbo	0.90	0.60	0.72

its content only, without reference to other documents or information. This is the setting in which technology-assisted review is typically applied as well, with aspects of responsiveness requiring information outside the "four corners" of the document handled by a separate workflow.

Our approach also assumes that the review protocol is not *too* long. While usable context lengths have been rapidly growing in LLMs, too long a prompt can potentially cause effects similar to the observed sensitivity of results to the example ordering in in-context learning [41].

Our prototypes, unlike human 1L reviewers, do not learn from senior attorney feedback as the review progresses. Such feedback can take the form of modifications to the review protocol, guidance on particular classes of examples conveyed in natural language, and/or corrections to 1L review decisions. An updated review protocol can naturally be accommodated in our approach, as can senior attorney comments. Feedback in the form of corrected review decisions would enable the use of few-shot learning, where labeled data can be used to automatically improve a prompt [39]. Another approach would be to pair in-context learning with Retrieval Augmented Generation (RAG) [37] and a knowledge base that is built from the review population.

Rather than simply accepting whatever feedback senior attorneys spontaneously provide, particular types of feedback could be sought through traditional active learning (eliciting labeled examples) or active approaches to eliciting text to add to prompts.

8 RESPONSIBLE AI CONSIDERATIONS

A number of problematic characteristics of LLM-based systems and generative AI in general have been identified [9]. Of most concern in our document review application are hallucinations [28, 43] and generative bias [54]:

- Hallucinations: Our prototype provides explanations for its responsiveness determinations. These explanations could include pointers to passages that do not actually exist in the document, or "reasoning" that is backed by hallucinated "facts".
- Bias: Language models trained from large corpora may reflect social biases in their training data [6, 26, 35, 44], including biases around gender, race, religion, disability, and other characteristics of people. Such biases could in theory affect both the ordinal rating assigned by our system to a document, as well as the explanations provided for that rating. If the score of the model is used to prioritize or filter which documents are reviewed, there is a risk of finding or failing to find relevant information to incriminate or

exonerate individuals or groups on the basis of their language patterns. Considering that the information retrieved by eDiscovery software is used to inform decisions in both criminal and civil investigations and litigations, there is potential risk in these systems to disproportionately impact the carriage and miscarriage of justice based on group identity.

On the other hand, the document review context provides a number of error correction processes that are not present in many applications of generative AI. Foremost is the fact that a senior attorney is always legally responsible for understanding the details of, and must attest in writing to, the adequacy of a document production. They and their team are responsible for ensuring that appropriate quality control procedures are in place to compensate for errors and biases in human review. While failures in an initial LLM-based review will likely be different from those in a first level review by attorneys, many of the concerns are the same.

At a larger scale, the legal system as a whole has many rules and procedures that may serve as safeguards for problems that may arise within the discovery process, including court rules and attorney rules of professional conduct. Legal teams are accustomed to identifying deficiencies in their processes, and checking for deficiencies in productions received from other parties. The adversarial structure of the legal system in common law countries is based on requiring defensible processes and providing recourse through challenges when deficiencies are identified.

We are also actively developing additional technical mitigations to some of these problems. For instance, citations generated by the LLM as part of its chain-of-thought process may in turn be externally validated, mitigating concerns about hallucinations.

Another social concern that has been raised is the impact of AI deployment on employment in the legal profession [33]. Here, the history of technology-assisted review is somewhat reassuring. According to the American Bar Association's Lawyer Population Survey, over the two decades from 2000 - 2019 (inclusive), during which eDiscovery technologies including supervised machine learning were increasingly deployed, the number of active attorneys in the United States grew by nearly 30%. The trend was positive in every year of that time period except the last one, 2019, which saw a modest decline of 1.7% [1]. This suggests that, while AI's potential impact on demand for labor remains a subject of legitimate concern, employment in the legal industry has historically been robust, with new roles for attorneys in society continually emerging.

9 CONCLUSION

In this work, we examined a method for guiding the technological assessment of documents for responsiveness to a request for production in civil litigation (eDiscovery). By pairing large language models with a set of instructions that one would normally have given to human reviewers (the *review protocol*), we have removed the need for substantial matter-specific labeling of training data by attorneys.

