ISSN 1936-5349 (print)
ISSN 1936-5357 (online)

HARVARD

JOHN M. OLIN CENTER FOR LAW, ECONOMICS, AND BUSINESS

Building a Better Lawyer

Experimental Evidence that AI Can Increase Legal Work Efficiency

Aileen Nielsen, Stavroula Skylaki, Milda Norkute, and Alexander Stremitzer

Discussion Paper No. 1111
10/2024
Harvard Law School
Cambridge, MA 02138

Building a Better Lawyer

Experimental Evidence that AI Can Increase Legal Work Efficiency

Aileen Nielsen, Stavroula Skylaki, Milda Norkute, and Alexander Stremitzer*

Abstract

Rapidly improving artificial intelligence (AI) technologies have created opportunities for human-machine cooperation in legal practice. We provide evidence from an experiment with law students (N=206) on the causal impact of machine assistance on the efficiency of legal task completion in a private law setting with natural language inputs and multidimensional AI outputs. We tested two forms of machine assistance: AI-generated summaries of legal complaints and AI-generated text highlighting within those complaints. Compared to no AI assistance, AI-generated highlighting reduced task completion time by 30% without any reduction in measured quality indicators. AI-generated summaries produced no change in performance metrics compared to no AI assistance. AI summaries and AI highlighting together improved efficiency but not as much as AI highlighting alone. Our results show that AI support can dramatically increase the efficiency of legal task completion, but finding the optimal form of AI assistance is a fine-tuning exercise. Currently, AI-generated highlighting is not readily available from state-of-the-art, consumer-facing large language models, but our work suggests that this capability should be prioritized in the development of legal AI products.

Key words
Artificial Intelligence (AI), Experimental Study, Law Students,
Human-AI Cooperation, Legal AI

*Corresponding author Aileen Nielsen, Visiting Assistant Professor, Harvard Law School, 1585 Massachusetts Avenue, Griswold 309, Cambridge, Massachusetts 02138, ainielsen@law.harvad.edu. Stavroula Skylaki, Director of Applied AI Research, Thomson Reuters Labs - Zug. Milda Norkute, Product and UX Lead, Thomson Reuters Labs - Zug, Alexander Stremitzer, Professor of Law, ETH Zurich, Center for Law and Economics. Data and software necessary to replicate the results of this article are available upon request from the corresponding author.

Introduction

The use of AI in courtrooms and law firms is increasing, yet very little is known about how AI products influence legal work products or processes (Sindar, 2022). In this work, we demonstrate the promise of a robust experimental approach to address empirical questions on the effects of AI in legal work. We study how the process and outputs of legal work may be impacted by new AI tools, a topic of particular importance since most new legal AI adoption is likely to reflect a "human in the loop" model, wherein AI does not directly determine outcomes but instead provides guidance to human decision makers (Mosqueira-Rey et al., 2022). We conduct an experiment studying the effect of AI assistance in the hands of law students (N=206). We measure the effects of two distinct forms of AI assistance - summarization and highlighting of complaint texts - to see how each influences law students' ability to identify text within a legal complaint and their judgments of the complaint quality.

We test all combinations of two forms of AI assistance: AI-generated text highlighting within a legal complaint document and an AI-generated one-sentence summary of the document. This focus on *specific forms of output* from an AI product distinguishes our work from other recent studies. Other recent studies have tested consumer-facing large language models (LLMs) as products rather than studying the influence of specific outputs or forms of outputs from such models. Our partnership with a firm currently using the studied AI model for commercial purposes allows us to test specific outputs thanks to this insider access to the model.

Measuring a range of quality indicators, we found no change in work quality resulting from the AI interventions. However, we also discovered that AI highlighting makes task completion about 30% faster, supporting the possibility that legal AI can make lawyers drastically more efficient without diminishing work quality. Given that consumer-facing LLMs do not readily provide access to such highlighting capacities (at the time of writing), this work suggests low-hanging fruit in producing additional AI outputs that would be useful to legal workers (Glassman et al. (2024)). Further, this work demonstrates the utility of studying specific forms of AI outputs rather than studying AI as a monolithic whole.

Several key generalizable insights can be gleaned from this work. We offer a template for defining and measuring the quality of legal work. We develop and offer to the community a reusable interface into which AI outputs can be inserted for behavioral testing. We look critically at *different AI outputs* rather than treating AI as an unbreakable whole. We proceed as follows. First, we survey previous work that measures the impact of AI assistance on legal tasks. We identify three key areas lacking knowledge in the existing literature: the effects of AI assistance in private legal practice, the comparative effects of different forms of AI assistance, and experimental methodologies that allow causal conclusions regarding the effects of legal AI tools. Next, we present our experimental design, including a description of the many variables we measured to assess the quality of the legal decision-making process and outcomes in the experimental task. We conclude with a discussion of the policy relevance of these findings and our experimental template more generally.

Background

We survey the existing state of knowledge regarding the use of AI[1] in the legal system,[2] identifying important, basic questions still left open in the literature. As the discussion below will show, previous studies have primarily looked at outcomes rather than processes, governmental actors rather than private actors, and unidimensional outputs rather than richer, multi-dimensional guidance. Further, as the discussion will show, the studies in this narrowly circumscribed domain have not uncovered compelling evidence of the benefits of deploying AI for human assistance in legal tasks,[3] which suggests that the near-future benefits of AI may be better fulfilled in other domains. From this review of existing work, we conclude that our study can expand existing knowledge by studying a private law setting and an AI tool that uses multi-dimensional outputs while measuring legal processes as well as legal outcomes.

Further, we observe in our review of the relevant scholarship that observational and experimental studies alike have often yielded null results, and that there are no clearly documented examples of AI-assistance improving outcomes in legal tasks. Some cases of introducing AI-assistance have even yielded normatively problematic outcomes, such as increased racial disparities. This motivates us to look beyond the criminal justice system and beyond unidimensional outputs in designing our study, to study domains in which upsides of AI assistance may be more accessible than they have proven in the especially challenging applications of criminal justice.

We recognize from the outset that studies of how AI has previously been introduced into the legal system - primarily in criminal adjudication - may not strongly indicate the likely results of an AI tool put into the hands of a private actor for private legal practice. Where government actors have incorporated AI, their goal has typically been to improve uniformity, accuracy, and fairness of criminal adjudication - all values typically irrelevant to strategic private actors, who will likely be more concerned with efficiency and performance than with fairness or consistency as end goals. Nonetheless, we survey the literature broadly to offer the reader not only a picture of the most recent and relevant studies in comparison to ours but also a holistic picture of the state of knowledge of how human legal decision-makers interact with AI.

---

[1] We do not rigorously define "AI" for this study or for the discussion of prior work. We use "AI" broadly to refer to data-driven machine assistance, no matter how simple or complex that assistance may be in its functionality, inputs, or outputs. Our use of "AI" therefore includes both large language models (LLMs) or linear regressions, where either is deployed to offer guidance to humans in task completion.

[2] We note a larger, adjacent literature looking to largely experimental studies of how AI tools might influence lay decision making about legal questions. We also note a larger, adjacent literature looking at the performance of algorithms as assessed by their outputs from input data alone and when not subject to human discretion. We find support for Megan Stevenson's observation that "There is ample evidence that [AI] *should* have beneficial effects" in law but that the evidence as to whether it does is either missing or pointing towards undesirable outcomes.

[3] The intended benefits of using AI in legal decisions will vary in different settings and may not necessarily be limited to correctness or efficiency. Nonetheless, even with respect to other objectives pursued when introducing AI to legal decision making, such as working towards racial equity or reduced crime, empirical studies suggest mixed or null results of introducing AI into the hands of human decision makers, as discussed *infra*.

Observational studies of risk assessment in the criminal justice system

One genre of studies focuses on the impact of AI on legal outcomes, leveraging observational data created by a jurisdiction and introducing a risk assessment tool into its criminal justice process. A 2019 meta-analysis by Viljoen et al. (2019) identified 22 such studies, around half of which were the results of white papers or policy papers rather than peer-reviewed journals. These 22 studies covered more than 1.4 million individuals subject to AI-assisted confinement decisions across 30 separate study locations. A meta-analysis of these 22 studies suggested a small but statistically significant reduction of restrictiveness in pre- and post-conviction placements resulting from the introduction of risk assessment tools as a human decision aid.

Despite these initial results, the authors ultimately concluded that the evidence that the use of risk assessment tools reduces restrictive placements was ultimately low because they identified methodological issues with the studies that suggested the possibility of bias. Specifically, the authors applied the Risk of Bias in Non-Randomised Studies (ROBINS-I) (Sterne et al., 2016) tool to each of the 22 studies[4] to assess the risk of statistical bias in the meta-analysis results introduced by each of the 22 studies. Viljoen et al. found that 11 of the 22 studies had a risk of bias (with an inter-rater agreement of .85). Once the problematic studies were removed from the meta-analysis, there was no statistically significant effect of AI assistance on reducing restrictive placements. Further, the researchers noted that even the directional effects in the remaining studies were heterogeneous, providing further reason to discount the initial statistical significance resulting from analysis of the entire 22 study set. Further, they noted a variety of confounds often present in the studies that were described as biased, including (1) the deployment of risk assessment tools *alongside other* interventions or political changes and (2) the *conceptually distinct* nature of the various risk assessment tools used, including fundamental differences such as relying on static versus dynamic characteristics.[5] Thus, the meta-analysis ultimately indicated that risk assessments achieved, at best, small or null effects in accomplishing their stated purpose of reducing restrictive placements. This finding is consistent with a mechanism of discarding algorithmic judgments, such as that identified in earlier work by Miller and Mahoney (2013), which used latent class analysis to estimate that roughly half of practitioners required to use risk assessment tools merely demonstrated "formal" compliance, often demonstrating decisions inconsistent with the risk assessment tool and with more recent findings such as Stevenson (2018) noting that effects of risk assessment decrease over time over their introduction, likely as judges resort back to their pre-tool work process.

Studies published since 2019 demonstrate a similarly insubstantial effect of algorithms deployed in real-world criminal justice systems. Stevenson and Doleac (2021) examined the use of a risk-scoring instrument provided to state court judges throughout Virginia in 2002. The risk instrument was implemented with the policy goal of diverting 25% of nonviolent offenders to non-carceral outcomes. The risk assessment was provided in all cases, but the judge retained

[4] ROBINS-I identifies seven domains that raise risk of introducing bias into the results of a reported study: confounding factors, selection of participants, classification of interventions, deviations from intended interventions, missing data, measurement of outcomes, and selective reporting), with these individual queries then used to calculate an overall rating of bias.

[5] Static characteristics are those that are not possible to change, such as past criminal history. Dynamic characteristics are those that can actively be changed, such as drug addiction status.

discretion as to sentencing. In studying both release outcomes and sentence lengths, Stevenson and Doleac found that the risk assessment ratings changed outcomes for defendants labeled low risk compared to those labeled high-risk, but that the ratings did so in heterogeneous ways suggesting that judges had alternative objectives or preferences. They deviated from risk score recommendations for some demographics: towards greater leniency for young defendants and less leniency for black defendants. As a result of such heterogeneous algorithmic treatment effects, the overall incarceration rate did not decline after the risk-scoring algorithm was introduced. This study provides evidence that AI integrated into human decisions may produce results quite different from the outputs of the AI system alone. Further, some of those results, suggesting differential treatment of algorithmic outputs by race, challenge the notion that fair algorithms will necessarily lead to fair decision-makers.

On the other hand, in a study of Travis County, Texas, Sloan et al. (2023) reached results that contradicted Stevenson and Doleac's and provided some evidence that AI tools in the hands of humans could achieve their stated policy objectives under appropriate conditions. In January 2013, Travis County changed its policies from not providing risk assessments at all to providing them for 75% of defendants. Using this sharp change as an opportunity for a regression discontinuity analysis, Sloan et al. (2023) found that, consistent with the county's stated policy, the introduction of the risk assessment tool led to less punitive outcomes, with increased rates of non-monetary bail and decreased rates of pre-trial detention. Also, in contrast to Stevenson and Doleac, Sloan et al. found no exacerbation[6] of racial disparities induced by the risk assessment instrument. In short, more recent literature replicates the ambiguities found in Viljoen et al.'s 2019 meta-analysis. It's unclear whether risk assessment instruments are normatively problematic vis-à-vis racial equity.

