Automating Accountability Mechanisms in the Judiciary System using Large Language Models

by

Ishana Shastri

B.S. Computer Science and Engineering, MIT, 2023

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2024

Authored by:	Ishana Shastri Department of Electrical Engineering and Computer Science May 17, 2024
Certified by:	Ashia Wilson Assistant Professor in EECS, Thesis Supervisor
Accepted by:	Katrina LaCurts Chair, Master of Engineering Thesis Committee

Automating Accountability Mechanisms in the Judiciary System using Large Language Models

by

Ishana Shastri

Submitted to the Department of Electrical Engineering and Computer Science
on May 17, 2024 in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

ABSTRACT

Holding the judicial system accountable often demands extensive effort from auditors who must meticulously sift through numerous disorganized legal case files to detect patterns of bias and systemic errors. For example, the high-profile investigation into the Curtis Flowers case took nine reporters a full year to assemble evidence about the prosecutor's history of selecting racially-biased juries. Large Language Models (LLMs) have the potential to automate and scale these accountability pipelines, especially given their demonstrated capabilities in both structured and unstructured document retrieval tasks. We present the first work elaborating on the opportunities and challenges of using LLMs to provide accountability in two legal domains: bias in jury selection for criminal trials and housing eviction cases. We find that while LLMs are well-suited for information extraction from eviction forms that have more structure, court transcripts present a unique challenge due to disfluencies in transcribed speech.

Thesis supervisor: Ashia Wilson
Title: Assistant Professor in EECS

Acknowledgments

Firstly, I would like to thank my thesis supervisor, Professor Ashia Wilson for the tremendous flexibility and support she has demonstrated in the past year. Additionally, I would like to express my sincere gratitude to Professor Barbara Engelhardt from Stanford University for also supervising this work and serving as an endless source of optimism and ideas throughout the past year. Both of my advisors have been steadfast in pushing the bounds of this project, connecting me to relevant collaborators, and being role models for strong and brilliant women in academia.

Next, I'd like to thank Shomik Jain for his contributions to this project and for helping me grow as a researcher in the past year. Along with that, I'd like to express my gratitude to all the members of the Wilson Lab for our insightful discussions and their commitment to growth and passion for their work.

This thesis also would not have been possible without our collaborations with Princeton's Eviction Lab, Professor Matthew Desmond, Georgetown Law's Professor Nicole Summers, and American Public Media's Will Craft.

I would also like to thank the course staff of 6.3950 [AI, Decision Making, and Society] and 6.8300 [Advances in Computer Vision] for allowing me to TA the courses and thus fund my graduate degree.

Finally, my thesis would not have been possible without the endless support from my family, friends, and my cat, Nemo. Through all the pivots and chaos, they were the pillars that kept me rooted and gave me faith in my work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The criminal legal system is known to be structurally biased in ways that amplify existing patterns of social inequality [1, 24, 25]. While most evidence of this state of affairs comes from the diligent work of reporters spotlighting egregious cases [14] and medium-scale studies by academics [13], not much is known about the frequency of bias that exists over large jurisdictions or that is exhibited across time by individual lawyers. While such bias may or may not be considered illegal, regular audits are known to improve adherence to standards by discouraging improper behavior through implied oversight [31]. Legal structures such as malpractice claims, the Department of Justice's Office for Professional Responsibility, and more recently, prosecutorial conduct committees, are in place to monitor these biases. However, these structures fail to impose consequences for biased behaviors, largely because they are unable to uncover and process the documents that track the biased actions. We argue that a substantial opportunity exists for AI to greatly enhance the existing accountability mechanisms by allowing lawyers and academics to more rapidly conduct audits and reform injustices.

There are a number of challenges to automating parts of the existing data-driven accountability mechanisms. First, data is difficult to access due to a lack of centralization and modernization. Clerk's offices hold onto printed casebooks until they reach their expiration date, and, while the information is made public, accessing court files can all too often require going to courthouses, pulling relevant files, and manually scanning them. A second concern is the difficulty in processing data. The kind of data collected by the legal system varies greatly. In criminal proceedings, most state courts require some kind of record keeping such as court transcripts, but the specifics of the level of documentation required varies from state to state, and the data is non-standard and incomplete in a variety of ways. Even within a jurisdiction, court data is largely unstructured and often poorly annotated and logged; building technologies robust enough to handle such high variance is a challenge. Even the conversion of court documents into fully readable formats remains a large problem in many parts of the legal system.

A key example of this data-driven accountability is an audit done by journalists working on the American Public Media (APM) Reports podcast "In The Dark," who spent a year gathering and analyzing court records in the Fifth Circuit Court District of Mississippi from 1992 through 2017 to build a dataset of peremptory strikes [14]. With this data, APM was able to uncover patterns of discrimination in a series of trials that culminated in a man's exoneration. According to one of the lead reporters, it took a team of several reporters almost a year to translate the court documents into a legible dataset [14]. The process consisted of showing up with a scanner to each courthouse, going through each courthouse's docket book and writing down the names of each trial, then pulling the relevant case files. The team then spent months entering information for every trial from their respective case files.

Large language models (LLMs) show promise in enhancing accountability through their abilities in information retrieval tasks. LLMs have successfully performed document summarization across various domains, such as news articles [42], books [10], electronic health records [35], and financial documents [26]. However, their application in the legal domain faces limitations, especially in tasks that require logical reasoning or precise calculations. For instance, LLMs struggle with totaling costs in financial documents [32] and can introduce errors by inaccurately contextualizing medical terms in clinical notes [27]. The state of Washington recently used LLMs to mine court-based evictions to identify biases in the eviction system, echoing the need to accurately extract age, gender, and race from court documents [34].

Our work is one of the first to use LLMs to improve accountability in two important court processes: criminal jury selection and eviction cases. Concretely, we make the following contributions:

- We define a taxonomy of information retrieval tasks tailored to unstructured legal documents that are needed for robust audits.

- We evaluate LLM performance across these tasks and surface several challenges, such as the difficulty in parsing transcribed speech in court transcripts, as well as opportunities, such as the use of LLMs to speed up the synthesis of eviction information.

- We contextualize how LLM accuracy and error metrics relate to real-world assessments of bias in the legal system.

- We call for investments on both the legal and technical sides to make automated accountability and auditing mechanisms more feasible.

1.1 Thesis Structure

This thesis is structured as follows:

Chapter 2 introduces the key legal processes investigated within this thesis as well as the foundations of machine learning relevant to the work.

Chapter 3 describes the previous work done in the fields of information retrieval, legal AI, and LLM reasoning.

Chapter 4 describes the data and the framework through which we've defined tasks for the purposes of this research problem.

Chapter 5 describes the experimental setup of our study.

Chapter 6 presents our results verifying our performance analysis and our framework's generalizability.

Chapter 7 provides insights into the challenges and limitations of LLMs uncovered during this process and details future work in this area.

Chapter 2

Case Studies and Opportunities

2.1 Legal Accountability

Accountability in the legal system is endlessly important to prevent exploitative practices and other biased behaviors from perpetuating injustices through the system. Accountability can be defined as *"the duty of a public decision maker to explain, legitimate, and justify a decision and to make amends where a decision causes injustice and harm"* [30]. Fine-grained accountability when tied to a repercussion at any level of the legal process can identify and remove legal practitioners acting in bad faith before their bias propagates too far. Even when not tied to a repercussion, auditing has been shown to improve performance on measured metrics by imposing a "surveillance state" that implicitly penalizes improper behavior in processes [31]. Without accountability of any sort, hundreds of unjust cases that have adverse effects on real people could slip under the radar. We focus on two legal processes that have explicit histories of biased and exploitative practices: criminal case jury selection and eviction.

2.2 Jury Selection

The process of jury selection holds paramount significance in ensuring the fair and impartial administration of justice in the US and other countries with common law systems. It serves as the cornerstone of the legal system, embodying the principles of diversity, transparency, and unbiased representation. However, the traditional methods of jury selection have come under scrutiny for their opacity and potential to perpetuate biases.

2.2.1 Jury Selection Process

In the US, the process of narrowing down a pool of potential jurors, or the *venire*, to the final list differs slightly across different jurisdictions and courts. However, the core procedure revolves around what is termed *voir dire*.

Initially, potential jurors are randomly sampled from a state's voter and motor vehicle registration lists. These potential jurors then proceed through the voir dire process, where they answer a verbal questionnaire aimed to make sure all jurors are impartial and capable of sitting on the jury for that case. Questions within the voir dire glean information such as if any health issues could potentially stand in the way of serving on the jury, if the potential juror has any pre-existing ties to the criminal justice system, and any demographic information about the potential juror. Procedurally, this questionnaire should inform prosecutors and the state defense if a potential juror could be partial to the case; however, doubts have been recently introduced as the demographic information can outweigh more relevant information in deciding whether a juror is struck.

In a case, either the prosecutor or the state defense can strike a juror. Strikes can be either *for cause*, in which legal basis for a juror's dismissal, such as bias or inability to understand the trial or effectively communicate with other jurors, is required to be given to the judge by the striker. Alternatively, a juror can be struck using a *peremptory strike*, where a lawyer may dismiss a certain number of jurors without reason.

2.2.2 Exploitative Practices

Due to their lack of transparency and supervision, these peremptory strikes are where biased and unjustified strikes are most salient. For example, prosecutors will oftentimes cite low intelligence as a “race-neutral" reason for striking people of color [4]. A sample from the APM dataset cited “He has an earring in his ear. I do not like to keep jurors, male jurors that wear earrings" as a valid reason for striking a black male from the jury. In practice, prosecutors often challenge jurors who seem biased in favor of the defendant, whereas defense attorneys excuse those who seem biased against their clients. However, because a prosecutor does not need to provide a reason unless specifically asked by the opposing party, peremptory strikes are an easy loophole through which jury manipulation occurs.

2.2.3 Legal Implications of Unfair Juries

What is legally mandated to be a jury of one's peers has loopholes that can be exploited, mostly at the cost of Black and Brown defendants. It has been demonstrated time and time again that a non-representative jury has the ability to alter the final disposition of a case. Specifically withholding people from a jury is a form of manipulation of what is oftentimes a high-stakes decision.

For example, in the infamous Curtis Flowers v. Mississippi case, Flowers was tried six times for murder, four of which resulted in convictions and the death sentence. However, these dispositions were later overturned for racial bias in the jury selection for his trials, which ended up having all-white juries non-representative of the jurisdiction. Flowers is the canonical example of the extent to which a biased jury can completely upend a case outcome and thus an innocent person's life. However, his case does not stand in isolation. The Groveland Four in 1949 faced a similar outcome when four innocent African American men were sentenced to death and life in prison for a crime they did not commit. Again, the all-white jury that the prosecutors selectively assembled resulted in a case outcome that only years later was overturned for racial bias. More recently, strikes against female [20] and Jewish [2] jurors in death row cases have come under question as instances of illegal bias, which could potentially lead to several high-profile overturned convictions.

Unjust actions like those during the Flowers case must be nipped in the bud to avoid unfair trials and unnecessary retroactive recourse.

2.2.4 Current Accountability Mechanisms

In the landmark 1986 Batson v. Kentucky case, the US Supreme Court ruled that peremptory strikes could not be used to exclude a potential juror solely on the basis of race, and later the ruling was extended to include gender and sexual orientation as well. This is currently the only form of accountability in the jury selection process, whereby the opposing party can challenge a peremptory strike if they are suspicious of biased striking. A challenge requires the suspected party to provide a race- and gender-neutral reason for why they struck the juror. These Batson challenges, however, are often deemed unsuccessful as more lenient judges can allow striking reasons that appear racially neutral but are clearly discriminatory. The ineffectiveness of Batson challenges to uphold legal accountability alone suggests that technological involvement or other more invasive methods may be promising for maintaining a fair and just jury selection process [6].

More generally, there are no concrete mechanisms for accountability. Journalists and social scientists have until now borne the brunt of the manual work of consolidating and parsing court documents into something comprehensible and useful. Several studies have taken place to better understand the dynamics of jury selection.

At a larger scale, there are no concretized mechanisms for accountability. Journalists and social scientists have until now borne the brunt of the manual work of consolidating and parsing court documents into something comprehensible and useful. Several studies have taken place to better understand the dynamics of jury selection. Most notably, and the study that contributed the data this thesis is founded on, is the American Public Media investigation into the Curtis Flowers case. APM collected 418 trial dockets and trial transcripts from the Mississippi court system spanning from 1992 to 2015 – the years around the Curtis Flowers case. A team of 9 reporters took nearly one year of visiting courthouses, manually pulling out court cases, hand scanning every page of their casefiles, and then reading all documents and assembling the dataset. This process, while unduly burdensome, allowed the reporters to get insights about Doug Evans, the main prosecutor of the Flowers cases, and learn bout how biased his peremptory challenges were. However, it also provides insight into the extent of manual labor necessary to get even a minor glimpse into the court process.