We conducted an experiment on a live production system and found that a review protocol-instructed language model achieves recall of 96% with precision of 60%, as evaluated on the 2L expert reviewer judgments. In comparison, the 1L human reviewer (also evaluated on 2L judgments) achieved higher precision (91%) but much lower recall (54%). Typically, at least 70-80% recall is accepted by the courts, which suggests that protocol-instructed LLMs are well-situated to meet one's legal obligations.

Overall, the outcomes underline promising potentials for adopting GPT-4 to enhance the course of eDiscovery by assisting legal teams to expedite document review tasks.

ACKNOWLEDGMENTS

The authors would like to acknowledge Sarah Green for her thoughtful commentary on an early draft of this manuscript.

REFERENCES

[1] American Bar Association. 2022. Legal Profession Statistics. https://www.americanbar.org/about_the_aba/profession_statistics/
[2] Bissan Audeh, Philippe Beaune, and Michel Beigbeder. 2013. Recall-Oriented Evaluation for Information Retrieval Systems. In *Multidisciplinary Information Retrieval*, Mihai Lupu, Evangelos Kanoulas, and Fernando Loizides (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 29–32.
[3] Jason R Baron, David D Lewis, and Douglas W Oard. 2006. TREC 2006 Legal Track Overview.. In *TREC*. Citeseer.
[4] Jason R Baron, Nathaniel W. Rollings, and Douglas W. Oard. 2023. Using ChatGPT for the FOIA Exemption 5 Deliberative Process Privilege. In *3rd International Workshop on Artificial Intelligence and Intelligent Assistance for Legal Professionals in the Digital Workplace (LegalAIIA)*.
[5] Jonathan Bart and Susan Earnst. 2002. Double sampling to estimate density and population trends in birds. *The Auk* 119, 1 (2002), 36–45.
[6] Christine Basta, Marta R Costa-Jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.08783* (2019).
[7] Colin B Begg and Robert A Greenes. 1983. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* (1983), 207–215.
[8] Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150* (2020).
[9] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 610–623.
[10] Michael D Berman, Courtney Ingraffia Barton, and Paul W Grimm. 2011. *Managing E-Discovery and ESI: From Pre-Litigation Through Trial.* ABA Section of Litigation.
[11] David C Blair and Melvin E Maron. 1985. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Commun. ACM* 28, 3 (1985), 289–299.
[12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
[13] Youngjin Chae and Thomas Davidson. 2023. Large language models for text classification: From zero-shot learning to fine-tuning. *Open Science Foundation* (2023).
[14] The Sedona Conference. 2017. Commentary on Proportionality in Electronic Discovery. *Sedona Conference Journal* 18 (2017), 141.
[15] 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*. ACM, Gold Coast Queensland Australia, 153–162. https://doi.org/10.1145/2600428.2609601
[16] Giorgio Maria Di Nunzio, Evangelos Kanoulas, and Prasenjit Majumder. 2022. Augmented intelligence in technology-assisted review systems (ALTARS 2022): Evaluation Metrics and Protocols for ediscovery and systematic review systems. In *European Conference on Information Retrieval*. Springer, 557–560.
[17] Giorgio Maria Di Nunzio, Evangelos Kanoulas, and Prasenjit Majumder. 2023. 2nd Workshop on Augmented Intelligence in Technology-Assisted Review Systems (ALTARS). In *European Conference on Information Retrieval*. Springer, 384–387.
[18] Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. 2023. Perspectives on Large Language Models for Relevance Judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval* (Taipei, Taiwan) *(ICTIR '23)*. Association for Computing Machinery, New York, NY, USA, 39–50. https://doi.org/10.1145/3578337.3605136
[19] Bhaskar Kumar Ghosh and Pranab Kumar Sen. 1991. *Handbook of sequential analysis.* CRC Press.

[20] Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, et al. 2023. LLMs Accelerate Annotation for Medical Information Extraction. In *Machine Learning for Health (ML4H)*. PMLR, 82–100.

[21] Lorraine Goeuriot, Liadh Kelly, Hanna Suominen, Aurélie Névéol, Aude Robert, Evangelos Kanoulas, Rene Spijker, Joao Palotti, and Guido Zuccon. 2017. CLEF 2017 eHealth evaluation lab overview. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11–14, 2017, Proceedings 8*. Springer, 291–303.