Despite the theoretical benefits of bringing AI into the legal system (Kleinberg et al. 2017), observational studies fail to support those optimistic predictions. In existing observational studies, the effects of introducing AI into the legal process amount to minimal upside, and may run counter to fundamental normative values, such as equality before the law.

Experimental studies of the criminal justice system

Experimental studies of the effects of legal AI tools on actors in the legal system are a smaller (but emerging) body of literature compared to the body of observational studies. Early experimental studies of other risk assessment instruments are also few and far between. Before the contemporary renewed interest in the impact of risk assessment tools in the hands of humans, Imai et al. (2021) note only two prior experimental studies, both many decades old, Ares et al. (1963) and Goldkamp and Gottfredson (1985).

The primary risk assessment tool in the U.S. criminal justice system is the Public Safety Assessment PSA). The first-ever experimental study of the pretrial PSA took place in Dane County, Wisconsin, for a 30-month period from 2017 to 2019. In the Dane County study, Imai et

[6] Of course, "no exacerbation" of disparities is not the same as an improved outcome.

al. (2021) found – on a preliminary basis[7] – that the PSA had little impact on judicial decision-making. The only impact they found was a tendency to exacerbate existing disparities in decision-making as between male and female defendants; on the other hand, they found no PSA-associated change in the influence of race on judicial decisions and likewise no change otherwise in the overall pattern of sentencing.

More recently, in a working paper, Danser et al. (2023) reported an experimental study of the PSA in Polk County, Iowa, conducted in 2018. The authors reported a significant change in judicial decision-making, specifically a shift in the distribution towards more lenient (non-monetary) and harsher (high monetary bonds) pre-trial decisions. However, puzzlingly, such changes did not correlate to changes in recidivism. Instead, the authors found suggestions of increased failure to appear after the deployment of the PSA, although they explained that there were several competing hypotheses to explain such a result. In short, the two existing randomized controlled trials of the PSA point to mixed results, finding the instrument may or may not change judicial decision-making and appeared to make no change to metrics of criminality (recidivism) or performance of the criminal justice system (such as failure to appear or incarceration rates).

In short, both observational and experimental studies of risk instruments suggest that the gains of algorithms have probably been minimal in practice in the criminal justice system. Further there is a risk, substantiated in some observational and experimental studies, that these instruments may even exacerbate some of the disparities experts otherwise hoped AI could cure (Yu, 2020). This mixed bag of null and normatively problematic outcomes necessarily raises the question of whether there is a role for AI in the legal system any time soon.[8]

One possible reason, among many, that risk assessment instruments seem to underperform their theoretical promise in criminal justice settings is that they make so little use of the available possibilities, mainly relying on a single, unidimensional output. As currently implemented, risk assessment tools are simple instruments with limited expressive value. In an entirely different experimental setup, Chohlas-Wood et al. (2021) imagined the possibility of a tool that could use demographic blindness to promote fairness. In a quasi-experimental study,[9] the authors developed and tested a tool that obscured identifying information about people and places. Prosecutors in a large U.S. city used this tool to make preliminary charging decisions (later finalized without the blinding treatment). The authors studied the outcome of charging decisions with and without AI blinding. Although in an entirely new context and task, similar to previously discussed studies, the authors reported no effect of the tool. They noted, however, that this apparent null effect could be a ceiling effect due to the lack of evidence of racial disparities in charging decisions even before the AI blinding treatment in that prosecutor's office, suggesting that potential upsides might be found in a different experimental setting.[10]

---

[7] The authors noted that as of the time of writing a sufficient period of time had not passed for full data collection and stressed the preliminary nature of their analyses.

[8] We leave to the side use cases in which AI would directly stand in for humans as being unlikely in the near future.

[9] The authors described the study in these terms, in part because administrative reasons limited their ability to randomize the time and place of treatment with their tool.

[10] The authors noted that future work should look to a use case where a racial disparity was documented prior to the use of the instrument to fairly test its capacity to reduce such disparities.

Experiments outside the criminal justice system

We first note a prior result that motivated our present study. Norkute et al. (2021) studied an industry tool designed to produce one-sentence summaries of legal complaints alongside two proposed means of highlighting text as a form of explanation of an AI output. Specifically, Norkute et al. studied an AI tool[11] that produced one-sentence summaries of legal complaint documents and studied two methods of explaining an AI's output, both forms of text highlighting. The authors studied (1) firstly, attention-based highlighting, in which the color of the highlighting was computed from values that the AI model itself generated and that constituted a numerical proxy for the relative importance or influence of individual words on the model's final outputs. Darker highlighting indicated a larger weight attributed to the word by the model in constructing the model's outputs, and lighter-colored highlighting indicated less weight attributed to a word. The authors also studied (2) secondly, abstractive highlighting, a model agnostic methodology in which the explanation highlighting features are generated after the summary is generated by looking at the initial input text and the output, highlighting any portions of the text that also appear in the one-sentence summary output.

In the case of attention-based highlighting - but not abstractive highlighting – Norkute et al. found a statistically significant decrease in the time that two specialized legally-trained employees needed to review AI-generated summaries for correctness and quality of the writing. The attention-based highlighting reduced the task completion time from 148 seconds to 83 seconds per complaint document, a 44% reduction, whereas abstractive highlighting failed to reduce task completion time. In semi-structured interviews with the participants, Norkute et al. found that attention-based highlighting enhanced the trust the participants expressed in the AI summaries. Norkute et al.'s results were one of the first instances of a substantial benefit of human use of AI in a legal task, while some prior work in non-legal domains had suggested that highlighting could decrease work quality (Ramírez et al., 2019). Norkute et al.'s results also show the importance of experimentally distinguishing helpful from unhelpful AI assistance, suggesting that improvements to legal work may result more from experimenting with the form and context of that assistance than from repeatedly deploying and testing similar but unimaginative legal AI tools.[12]

Norkute et al. (2021) identified promising evidence that AI can improve outcomes in legal work, but those existing results have significant limitations. The limited sample size (two attorneys) and the task at hand (producing one-sentence summaries of complaints for a proprietary legal subscription service) limit the external validity. Further, Norkute et al. could not control for task performance quality as the participants were required to turn in only work of acceptable quality and no numerical quality rating was used to permit finer tuned quality measurements.

We took this initial work as a step to build in the present study. We also ran a parallel study (Nielsen et al., 2023) with the same experimental interface and participants. In that study we

[11] The same AI tool we study in this work.

[12] Cf the risk assessment story, in which similar tools with similar goals have been repeatedly deployed despite little empirical basis for using them.

looked at the specific process of how research participants read a legal document. We found that the same attention-highlighting methodology studied by Norkute at al. did not lead to changes in attention allocation. That is, attention-based highlighting does not distort how humans process a legal document by moving their attention towards segments of a legal complaint that are highlighted (to the detriment of other portions). By contrast, the authors found the proportional amount of time participants spent within different sections of a legal document was the same with or without highlighting. We found evidence for one of the legal documents that the highlighting reduced the spatio-temporal distribution of user attention. These results are promising because they show (1) that highlighting need not distort human judgment about what was important within a document, and (2) some potential in reducing the variance across a group of readers in how they worked through a legal document.

We note that the past year and the rise of consumer-facing large language models (LLMs) has brought about the possibility for far more accessible experimentation with AI in legal applications. Both Blaire-Stanek et al. (2023) and Choi et al. (2023) reported passing (or even excellent) results of such consumer-facing LLMs in blind grading and in comparison to law students. These studies and similar are complementary to our work but study the distinctive question of how well AI could *replace* rather than *assist* a human performing legal work, finding that LLMs can do so in some minimally plausible way when it comes to taking law school exams. More recent studies have begun to address the question of human-AI cooperation directly. Choi and Schwarcz (2024) found that LLMs enhanced the performance of undergraduates and law students at the University of Minnesota on exams in some cases: more for simple, multiple-choice questions than for complex issue spotters and more for lower-performing students than for higher-performing students. They hypothesized that AI may produce an "equalizing effect" on law exams, or in the practice of law more generally, such that the AI helps low performers more, and may even hurt high performers. In a still more recent study, Choi et al. (2024) studied University of Minnesota law students on a series of four realistic drafting assignments with or without the use of an assistive LLM and following training on best practices when using LLMs for legal tasks. Choi et al. found only small or null changes in work quality on these tasks but found that LLMs did produce increases in speed, sometimes substantial ones. Finally, Chien and Kim (2024) report the results of a field study of practicing attorneys using LLMs to provide legal services, finding that legal professionals self-reported higher rates of productivity with access to an LLM and training in how to use the tool.

Though predating these studies of law students and legal professionals, and using an older AI tool, this current study nonetheless contributes to the existing literature in several ways. We use a larger (N = 206) study pool than that of Choi and Schwarcz (N = 48) and that of Choi et al (N = 59), and unlike Chien and Kim (N = 202) we use observed rather than self-reported measures of work quality and productivity. In contrast to all three studies, we use a fully randomized between-subjects design so that we can attribute full causality of differences to the ML intervention rather than having to make assumptions about ordering effects for a within-subjects study. In contrast to the other studies, we employ a tool that has already previously been used in a commercial legal application. This model is in actual commercial use. The legal editorial tool is used by a team of editors who monitor and collect new court cases and perform various editorial tasks. Further, and in contrast to the consumer-facing LLMs studied elsewhere, we rely on a model that followed many ethical best practices, such as training on data of known

provenance and without copyright-infringement (Bender and Friedman (2018)). Further distinguishing our study from other existing studies, we use an experimental interface that permits direct observations of and measurement of participant behavioral, permitting us some insights into how use of legal AI affects *process* as well as *outcomes*.

Concluding observations

From empirical and experimental studies of human-AI cooperation in legal decision making, we note three gaps in the existing literature. First, studies of the effects of AI assistance on legal tasks have almost exclusively focused on government actors in the criminal justice system. Second, variation in the kinds of AI assistance provided to legal decision makers has been little explored, despite strong evidence that the implementation and presentation of AI assistance can matter tremendously. Finally, assessments of AI influence have been limited to decision outcomes in narrow decision tasks in the criminal justice system and have not included procedural or behavioral observations. Two prior studies (Norkute et al. 2021; Nielsen et al. 2023) suggest a promising opportunity to fill in this gap in the literature and support an optimistic view about the role of AI in legal work.

Overview of Experimental Approach

For a variety of reasons like attorney-client privilege, much of the work product from private law is closed off to researchers. Private legal practice discloses few of the kinds of data artifacts needed for observational studies. For this reason, an experimental approach is particularly valuable for understanding the likely effects of AI in the private practice of law.
Our experiment was designed with two goals in mind to assure robust experimental conclusions. First, we wanted to study a legal task that had clear analogies in the real world. Second, we wanted to study a task in which there were some clear and objective measures of performance such that we could draw conclusions as to whether an AI tool was affecting work quality. With these goals in mind, we settled on a task that involves reading a complaint, identifying some objective textual details or assertions in that document, and drawing subjective conclusions as to the quality of that complaint document. Reading a complaint document and recalling or retrieving information from that complaint is analogous to many real-world legal work experiences, particularly those of junior law clerks or junior law firm associates, in which a junior worker might be the first to review a legal complaint and determine what level of response or review is appropriate for that document. One can likewise imagine that such a junior worker might be quizzed about the factual elements or legal arguments within the complaint under time-pressure, such as when asked by a senior partner to quickly review a document or when she faces stiff time constraints in the course of her work.