For example, the Jury Sunshine Project at Wake Forest University conducted a similar case study [39] in which they collected felony trials from 2011 in 100 of North Carolina's counties. A large team of undergraduate students, law students, and law librarians manually traveled to clerk's offices across North Carolina from 2013 to 2015 and took notes on the 1,306 trials they were able to access. While they were able to conclude that non-white jurors were excluded in a "persistent" and "predictable" way, their study focused far more on the lack of accessibility of public records. According to them, there exists "no vantage point from which one might see the whole of jury selection, rather than the selection of a single jury" and "poor access to records is the single largest reason why jury selection cannot break out of the litigator's framework to become a normal topic for political debate." [39]

However, current accountability mechanisms largely surround high-profile cases or suspected districts. No progress has been made to provide accountability across jurisdictions en masse in the US. A pipeline that allows for centralized live tracking of legal patterns could potentially be a solution for auditing biased behaviors; however, movement in that area is severely limited by a lack of data access. Furthermore, from the justice perspective, the everyday person who faces biased behaviors has no form of recourse without proper investigation with current mechanisms. Only those lucky enough to become a high-profile case have the potential for their court process to be audited, but the broader impacts of the unseen cases

That is, *data accessibility* and *data processing* are the two main limiting factors in creating a pipeline that allows for auditing for biased legal patterns.

2.3 Eviction

Similar to jury selection, the process of eviction is both a complicated and opaque process outside of the court.

2.3.1 Eviction Process

Evictions begin with a landlord issuing an eviction notice, commonly called a Notice To Quit (NTQ), to a tenant demanding removal from the property. There are three grounds on which a landlord could evict a tenant. The first, and most popular, is standard non-payment of rent, which warrants a 14-day eviction notice. The second is an at-fault eviction, in which the landlord claims the tenant has violated the terms of the lease or otherwise caused disruptions and damages to the property. The final case type is a no-fault eviction, in which a landlord does not claim any faults by the tenant but is asking for repossession of the property. No-fault evictions are commonly used to vacate properties in order to gentrify cities [17]. Both fault and no-fault evictions tend to warrant 30-day eviction notices.

Following the NTQ issuance, the tenant either vacates without disputing the notice, stops doing the behavior that resulted in the fault, if applicable, or is summoned to court. If the latter, a landlord will issue a Summary Process Summons and Complaint (S&C) to the tenant, informing them of the legal action being taken. If a tenant meets the landlord in court, several outcomes could occur – either the tenant wins and no actions are taken or the landlord wins and forcibly removes the tenant from the property. From recent work in understanding patterns in eviction pathways, it known that corporate landlords and landlords with legal representation in court account for the majority of forced-move-out settlements [33].

2.3.2 Exploitative Practices

There are several situations in which evictions are deserving of additional scrutiny. First, a *constructive*, or “self-help" eviction occurs when a landlord intentionally does not keep the property up to livable standards. Prevention of electricity or heating access, coercion, and creating a hostile environment fall under this type of eviction. The tenant must successfully prove this in court using counterclaims to defend their non-payment of rent. However, tenants are usually not represented by a counsel in court, making this process difficult, especially for those with disabilities, language barriers, or who are otherwise unable to make a strong case for themselves in court. An exploitative landlord also has the power to make living conditions so unbearable that a tenant voluntarily evicts themselves rather than having to go through a formal process. A study conducted in Milwaukee from 2009 to 2011 showed that only approximately 24% of evictions during the time period went through the formal court system.

Similarly, landlords may conduct *retaliatory* evictions where they evict a tenant as a consequence of complaining about poor living conditions. Finally, landlords may exploit tenants' lack of legal knowledge in order to intimidate them during the eviction process.

All of these methods would require strict documentation in order to potentially catch harmful patterns; however, a complete view of evictions often requires access to documents beyond the court files. In particular, surveys, including ones from the American Housing Survey (AHS), which collect information on formal and informal evictions, offer a more in-depth understanding of true nature of evictions and their impact. Another example is the Milwaukee Area Renters Study (MARS), which developed a survey to more closely examine reasons for previous evictions. To summarize, shedding light on evictions requires gathering housing and court documentation, foreclosures and mortgages, and tracing individual connections between landlords and the properties that they own through different landlord corporations.

2.3.3 Legal Implications of Unfair Evictions

Being evicted can make it considerably more difficult to find future housing and other opportunities. Furthermore, bias in eviction decisions maintain structural inequalities that only further perpetuate housing stability. For example, Desmond and Gershenson [17] write that large families and tenants with children, low-income tenants, and tenants in disadvantaged neighborhoods are disproportionately affected by evictions. Additionally, in Milwaukee's Black inner-city neighborhoods from 2009 to 2011, Black renters were twice as likely to be evicted through the courts than white renters, and female renters were more than twice as likely to be evicted as male renters [16, 18]. Unfair evictions maintain structural inequalities that only further perpetuate housing stability. For example, Desmond & Gersherson (2016) write that large families and tenants with children, low-income tenants and tenants in disadvantaged neighborhoods are disproportionately affected by evictions.

2.3.4 Current Accountability Mechanisms

Similar to the accountability in jury selection, most current mechanisms are retroactive, occurring after an eviction case has concluded and a tenant files for wrongful eviction. Currently, there are few proactive measures to prevent constructive eviction, except when a tenant is able to present a strong counterclaim during the eviction proceedings.

In addition, there are no large-scale accountability procedures in place. Legal academics such as the Eviction Lab at Princeton University have been able to construct large datasets throughout the years through vigorous effort by large teams [21]. The Eviction Lab has been at the forefront of accountability work in eviction thus far, combining public eviction records with census data and proprietary individual records to assemble the largest dataset covering all 50 US states [21]. Other examples of academic efforts to track evictions include researchers at Georgetown Law and MIT's Department of Urban Studies and Planning, who have conducted a similar process of manually pulling eviction files from the Boston Housing

Court, scanning them, and hand-coding them over several years [33]. Both efforts entail the resource-intensive step of aggregating large numbers of documents and assembling large teams to create datasets from them to analyze for bias.

2.4 Difficulty of Reform

The legal system is one of the most pertinent places where accountability and transparency need to exist, yet do not. The implications of biased judgments and opaqueness in the court system have the potential to affect people's lives and futures.

2.4.1 Prosecutorial Misconduct

Prosecutorial misconduct is defined as when a prosecutor intentionally breaks a law or commits an action that is otherwise against the lawyer's code of ethics during active prosecutions. Existing auditing mechanisms are oftentimes unsuccessful in both identifying perpetrators and enforcing repercussions. The Innocence Project's Nina Morrison explains, "[We] have found documents or notes hidden in a prosecutor's *case file containing information that would have directly supported our client's innocence defense, but which was held back by the prosecutor at trial and kept hidden for decades*. And in other cases, credible leads to suppressed evidence can't be pursued because the original files are destroyed, or witnesses have died or gone missing...*It is very difficult to find proof of misconduct that by definition is designed to stay hidden* – especially when prosecutors hold so much power to control access to what's in their files and to witnesses...Often, the bar discipline committees that are charged with investigating these cases are overwhelmed with other cases, lack expertise in criminal law or, in some cases, are biased in favor of prosecutors and give them every benefit of the doubt" [41]. This only further supplements our work that casefiles mask much of the important processes that occur in legal practices, and being able to parse casefiles before their expiration dates can help to uncover legal biases.

2.4.2 Unstructured Data

Unstructured data adds another layer of complexity to the accountability pipelines. Most legal documents, especially those relevant to everyday, non-high-profile cases, are hastily compiled and thrown into a clerk's office to remain unkempt and disorganized for years until the case file reaches its legally mandated retention period. These documents are often unstructured, which for our purposes is defined as having a non-standardized form of how the information is stored (e.g. a mixture of tables, checkboxes, and plaintext), and a non-standardized form of how the information is presented (e.g. a mixture of handwritten and typed material).

2.5 Large Language Models

With the recent surge in popularity of Large Language Models (LLMs) such as OpenAI's GPT series, Anthropic's Claude, and Google's Gemini, many questions have arisen regarding the application of this technology to automate and support different domains. It is well known that novel advancements in AI have allowed for innovation in diverse domains, from robotics to workforce automation to the legal sector. In the legal domain, however, most of the progress to date lies in using AI to support lawyers by providing recommendations for future steps, conducting contract analysis, explaining legal concepts, or serving as a legal knowledge bank.

2.5.1 Fundamentals

LLMs are designed to understand, generate, and manipulate human language with high accuracy. Based primarily on the transformer architecture, LLMs have allowed for sophisticated text processing and generative capabilities. Transformers use self-attention mechanisms to parse the entire context of an input simultaneously in order to form better long-range dependencies between language elements as compared to previous versions of language models such as RNNs and LSTMs [36].

2.5.2 Emergent Behaviors

Due to the high number of learnable parameters in a LLM (e.g. GPT3 has 1.76 trillion parameters), LLMs are able to be highly performant in language generation and contextual understanding, but also this concept of *emergent behaviors*. These behaviors, initially introduced by Wei et al., refer to unpredictable capabilities in downstream tasks that are not explicitly coded for in the model's architecture or learned during training. Some notable emergent behaviors in LLMs include creative abilities, contextual and logical abilities, and few-shot learning.

Chapter 3

Related Work

The incredible proliferation of data-driven technologies across the U.S. criminal legal system is well-documented [5, 37]. These technologies have mostly prioritized the risk management of crimes, and work from the algorithmic fairness community has largely focused on ensuring these risk management tools satisfy technocratic notions such as accuracy and unbiasedness [7, 8, 12, 19].

Our work aligns with those who call to reimagine how AI systems are used in legal contexts. Specifically, as several works have pointed out, a more substantive understanding of what it means for AI to benefit carceral contexts would go from ensuring the measurement of defendants' pathologies and deficiencies are “fair” and “accurate" to serving decarceral ends [5]. As Chelsea Barbaras further describes: *“an abolitionist re-imagining of AI in criminal law would require shifting away from measuring criminal behavior and towards understanding processes of criminalization, from supporting law and order towards increasing community safety and self-determination, and from surveilling risky populations towards holding accountable state officials.”* [5]. Our work seeks to hold state officials accountable.

Much has been made about the new generative AI technologies, including LLMs [9], and a growing amount of research has explored using LLMs for information retrieval from legal documents. Many of these works focus on automating manually-arduous tasks currently performed by lawyers, such as legal contract review [23] and case summarization [3]. To this end, Guha et al. [22] built a benchmark dataset to measure six different types of legal reasoning: issue-spotting, rule-recall, rule-conclusion, rule-application, interpretation, and rhetorical-understanding. They found that GPT-4 has the strongest overall performance among 20 different LLMs, and that rule-recall and interpretation were the most difficult tasks. This corresponds with the high error rate discovered by Dahl et al. [15] for tasks related to the recall of various information from legal cases that are in the public domain (not from case files).

Only a few studies have explored using LLMs for accountability or transparency in the legal domain. Chien and Kim [11] focus on the potential for LLMs to make legal processes and information more accessible to low-end consumers. Specifically, they developed a GPT-powered chatbot to extract information on Arizona's eviction rules and provide guidance to users on eviction forms and procedures. Pereira et al. [29] investigate the ability of GPT-4 to streamline the processing of Brazilian audit cases. In a pipeline similar to ours, they start with raw case documents and attempt to determine the allegations made as well as the legal admissibility and plausibility of the case. While they hope to assist audit courts in speeding the processing of cases, we differ in our goal of providing accountability of court cases after their resolution. We contribute a novel application of LLMs for automating accountability mechanisms in how cases are adjudicated.

Chapter 4

Data and Methods

4.1 Datasets

We sourced two pre-assembled datasets for this work: one to investigate jury selection in criminal cases and one for eviction cases.

4.1.1 Jury Selection Dataset

We used the aforementioned public dataset of Mississippi criminal trials published by American Public Media (APM) Reports [14], which consists of scanned court records. In particular, the dataset consists of information about 305 criminal trials from 1992 to 2017 in Mississippi's Fifth Circuit Court District.

The jury selection information for each case was present in either (1) a court transcript of the jury selection process or (2) a jury strike sheet.[1] Court transcripts include the voir dire questionnaire process for each juror, as well as the final jury roll call. Both sections of the transcript possibly reveal which jurors were chosen and their gender (based on identifiers like 'Mr.' or 'Mrs.' and other pronouns used to refer to them in the transcript). Jury strike

[1] The raw case files are available at: https://features.apmreports.org/in-the-dark/season-two/source-notes.html

sheets[2] include a complete list of all summoned jurors as well as demarcations for who was struck or chosen. Sometimes, the prosecutor for the case would also include handwritten notes about why the juror was struck and their race and gender, usually coded as 'W' or 'B' for White/Black, and 'M' or 'F' for Male/Female. APM manually reviewed each of these files in order to compile an aggregated dataset[3] of jurors, whether they were selected or struck, and demographic information (if present).

For our experiments, we focus on the feasibility of automating information retrieval from court transcripts. We chose this because information retrieval from court transcripts is more generalizeable to other legal accountability mechanisms and because there are Optical Character Recognition (OCR) bottlenecks to extracting handwritten annotations from the strike sheet (c.f. Discussion). APM coded information about the voir dire questionnaire, the juror demographics, and the trial specifics for all trials. However, we still limit our analysis to the 75 cases that have both a transcript and jury strike sheet for cross-referencing purposes with the anonymized APM dataset of jurors. Among these, we further focus on the 50 cases with a final jury roll call in order to be able to test the effect of fine-tuning (c.f. Experimental Details). We used Adobe Acrobat's OCR technology to extract the text from the scanned transcripts, which ranged anywhere from 15 pages to 400 pages.