[22] Maura R Grossman, Gordon V Cormack, Bruce Hedin, and Douglas W Oard. 2011. Overview of the TREC 2011 Legal Track. In *The Twentieth Text REtrieval Conference (TREC 2011) Proceedings*. NIST.

[23] Michael J Harper, Michael A McCarthy, Rodney Van Der Ree, and Julian C Fox. 2004. Overcoming bias in ground-based surveys of hollow-bearing trees using double-sampling. *Forest Ecology and Management* 190, 2-3 (2004), 291–300.

[24] Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2023. AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators. arXiv:2303.16854 [cs.CL]

[25] MA Hidiroglou. 2001. Double sampling. *Survey methodology* 27, 2 (2001), 143–154.

[26] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. *arXiv preprint arXiv:2005.00813* (2020).

[27] Jessica Jaquez. 2013. *Towards Scalable E-Discovery using Content-based Hierarchical File Clustering*. Ph. D. Dissertation. John Jay College of Criminal Justice.

[28] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.

[29] Sachindra Joshi, Danish Contractor, Kenney Ng, Prasad M Deshpande, and Thomas Hampp. 2011. Auto-grouping emails for faster e-discovery. *Proceedings of the VLDB Endowment* 4, 12 (2011), 1284–1294.

[30] Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. 2018. CLEF 2018 technologically assisted reviews in empirical medicine overview. In *CEUR workshop proceedings*, Vol. 2125.

[31] Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. 2019. CLEF 2019 technology assisted reviews in empirical medicine overview. In *CEUR workshop proceedings*, Vol. 2380. 250.

[32] Davinder Kaur, Suleyman Uslu, Kaley J Rittichier, and Arjan Durresi. 2022. Trustworthy artificial intelligence: a review. *ACM Computing Surveys (CSUR)* 55, 2 (2022), 1–38.

[33] Sara Khan and Elizabeth Powers. 2023. Efficiency, Ethics, and Algorithms: The Implications of AI on the Legal Profession and the ABA Model Rules. *SSRN Electronic Journal* (28 May 2023). https://doi.org/10.2139/ssrn.4461276

[34] Shannon H Kitzer. 2018. Garbage in, Garbage out: Is Seed Set Disclosure a Necessary Check on Technology-Assisted Review and Should Courts Require Disclosure? *U. Ill. JL Tech. & Pol'y* (2018), 197.

[35] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337* (2019).

[36] Andrew K. Lampinen, Ishita Dasgupta, Stephanie C. Y. Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L. McClelland, Jane X. Wang, and Felix Hill. 2022. Can language models learn from explanations in context? http://arxiv.org/abs/2204.02329 arXiv:2204.02329 [cs].

[37] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 9459–9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[38] Jiyi Li. 2024. A Comparative Study on Annotation Quality of Crowdsourcing and LLM via Label Aggregation. *arXiv preprint arXiv:2401.09760* (2024).

[39] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Comput. Surveys* 55, 9 (2023), 1–35.

[40] Sharon L Lohr. 2022. *Sampling: design and analysis* (3rd ed.). CRC press.

[41] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 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 (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 8086–8098. https://doi.org/10.18653/v1/2022.acl-long.556

[42] Reza Yousefi Maragheh, Chenhao Fang, Charan Chand Irugu, Parth Parikh, Jason Cho, Jianpeng Xu, Saranyan Sukumar, Malay Patel, Evren Korpeoglu, Sushant Kumar, et al. 2023. LLM-TAKE: Theme-Aware Keyword Extraction Using Large Language Models. In *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 4318–4324.

[43] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173

[44] Alfonso Min. 2023. Artifical Intelligence and Bias: Challenges, Implications, and Remedies. *Journal of Social Research* 2 (10 2023), 3808–3817. https://doi.org/10.55324/josr.v2i11.1477

[45] Douglas W Oard, Jason R Baron, Bruce Hedin, David D Lewis, and Stephen Tomlinson. 2010. Evaluation of information retrieval for E-discovery. *Artificial Intelligence and Law* 18 (2010), 347–386.