We also began our experimental design with a tool already in hand that we wanted to test. The tool has already been in use for several years by our industry partner to assist attorneys in generating summaries of legal complaint documents. Details about the creation of the neural network model can be found in Norkute et al. (2021). This tool is a real use case to supply a real legal service and therefore per se interesting. As described in the previous section, Norkute et al. found that this highlighting could increase the speed of work for two expert attorneys in the

specialized task of reviewing the AI generated summaries. We now seek to test the effects of this tool in a less specialized, larger population and on a more representative legal task: Testing the performance of a less specialized task by a less specialized pool of workers.

In this pre-registered study[13] conducted in an academic-industry collaboration, we measure the effects of legal AI assistance with a holistic rubric of indicators related to legal work quality. We acknowledge and embrace the importance of full academic independence in a study such as this. Prior to beginning work on the experimental design, we agreed on a process to ensure full academic freedom in the design, analysis, and publication of this research, and our industry partner has fully honored that agreement.[14]

Design and Procedure

In this experiment, we presented the AI outputs in a custom-built web interface.[15] We conducted the experiment online across a variety of law schools, with rankings as high as top-5 to the mid-80s in national law school ranking reports, with students recruited via a snowball sampling methodology and able to participate in the experiment on their own computers and at a time of their own choosing. They were solicited by course instructors, by emails from law school administrators, and by heads of law school student organizations. In the following explanation of our experimental design, we first describe the structure and appearance of the interface that was presented to research participants. We then describe the overall flow of the experiment and the between-subjects factorial design. We then describe our process for selecting specific complaint documents (and accompanying AI assistance) for the experimental task and describe and define our experimental outcomes of interest. Finally, we present a detailed description of the experimental procedure and the analytic methods applied to the experimental data.

Interface

We used a proprietary online interface that we designed and built for this experiment, with the interface modeled on a prior design by Norkute et al. (2021). The interface was programmed in React.js and accessible with any standard web browser by anyone who had an access code, which could be obtained by signing up for the experiment with a whitelisted law school email

[13] The pre-registration document is available on Open Science at https://osf.io/4nf9q. The study was reviewed and approached by the ETH Ethics Commission. Note that the URL is currently closed to the public but a copy of the preregistration document is also available here: https://docs.google.com/document/d/1uU_ILFoJ4icIPsQSHVrhanCY9YNDpDoW99nGyFPPi2o/edit?usp=sharing.

[14] Prior to starting this work, our institutions contracted for full academic freedom to design the experiment, analyze the data, and publish the research results. The cost of the experimental study (implementation of the proprietary interface and compensation for participants) was funded by ETH, while Thomson Reuters contributed access to the proprietary tool as well as the participation and expertise of two of the authors of this work. A copy of our collaboration agreement is available upon reasonable request.

[15] Access to using the interface to observe the experimental circumstances is available upon reasonable request. Access to the source code to implement the interface is likewise available upon reasonable request.

address. [16] We hosted the interface on ETH Zurich's servers, and protected the experiment from ineligible participants using single-use access tokens that were only dispensed to eligible participants.

The interface, shown in Figure 1, displayed the text of the legal complaint. In a column to the right of the complaint text, the questions about the text were presented. As will be described infra, these questions asked both about information that could be identified directly in the text and about participants' subjective assessment of the complaint quality. Where participants were assigned to an experimental treatment that included access to AI highlighting or summary, these were displayed as they are shown in Figure 1. The summary was displayed in the upper right-hand corner in a box with a bold and colored outline. The highlighting was displayed directly on the text, with a legend in the upper left-hand corner. When such features were not available to a participant given her experimental treatment, the interface was adjusted accordingly to preserve the appropriate aesthetics of the interface. When highlighting was not present, the color bar legend in the upper left-hand corner was absent. When a summary was not present, the question column was slightly elongated to fill in that area. Participants were able to scroll in both the legal complaint window (left column) and the question window (right column) independently, and they could use a "Next" button displayed in the lower right-hand corner to advance to the next text (but only after they had responded to all questions).

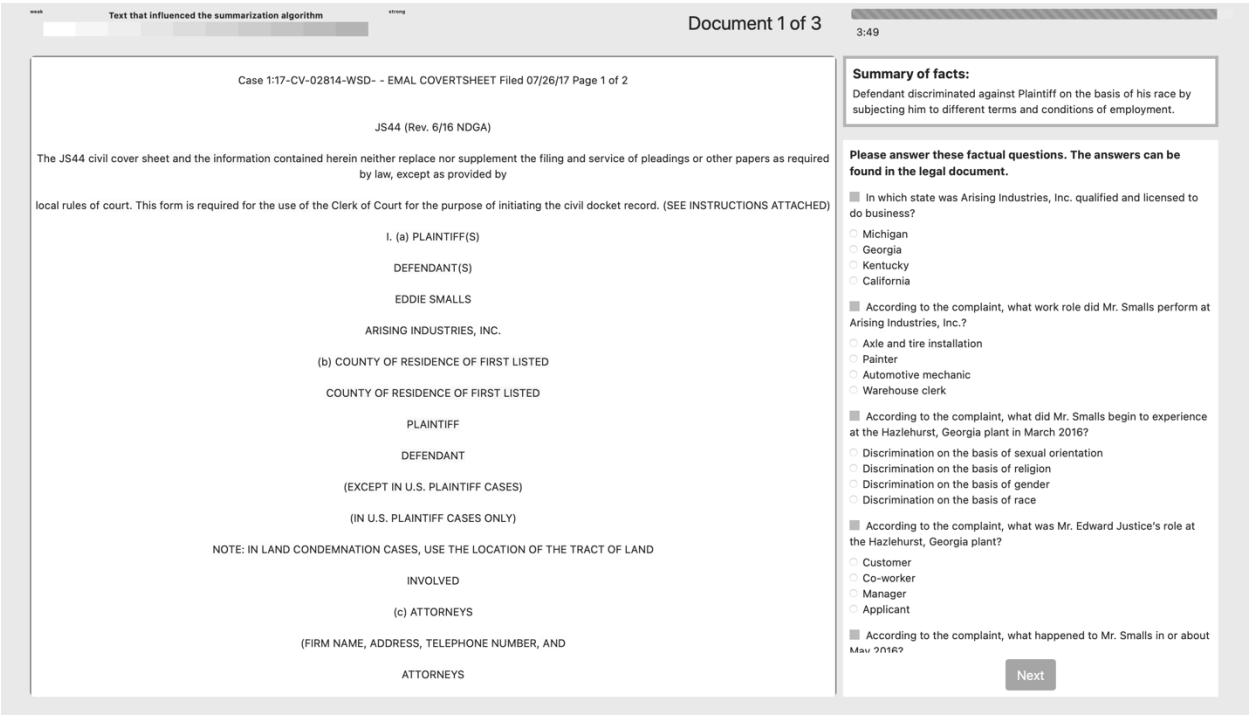


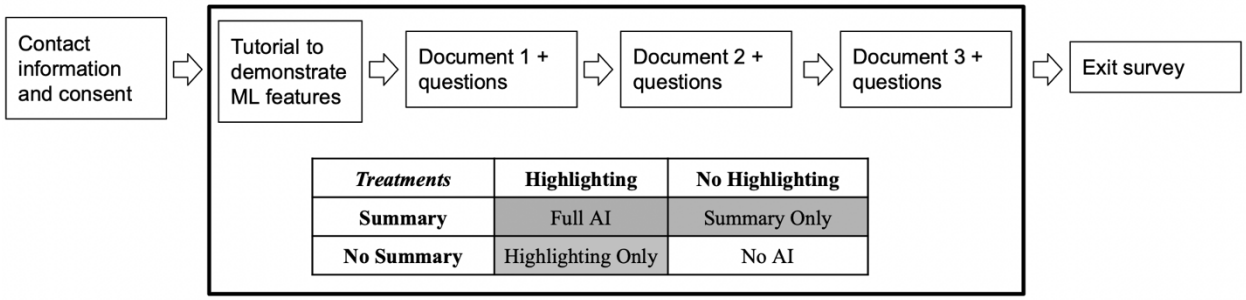
Figure 1: A screenshot of the experimental interface presents an example interface that includes both AI-generated highlighting (shading of text in left window) and AI-generated summary (sentence in box in upper right-hand corner of screen).

[16] Identifying information was kept in a data silo separate from the experimental data so that no identifying information was available to the researchers. The email address was collected to verify eligibility and to provide compensation. We whitelisted email addresses only at law schools where we were actively recruiting.

Participants responded to the multiple-choice questions by clicking on the circular button to the left of their chosen text. They answered a freeform text question by typing into a freeform text input window. They could advance to the next document only after answering all questions and then clicking on the "Next" button, which was inactive until all questions were completed.

Experiment design

The experimental interface followed an experimental flow that included informed consent, a tutorial[17] explaining to participants how the interface worked and what they would be asked to do, three legal complaints with associated questions, and a demographic exit survey. Within that flow, participants were assigned on a between-subjects randomized basis to any of four treatments resulting from a 2 x 2 factorial design,[18] in which the availability of the AI-summary and the AI-highlighting were permuted, creating four experimental treatments: No AI, Full AI, Summary Only, and Highlighting Only. The experimental flow is shown in Figure 2.



Treatments	Highlighting	No Highlighting
Summary	Full AI	Summary Only
No Summary	Highlighting Only	No AI

Figure 2: The experimental flow included informed consent, a tutorial customized to a participant's assigned experimental treatment, three consecutive self-contained tasks of reading a legal complaint and answering some questions, and an exit survey.

When reviewing each of the three legal complaint documents, participants had access to the text of the complaint and questions about the complaint; in other words, they did not need to review the complaint before seeing the questions, nor did they lose access to the complaint document when answering the questions. Further participants always had access to whichever forms of AI

[17] Screenshots of the tutorial are available here https://docs.google.com/document/d/1crBYsmSxijGx54nWh0rBZs07Ifc2Wv9exXKYB34OY40/edit?usp=sharing for the case of the Full AI treatment. Screenshots that show the pattern of highlighting for all the documents, as well as the location of the document text that correlated with correct answers to textual questions, are available here: https://drive.google.com/drive/folders/1woq81EnhNIpi9SQoYqzH5r6oS-vIg4wf?usp=sharing.

[18] The third task was subject to a 2 x 2 x 2 between-subjects design. In the text of the third complaint, we manipulated the name of the plaintiff in the third complaint as between "Thomas Stephens" (the original name in the complaint document) and "Juan Lopez" to look for potential effects of the name manipulation. However, we found no differences in the experimental populations treated with the two names, and subsequent testing on two pilot participants suggest that the manipulation likely failed in the experiment, as both pilot participants were unable to recall or even guess at the plaintiff's name.

assistance they had been experimentally assigned. All participants saw the same three legal complaint documents in the same order.

Our industry partner provided 100 randomly selected complaints along with the AI model outputs for those complaints and with quality ratings of the summaries as independently assessed by two expert attorneys. We limited our selection of complaints to those for which the model produced an output that was judged publication-ready by both of two attorney raters. Therefore, by design, we study in this experiment the effects of AI assistance, when that AI assistance is of a presumptively good standard. However, this selection is already representative of most outputs from the AI tool if not of all outputs. 58 out of 100 documents were rated by both attorneys as publication-ready, while 91 out of 100 were rated as such by at least one attorney reviewer. In contrast to our choice, we recognize that the growing body of literature on algorithmic aversion looks to the impact on human users of algorithmic mistakes. This is a key area for future investigation and is also important for understanding how AI assistance will affect real world legal practice (Dietvorst, 2015).

The two experimentally manipulated variables were the availability of two forms of AI assistance: a summary and highlighting. The summary was a one-sentence description of the facts in a legal complaint. The three summaries are provided in full below in the order corresponding to the task sequence.

- Defendant discriminated against Plaintiff on the basis of his race by subjecting him to different terms and conditions of employment.
- Defendants breached the operating agreement by entering into a mortgage and allowing new members into the company without consent.
- Defendant wrongfully attempted to collect a debt allegedly owed by plaintiff by providing him with false and misleading information.