4.1.2 Eviction Dataset

For our eviction analysis, we used a dataset of 107 case files from the Boston Housing Court in 2013 assembled by Summers and Steil [33]. Each case file included a Notice To Quit, the Summons and Complaint, counterclaims from the tenant, court order information, and information about the ultimate case disposition. Summers and Steil scanned each of these files from the courthouse, and then manually reviewed each file in order to compile an aggregated dataset of information about each case. We focus on automating information

[2] Figure 4.1 shows example strike sheets.

[3] APM's aggregated dataset of jurors is available at: https://github.com/APM-Reports/jury-data

retrieval from the S&C[4], which includes key legally mandated information about the eviction such as the landlord and tenant representation status, the grounds for eviction, and the amount requested to be paid if relevant. From each case book, we extracted the first page of the S&C and used Microsoft's Azure OCR Model[5] to extract the text from the scanned documents.

4.2 Accountability-Related Retrieval Tasks

We tested a variety of information retrieval tasks related to accountability in both jury selection and eviction. We chose these tasks for their potential to automate parts of the real-world pipelines used by APM and Summers and Steil. Given the different challenges we encountered, we categorize tasks into varying levels of "difficulty" (Table 4.1). We define easy tasks as the retrieval of information directly stated in the document, where no logical or legal reasoning is required. Medium tasks involve some logical reasoning, such as the synthesis or categorization of information. Hard tasks involve the additional component of legal knowledge or reasoning. We also separate out tasks that required parsing human handwriting into their own category because of the OCR bottleneck to analyzing this information. Table 4.1 specifies the tasks by level of difficulty that we chose for each domain.

All tasks were motivated by their relation to accountability in their respective domain. We framed our experiments in this manner to tether our results to potential real-world applications of using LLMs for retrieval of information that influences different conclusions. Applying our framework to our two domains results in the following task definitions.

[4] Figure 4.2 shows an example S&C.

[5] We originally tried Adobe Acrobat OCR on these documents as well, but found Microsoft Azure Document Intelligence to have better performance.

Complexity	**Logical Reasoning**	**Legal Reasoning**	**Handwritten Information**	**Jury Selection**	**Eviction**
Easy	✗	✗	✗	Selected Juror Names	Zip Code
Medium	✓	✗	✗	Jury Gender Composition	Case Type
Hard	✓	✓	✗	Batson Challenges	Landlord Representation
Other (OCR)	–	–	✓	Jury Race Composition	Case Disposition

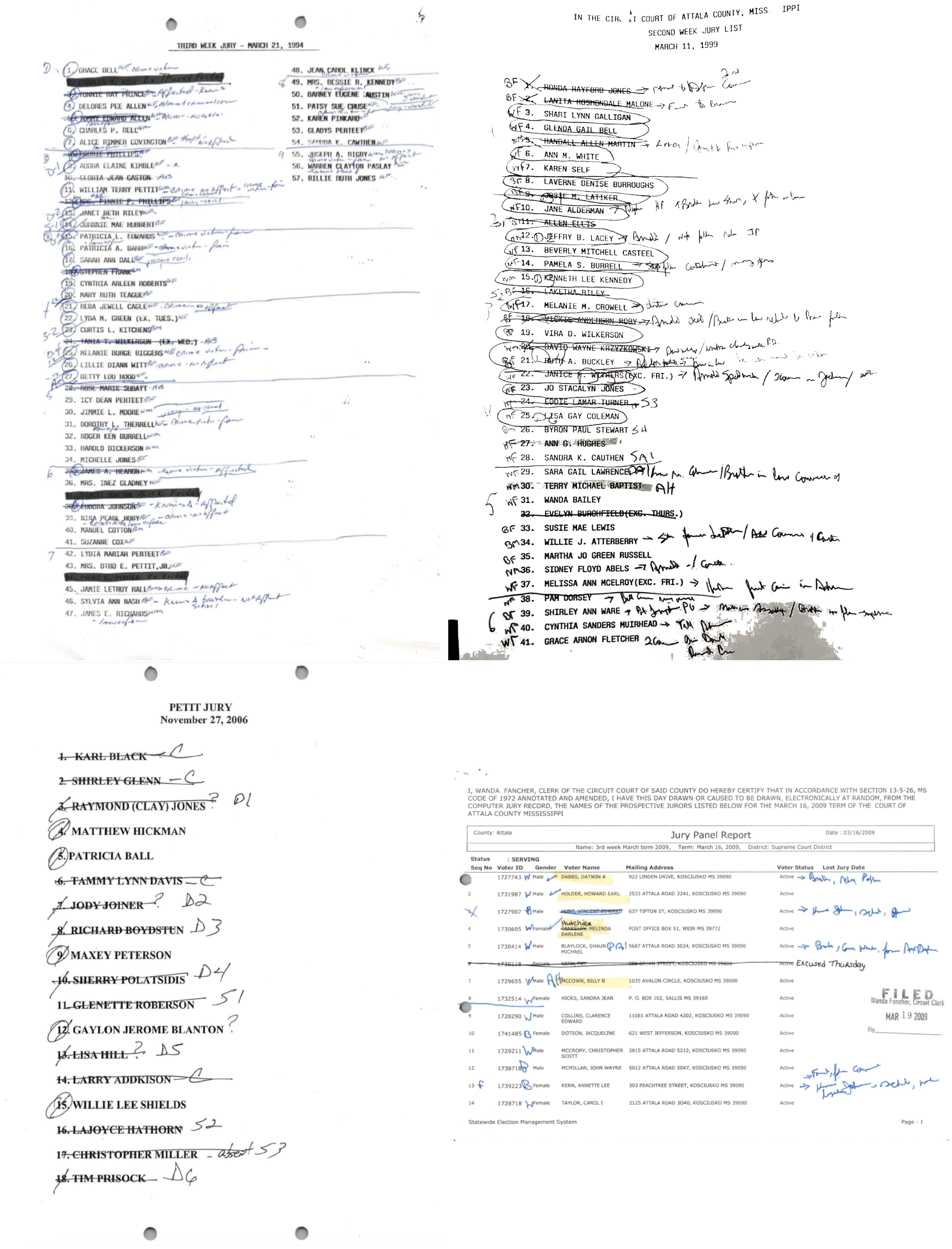
Table 4.1: Task complexity definitions and examples for each domain.

4.2.1 Jury Selection Tasks

For jury selection, we chose extracting the names of selected jurors as the easy task, gender composition as the medium task, and Batson challenges as the hard task. Extracting the names of the final venire of jurors selected to serve on a case is easy because it is always present at the end of the voir dire transcript as part of a final roll call by the judge. The medium task of determining the jury's gender composition requires knowing which jurors were selected (the easy task), finding their gender (specified through pronouns or prefixes like 'Mr.' or 'Mrs.'), and outputting the final count of female and male jurors. The hard task of determining whether a Batson challenge occurred during a case and from which party requires the LLM to understand the legal context of what constitutes a Batson violation. A Batson challenge happens when the opposing party objects to the use of a peremptory strike on the basis that race was used to exclude the juror. We additionally classify the ability to parse racial demarcations from the jury strike sheets as an OCR task.

4.2.2 Eviction Tasks

For eviction cases, we chose extracting the zip code as the easy task, the case type as the medium task, and the landlord's representation status as the hard task. The zip code of the property is directly specified on the S&C and is important for unveiling spatial patterns in discriminatory evictions. The medium task of determining the case type involves classifying the listed reasons for eviction into one of the following categories: "Fault", "No Fault", or "Non-Payment of Rent". This is often not directly stated on the form but can be inferred from simple logical reasoning (e.g., "Damage to the property" would constitute a fault eviction). The hard task of determining the landlord's representation status (whether they were represented by an attorney) requires familiarity with legal signatures. Sometimes, this can be inferred if the signatory name differs from the landlord's name, but in other cases, this can be deduced from whether the printed signature contains an 'Esquire' or 'Esq.' suffix or if an attorney registration number is present. The key motions during the trial and the final disposition of the case are found on the Docket Entries sheet of the casebook (example shown in Figure 4.3), which we classify as an OCR task. All these pieces of information are crucial for accountability into which types of landlords evict more and what reasons they claim for eviction [33].



THIRD WEEK JURY - MARCH 21, 1994

D-1 1. GRACE BELL WF crime victim
[redacted]
~~3. TONNIE RAY PRINCE~~ BM - Affected - Knows
4. DELORES PEE ALLEN WF
~~5. TOMMY EDWARD ALLEN~~ WM - crime - no affect
6. CHARLES P. BELL WM
7. ALICE RIMMER COVINGTON BF - theft - no affect
~~8. BOBBIE PHILLIPS~~ BF
D-1 9. AUDRA ELAINE KIMBLE WF
~~10. GLORIA JEAN GASTON~~ - AB
11. WILLIAM TERRY PETTIT WM - crime no affect - crime victim - fam
~~12. MRS. PINNIE P. PHILLIPS~~ BF jury - conv
D-2 13. JANET BETH RILEY WF
S-1 14. JOHNNIE MAE HUBBERT BF
D-3 15. PATRICIA L. EDWARDS WF - crime victim - fam
16. PATRICIA A. BARR WF - crime victim - fam
17. SARAH ANN DALL BF jury - conv
~~18. STEPHEN FRANK~~ WM
19. CYNTHIA ARLEEN ROBERTS WF
20. MARY RUTH TEAGUE BF
21. REBA JEWELL CAGLE WF . crime - no affect
22. LYDA M. GREEN (EX. TUES.) WF
S-2 23. CURTIS L. KITCHENS BM
~~24. TANIA T. WILKERSON (EX. WED.)~~ - AB
D-4 25. MELANIE BURGE BIGGERS WF crime victim - fam -
26. LILLIE DIANN WITT BF . crime - no affect
27. BETTY LOU HOOD WF
~~28. ROSE MARIE SUBATT~~ - AB
29. ICY DEAN PERTEET BF
30. JIMMIE L. MOORE WM jury - no conv
31. DOROTHY L. THERRELL WF - crime victim - fam
32. ROGER KEN BURRELL WM
33. HAROLD DICKERSON WM
34. MICHELLE JONES BF
~~35. JAMES A. HEARON~~ WM - knows victim - affected
36. MRS. INEZ GLADNEY WF
[redacted]
~~38. EUDORA JOHNSON~~ BF - knows Δ - affected
39. NINA PEARL ROBY BF - crime - no affect - related to law enforce
40. MANUEL COTTON BM
41. SUZANNE COX WF
42. LYDIA MARIAH PERTEET BF
43. MRS. OTHO E. PETTIT, JR. WF
[redacted]
45. JAMIE LETROY HALL BM crime - no affect
46. SYLVIA ANN NASH BF - knows Δ brother - school - not affect
47. JAMES E. RICHARDS WM - law enforcement
48. JEAN CAROL KLINCK WF
49. MRS. BESSIE R. KENNEDY BF
50. BARNEY EUGENE AUSTIN WM
51. PATSY SUE CRUSE WF
52. KAREN PINKARD WF
53. GLADYS PERTEET BF
54. SANDRA K. CAWTHEN WF
55. JOSEPH A. RIGBY WM
56. WARREN CLAYTON PASLAY
57. BILLIE RUTH JONES WF

IN THE CIRC IT COURT OF ATTALA COUNTY, MISS IPPI
SECOND WEEK JURY LIST
MARCH 11, 1999

BF ~~1. RONDA RAYFORD JONES~~ →
BF ~~2. LANITA ROSHONDALE MALONE~~ →
WF 3. SHARI LYNN GALLIGAN
WF 4. GLENDA GAIL BELL
WM ~~5. RANDALL ALLEN MARTIN~~ → Error
WF 6. ANN M. WHITE
WF 7. KAREN SELF
BF 8. LAVERNE DENISE BURROUGHS
BF ~~9. JOSIE M. LATIKER~~
WF 10. JANE ALDERMAN →
BM ~~11. ALLEN ELLIS~~
WM 12. JEFFRY B. LACEY →
WF 13. BEVERLY MITCHELL CASTEEL
WF 14. PAMELA S. BURRELL →
WM 15. KENNETH LEE KENNEDY
S2 BF ~~16. LAKETHA RILEY~~
WF 17. MELANIE M. CROWELL →
BF ~~18. VICKIE ANDERSON ROBY~~ →
BF 19. VIRA D. WILKERSON
WM ~~20. DAVID WAYNE KRZYZKOWSKI~~ →
BF 21. ~~RUTH A. BUCKLEY~~ →
WF 22. JANICE M. WITHERS(EXC. FRI.) →
WF 23. JO STACALYN JONES
WM ~~24. EDDIE LAMAR TURNER~~ S3
WF 25. LISA GAY COLEMAN
BM 26. BYRON PAUL STEWART S4
WF ~~27. ANN G. HUGHES~~
WF 28. SANDRA K. CAUTHEN SA1
WF 29. SARA GAIL LAWRENCE
WM 30. TERRY MICHAEL BAPTIST AH
WF 31. WANDA BAILEY
~~32. EVELYN BURCHFIELD(EXC. THURS.)~~
BF 33. SUSIE MAE LEWIS
BM 34. WILLIE J. ATTERBERRY →
BF 35. MARTHA JO GREEN RUSSELL
WM 36. SIDNEY FLOYD ABELS →
WF 37. MELISSA ANN MCELROY(EXC. FRI.) →
WF 38. PAM DORSEY →
BF 39. SHIRLEY ANN WARE →
WF 40. CYNTHIA SANDERS MUIRHEAD →
WF 41. GRACE ARNON FLETCHER