[46] OpenAI. 2023. GPT-4 Technical Report. https://doi.org/10.48550/arXiv.2303.08774 arXiv:2303.08774 [cs].

[47] OpenAI. 2023. New models and developer products announced at DevDay. https://openai.com/blog/new-models-and-developer-products-announced-at-devday

[48] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 27730–27744. https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf

[49] Nicholas M. Pace and Laura Zakaras. 2012. *Where the Money Goes: Understanding Litigant Expenditures for Producing Electronic Discovery*. RAND Corporation, Santa Monica, CA.

[50] Sumit Pai, Sounak Lahiri, Ujjwal Kumar, Krishanu Baksi, Elijah Soba, Michael Suesserman, Nirmala Pudota, Jon Foster, Edward Bowen, and Sanmitra Bhattacharya. 2023. Exploration of Open Large Language Models for eDiscovery. *Proceedings of the Natural Legal Language Processing Workshop 2023* (2023). https://api.semanticscholar.org/CorpusID:265607975

[51] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language Models are Unsupervised Multitask Learners. (2018).

[52] Adam Roegiest, Gordon V Cormack, Charles LA Clarke, and Maura R Grossman. 2015. Impact of surrogate assessments on high-recall retrieval. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 555–564.

[53] Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng-Xin Yong, Harshit Pandey, Michael Mckenna, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *ICLR 2022 - Tenth International Conference on Learning Representations*. Online, Unknown Region. https://inria.hal.science/hal-03540072

[54] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2021. Societal Biases in Language Generation: Progress and Challenges. *CoRR* abs/2105.04054 (2021). arXiv:2105.04054 https://arxiv.org/abs/2105.04054

[55] C. J. Van Rijsbergen. 1979. *Information Retrieval* (2nd ed.). Butterworths.

[56] Veniamin Veselovsky, Manoel Horta Ribeiro, Philip Cozzolino, Andrew Gordon, David Rothschild, and Robert West. 2023. Prevalence and prevention of large language model use in crowd work. *arXiv preprint arXiv:2310.15683* (2023).

[57] Ellen M Voorhees. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management* (2000).

[58] Chaofan Wang, Samuel Kernan Freire, Mo Zhang, Jing Wei, Jorge Goncalves, Vassilis Kostakos, Zhanna Sarsenbayeva, Christina Schneegass, Alessandro Bozzon, and Evangelos Niforatos. 2023. Safeguarding Crowdsourcing Surveys from ChatGPT with Prompt Injection. *arXiv preprint arXiv:2306.08833* (2023).

[59] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 24824–24837. https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

[60] Paul Weiner and Denise Backhouse. 2019. " Transparency,"" Discovery-on-Discovery" Type Disclosures, and Party-Opponent Validation in eDiscovery. *Labor Law Journal* 70, 3 (2019), 212–219.

[61] Harold G Wilm, David F Costello, and Graydon E Klipple. 1944. Estimating forage yield by the double-sampling method. *Journal of the American Society of*

Agronomy 36, 3 (1944).

[62] Tongshuang Wu, Haiyi Zhu, Maya Albayrak, Alexis Axon, Amanda Bertsch, Wenxing Deng, Ziqi Ding, Bill Guo, Sireesh Gururaja, Tzu-Sheng Kuo, et al. 2023. Llms as workers in human-computational algorithms? replicating crowdsourcing pipelines with llms. *arXiv preprint arXiv:2307.10168* (2023).

[63] Eugene Yang, David D Lewis, and Ophir Frieder. 2019. Text retrieval priors for Bayesian logistic regression. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1045–1048.

[64] 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*. 1–10.

[65] Eugene Yang, David D. Lewis, and Ophir Frieder. 2021. TAR on Social Media: A Framework for Online Content Moderation. In *Proceedings of Design of Experimental Search & Information REtrieval Systems (DESIRES)*. https://arxiv.org/abs/2108.12752

[66] Eugene Yang, Sean MacAvaney, David D Lewis, and Ophir Frieder. 2022. Goldilocks: Just-right tuning of bert for technology-assisted review. In *European Conference on Information Retrieval*. Springer, 502–517.

[67] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori Hashimoto. 2023. Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics* 12 (2023), 39–57. https://api.semanticscholar.org/CorpusID:256416014