The highlighting consisted of blue shading of varying darkness that corresponded to a numerical value generated for each word of the complaint as a function of running the text inputs through the model to output the one-sentence summary. The darker the highlighting, the larger the numerical attention value generated for that word, which in turn indicated that the word had been more influential in shaping the model's outputs in constructing the one-sentence summary of the complaint. As described supra, Norkute et al (2021) had previously found that this technique of highlighting improved subjective measures of trust in the AI sentence outputs and led to faster task completion in a small expert sample. We deployed the same methodology for calculating and displaying the attention-based score, which can be seen in their full visualization in the online appendix.[19] As described in the Appendix, we generated the questions to be answered for each task independently of this highlighting.

[19] The full complaint text and accompanying highlighting can be seen here: https://drive.google.com/drive/folders/1woq81EnhNIpi9SQoYqzH5r6oS-vIg4wf?usp=sharing.

Measures

Defining the quality of legal work is challenging, and there is, to our knowledge, no established set of metrics for either experimental or industry purposes (Linna, 2020). We defined a set of task outcomes with two goals in mind. First, we wanted objective indicators of performance such that improving performance was monotonically related to work quality. Second, we wanted measures that invited subjective opinions to gain insights into how the exercise of judgment might be affected by the presence of AI assistance. In other words, we sought indicators about discretionary judgment to address whether the presence of AI assistance might have an impact on discretionary judgments when it shouldn't (Kahneman et al., 2021). Finally, and to complement these two sets of measures, which would both provide indicators of outcomes, we also defined a separate set of indicators related to process as implemented through measurements of behavior.

Work quality measures

With respect to objective performance indicators, we developed textual questions with clear correct or incorrect answers to gauge participants' effectiveness at recalling or retrieving information.[20] These questions related to the factual and legal allegations in the complaint (e.g., *In which state does Portfolio Recovery Associates have an address for service? Which actions of Portfolio can be characterized as unfair practices under 15 U.S. Code § 1692f?).* The questions were of varying degrees of complexity and could be quite detailed. See the Appendix for the entire set of questions.

We targeted the exercise of professional judgment with two questions posed to participants as the last questions for each complaint (*Based on your best assessment of the case given the information provided in the document, should the plaintiff prevail on the merits? [Strongly disagree…Strongly agree]; (2) Please explain the reasons for your assessment of the merits of the plaintiff's argument. [freeform input])*.

Given that the presence of AI assistance is not legally relevant to the quality of a legal complaint, any shift in the subjective assessment of a document's argument on the basis of experimental treatment assignment would suggest a distortionary character of the AI relative to not having AI assistance. We anticipated detecting any such distortion through the multiple-choice question. In the case of the freeform text input, we envisioned using this to see what amount of effort and expression participants invested in providing an explanation as well as what the quality of that explanation might be.

[20] Full question text, response choices, and designated correct answer information are available in the Appendix.

Behavioral measures

By programming our proprietary web interface to record some behavioral data observable in a web browser, we collected indicia of how participants behaved during the experiment. We recorded the time that participants began and ended each individual document's task as well as the timestamps for answering each question and for each mouse click. We also recorded participants' scrolling behavior in 1 second increments; every second we recorded the topmost and bottommost text that was visible to the participant. In this way we could get at basic procedural questions, such as whether participants revisited questions or whether they revisited portions of the text after their initial read. In another paper, we also report on the allocation of attention and potential effects of the AI assistance on the allocation of attention (Nielsen et al. 2023).

Overview of relationship between experimental measures and legal work quality

The variables we track in this experiment are highly heterogeneous and may not obviously cohere as a test of the quality of legal work. We therefore now provide an overview of the ensemble of metrics to describe how these variables offer a holistic assessment of the quality of legal work. The outcomes we examine in this work, as an ensemble, might offer a potential template upon which to build in future work measuring the effects of AI assistance (or other interventions) on legal work processes and outcomes.

The raw experimental measures and their position in a larger conceptual umbrella delineating aspects of legal work quality are presented in Table 1. As above, there are three categories of variables, but these are not the same; these conceptual categories go to the aspect of quality measured. In contrast, the categories presented supra go to the structure of the task and interface; here we seek to show the conceptual groupings.

Category	Recorded data	Variable for analysis	Motivation/rationale
Objective performance measures	Textual question responses	Correctness/accuracy	Mastery of case facts and quick text inspection are key in case-law driven legal systems (Gardner, 2021; Cordon, 2011).
	Task completion time	Task completion time	Reducing legal work time can lower costs and increase access to justice (Moran, 2021; Jain, 2022).
Exercise of professional judgment	Numerical quality rating	Numerical quality rating	Indicates whether AI assistance affects judgment quality. A shift in ratings between treatments suggests AI-induced bias, requiring normative assessment (Kahneman et al., 2021).

	Freeform text explanation of quality rating	Character count	Proxy for participant effort. A shift suggests AI-induced behavioral bias.
		Automatically assessed grade level	Proxy for effort and professional standard adherence. Higher reading level suggests more precise language (Killian, 2019).
		Extractiveness of response	Measures text overlap with complaint document. More overlap indicates less effort in rephrasing, akin to student essays expressing original thoughts.
Exercise of professional responsibility	Scrolling position in document	Binary indicator of scrolling until end of document	Checks whether participants scrolled through the document to provide an informed opinion on the complaint's quality (American Bar Association, 1983).

Table 1: The variables of interest are computed from the recorded measures, sometimes with an additional layer of analysis to create the performance measure.

Experimental Procedure

Data was collected via snowball sampling from January to April[21] of 2022, with a total of 206 participants recruited from nine U.S. law schools, ranging from a law school ranked in the 70s nationally to a law school ranked in the top 5. Participants were contacted via emails[22] distributed by law student organizations or by course instructors. Participants who wished to participate clicked on a signup link and provided their email, which had to match a whitelisted template for participating law schools. Eligible participants then received an email with a personalized link to the experimental interface. We also invited participants to share information with their eligible classmates after they completed the experiment, in exchange for a $5 Amazon voucher.

There was no time limit[23] for completing the experiment; participants were free to go as quickly or slowly as they wished, and they could use a personalized experiment link at any time after sign-up (in practice all participants did so shortly after signing up). Compensation for participating was $20 for all participants regardless of how quickly or correctly they completed the experiment.[24]

[21] We pre-registered an intention to collect data until August 2022. We left the experiment open until that time; however, we did not have any new recruiting opportunities or obtain any new data points after April 2022.

[22] The email template that was used for this recruitment is provided in the Appendix.

[23] We pre-registered a minimum time requirement of 5 minutes for inclusion in the data analysis, but we did not enforce a minimum time for participants to receive compensation for completing the experiment.

[24] We feared that this might lead to low quality data as would be indicated by low levels of accuracy on the textual questions or extremely short task completion times, but this was not the case. During the course of the experiment, we monitored the quality of participation to prepare for the possibility that it would be necessary to adopt performance-based compensation, if participant effort was too low. However, this proved not to be the case, and so compensation was flat and not performance-based for all participants.

Once participants completed the experiment, compensation was manually emailed to them by a backend software support specialist at ETH who did not look at experimental data or participated in the experimental analysis.

Analytical Methods

We used a between-subject-design so that each participant only saw one. All data analysis was conducted with the R statistical package. We applied a number of standard statistical tests when looking at population means and distributions in numerical data. We rely mainly on the non-parametric Wilcoxon rank-sum test to compare medians of distributions. We also apply parametric two-sided t-tests to compare means, for robustness checks in pairwise comparisons. To test for equivalence of two populations, we apply two one-sided t-tests (TOST) with a pre-registered medium effect size (Cohen's d = 0.5), as implemented in the R TOSTER package (Lakens, 2017). To compare distributions, we apply the Kolmogorov-Smirnov test, as implemented in the R stats package, as well as the chi-squared test as implemented in base R. The standard errors reported for linear and logistic regressions are clustered by participant with the R sandwich (Zeilis et al., 2023) and lmtest (Hothort, 2022) packages. The reported results of ordinal logistic regression are computed with the R ordinal package (Christensen, 2023).

In analyzing the freeform text responses, we applied measures of textual complexity and of extractiveness of a text sample with respect to the legal documents. To calculate sentence complexity, we computed the Flesch-Kincaid grade level (Kincaid et al., 1975) score as implemented in the textstat Python module. To calculate extractiveness of the text sample, we computed the ROUGE-1 score of the freeform text with respect to its correlated complaint document using the huggingface Python implementation of ROUGE-1 (Lin, 2004).

Our pre-registration included analyses but no directed hypotheses due to the difficulties we encountered in running pilot studies.[25] In reporting the results, we distinguish between analyses that were announced in the pre-registration and those that were not, with the latter labeled as post hoc.

[25] Law students are a scarce participant population to access. We therefore attempted to conduct pilot studies with Amazon Mechanical Turk, but the task appears to have been too difficult for a meaningful experiment. mTurk workers completed the task with very low rates of accuracy. Even with performance-based compensation, accuracy rates were low and behavioral data suggested low effort (very fast completion times, minimal scrolling within the document).

Results

Participants

All 206 participants passed the pre-registered inclusion criteria.[26] The mean age of participants was 25.2 years, with a standard deviation of 2.8 years. 69% of participants identified as women, 26% as men, and 2% as non-binary, with 3% declining to answer. 64% of participants identified as white, 19% as Asian, 8% as Black,[27] and 12% as Hispanic or Latino.[28]

The participating law schools were located in a variety of geographic regions across the country. The rankings of the law schools varied from a top 5 institution as the highest ranking to an institution with a ranking in the 70s.[29] 93% of participants were pursuing a J.D. degree, with the rest pursuing a master's degree. Of the J.D. degree participants, 49% were 1Ls, 26% were 2Ls, and 21% were 3Ls. Exploratory data analysis did not show any effect of demographic category or year of study in the experimental measures.

Task completion time

We begin by considering the task completion time, with mean task completion times shown in Figure 3.[30] The mean times to complete the experiment ranged from under 16 minutes (941 s) for the Highlighting Only group to over 22 minutes (1,338 s) for the No AI treatment, reflecting a 29.7% reduction in task completion time for the Highlighting Only group.[31] This difference between Highlighting Only and No AI was significant, (post hoc, Wilcoxon, $p < .01$; ttest, $p = .01$) as was the difference between Highlighting Only and Full AI (post hoc ,Wilcoxon , $p < .1$; ttest, $p < .1$). The difference between the time to completion in the No AI and Summary Only

[26] The criteria for inclusion were two-fold: (1) taking a total of at least five minutes to complete the task and (2) scoring an accuracy level on textual questions of at least one standard deviation above the 25% accuracy value than would be achieved in expectation through random guessing. Further, 206 students began the experiment, and 206 students completed the experiment. We did not record any students dropping out part way through the experiment.

[27] Participants could select more than one ethnic affiliation.

[28] The sample included a larger portion of women and slightly larger portion of white students than was reported in the publicly available demographic information about the sampled law schools, although we did not have information regarding the demographic qualities of the specific students exposed to solicitations to participate in our experiment (often this would have been a specific class or student group, which might have different demographic statistics than the student body at large).

[29] The effects we describe persisted on a within-institution basis as assessed through a qualitative assessment; we did not perform a statistical analysis due to the small sample size from each institution.

[30] Plots of the distributions of task completion time are in Figure 1.

[31] We measured the time to complete the experiment as the sum of the times necessary to complete the set of questions for each of the three documents; in other words, a participant's completion time was computed as the sum of the time to review and answer all questions for document 1 and for document 2 and for document 3. In this way, we excluded any time that would have reflected the consent screen, the exit survey, or transitioning between documents.

treatments was not statistically significant (post hoc, Wilcoxon, p = 1; ttest, p = .8). All pairwise comparisons are reported in Table 2.[32]

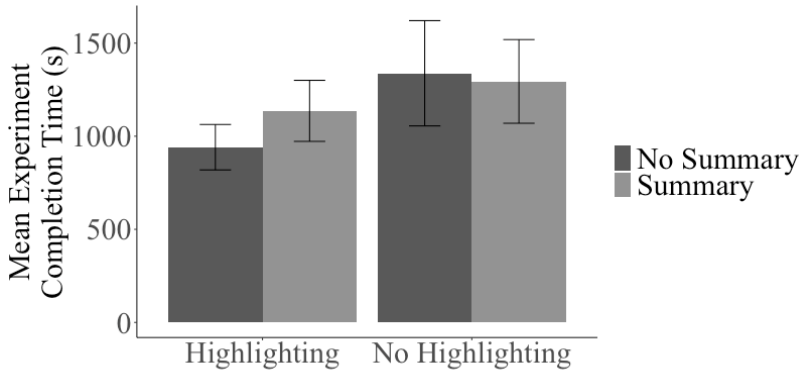


Figure 3: The mean time to complete experiment varied significantly by experimental treatment. Error bars indicate +/- 1 standard error.