PETIT JURY
November 27, 2006

~~1. KARL BLACK~~ - C
~~2. SHIRLEY GLENN~~ - C
~~3. RAYMOND (CLAY) JONES~~ ? D1
4. MATTHEW HICKMAN
5. PATRICIA BALL
~~6. TAMMY LYNN DAVIS~~ - C
~~7. JODY JOINER~~ ? D2
~~8. RICHARD BOYDSTUN~~ D3
9. MAXEY PETERSON
~~10. SHERRY POLATSIDIS~~ D4
~~11. GLENETTE ROBERSON~~ S1
12. GAYLON JEROME BLANTON ?
~~13. LISA HILL~~ ? D5
~~14. LARRY ADDKISON~~ - C
15. WILLIE LEE SHIELDS
~~16. LAJOYCE HATHORN~~ S2
~~17. CHRISTOPHER MILLER~~ - absent S3
~~18. TIM PRISOCK~~ - D6

I, WANDA FANCHER, CLERK OF THE CIRCUIT COURT OF SAID COUNTY DO HEREBY CERTIFY THAT IN ACCORDANCE WITH SECTION 13-5-26, MS CODE OF 1972 ANNOTATED AND AMENDED, I HAVE THIS DAY DRAWN OR CAUSED TO BE DRAWN, ELECTRONICALLY AT RANDOM, FROM THE COMPUTER JURY RECORD, THE NAMES OF THE PROSPECTIVE JURORS LISTED BELOW FOR THE MARCH 16, 2009 TERM OF THE COURT OF ATTALA COUNTY MISSISSIPPI

County: Attala
Jury Panel Report
Date : 03/16/2009
Name: 3rd week March term 2009, Term: March 16, 2009, District: Supreme Court District

Status : SERVING

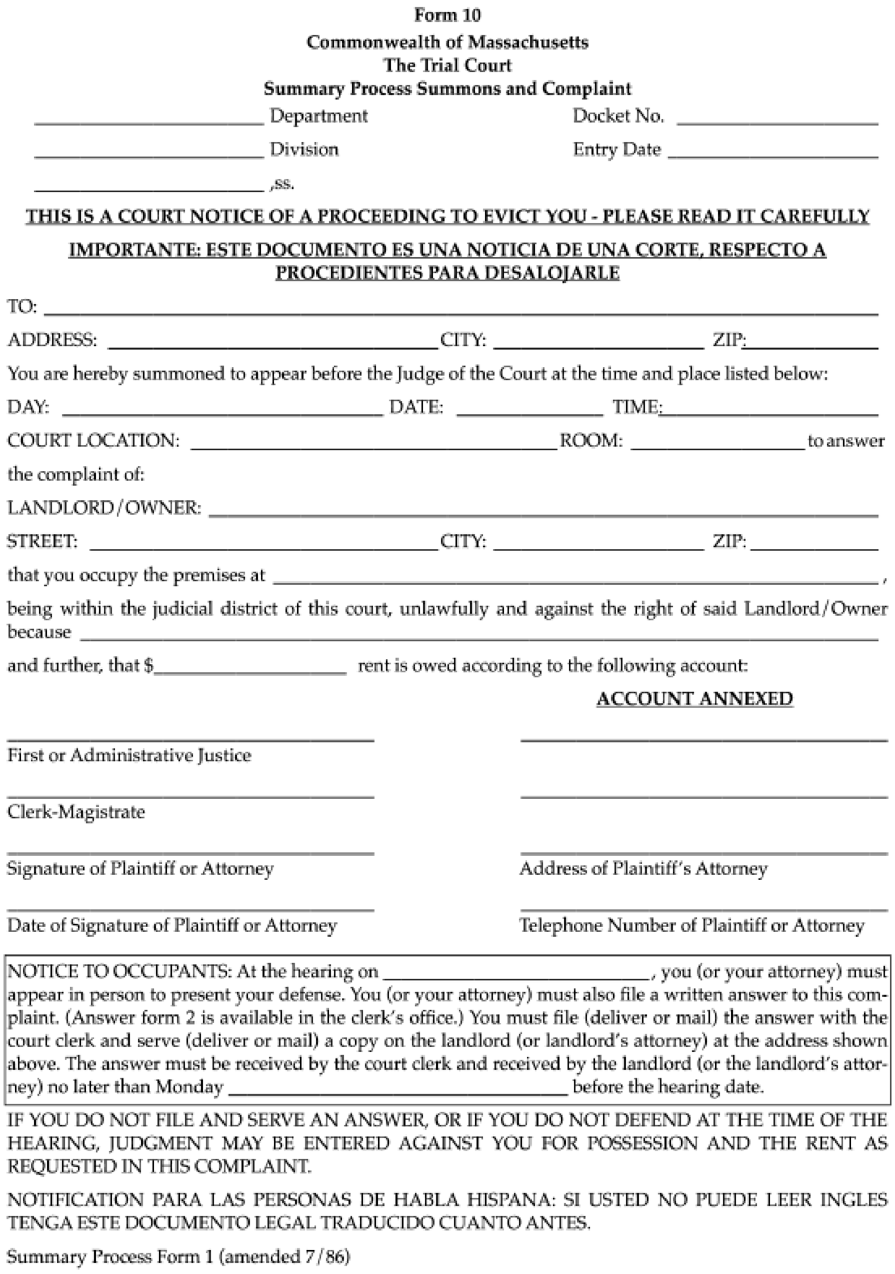
Seq No	Voter ID	Gender	Voter Name	Mailing Address	Voter Status	Last Jury Date
	1727743	W Male	DABBS, DATWIN A	922 LINDEN DRIVE, KOSCIUSKO MS 39090	Active →	
2	1731987	W Male	HOLDER, HOWARD EARL	2533 ATTALA ROAD 3241, KOSCIUSKO MS 39090	Active	
3	1727907	B Male	~~HUNT, VINCENT EDWARD~~	637 TIPTON ST, KOSCIUSKO MS 39090	Active →	
4	1730605	W Female	Hutchison ~~TANKSLEY,~~ MELINDA DARLENE	POST OFFICE BOX 51, WEIR MS 39772	Active	
5	1730414	W Male	BLAYLOCK, SHAUN MICHAEL PA1	5687 ATTALA ROAD 3024, KOSCIUSKO MS 39090	Active →	
~~6~~	~~1730118~~	~~Female~~	~~KERN, FAY~~	~~200 BRYAN STREET, KOSCIUSKO MS 39090~~	~~Active~~ Excused Thursday	
7	1729655	W Male	Alt MCCOWN, BILLY B	1035 AVALON CIRCLE, KOSCIUSKO MS 39090	Active	
8	1732514	W Female	HICKS, SANDRA JEAN	P. O. BOX 102, SALLIS MS 39160	Active	
9	1728290	W Male	COLLINS, CLARENCE EDWARD	11081 ATTALA ROAD 4202, KOSCIUSKO MS 39090	Active	
10	1741485	B Female	DOTSON, JACQUELINE	621 WEST JEFFERSON, KOSCIUSKO MS 39090	Active	
11	1729211	W Male	MCCRORY, CHRISTOPHER SCOTT	2815 ATTALA ROAD 5212, KOSCIUSKO MS 39090	Active	
12	1738718	B Male	MCMILLAN, JOHN WAYNE	5012 ATTALA ROAD 5047, KOSCIUSKO MS 39090	Active	
13 f	1739223	B Female	KERN, ANNETTE LEE	303 PEACHTREE STREET, KOSCIUSKO MS 39090	Active →	
14	1728718	W Female	TAYLOR, CAROL I	2125 ATTALA ROAD 3040, KOSCIUSKO MS 39090	Active	

FILED
Wanda Fancher, Circuit Clerk
MAR 19 2009
By________

Statewide Election Management System

Page : 1

Figure 4.1: Example strike sheets showing the variance in note-taking that occurs to document juror demographics and strike status. Common demarcations include 'W'/'B' for race, 'F'/'M' for gender, SX/DX for state and defense strikes, and 'C' for for-cause strikes.



Form 10
Commonwealth of Massachusetts
The Trial Court
Summary Process Summons and Complaint

________________ Department Docket No. ______________

________________ Division Entry Date ______________

________________ ,ss.

THIS IS A COURT NOTICE OF A PROCEEDING TO EVICT YOU - PLEASE READ IT CAREFULLY

IMPORTANTE: ESTE DOCUMENTO ES UNA NOTICIA DE UNA CORTE, RESPECTO A PROCEDIENTES PARA DESALOJARLE

TO: ____

ADDRESS: ____________________ CITY: ______________ ZIP: __________

You are hereby summoned to appear before the Judge of the Court at the time and place listed below:

DAY: ____________________ DATE: __________ TIME: ______________

COURT LOCATION: ______________________ ROOM: ____________ to answer the complaint of:

LANDLORD/OWNER: ____

STREET: ____________________ CITY: ______________ ZIP: __________

that you occupy the premises at ____ ,

being within the judicial district of this court, unlawfully and against the right of said Landlord/Owner because ____

and further, that $______________ rent is owed according to the following account:

ACCOUNT ANNEXED

______________________	______________________
First or Administrative Justice	
______________________	______________________
Clerk-Magistrate	
______________________	______________________
Signature of Plaintiff or Attorney	Address of Plaintiff's Attorney
______________________	______________________
Date of Signature of Plaintiff or Attorney	Telephone Number of Plaintiff or Attorney

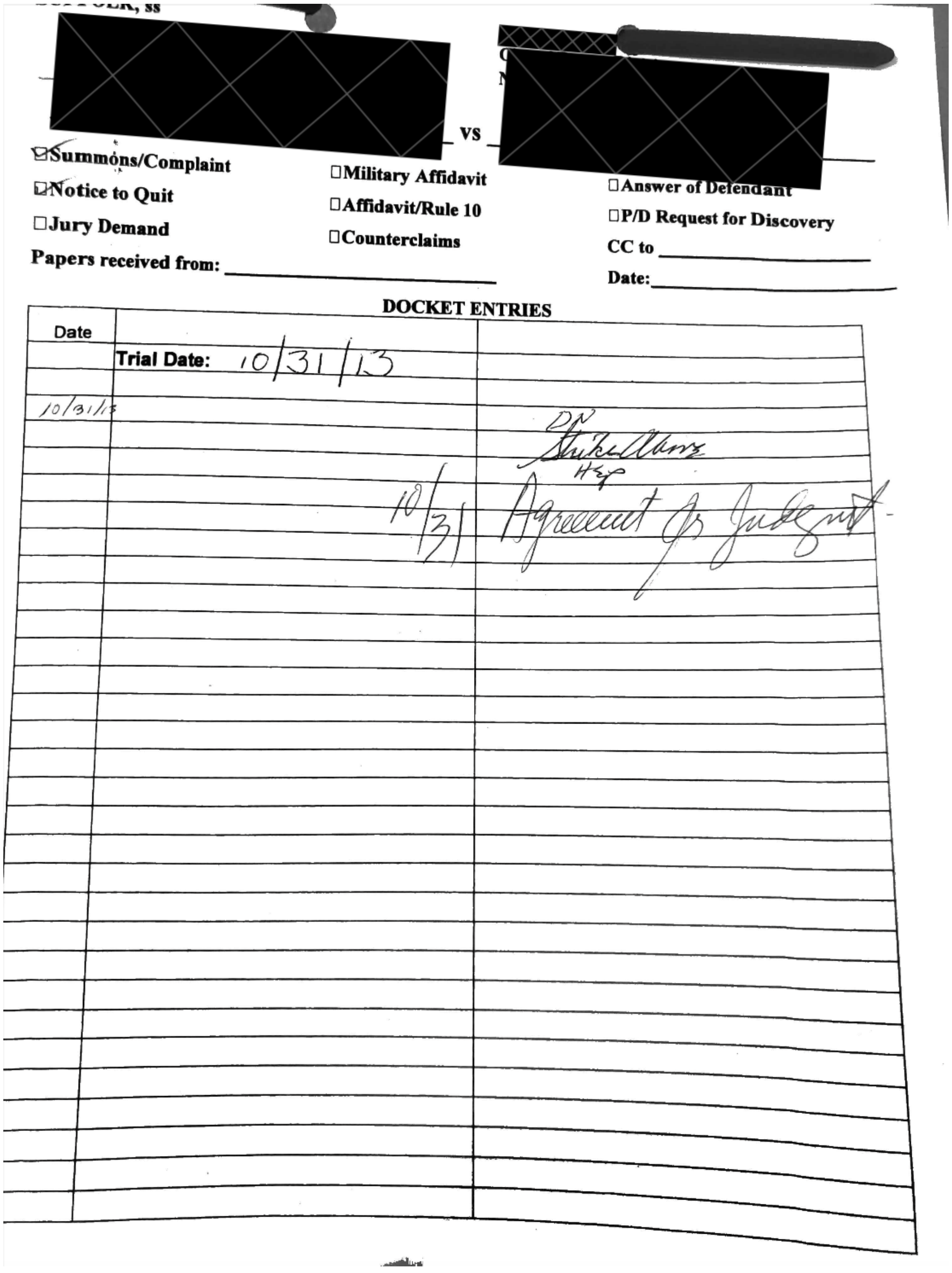
NOTICE TO OCCUPANTS: At the hearing on ________________, you (or your attorney) must appear in person to present your defense. You (or your attorney) must also file a written answer to this complaint. (Answer form 2 is available in the clerk's office.) You must file (deliver or mail) the answer with the court clerk and serve (deliver or mail) a copy on the landlord (or landlord's attorney) at the address shown above. The answer must be received by the court clerk and received by the landlord (or the landlord's attorney) no later than Monday ______________________ before the hearing date.

IF YOU DO NOT FILE AND SERVE AN ANSWER, OR IF YOU DO NOT DEFEND AT THE TIME OF THE HEARING, JUDGMENT MAY BE ENTERED AGAINST YOU FOR POSSESSION AND THE RENT AS REQUESTED IN THIS COMPLAINT.

NOTIFICATION PARA LAS PERSONAS DE HABLA HISPANA: SI USTED NO PUEDE LEER INGLES TENGA ESTE DOCUMENTO LEGAL TRADUCIDO CUANTO ANTES.

Summary Process Form 1 (amended 7/86)

Figure 4.2: Example Summary Process Summons and Complaint issued by the landlord to call the tenant to court and inform them of the grounds of eviction.



SUFFOLK, ss

[redacted] vs [redacted]

[redacted]

☑ Summons/Complaint
☑ Notice to Quit
□ Jury Demand

□ Military Affidavit
□ Affidavit/Rule 10
□ Counterclaims

□ Answer of Defendant
□ P/D Request for Discovery

Papers received from: ____________

CC to ____________

Date: ____________

DOCKET ENTRIES

Date		
	Trial Date: 10/31/13	
10/31/13		DN
	10/31	Agreement for Judgment

Figure 4.3: Example docket entry page including the final disposition (Agreement for Judgement) of an eviction case. The variability in handwriting and format of this page makes it difficult to automatically extract information.

Chapter 5

Experimental Setup

We performed all our experiments using OpenAI's GPT-4 Turbo model (`gpt-4-turbo-2024-04-09`). We chose this model because prior work found it to have the best performance in the legal domain [22, 29]. We use a zero-shot prompt structure for our main analysis. We also experimented with two-shot prompting and sequential prompting for the more difficult tasks, as elaborated on in Table 6.6. Table 5.1 specifies the exact prompt we used for each task, which we arrived at after testing variations on a few documents. For each case document and task, we ran the query 5 times.

Performance Measures.

Our primary metric for all tasks is accuracy: the percentage of fully correct responses to the prompt over all cases and iterations. For the jury gender composition task, we also report the absolute error: the sum of the differences in predicted and actual counts of male and female jurors. We further conduct downstream impact tests to check how well the LLM predictions might be able to answer accountability questions in each domain.

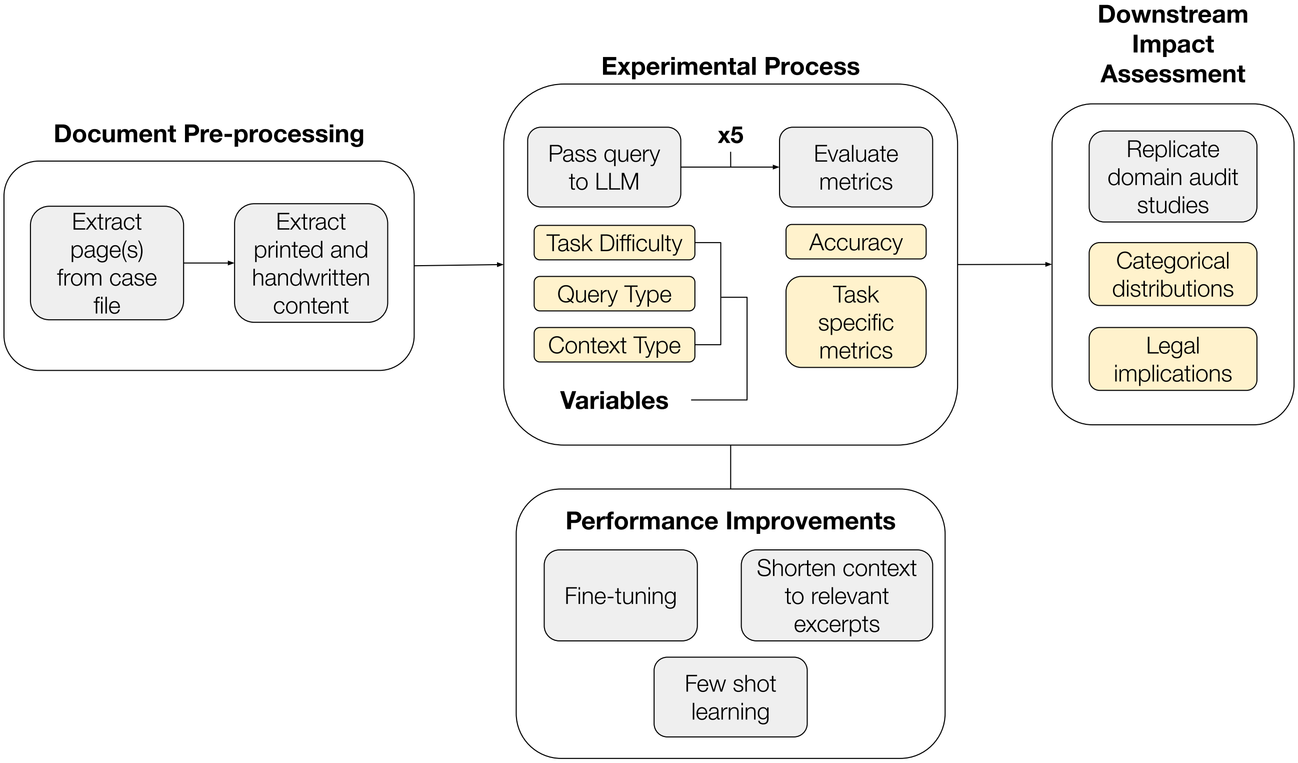


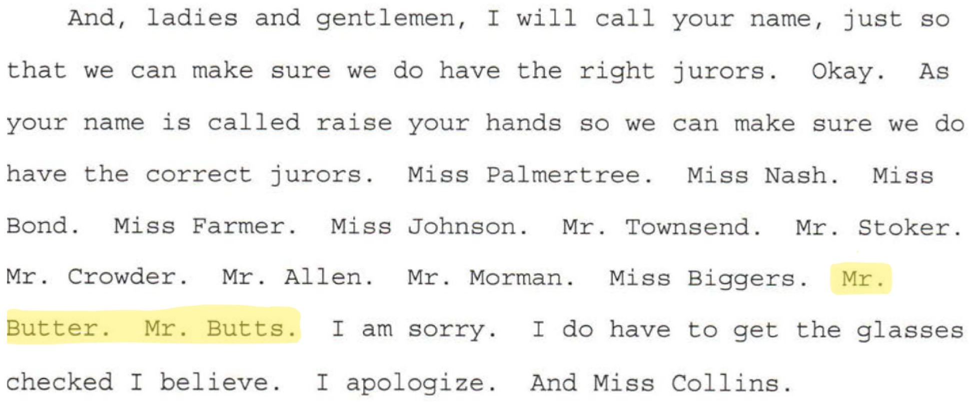
Figure 5.1: Overview of experimental pipeline from data processing through final evaluations.

Jury Selection Transcript Excerpts and Fine-Tuning.

In order to improve jury selection performance, we experimented with shorter input lengths and fine-tuning. Instead of using the full court transcript, which ranged from 15 to 400 pages, we tried the easy and medium tasks with just the excerpt of the final jury roll call (around 1 page). These excerpts could be extracted in an automated pipeline through keyword searches. Querying the excerpt is still non-trivial because judges will often repeat or mispronounce names in the roll call, as well as refer to jurors with the wrong pronouns. We also experiment with fine-tuning on the excerpts[1] for the medium task of jury gender composition. Fine-tuning was done using `gpt-3.5-turbo-0125`, and results were averaged over ten iterations of a 60/40 train-test split[2].

[1] GPT does not currently support fine-tuning on large documents which is why we did not try this for the full-length transcripts or for the hard task.

[2] In our sample of 50 cases, this corresponds to 30 cases in the training set and 20 cases in the test set.



And, ladies and gentlemen, I will call your name, just so that we can make sure we do have the right jurors. Okay. As your name is called raise your hands so we can make sure we do have the correct jurors. Miss Palmertree. Miss Nash. Miss Bond. Miss Farmer. Miss Johnson. Mr. Townsend. Mr. Stoker. Mr. Crowder. Mr. Allen. Mr. Morman. Miss Biggers. Mr. Butter. Mr. Butts. I am sorry. I do have to get the glasses checked I believe. I apologize. And Miss Collins.

Figure 5.2: Example of a jury selection voir dire transcript excerpt. We extracted these excerpts of the final jury roll call in order to improve performance on the tasks of extracting selected juror names and determining the jury's gender composition. The highlighted segment displays a disfluency that causes the model to miscount jurors.

Task	Zero-shot Prompt
Selected Juror Names	Can you give me the names of the final list of jurors that were selected to serve on this case as a comma separated list? Include alternate jurors, if present. Do not output any other text, explanations, or annotations, and make sure to give the juror names.
Jury Gender Composition	Count the number of female and male jurors that were chosen to serve on this case. Female jurors are denoted using Ms. in the transcript, and males using Mr. Only count jurors that were chosen to serve on the jury or serve as alternates. Only output the number as a comma separated list. There should only be 12-14 jurors total. Do not output any other text, explanations, or annotations.
Batson Challenges	Can you output if there was a Batson challenge claim made by the defense and state respectively? A Batson challenge happens when a party objects the opposing party's peremptory challenge on grounds that it was used to exclude a potential juror based on race, ethnicity, or sex. Output as a comma-separated value with 'Yes' or 'No' for each of the parties. Do not output any text, explanations, or annotations.
Zip Code	Can you give me the zip code of the property in this case? Do not output any other text, explanations, or annotations.
Landlord Type	Can you give me the type of landlord involved in this case? Output either 'Corporation', 'Individual', or 'Boston Housing Authority'. Do not output any other text, explanations, or annotations.
Eviction Case Type	Can you give me the type of eviction case this is? Output either 'Nonpayment', 'Fault', or 'No', if there is no fault. Do not output any other text, explanations, or annotations.
Landlord Representation	Can you output if the landlord was represented by a counsel. Representation by counsel is if an attorney signed a pleading or court document on behalf of the landlord/tenant at any point during the case. Do not output any other text, explanations, or annotations. If the information is not there, output 'N/A'.

Table 5.1: Zero-shot Prompts

Chapter 6

Results and Challenges

We begin by presenting the overall performance across the task gradations and then investigating our more fine-grained experiments on the downstream impact tests. We then discuss avenues to improve performance.

6.1 Performance Across Task Difficulties

Overall, we find that as the difficulty of the task increases, the performance across the documents for both jury selection and eviction tasks decreases. We consider the relationship between performance and our task stratification for both domains using zero-shot prompting (Figure 6.1). Table 5.1 in the Appendix lists the prompts used for all experiments. LLMs perform the best when solving the easy tasks in both domains, which is expected as the task merely demands a simple search and return scheme. However, LLMs struggle when logical reasoning is required, as in the medium and hard tasks in both domains. The variance of the model's performance is directly related to the task difficulty as seen with the 95% CI (c.f. Figure 6.1).

Domain	# Cases	Input Type	Two-Shot	Fine-Tuning	Easy	Medium	Hard
Jury Selection	50	Full Transcript	×	×	81.6 ± 4.8	3.6 ± 2.3	23.2 ± 5.2
		Full Transcript	✓	×	94.8 ± 2.8	18.4 ± 4.8	76.8 ± 5.2
		Transcript Excerpt	×	×	100.0 ± 0.0	23.8 ± 5.3	–
		Transcript Excerpt	×	✓	–	34.0 ± 6.6	–
Eviction	107	Summons & Complaint	×	×	95.9 ± 1.7	89.9 ± 2.6	70.7 ± 3.8

Table 6.1: Accuracy with 95% CI by domain, experiment type, and task complexity. Computed over all iterations (5 per case).