The pattern of timing is interesting not only because Highlighting Only significantly reduces the completion time but also because Highlighting Only represents an interior solution in the space of AI assistance tested in this experiment. Highlighting Only reduces task completion time relative to No AI but also relative to Full AI. This relative slowness of Full AI is not due to the summary being detrimental–completion times for the Summary Only group were statistically equivalent[33] to completion times for the No AI treatment. Rather we here experimentally demonstrate the importance of calibrating the amount and form of AI assistance provided during human-machine cooperation to complete a legal task.

[32] We chose a non-parametric test as our primary statistical criterion due to our expectation that the distribution of outcomes need not match the distribution assumed for a t-test. The t-test was used as a robustness check.

[33] A two-one sided t-test confirms that Summary Only and No AI are statistically equivalent ($p < .05$, Cohen's $d = .5$).

	Mean (SD)	Highlighting	Summary	Full ML
No ML	1338	<.01*	1	.4
	(948)	(<.01*)	(.8)	(.2)
Highlighting	933		<.001**	<.1'
	(468)		<.01*	<.1'
Summary	1294			.3
	(818)			(.3)
Full ML	1136			
	(580)			

Table 2: Post hoc pairwise comparisons of mean time to task completion. The Highlighting Only treatment reduces task completion time significantly relative to all other treatments. Primary indicators report significance levels for Wilcoxon rank sum tests. Indicators in parentheses report significance levels for t-tests. 'p < .1, * p < .01, ** p < .001.

Of course, speeding up work may occur for a variety of reasons, including carelessness or overconfidence. It could be that the treatment that enabled participants to work the fastest did so at the expense of work quality. If a mechanism of lower effort or lower care explains the lower task completion time for the Highlighting Only treatment, we would expect to see that the speed gains came at the price of reduced work quality. We therefore next turn to accuracy in the recall and information retrieval textual questions to see whether the timing results reflect a time-accuracy trade-off.

Accuracy

The mean accuracy in responding to the textual questions was strikingly similar (and high[34]) across treatment groups, ranging from M = 85.4% (SD = 14.6%) in the Full AI condition to M = 87.7% (SD = 12.5%) in the Summary Only condition. The difference between the highest and lowest performing groups was not statistically significant (Wilcoxon, p = .7; ttest, p = .8). Indeed, the two groups were statistically equivalent as assessed by a TOST analysis for a medium effect size (post hoc,[35] p < .01, Cohen's d = .5). The AI treatments therefore had had no impact on the mean accuracy of responses to the textual questions.

Even if mean accuracy did not vary across treatments it could still be that potential differences in the distribution might raise concerns about AI-assisted work quality. We therefore tested for

[34] We consider the accuracy high given that compensation was flat; there was no financial incentive to gain a high accuracy and yet participants nonetheless took substantial amounts of time to locate correct answers.

[35] All equivalence tests are post hoc as we neglected to explicitly mention these tests in our pre-registration. We did, however, account for the possibility of equivalence in our open-ended listing of potential outcomes.

differences in the distributions of accuracy scores[36] using Kolmogorov-Smirnov tests. However, there were no statistically significant differences in the accuracy score distributions across treatments. In all pairwise comparisons, the distributions were not statistically different (post hoc, $p > .2$ for all pairwise comparisons).[37] Likewise, chi-square tests for the rates of achieving the highest levels of accuracy (at least 95% accuracy, or, no more than 1 question wrong) indicated no difference in the rate of achieving very high levels of accuracy across the AI treatments (post hoc, $p > .8$ in all cases).

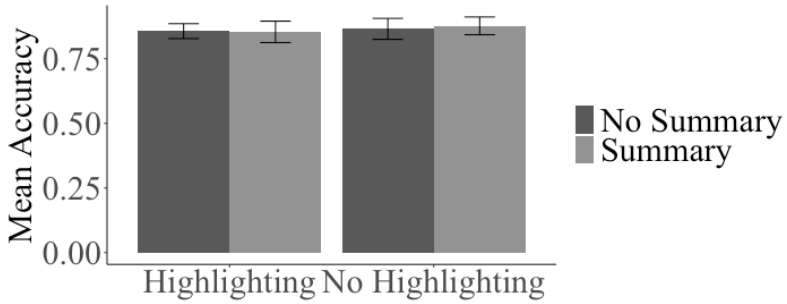


Figure 4: Mean accuracy across treatment groups was statistically equivalent.

The results for accuracy suggest that both the mean accuracy and the distribution of accuracy within treatment groups did not vary in response to the experimental treatment. In view of the substantial speed-up we see with Highlighting Only, the statistical equivalence in mean accuracy and in distributions of accuracy seem promising. Of course, it is possible that AI that speeds participants up distorts their performance in ways not tied to accuracy; we therefore next consider indicators associated with discretionary judgment and professional responsibility.

Discretionary judgment and professional responsibility

The last two questions on each task consisted of a Likert scale rating of the quality of the legal complaint and a freeform description to explain the quality rating. We use this data to evaluate discretionary judgment and professional responsibility, as described in Table 1. We find no difference in exercise of judgment or responsibility across treatments.

We first analyze the numerical quality ratings using a linear regression. We regress quality ratings against the document identifier and against the interaction of the two treatment variables. We find significant effects only on the document identifiers, consistent with participants finding differences in quality as among the three legal complaints. However, results do not exhibit any

[36] Distribution plots for accuracy are shown in Figure A2.

[37] Plots of the distributions of accuracy are available in the appendix. We inspected distributions in part due to inspiration from Stevenson and Doleac's work, where the authors found changes in the distribution of sentencing leniency even as the overall rate of incarceration remained unchanged.

AI-induced bias in participants' assessment of the complaint document. Full results of the regression are reported in Table 3.[38]

	Quality Rating	Character Count	Complexity Score	Extractiveness Score
Document 1	**4.35*** ** (.08)	**161.8*** ** (11.3)	**10.04 *** ** (.48)	**.17 *** ** (.01)
Document 2	**3.87*** ** (.08)	**120.1*** ** (11.3)	**9.84 *** ** (.48)	**.17 *** ** (.01)
Document 3	**3.87*** ** (.08)	**116.9*** ** (11.3)	**8.96 *** ** (.48)	**.15 *** ** (.01)
Highlighting	-.07 (.09)	-6.1 (12.6)	-.38 (.53)	.009 (.01)
Summary	-.09 (.10)	13.3 (12.9)	-.19 (.55)	.002 (.01)
Highlighting * Summary	.13 (.14)	-20.4 (17.7)	.49 (.75)	-.02 (.02)

' p < .05, * p < .01, ** p < .001, *** p < .0001

Table 3: The regression analyses for professional judgment and responsibility ratings uniformly show no effect of the experimental manipulations.

We next consider three effort proxies as described in Table 1: character count, complexity, and extractiveness. As shown in Table 3, the patterns for these effort indicators are the same as the pattern for the numerical quality rating. Across all three effort proxies, we see that the experimental AI treatments did not systematically bias participants' chosen effort level.

Character count represents the number of alpha-numeric characters in a response. Participants were required to write a minimum of 20 characters to advance to the next task, but as Table 3 shows, they wrote more, with a mean count of around 160 characters. This indicates that participants took the task seriously, providing what they regarded as a meaningful response rather than merely the bare minimum. Key to our results we find that participants invested the same amount of effort across treatments, as measured by the character count proxy. This is significant because it shows that participants in the Highlighting-Only treatments did not save time by cutting down on their freeform explanations. Similar results obtain with respect to the sentence complexity ratings and extractiveness ratings. This suggests equal effort investment or work quality by participants across treatments. This provides additional evidence that participants in the Highlighting Only treatments did not reduce their task completion time by reducing effort or quality in the freeform text inputs. We interpret this as suggesting that participants showed the same sense of professional responsibility across treatments.

[38] Column 1 of Table 3 reports the results of a linear regression for a Likert scale judgment variable. We have also applied an ordinal logistic regression, which yields the same results of non-significance of the experimental treatments. Those results are reported in Table A1 in the Appendix.

Finally, we consider the exercise of professional responsibility by examining the reading process itself. We measure whether participants read to the end of a document before giving an assessment regarding the quality of that document.[39] Applying logistic regression analysis, we find no treatment effects on this measure of procedural quality as shown in Table 4. Participants' exercise of some minimal level of professional responsibility – in this case by looking at the whole legal complaint before offering a quality assessment – was unaffected by the presence of either form of AI assistance.

Variable	Estimate
Document 1	**-0.37** . (.22)
Document 2	**-1.24** *** (.25)
Document 3	**0.61** ** (.22)
Highlighting	**-0.08** (.25)
Summary	**-0.15** (.26)
Highlighting * Summary	**-0.05** (.36)
' p < .05, * p < .01, ** p < .001, *** p < .0001	

Table 4: Logistic regression analyses for odds likelihood ratio of reading legal complaint document through to completion.

Discussion

We find that the right kind of AI assistance can speed up the average rate of task completion on a realistic legal task while not affecting the means or distributions of measurements of the quality of legal work and process. Further, we document the possibility of too much AI, showing that too much assistance can reduce the benefits of human-machine cooperation in legal work.

[39] It's possible that the language and structure of a complaint is so formulaic that it is unnecessary to read to the end of a document. Likewise, it's possible that participants read to the effective end of a document but did not scroll to see structural elements such as a signature or formatting at the end of the document. We account for these possibilities by defining the end of the document as the maximum position with the complaint to which at least 10% of participants had scrolled. In two cases, more than 50% of participants had scrolled to that maximum position and in one case roughly 15% of participants had scrolled to that maximum position. We also examined these empirically defined ending positions and found these were reasonably the end of the documents. Further, the results presented here were robust to different cutoffs for defining the end of the document.

The fact that the Highlighting Only was so helpful may seem foreordained to a reader with the benefit of hindsight but was in fact surprising. The highlighting could easily have proved distracting, at least when the highlighted content did not overlap with a question. The highlighting being helpful was especially unexpected considering it was not trained with the intention of being helpful to humans, but rather is a byproduct of training the AI to produce one sentence summaries. It is not entirely surprising that the highlighting points to useful information, but this was far from guaranteed.

Our study also shows an instance, in which repurposing AI outputs – a practice sometimes regarded as potentially troubling (Nielsen, 2021) - worked. We took an AI product trained for one task and one population (to assist expert attorney employees in producing publishable one sentence summaries of complaint documents) and tested it on another task and another population.

Of course, an optimistic interpretation of the benefits of AI in the legal system depends on the external validity of our results. In understanding how likely these results are to generalize, it would be helpful to understand why Highlighting Only speeds up work without sacrificing quality. While our experiment does not definitively answer this question, we present some circumstantial evidence that tends to rule out possibilities relating to the information content of the highlighting. We also present circumstantial evidence that highlighting did not work through making participants overconfident, suggesting that the highlighting likewise does not work through an exclusively affective mechanism. Likely the mechanism involves both cognitive and affective mechanism.

Highlighting did not change the distribution of participants' answers

One plausible mechanism for the highlighting to speed up participant performance without any trade-off in quality could be that the highlighting contained answers to questions and therefore was a useful source of information. If this were the case, we would expect that participants with highlighting performed better than those without highlighting on questions for which the response was highlighted. However, this was not the case, as we confirmed with a chi-square distribution test in which we looked at whether participants with access to highlighting performed better than participants without highlighting on questions for which the correct response was highlighted, but we found no statistical differences between the groups (chi-square test, post hoc, $p > 0.05$ in all cases).