6.1.1 Easy Tasks.

For the tasks that require no reasoning, the model achieves consistently good performance. We track two different errors that occur at this level of task difficulty. The first, which is only observed in jury selection cases, is *incomplete recall*, where the model "forgets" part of the answer to the task. The other error, present in both jury and eviction cases, is an incorrect output, such as outputting the juror IDs rather than their names, or the zip code of the landlord's office instead of the property.

6.1.2 Medium Tasks.

When the tasks demand logical reasoning, we see a drop off in performance, especially for the jury selection case[1]. We also observe more complex failure points for these tasks. The primary failure point in the jury selection task of gender aggregation is related to the speaking patterns during the roll call. For example, sometimes a prosecutor will misread a name and proceed to correct it, or call the same person twice (Figure 5.2 shows an example of this). These prose patterns and mistakes have no bearing on gender aggregation but may introduce errors if logic is not used to pick up on these mistakes. For the eviction task of extracting the case type, the primary failure point is if rent is owed while another reason for eviction is listed. For example, in fault cases, a missed payment amount may be included along with

[1] We elaborate more on potential hypotheses for the domain discrepancies in the Discussion.

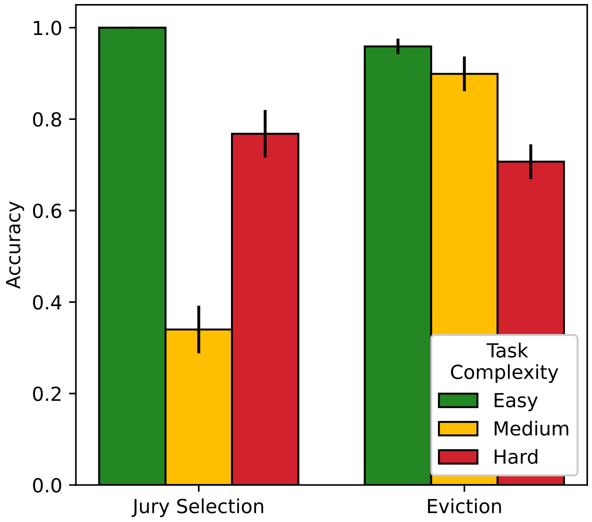


Figure 6.1: Accuracy by Task Complexity. Jury Selection results from the best experimental types (c.f. Table 6.1). Error bars represent 95% CI over all cases and iterations (five per case).

the descriptions of tenant fault; however, the fault would take priority as that is the grounds for eviction.

6.1.3 Hard Tasks.

The hard tasks had differing performances across the two domains. In the jury case, the accuracy was better than the medium task accuracy in the zero-shot case, though it was still lower than the easy case standing at 23.2%. We elaborate more on the two-shot performance in the Few-Shot Learning section. However, for the eviction documents, we observe a strictly decreasing accuracy as difficulty increases, with the most difficult task achieving 70.7% accuracy. The performance in the eviction case is expected, since we hypothesized that requiring both legal and logical reasoning would be more complex for the model. There were also no observable patterns of errors in the jury selection case, while the main source of error in the eviction case was outputting that “No information could be found,” when more logical reasoning was required to understand whether or not a landlord was represented.

6.1.4 Confabulation.

When checking the model to determine the distribution of errors, we also tracked if model hallucinations were an issue, especially since [15] noted high rates of hallucination for high-complexity legal queries. However, we found only one instance of a hallucinated response across all the trials of all experiments, suggesting that information extraction queries do not suffer from similar queries without any context provided.

6.2 Improving Jury Selection Performance

We ran additional experiments to explore different avenues of improving performance on the jury selection tasks. We primarily focus on the medium jury selection task given that it had the worst performance across all tested tasks.

Context Type	Query Type	# Cases	Precision	Recall	F1
Transcript	Zero Shot	75	0.83 ± 0.023	0.837 ± 0.023	0.83
Transcript	Two Shot	75	**0.92 ± 0.016**	**0.92 ± 0.046**	**0.92**
Transcript	Zero Shot	50	0.82 ± 0.037	0.82 ± 0.037	0.82
Transcript	Two Shot	50	0.95 ± 0.014	0.95 ± 0.014	0.95
Excerpt	Zero Shot	50	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0**

Table 6.2: Average metrics for the jury selection easy task of name extraction by context type, query type, and the number of cases across 5 iterations per case. Reported with standard error across the number of cases.

6.2.1 Few-shot Learning.

We investigated whether few-shot learning improved performance by providing examples to the model of the logic required to accomplish the task. We only ran the few-shot experiments for the jury selection tasks as the performance across the board was worse than the eviction cases performance, and as the context length for the jury selection tasks is sufficiently long

Context	Query Type	# Cases	Female Error	Male Error	Total Error	Accuracy
Transcript	Zero Shot	75	2.08 ± 0.154	2.12 ± 0.154	4.20 ± 5.42	0.269
Transcript	Two Shot	75	**1.62 ± 0.139**	**1.55 ± 0.131**	**3.18 ± 0.251**	**0.144**
Transcript	Zero Shot	50	2.07 ± 0.172	2.01 ± 0.171	4.11 ± 4.67	0.360
Transcript	Two Shot	50	1.28 ± 0.134	1.33 ± 0.141	2.61 ± 0.263	0.184
Excerpt	Zero Shot	50	1.08 ± 0.114	1.09 ± 0.133	2.17 ± 0.233	0.238
Excerpt	Fine-Tuned	50	**0.63 ± 0.028**	**0.77 ± 0.04**	**1.40 ± 0.053**	**0.340**

Table 6.3: Average metrics for the jury selection medium task of gender aggregation by context type, query type, and the number of cases across 5 iterations per case. Reported with standard error across the number of cases

Context Type	Query Type	# Cases	Defense Accuracy	State Accuracy	Accuracy
Transcript	Zero Shot	75	0.669 ± 0.035	0.464 ± 0.050	0.235
Transcript	Two Shot	75	**0.837 ± 0.026**	**0.872 ± 0.012**	**0.755**
Transcript	Zero Shot	50	0.684 ± 0.045	0.460 ± 0.062	0.232
Transcript	Two Shot	50	**0.848 ± 0.032**	**0.892 ± 0.02**	**0.768**

Table 6.4: Average metrics for the jury selection hard task of identifying Batson challenges, by context type, query type, and the number of cases across 5 iterations per case. Reported with standard error across the number of cases.

enough to benefit from more concise examples. We found that, for all three task levels, the performance improved with two-shot learning compared to zero-shot learning. However, we observe that the hard task had the largest jump in performance, from 23.2% to 76.8%. This result suggests that legal contextual information is easier to inject into the model than other more nuanced forms of logic. We also experiment with other types of prompting, as described in Section 6.6, but we find that two-shot prompting persists as the most performant query type.

6.2.2 Performance Across Document Length.

Between the full transcripts and the excerpts without fine-tuning or few-shot learning according to the 0-1 accuracy, we see an improvement in accuracy from 3.6% to 23.8%. However, if we examine the total absolute errors (Figure 6.2), we see improvements in the performance as well as reduced variance across the documents. These results are expected since shortening the query context to only the section of the text with the final roll call simplifies the synthesis that needs to occur and removes the possibility of extraneous extractions. However, shortening the excerpt does not have any substantial bearing on the ease of logical reasoning.

6.2.3 Fine Tuning.

The performance improvements between the full documents and the excerpts suggested that fine-tuning would help the model learn the logic patterns to carry out the jury gender aggregation task. In fine tuning on the 50 selected jury roll call excerpts, we observed an increase of the overall accuracy from 23.8% without fine tuning to 34% with fine tuning. Fine tuning for the medium-level logic was not successful in substantially improving the model's overall accuracy. However, the average total absolute error and variance reduced by 35.7% and 94.9% respectively after fine tuning (Figure 6.3). This suggests that the primary benefit of fine tuning for this task is in the reduction of variance, while eliminating errors in logical deduction may require more task-specific engineering.

6.3 Downstream Impact Tests

To contextualize the performance of our queries, we generated downstream impact tests for both jury selection and eviction. These tasks aim to demonstrate the impact of the model's performance on tasks and questions that accountability work would ask. We framed our downstream impact tests in the form of auditing questions, meaning tests that would be useful for researchers and journalists to use to understand the legal system better.

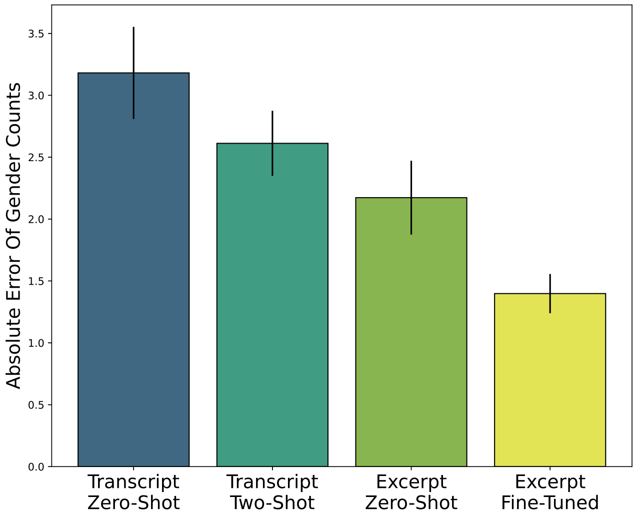


Figure 6.2: Absolute Error of Jury Gender Aggregation by Query Type (c.f. Table 6.3). Error bars represent standard error over 50 samples.

6.3.1 Jury Selection.

In their original white paper, APM Reports wanted to understand the female-to-male selection ratios on the jury and how they differed across specific prosecutors. We replicated that study for our downstream impact test to see if the errors made by the model impacted the broader conclusions one could draw about the dataset. In doing this, we hope to elucidate the real-world implications of these models being added to accountability pipelines.

We find that, when using the LLM to extract information from the corpus of documents, the big-picture conclusions are generally not affected. However, as the question becomes increasingly fine-grained, the resultant answers are affected differently. The overall Female:Male ratio across all cases decreased from 1.49 to 1.18, meaning the model still predicted female-dominance, but the margin of dominance decreased substantially. We found large changes in the ranking of most female-biased juries by county and prosecutor (Tables 6.5 and 6.6). We see that, by county, the ranking changed for four of the seven counties, whereas for prosecutors, the ranking changed for all but one prosecutor. Of the 75 total jury cases, the dominant gender of the jury was flipped in 20 of them.

County	Original F:M Ratio
Attala	2.794
Montgomery	2.115
Winston	2.082
Grenada	1.642
Webster	1.605
Carroll	1.222
Choctaw	1.165

County	Inferred F:M Ratio
Attala	2.391
Montgomery	1.468
Grenada ↑	1.299
Winston ↓	1.215
Choctaw ↑	1.054
Carroll	1.037
Webster ↓	1.027

Table 6.5: Original and inferred rankings of counties based on the average female-to-male ratio of jury composition.

Prosecutor	Original F:M Ratio
Greg Meyer	3.550
Mickey Mallette	2.859
Michael Howie	2.143
Susan Denley	2.088
Walter Bleck	2.058
Doug Evans	1.967
Kevin Horan	1.611

Prosecutor	Inferred F:M Ratio
Doug Evans ↑	2.180
Susan Denley ↑	1.654
Mickey Mallette ↓	1.494
Greg Meyer ↓	1.383
Walter Bleck	1.263
Michael Howie ↓	1.253
Clyde Hill ↑	1.232

Table 6.6: Original and inferred rankings of prosecutors based on the average female-to-male ratio of jury composition.

While the overall gender dominance of juries seems to not have been changed using the LLM's outputs, it is evident that the errors introduced can substantially alter the outcomes of a potential audit. For example, Doug Evans, who was originally ranked as the 6th most female-biased prosecutor, was moved to 1st using the LLM extractions. A prosecutor that was not included in the original top 7 was introduced in the inferred ranking. If applied to a larger dataset that included more jurisdictions' and states' documents, these effects would be further exacerbated, making it difficult to directly rely on LLMs for accountability in this scenario.

6.3.2 Eviction.

Similarly, Summers & Steil wanted to understand the breakdown of characteristics most influential in determining which cases led to forced tenant move-outs. We replicate this analysis to see if incorrect retrieval or misclassification of the case characteristics impacted the legal conclusions drawn about eviction pathways. We focused specifically on the distribution of the different eviction reasons and landlord representations resulting in a forced tenant move-out.

In the original dataset, 72.2% of forced move-outs came from nonpayment cases, while 11.1% came from fault cases and 16.7% came from no-fault cases. Using the LLM inferences, the distribution became 88.9%, 11.1%, and 0% respectively. We also find that in the original dataset, 43.6% of forced move-outs came from individual landlords, while 51.2% came from corporate landlords and 5.1% came from the Boston Housing Authority. Using the LLM inferences, the percentages become 43.6%, 48.7%, and 7.7% respectively The percentages were unchanged for the representation status of the landlord – 66.6% were represented. Figure 6.3 visualizes the change in distributions for all three features. These modifications fundamentally do not change the dominant narrative about this data – the majority of forced move-outs still come from nonpayment cases, corporate landlords, and landlords that are represented by a counsel. However, the lack of any predicted no-fault cases could mask the effects of these types of cases in eviction pathways. This suggests that while eviction data may lend itself better than jury selection data to opportunities using LLMs to extract information, there are still downstream ramifications to be wary of.

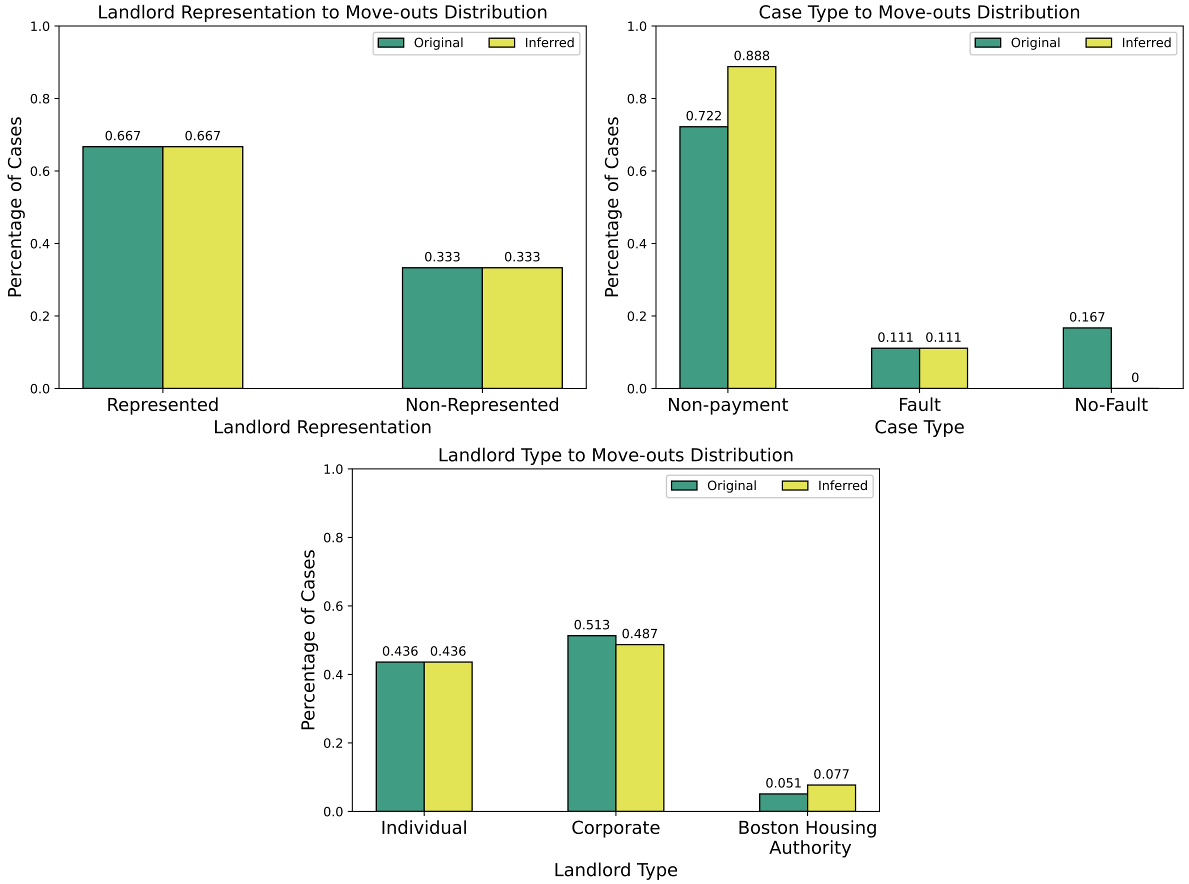


Figure 6.3: Eviction impact tests showing the change in distributions of key case features in the original dataset and according to the predictions by the LLMs.

6.4 Additional Experiments

6.4.1 Correlations

In order to determine the relationships between the performance on the different tasks, we calculated simple linear regressions on different features as follows. We conduct these experiments on all 75 case documents for the jury selection domain.

Inferred vs. Original Medium Performance

To ensure that the model wasn't randomly guessing the gender aggregation, we regressed the inferred counts of females and males on the jury against their true counts. We perform this regression for both the zero shot and two shot queries, resulting in Figure 6.4. We observe that in both cases, the two shot regression ($r^2 = 0.538$ and 0.608) was significantly more correlated than the zero shot parallel ($r^2 = 0.213$ and 0.283). We observe a slight increase from the female to male regressions, which is expected since the male prediction performs slightly better (Table 6.3).

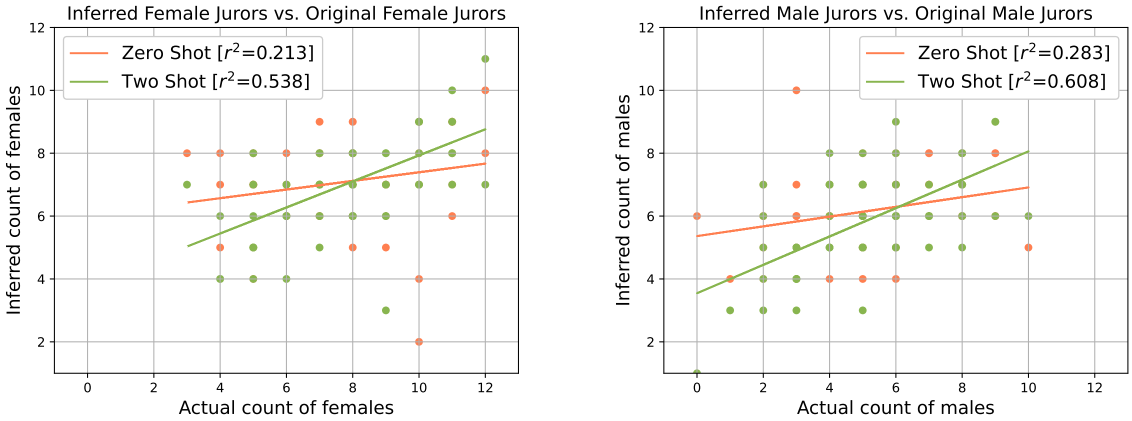


Figure 6.4: Correlations between the true gender counts and inferred gender counts in the medium task. Split by gender since we observed a general performance gap across the genders.

Easy Task vs. Medium Task

The jury selection easy and medium tasks had an inherent logical relationship since the medium task asked for a count of the genders of all the selected jurors, which would require first completing the easy task of extracting the final list of jurors before aggregating their genders. Thus, we hypothesized that cases with better performance on the easy task would thus have better performance on the medium task. From the regressions in Figure 6.5 we see a negligible relationship between both precision and recall of the name extraction and the gender aggregation. Even if the model performs with perfect precision and recall on a case, it does not necessarily perform better on the gender aggregation. That is, the model struggles primarily during the logical step between extracting the names and aggregating their genders.

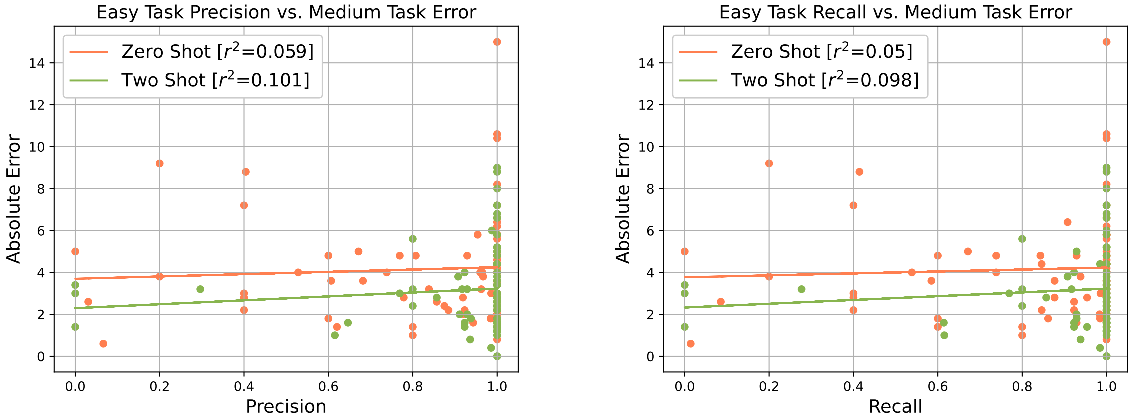


Figure 6.5: Correlations between the jury selection easy task performance and medium task performance. Split by precision and recall for both the zero shot and few shot cases.

Transcript Length vs. Medium Task

The last regression of the absolute error in the medium task against the number of transcript pages motivated the fine-tuning experiments in 6.2.3. We suspect that the variables are only slightly positively correlated ($r^2 = 0.039$ for zero-shot, $r^2 = 0.188$ for two-shot) because of the heavily left-skewed distribution of pages.

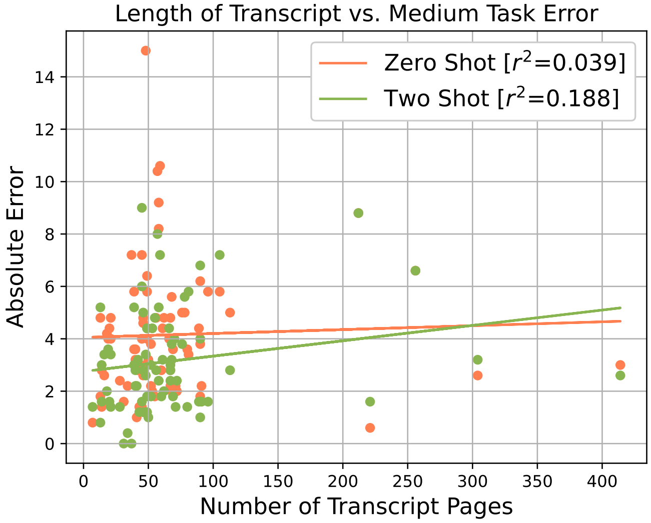


Figure 6.6: Correlations between the number of transcript pages and the medium task performance.

6.5 Optical Character Recognition Analysis

As mentioned in 2.4.2, a large bottleneck for mass aggregation and synthesis of legal data is the unstructuredness of data. We found that a lot of important information that would be beneficial to automatically extract took the form of handwritten notes in and along the margins of the printed documents. With current OCR technologies, it is nearly impossible to reliably convert these notes into readable text as there is a lot of variance in both the handwriting and the short-hand language that the notes are oftentimes written in. We conducted a brief analysis of Microsoft's Azure Document Intelligence OCR technology on a few varying strike sheets.

We randomly sampled 10 strike sheets and manually reviewed Azure's performance. Manual review was necessary as even the state-of-the-art model's outputs are unclean and difficult to parse algorithmically. Note that recall is defined by the proportion of demarcations detected by the model and precision is defined by the number of correctly identified demarcations. We observe that while the detection of demarcations is relatively high, the accuracy of the detections is incorrect (Table 6.7). Qualitatively, we also observe that of the race demarcations ('W' and 'B'), it is more common for the 'B' demarcations to be incorrectly identified. While this is more related to the physical characteristics of the demarcation and not any source of real bias between the two letters, it is important to recognize the potential downstream impacts of incorrectly detecting black compared to white jurors.

Context Type	# Cases	Precision	Recall	F1
Strike Sheet	10	0.551 ± 0.224	0.879 ± 0.167	0.678

Table 6.7: Average metrics for the optical character recognition of race and gender demarcations on strike sheets. Reported with standard deviation.

6.6 Step-by-step Prompting

We explored the feasibility of using step-by-step prompting, where the query enumerates the steps the LLM would need to undergo to arrive at the correct answer. An example prompt for the medium jury selection task is below.

```
Extract the final list of jurors chosen to serve on this case.

Using that list, count the number of female and male jurors.  Female jurors are denoted using Ms.  in the transcript, and males using Mr.

Only count jurors that were chosen to serve on the jury or serve as alternates.

Only output the number as a comma-separated list.  Do not output any other text or annotations.
```

We randomly selected 20 documents from the full list of 75 cases to conduct this experiment. We compare the performance of the three types of prompting in 6.8. We observe that the two shot query exceeds the other query types in minimizing error and standard error and maximizing accuracy.

Prompt Type	Female Error	Male Error	Total Error	Accuracy
Zero-Shot	1.877 ± 0.071	2.170 ± 0.0718	4.048 ± 0.133	0.1
Two-Shot	**1.613 ± 0.044**	**1.632 ± 0.039**	**3.245 ± 0.079**	**0.15**
Step-by-step	2.00 ± 0.073	2.173 ± 0.079	4.173 ± 0.147	0.06

Table 6.8: Average metrics for the medium jury task using different prompt types aggregated over 20 randomly sampled cases over all iterations (5 per case). Reported with standard error across the number of cases.

Chapter 7

Discussion

Our results indicate the feasibility of using LLMs to automate accountability pipelines in legal documents. However, we uncover several challenges that require further investigation before the wide-scale application of these methods in legal auditing.

How do limitations in logical and legal reasoning complicate information retrieval tasks for legal accountability?