We also considered a broader version of this mechanism, in which highlighting in some other way affected the content of salient text to participants. Even if participants with access to highlighting did not do better on questions for which the correct response was highlighted, it could be that highlighting conveyed information or otherwise changed the information most salient to participants. If this were the case, we might expect the pattern of wrong responses to be different as between participants who had access to highlighting and those who didn't. Again, however, we found no difference in the distribution of responses, correct or incorrect, to any individual question among the different experimental treatments (chi-square test, post hoc, p >

0.05 in all cases). Examples of distributions for individual questions are shown in Figures A6 and A7 for the case of No AI and Highlighting Only.

In short, there is no statistical difference we observe in how participants answered the multiple-choice questions. Not only is accuracy the same across experimental treatments, but the pattern of response choices (correct or incorrect) is also statistically indistinguishable.

Highlighting did not lead participants to be overconfident in their performance

As we were collecting data, we performed exploratory data analysis that might lead us to pre-register further analyses or monitor for any potential reasons to pause or redesign the experiment.[40] We introduced a subjective self-assessment to the exit survey that includes data from approximately ⅓ of the total participant pool (N = 68). Participants were asked to rate, on a 5-point Likert scale, how strongly they agreed that "I performed well on the tasks in this experiment." With this additional question, we sought to understand whether an affective rather than informational mechanism might explain the efficiency gain of the Highlighting Only condition.

In the subjective performance assessments, there was no statistically significant difference in the mean rating among the four treatment groups, which ranged from M = 3.5 (SD = 1.2) in the case of No AI to a mean of M = 4.1 (SD = 1.0) in the case of Full AI (post hoc, W = 171, p = .2). The full distributions for self-reported performance assessments are shown in Figure A5.

Participants had a good sense of their performance on the task. We regressed accuracy on both the experimental treatments (previously found to have a null effect) and subjective performance.[41] Subjective performance was predictive of accuracy (post hoc, $\beta = .68$, SD = .30, $p < .05$), while the experimental treatments again showed a null effect. We interpret these findings as suggesting that participants had a good sense of whether they had located the correct answers to questions regardless of the experimental treatment to which they were assigned.[42] If there had been differences among ML treatments in the effect of subjective performance, this could have suggested an affective mechanism: for example, perhaps the Highlighting Only treatment biased participants to be more confident relative to the rest of the subject pool. However, we found no evidence for this or other affective mechanisms that might work through biasing confidence.

On the other hand, we speculate that participants took however much time was necessary to achieve high performance because in this task, they had real-time feedback in that they could easily know whether they had found the relevant complaint text. The time that was ultimately necessary was systematically different depending on the AI treatment. If this interpretation is correct, we expect that an experiment with sufficiently constraining exogenous time pressure

[40] We indicated this intentional in our initial pre-registration document. With this modification we did not deviate from our initial procedure other than in adding one additional question to our exit survey.

[41] We did not include an interaction term in the regression due to the small number of participants.

[42] Given that the correct answers were designed to be clear from the text of the complaint, this is unsurprising. Participants should have known whether they had or had not located the relevant portion of text in the complaint.

would bring about accuracy differentials under time pressure that appear in this paper as time differentials without time pressure. In other words, we would expect, under time pressure, that participants with highlighting should demonstrate higher accuracy than participants without highlighting.

Limitations

We highlight the novel contribution of our results in pointing the way towards a positive outcome - one that could increase the efficiency of legal work without any identifiable costs. However, the appropriate degree of optimism about these experimental results depends on the expected external validity of a sample of around 200 law students and of the experimental task.

Consider first that the sample is small relative to the effect sizes that may matter to some readers. Even with this small sample, the efficiency gain was easily detected because it is so substantial. It could be, however, that the null effect of the AI manipulation in this experiment is a result of the small sample size rather than a true null. Our sample permits the construction of confidence intervals (reported supra) on any such differences. Thus, where these potential differences matter can already be assessed by any reader who brings a normative prior as to the acceptable effect sizes of any such differences. Further, it's possible that there aren't any trade-offs at all - it's possible that the Highlighting Only treatment might also produce an as-yet undetected improvement in accuracy. Even if there are trade-offs, our results show the bounds as to these trade-offs and suggest that the Highlighting Only treatment is a good option.

Next consider what limitations may be imposed by the sample population, law students. It has previously been shown that results from law students may not replicate on attorneys, judges, or other professionals involved in the administration of law. Both Spamann and Klöhn (2022) and Kahan et al. (2016) found that law students and judges do not behave similarly in legal task experiments. On the other hand, the experimental task in this study is different from those of the previous works. The task of information recall or retrieval may be less amenable to prejudice or affect, a mechanism that seemed to explain some of the differences identified in the prior study, suggesting one key distinction as between this experiment and studies showing that law student behavior does not generalize to legal professionals. In prior studies, ideology or other legally irrelevant factors affected judges and students differently in the experimental task, but they were also asked to make ultimate judgments rather than to identify facts or otherwise engage in smaller procedural steps that involved less discretion. Further, previous studies consisted of tasks (making case judgments) more often performed by senior legal professionals, whereas our task focused on tasks more likely to be performed by junior legal professionals, such as law clerks or junior law firm associates. Future methodological work should look to whether the tasks we study here might be ones in which law students behave similarly to legal professionals.

Another factor limiting external validity is that participants may not have faced the same time pressure as working professionals would. Judges, attorneys, and even law clerks work under heavy time pressure and are often said to face overwhelming caseloads (Judicial Council of California, 2020). The law student participants in this study took the time necessary to achieve high levels of accuracy, despite no monetary incentive to do so. In other words, it appeared that their drive to do good work could often be satisfied despite any constraints imposed on them by

law school coursework or other family obligations; it's not so clear that practicing attorneys or judicial clerks have this same luxury in all cases. Future work could test the AI assistance under exogenously time constrained circumstances to see whether the advantages of Highlighting Only persist under strong time pressure.[43]

In this experiment, we looked at short-term outcomes during a single experimental session. It's possible that the effects we find might diminish or qualitatively change for users who adopted the AI assistance into their everyday legal practice. There is precedent for this in the literature, as with Stevenson and Doleac's (2021) findings, discussed supra, that the impact of a risk instrument tool diminished over time. On the other hand, there are reasons to imagine that the effects would remain steady and or might even grow. Users of the tool could come to trust the tool more, or alternatively become more adept at knowing its strengths and weaknesses, producing even greater efficiency gains over time compared to their first experience. A distinction between our tool and the risk instrument tool studied in Stevenson and Doleac's work is that for our task and tool of interest it is unlikely that participants would have strong priors or ideological preferences that would systematically contravene the AI assistance, in contrast to the case of a risk instrument where judges might systematically prefer leniency or systematically have priors that make them inclined to disagree with the risk assessment tool. Future work could look to the long-term outcomes of access to highlighting to confirm that the efficiency gains we identify would indeed be long-lasting.

Our conclusions should be interpreted as a demonstration of possibility rather than a guarantee of success of AI for enhancing legal work efficiency. Our conclusions are also specific to the task we studied. We do not argue that highlighting is a better intervention for all legal outcomes; we show only that highlighting enhances efficiency without introducing distortions into legal work for the specific task studied here. It is quite likely that for other legal tasks, another form of AI assistance would be superior. For example, one could imagine that the summary but not the highlighting would speed participants up if they were tasked with identifying relevant precedent for a given set of facts, another common but distinct legal task often performed by junior attorneys.

In a time of rapid AI proliferation combined with a relative lack of expertise as to the best ways to use AI, the form of AI assistance we study here, as a tool rather than a replacement for human decision makers, is the most likely form we will see (and already do see) adapted in legal work. In the absence of deliberate policy change, legal AI tools that speed up legal work without otherwise biasing the quality of that work are likely preferable to AI tools with unknown effects or with known distortions. Thus, even with the limitations discussed here, our findings suggest reason for optimism and rapid adoption of forms of AI assistance known to enhance efficiency and therefore likely to prove helpful in increasing access to legal work and to the justice system.

[43] In a pilot study of law students with exogenous time pressure, the time pressure did not eliminate time differences between treatments. Results available from authors upon request.

Conclusion

To date, the empirical literature on AI assistance in legal applications has mostly been limited to a focus on use cases by government actors working in criminal justice with unidimensional risk-scoring instruments. Existing studies have not demonstrated consistent or significant gains from algorithmic assistance, despite obvious theoretical benefits of AI in the legal system. We demonstrate a substantial gain in legal work efficiency without loss of work quality and without substantial training for human users. The effect is an interior solution, that is one in which the right kind of AI assistance, and not more, proves most useful within the set of possibilities we test. In proposing, measuring, and analyzing a range of metrics to assess the impact of the AI tool in the hands of humans, we put forward a simple rubric for evaluating AI assistance in legal tasks upon which future research can build.

Bibliography

American Bar Association. (1983). "Model Rules of Professional Conduct: Rule 1.3 - Diligence."https://www.americanbar.org/groups/professional_responsibility/publications/model_rules_of_professional_conduct/rule_1_3_diligence/

Nielsen Aileen, Skylaki Stavroula, Norkute Milda and Stremitzer Alexander, (2023). "Effects of XAI on Legal Process." In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law (ICAIL 2023)* (pp. 1-5). ACM. https://doi.org/10.1145/3594536.3595128

Ayres, C. E., Rankin, A., & Sturz, H. (1963). "The Manhattan Bail Project: An Interim Report on the Use of Pretrial Parole." *New York University Law Review*, 38, 67-95.

Bender Emily M. and Friedman Batya. (2018). "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science." *Transactions of the Association for Computational Linguistics, 6*, 587–604. DOI:10.1162/tacl_a_00041

Blair-Stanek Andrew, Carstens Anne-Marie, Goldberg Daniel S., Graber Mark, Gray David C. and Stearns Maxwell S. (2023). "GPT-4's Law School Grades: Con law C, crim C-, law & econ C, partnership tax B, property B-, tax B." *SSRN*. http://dx.doi.org/10.2139/ssrn.4443471

Chien Coleen V. and Kim Miriam. (2024). "Generative AI and legal aid: Results From a Field Study and 100 Use Cases to Bridge the Access to Justice Gap." *UC Berkeley Public Law Research Paper*, forthcoming. *Loyola of Los Angeles Law Review*, forthcoming. SSRN. https://ssrn.com/abstract=4733061

Chohlas-Wood Alex, Nudell Joe, Yao Keniel, Lin Zhiyuan. (Jerry), Nyarko Julian, and Goel Sharad (2021). "Blind Justice: Algorithmically Masking Race in Charging Decisions." In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 35–45). Association for Computing Machinery. https://doi.org/10.1145/3461702.3462524

Choi Jonathan H. and Schwarcz Daniel. (2023). "AI Assistance in Legal Analysis: An Empirical Study." *Journal of Legal Education*, 73, forthcoming. https://ssrn.com/abstract=4539836

Choi Jonathan H., Monahan Amy and Schwarcz Daniel. (2023). "Lawyering In the Age of Artificial Intelligence." *Minnesota Law Review, 109*, forthcoming 2024. Minnesota Legal Studies Research Paper No. 23-31. https://ssrn.com/abstract=4626276

Choi, Jonathan H., Hickman Kristin E., Monahan Amy, and Schwarcz Daniel. (2023). "ChatGPT Goes to Law School." *SSRN*, 4335905. Accessed March 10, 2023. https://ssrn.com/abstract=4335905

Christensen Rune Haubo Bojesen (2023). *Regression Models for Ordinal Data*. https://cran.r-project.org/web/packages/ordinal/ordinal.pdf

Cordon Matthew C. (2011). “Task Mastery in Legal Research Instruction.” *Law Library Journal*, 103(30), 395-400.

Danser, R., Greiner, D. J., Halen, R., Griffin, C. L., & Stubenberg, M. (2023). Randomized control trial evaluation of the implementation of the PSA-DMF system in Polk County, IA. Manuscript in preparation. [Manuscript on file with author; draft dated March 31, 2023].

Dietvorst Berkeley J., Simmons Joseph P. and Massey Cade. (2015). “Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err.” *Journal of Experimental Psychology: General*, 144(1), 114–126. https://doi.org/10.1037/xge0000033

Gardner James A. (2021). “Transmission of Mastery.” *Buffalo Law Review*, 69, 55-68.

Glassman Elena L., Gu Ziwei and Kummerfeld Jonathan K. (2024). *AI-resilient Interfaces [Working draft]*. 1(1), 17 pages. https://arxiv.org/abs/2405.08447

Goldkamp John S. and Gottfredson Michael R. (1985). *Policy Guidelines for Bail: An Experiment in Court Reform*. Temple University Press.

Imai Kosuke, Jiang Zhichao, Greiner James, Halen Ryan and Shin Sooahn. (2021). “Experimental Evaluation of Algorithm-assisted Human Decision-making: Application to Pretrial Public Safety Assessment.” *Arxiv*. https://arxiv.org/abs/2012.02845

Judicial Council of California. (2020, January). “Fact Sheet: Judicial Workload Assessment.” https://www.courts.ca.gov/documents/cjwa.pdf

Kahan Dan M., Hoffman David, Evans Danieli, Devins Neal, Lucci Eugene and Cheng Katherine. (2016). “"Ideology" or "Situation Sense"? An Experimental Investigation of Motivated Reasoning and Professional Judgment.” *Faculty Publications*, 1801. https://scholarship.law.wm.edu/facpubs/1801

Kahneman Daniel, Sibony Olivier and Sunstein Cass (2021). *Noise: A flaw in human judgment*. Random House.

Killian Alex. (2019). “Towards an Automatic Measure of Effort in Writing [Master's thesis, Computer Science].” *ProQuest Dissertations Publishing*. https://www.proquest.com/openview/f30a5a651c6ea25fdb50faf224a34e29/1?pq-origsite=gscholar&cbl=18750&diss=y

Kincaid J. Peter, Fishburne Robert P., Rogers Richard L. and Chissom Brad S. (1975). “Derivation of New Readability Formulas (Automated readability index, fog count, and Flesch reading ease formula) for Navy Enlisted Personnel.” *Research Branch Report*, 8-75. Institute for Simulation and Training. https://stars.library.ucf.edu/istlibrary/56

Kleinberg John, Lakkaraju Himabindu, Leskovec Jure, Ludwig Jens and Mullainathan Sendhil. (2017). "Human Decisions and Machine Predictions." *The Quarterly Journal of Economics*, 133, 237-293. https://doi.org/10.1093/qje/qjx032

Lakens Daniel. (2017). "Equivalence Tests: A Practical Primer for T Tests, Correlations, and Meta-analyses." *Social Psychological and Personality Science*, 8, 355-362. https://doi.org/10.1177/1948550617697177

Lin Chin-Yew. (2004). "ROUGE: A Package for Automatic Evaluation of Summaries." In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)* (pp. 25-26). Barcelona, Spain, July 25-26, 2004.

Linna Daniel. (2020). "Evaluating Legal Services: The Need for a Quality Movement and Standard Measures of Quality and Value." In *R. Vogl (Ed.), Research Handbook on Big Data Law*. https://www.legaltechlever.com/wp-content/uploads/sites/151/2020/03/Linna-Evaluating-Legal-Services-Quality-Value-2020-03-12.pdf

Miller Joel and Maloney Carrie. (2013). "Practitioner Compliance With Risk/Needs Assessment Tools: A Theoretical and Empirical Assessment." *Criminal Justice and Behavior*, 40, 716-736.
Moran Lyle. (2021, August 31). "Court Backlogs Have Increased By an Average of One-third During the Pandemic, New Report Finds." ABA Journal. https://www.abajournal.com/news/article/many-state-and-local-courts-have-seen-case-backlogs-rise-during-the-pandemic-new-report-finds

Mosqueira-Rey Eduardo, Hernández-Pereira Elena, Alonso-Ríos David, Bobes-Bascarán Jose and Fernández-Leal Ángel. (2022). "Human-in-the-loop Machine Learning: A State of the Art." *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-022-10246-w

Nielsen Aileen. (2021). *Practical Fairness*. O'Reilly Media.

Norkute Milda, Herger Nadia, Michalak Leszek, Mulder Andrew and Gao Sally. (2021). "Towards explainable AI: Assessing the usefulness and impact of added explainability features in legal document summarization." In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI EA '21)* (Article 53, pp. 1–7). Association for Computing Machinery. https://doi.org/10.1145/3411763.3443441

Ramírez Jorge, Baez Marcos, Casati Fabio and Benatallah Boualem. (2019). "Understanding the impact of highlighting in crowdsourcing tasks." In *The Seventh AAAI Conference on Human Computation and Crowdsourcing*. https://ojs.aaai.org/HCOMP/article/download

Sloan Carly Will, Naufal George and Caspers Heather. (2023). "The Effect of Risk Assessment Scores on Judicial Behavior and Defendant Outcomes." *Journal of Human Resources*, 0221-11470R3. https://doi.org/10.3368/jhr.0221-11470R3

Spamann Holger and Klöhn Lars. (2022). "Can Law Students Be Used to Study Judicial Decision-making Experimentally?" [Manuscript in preparation]. Draft paper.

Stevenson Megan T. (2018). "Assessing Risk Assessment in Action." *Minnesota Law Review*, 103, 303-384.

Stevenson Megan T. and Doleac Jenifer L. (2022, September 29). "Algorithmic Risk Assessment in the Hands of Humans." *SSRN*. https://dx.doi.org/10.2139/ssrn.3489440

Sundar Sindhu. (2022, Oct. 20). "As Big Law Looks to Tech to Help Draft Legal Briefs and Log Billing Hours, Top Firms Like Orrick and Shearman & Sterling are Using AI to Fast-Track Work on M&A Deals." *Business Insider*. https://www.businessinsider.com/big-law-firms-using-ai-fast-track-ma-deals-work-2022-10.

Torsten Hothorn, Zeileis Akim. Farebrother Richard W., Cummins Clint, Giovanni Millo and Mitchell David. (2022). *Sandwich: Testing Linear Regression Models*. https://cran.r-project.org/web/packages/lmtest/lmtest.pdf

Viljoen Jodi L., Jonnson Melissa R., Cochrane Dana M., Vargen Lee M., and Vincent Gina M. (2019). "Impact of Risk Assessment Instruments on Rates of Pretrial Detention, Postconviction Placements, and Release: A Systematic Review and Meta-analysis." *Law and Human Behavior*, 43(5), 397. DOI: 10.1037/lhb0000344

Whiting Penny, Savović Jelena, Higgins Julian P., Caldwell, D. M., Reeves, B. C., Shea, B., Davies, P., Kleijnen Jos, Churchill Rachel and ROBIS group. (2016). "ROBIS: A New Tool to Assess Risk of Bias in Systematic Reviews Was Developed." *Journal of Clinical Epidemiology*, 69, 225-234. https://doi.org/10.1016/j.jclinepi.2015.06.005

Yu Alan. (2020, February 20). "Can Algorithms Help Judges Make Fair Decisions? Is Taking Away the Human Factor the Key to More Just Rulings?" *WHYY*. https://whyy.org/segments/can-algorithms-help-judges-make-fair-decisions/

Zeileis Achim, Lumley Thomas, Graham Nathaniel and Koell Susanne. (2023). "Sandwich: Robust Covariance Matrix Estimators." https://cran.r-project.org/web/packages/sandwich/sandwich.pdf

Appendix

Selection of complaint documents for experiment

We selected three legal complaint documents of varying length and legal complexity. In order of task presentation, the legal complaint documents concerned an employment discrimination lawsuit, fraud in an LLC, and a claim for attorney's fees under the Fair Credit Reporting Act. When selecting legal complaint documents, we faced several constraints as they related to the provision of a sample from which to choose and as they related with respect to the number of complaints we could hope to test in the experiment. We describe those limitations and our selection priorities and process here.

We were limited to the complaints provided to us by our industry partner because we did not have direct access to the AI tool but rather only to its inputs and outputs. We therefore cannot and do not put forward any assertions regarding the quality of the model outputs generally. As reported by Norkute et al. (2021), in a blinded study conducted by expert attorney labeling, 75% of the one-sentence summaries generated by the model were judged acceptable for publication, as compared to 88% of those created by humans. For purposes of this experiment, having access to the algorithm or only to its inputs and outputs is indistinguishable. Our industry partner provided 100 randomly selected legal complaint documents along with the corresponding AI-generated summaries and highlighting. These same documents had been reviewed by two attorney employee raters who independently checked whether the AI summaries were publication ready.

We also faced constraints due to the difficulty of recruiting law students for experimental studies. We designed the experiment to minimize variation in the experiment that might otherwise arise from experimental permutations in ordering or arrangement of texts rather than from the AI features, to minimize the variation in outcomes that would not directly relate to AI feature availability. We also needed to keep the length of time of the experiment relatively short to mitigate concerns about difficulties in recruiting students. Small initial pilot tests led us to selecting three documents as the right number for a 20-minute experiment.

We limited our selection of complaints to those for which the model produced an output that was judged publication-ready by both of two attorney raters. Therefore, by design, we study in this experiment the effects of AI assistance, when that AI assistance is of a presumptively good standard. However, this selection is already representative of most outputs from the AI tool if not of all outputs. 58 out of 100 documents were rated by both attorneys as publication-ready, while 91 out of 100 were rated as such by at least one attorney reviewer. In contrast to our choice, we recognize that the growing body of literature on algorithmic aversion looks to the impact on human users of algorithmic mistakes. This is a key area for future investigation and is also important for understanding how AI assistance will affect real world legal practice (Dietvorst, 2015).

Winnowing the cases based on the publication-ready rating left 58 complaints for review. Next, without knowledge of the content of the AI summaries or highlighting, one of us manually

reviewed 22 cases.[44] The author reviewed cases looking for complaints that included good quality optical character recognition(that is, cases in which the conversion of pdf images into plain text was of good quality), a range of legal topics, a maximum word count consistent with the maximum length of the AI model, and well delineated and sufficiently complex facts and law so that there would be enough content for question generation. This was done manually but could have been an automated process based on character count. In the interests of quality, of user experience, and of conforming to our experimental remit to study well-performing AI, we further limited ourselves to reviewing documents that were not so long that the inputs exceeded the input capacity of the algorithm. If we had included lengthier complaint documents, this would have resulted in highlighting that did not run the full length of the document and therefore would not follow our self-appointed task of testing the impact of well performing AI assistance. Of the 22 cases that were manually reviewed, 3 cases were discarded because their length ran longer than the model's maximum input length.

Of these 22 cases reviewed by an author, 8 documents were selected and assigned to several legal student research assistants for review and question generation as representing a range of difficulties and topic areas.[45] The research assistants were instructed to draft multiple choice questions of varying levels of difficulty regarding what they identified as the key facts or legal assertions in the document. The questions were generated without knowledge that AI summaries or highlighting existed and without knowledge of the experiment design and hypotheses. The questions were then reviewed by the researchers, who remained blind to the content of the AI summaries or highlighting when selecting[46] the questions and documents for inclusion. The researchers selected the documents where the question quality from the research assistants was highest.

Ultimately, we selected three complaints that covered a variety of legal topics, had questions of varying difficulty level, and for which the question-answers were well-distributed throughout the text of the document. The three legal complaints covered (in order presented) topics of employment discrimination on the basis of race, fraud by one partner in a limited liability company (LLC) perpetrated against the other partners in conveying away real property held by the LLC, and recovery of attorney's fees for a previously litigated violation of the Fair Credit Reporting Act (FCRA).[47] The documents were chosen without any pilot data about performance on the questions and without any knowledge of the content of the summary or highlighting for the documents.

[44] Review of 22 randomly selected cases from the subset of cases rated as having an acceptable summary yielded more candidate complaints than could be included in the survey. We therefore judged this to be adequate for candidate selection and question generation.