We add to prior research that distinguishes between the complexity of information retrieval tasks, but we differ in our motivation for these capabilities by highlighting their relevance to legal accountability. The logical reasoning in our tasks involves the synthesis or categorization of information from documents – capabilities that have been shown to have success in many domains. However, the limitations that we observe indicate the need to orient benchmarks around more real-world tasks. For example, our jury gender composition query highlights the difficulty of a simple counting task – a capability for which models demonstrate high scores on toy datasets [40] that do not directly translate to real-world legal documents. Similarly, we find that limitations in legal reasoning may require the need for more training in the legal context. The improved performance for our few-shot and fine-tuning experiments indicates that LLMs may be able to learn the legal knowledge required for accountability-related tasks.

Why is the performance better for eviction than jury selection?

Several reasons may contribute to the better out-of-the-box LLM performance for eviction tasks. While the eviction Summons and Complaint forms are only one page, compared to the much longer court transcripts, we still observed a performance gap when compared to transcript excerpts that are only a few paragraphs. This suggests that the *structure* of the eviction forms may be the primary reason for better performance. These forms may be more similar to the internet-scraped data that LLMs are trained on than the court transcripts of human speech. In our results, we highlight examples of how transcribed human speech often contains *disfluencies* such as pauses, filler words, and repeated phrases. We find that these disfluencies, while easy for a human to parse, can severely complicate understanding by LLMs. These results demonstrate how training on internet-scraped data can limit out-of-the-box LLM functionality on meaningful real-world tasks.

What technical investments are required to make automation of judicial accountability mechanisms more feasible?

For model developers, more reflection on the types of data used for training and benchmarking needs to occur. Investment into more training on the types of unstructured data that we highlight could drastically improve the out-of-the-box performance of these tools for accountability use cases. However, this still requires a collection of labeled examples across various accountability-related tasks. The investment by law firms into collecting data for the types of tasks paralegals have traditionally performed (e.g., case summarization) may be helpful, but even these tasks may not directly translate into accountability use cases. An additional barrier to collecting this labeled data comes from the bottleneck of data access and centralization.

In order for a mechanism using LLMs to be deployed in the US alone, it would require scalability and generalizability across court files nationwide. This means that the model would need to be robust against the variance in possible formats and patterns of representing the information, implying that more data from different jurisdictions and across wider spans of time would be necessary to collect. The demographic data collection conducted by APM was also incredibly limited, only investigating Male-Female and Black-White analyses. However, to be more meaningful, further work should also encode all protected groups outside these binaries.

A large bottleneck for mass aggregation and synthesis of legal data is the unstructuredness of data. Explanatory pieces of important information that take the form of handwritten notes in and along the margins of the printed documents would be beneficial to automatically extract to further the accountability claims. With current OCR technologies, it is nearly impossible to reliably convert these notes into readable text as there is a lot of variance in both the handwriting and the short-hand language that the notes are oftentimes written in. Even state-of-the-art models fail to recognize legal notes accurately, motivating further training of both vision and language models on representative real-world data.

What legal investments are required to make automation of judicial accountability mechanisms more feasible?

There are several regulatory solutions that can help lay the groundwork for the aforementioned technical investments to be made. First, there exists a massive data accessibility problem. As described in Section 1, legal data cannot be accessed without submitting a request to the courthouse and getting approval from the clerk to hand pull casefiles. Summers and Steil share that attorneys registered with the Massachusetts Board of Bar Overseers (BBO) can access all cases virtually through an electronic portal. However, while the limitation of electronic access to registered state attorneys allows for more fine-grained jurisdictional control and certainty of ethical use of the files, it does raise several concerns from the legal accountability viewpoint.

First, as long as limitations like these are upheld, substantial increases in legal fairness will continue to be unobtainable. This means that pro se litigants and those who cannot afford legal representation have a severe disadvantage to those with a state lawyer to understand previous cases and prepare to defend themselves in court. Additionally, legal technological reform is substantially complicated if digitized case files are not accessible to the public through any means. A system similar to the Institutional Review Board (IRB) for human research subjects could be beneficial to implement in order to maintain a review process before accessing, but open up access to the public at large.

Another regulatory resolution that could be applied is standardization of case documents. While there is significantly more push back from legal practitioners to standardize a format across jurisdictions, states, and even firms, standardization would alleviate many legal shortcomings.

How can LLMs be pipelined into an automated legal accountability mechanism?

From our findings, we conclude that optimal performance comes from refined excerpts and cleaner, more structured data. Legal documents could benefit from existing mechanisms such as Retrieval Augmented Generation (RAG), which automatically identifies the most relevant parts of the provided context to then be used to answer the query. This would allow entire casebooks to be processed automatically instead of needing to isolate the most relevant pages and forms. It is also possible for LLMs in their current state to be used in conjunction with legal practitioners to assist with easier tasks that have less downstream impact. However, our case studies suggest that the bulk of processing labor and time comes from the medium and hard tasks where contextual reasoning is necessary. Additionally, the overhead of a human needing to feed logic into the model and monitor for errors questions the scalability of this technology.

Our findings also pose outstanding questions about the end users in the legal space who are willing to adopt this technology into their pipelines. More high-stakes legal work would regard each individual error in information extraction as more costly than their lower impact counterparts. We must continue to collaborate with social scientists and legal practitioners in order to understand the hesitations in employing LLMs to assist their work.

Who can potentially be harmfully affected by this technology?

We emphasize the potential disparate impacts of deploying these technologies out-of-the-box without further investigation into the failure points and direct implications. Higher profile cases and crimes that interface with wealthier people, such as white-collar crimes, generally have better-maintained documentation. This is due to the level of resources that private lawyers and investigative agencies such as the FBI and SEC possess. However, minorities are disproportionately represented by public def enders, who tend to be underfunded and have notably fewer resources [28], leading to poorer keeping of casefiles. However, from our experiments, we determined that automation pipelines such as LLMs are likely to perform better on well-maintained and structured data and less likely to retain performance in messier data structures. This means that, without deeper insights, wealthier people are more likely to be integrated into automated accountability pipelines, while minorities and those who would benefit the most from more accountability still face higher barriers.

We call on members of both the technical and legal communities to invest in solutions that can bring more accountability to the judicial system. By focusing efforts on legal data centralization, training models on unstructured data, and remaining vigilant about the implications of different classifications of errors, information extraction using LLMs can help automate and make judicial accountability more accessible.

Bibliography

[1] Laura I Appleman. Nickel and dimed into incarceration: Cash-register justice in the criminal system. *BCL Rev.*, 57:1483, 2016.

[2] Tim Arango. Exckusion of jewish jurors prompts review of california death row cases, 2024. URL https://www.nytimes.com/2024/05/13/us/california-oakland-death-penalty-jewish-jurors.html?smid=nytcore-ios-share&referringSource=articleShare&sgrp=c-cb.

[3] Elliott Ash, Aniket Kesari, Suresh Naidu, Lena Song, and Dominik Stammbach. Translating legalese: Enhancing public understanding of court opinions with legal summarizers. In *Proceedings of the Symposium on Computer Science and Law*, CSLAW '24, page 136–157, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703331. doi:10.1145/3614407.3643700.

[4] Larond Baker, Salvador Mungia, Jeffery Robinson, Lila Silverstein, and Nancy Talner. Fixing batson. *Litigation*, 48(4), 2022.

[5] Chelsea Barabas. Beyond bias: Re-imagining the terms of" ethical ai" in criminal law. *Geo. JL & Mod. Critical Race Persp.*, 12:83, 2020.

[6] Mark W Bennett. Unraveling the gordian knot of implicit bias in jury selection: The problems of judge-dominated voir dire, the failed promise of batson, and proposed solutions. *Harv. L. & Pol'y Rev.*, 4:149, 2010.

[7] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

[8] Sarah Brayne. Big data surveillance: The case of policing. *American sociological review*, 82(5):977–1008, 2017.

[9] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[10] Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. Booookscore: A systematic exploration of book-length summarization in the era of llms. In *The Twelfth International Conference on Learning Representations*, 2023.

[11] Colleen V Chien and Miriam Kim. How generative ai can help address the access to justice gap through the courts. *Loyola of Los Angeles Law Review, Forthcoming*, 2024.

[12] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[13] Alma Cohen and Crystal S Yang. Judicial politics and sentencing decisions. *American Economic Journal: Economic Policy*, 11(1):160–191, 2019.

[14] Will Craft. Peremptory strikes in mississippi's fifth circuit court district. https://features.apmreports.org/files/peremptory_strike _methodology.pdf, 2018.

[15] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. Large legal fictions: Profiling legal hallucinations in large language models. *arXiv preprint arXiv:2401.01301*, 2024.

[16] Matthew Desmond. Eviction and the reproduction of urban poverty. *American Journal of Sociology*, 118(1):88–133, 2012.

[17] Matthew Desmond and Carl Gershenson. Who gets evicted? assessing individual, neighborhood, and network factors. *Social Science Research*, 2016.

[18] Matthew Desmond and Rachel Tolbert Kimbro. Eviction's fallout: housing, hardship, and health. *Social Forces*, 94(1):295–324, 2015.

[19] Sharad Goel, Ravi Shroff, Jennifer Skeem, and Christopher Slobogin. The accuracy, equity, and jurisprudence of criminal risk assessment. In *Research handbook on big data law*, pages 9–28. Edward Elgar Publishing, 2021.

[20] Kayla Goggin. Death row inmate claiming gender discrimination on jury rejected by supreme court, 2024. URL https://www.courthousenews.com/death-row-inmate-claiming-gender-discrimination-on-jury-rejected-by-supreme-court/.

[21] Ashley Gromis, Ian Fellows, James Hendrickson, Lavar Edmons, Lillian Leung, Adam Porton, and Matthew Desmond. Estimating eviction prevalence across the united states.

[22] Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[23] Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. Cuad: An expert-annotated nlp dataset for legal contract review. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[24] Rebecca C Hetey and Jennifer L Eberhardt. The numbers don't speak for themselves: Racial disparities and the persistence of inequality in the criminal justice system. *Current Directions in Psychological Science*, 27(3):183–187, 2018.

[25] Elizabeth Hinton, LaShae Henderson, and Cindy Reed. An unjust burden: The disparate treatment of black americans in the criminal justice system. *Vera Institute of Justice*, 1(1):1–20, 2018.

[26] Alex Kim, Maximilian Muhn, and Valeri V Nikolaev. Bloated disclosures: can chatgpt help investors process information? *Chicago Booth Research Paper*, (23-07):2023–59, 2024.

[27] Niklas Mannhardt, Elizabeth Bondi-Kelly, Barbara Lam, Chloe O'Connell, Mercy Asiedu, Hussein Mozannar, Monica Agrawal, Alejandro Buendia, Tatiana Urman, Irbaz B Riaz, et al. Impact of large language model assistance on patients reading clinical notes: A mixed-methods study. *arXiv preprint arXiv:2401.09637*, 2024.

[28] Rebecca Marcus. Racism in our courts: The underfunding of public defenders and its disproportionate impact upon racial minorities. *UC Law Constitutional Quarterly*, 1994.

[29] Jayr Pereira, Andre Assumpcao, Julio Trecenti, Luiz Airosa, Caio Lente, Jhonatan Cléto, Guilherme Dobins, Rodrigo Nogueira, Luis Mitchell, and Roberto Lotufo. Inacia: Integrating large language models in brazilian audit courts: Opportunities and challenges. *ACM Digital Government: Research and Practice*, mar 2024. doi:10.1145/3652951. URL https://doi.org/10.1145/3652951.

[30] . R. Sudarshan Rajeev Dhavan and Salman Khurshid. *Judges and Accountability*. Sweet Maxwell, Ltd., 1985.

[31] Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 429–435, 2019.

[32] Pragya Srivastava, Manuj Malik, and Tanuja Ganu. Assessing llms' mathematical rea-

soning in financial document question answering. *arXiv preprint arXiv:2402.11194*, 2024.

[33] Nicole Summers and Justin Steil. Pathways to eviction, 2024.

[34] Tim Thomas, Alex Ramiller, Cheng Ren, and Ott Toomet. Toward a national eviction data collection strategy using natural language processing. *Cityscape*, 26(1):241–260, 2024.

[35] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, et al. Clinical text summarization: Adapting large language models can outperform human experts. *Research Square*, 2023.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[37] Jackie Wang. *Carceral capitalism*, volume 21. MIT Press, 2018.

[38] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022.

[39] Ronald F Wright, Kami Chavis, and Gregory S Parks. The jury sunshine project: Jury selection data as a political issue. *U. Ill. L. Rev.*, page 1407, 2018.

[40] Yiran Wu, Feiran Jia, Shaokun Zhang, Hangyu Li, Erkang Zhu, Yue Wang, Yin Tat Lee, Richard Peng, Qingyun Wu, and Chi Wang. Mathchat: Converse to tackle challenging math problems with llm agents. *ICLR Workshop on Large Language Model (LLM) Agents*, 2024.

[41] Emma Zack. Why holding prosecutors accountable is so difficult, April 23 2020. URL https://innocenceproject.org/why-holding-prosecutors-accountable-is-so-difficult/. Innocence Project.

[42] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57, 2024.