[45] The spreadsheet maintained by the author during complaint review and the text of the rejected complaints are available upon reasonable request.

[46] The researchers sought to ensure that questions were not redundant and had varying levels of difficulty in recall or retrieval of information. The questions were selected based on discretionary judgment and were not studied with pilots or with feedback from third parties.

[47] These three legal complaint documents are available in pdf form here.

Distribution of Task Time and Textual Question Accuracy

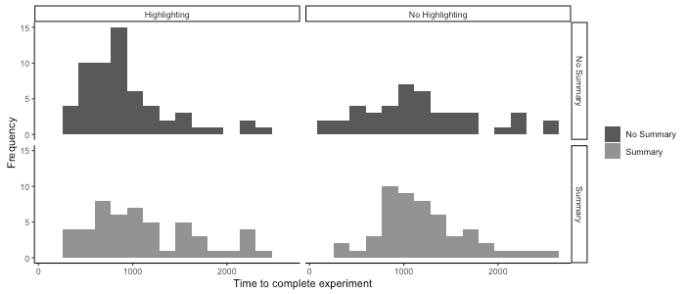


Figure A1: The distribution of task completion time for the four experimental treatments[48]

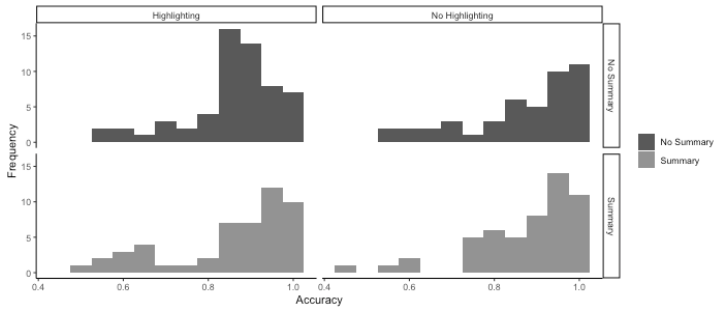


Figure A2: The distribution of accuracy for the four experimental treatments

[48] Three participants who took more than 60 minutes to complete the experiment were removed from the visualization to reduce the spread along the x-axis but were not otherwise removed from the data analysis.

Subjective Performance Assessments versus Accuracy

Mean accuracy per self-reported performance bin (error bars are +/- 1 standard error)

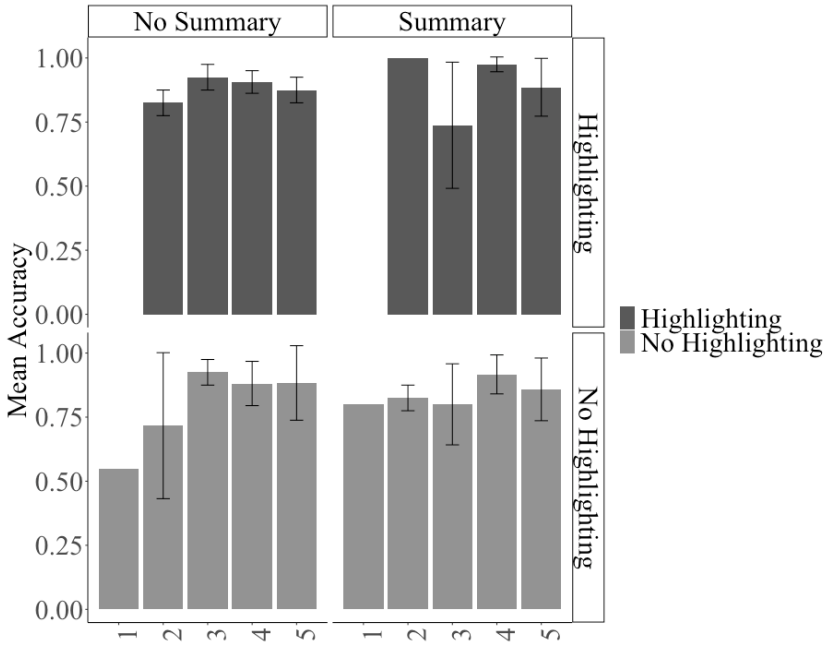


Figure A3: Mean accuracy for each experimental treatment for each value of self-reported performance (errors bars are +/- standard error)

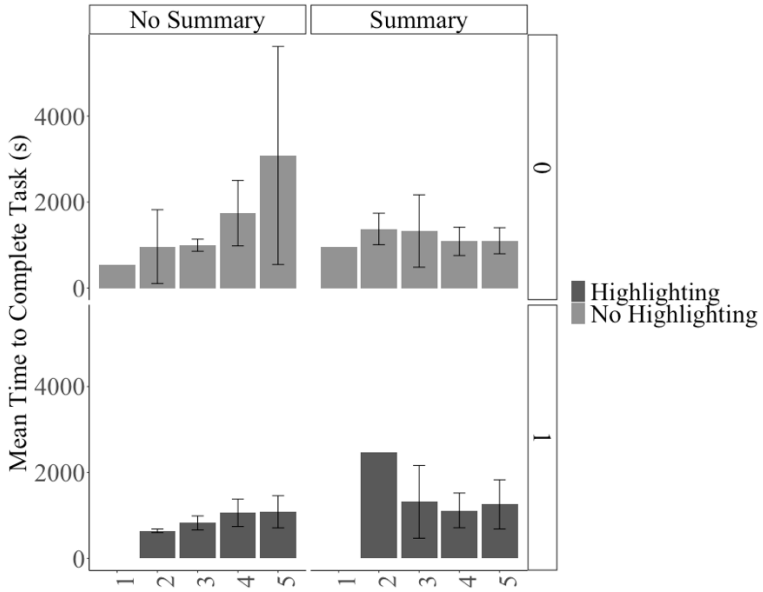


Figure A4: Mean task completion time for each experimental treatment for each value of self-reported performance (errors bars are +/- standard error)

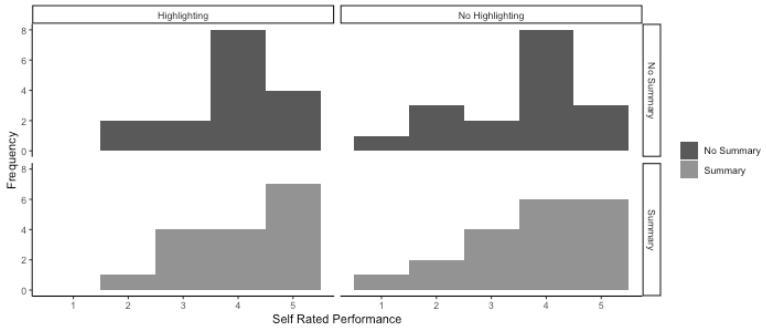


Figure A5: Distribution of self-reported performance ratings for the four experimental treatments

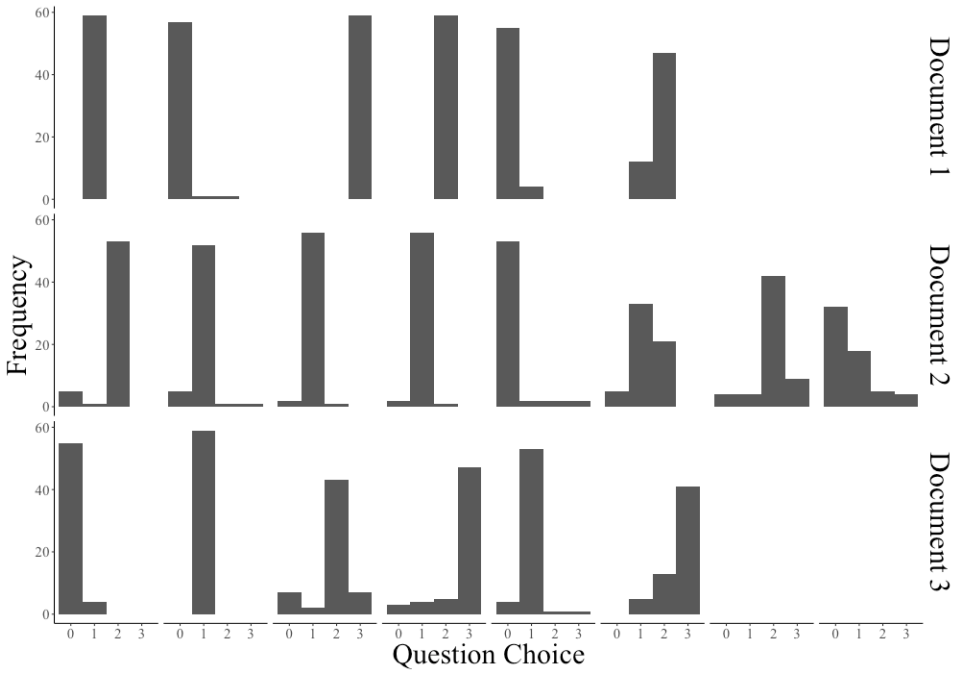


Figure A6: Distribution of responses for No AI treatment

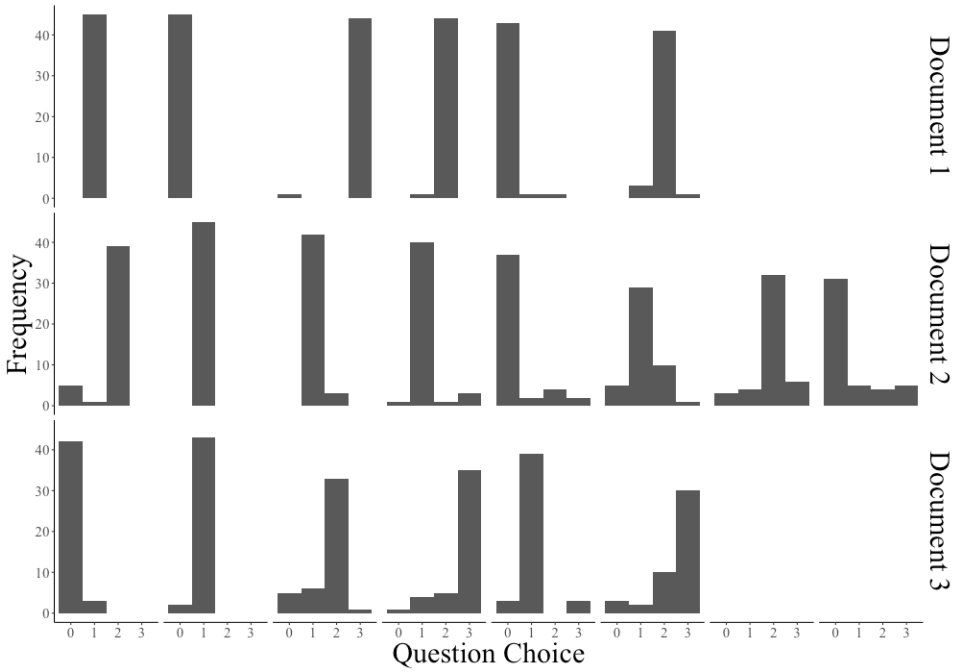


Figure A7: Distribution of answers for Highlighting Only treatment

Variable	Estimate
Document 2	**-1.42** *** (.20)
Document 3	**-1.39** *** (.20)
Highlighting	**-0.13** (.22)
Summary	**-0.22** (.23)
Highlighting * Summary	**.37** (.32)
' p < .05, * p < .01, ** p < .001, *** p < .0001	

Table A1: Logistic ordinal regression for the subjective complaint quality ratings shows significant predictive value only of the document identifier, confirming the results presented in Table 3 with a standard linear regression.

Recruitment Email Template

Subject line: Get paid to try out some new legal tech!

Dear Student,

We are writing to invite you to participate in a well-compensated legal technology experiment ($20 for 20 minutes). Researchers at ETH Zurich and Thomson Reuters (Westlaw) are actively recruiting law students to participate in this experiment regarding the use of proprietary technology for processing legal complaints.

The experiment will take about 20 minutes in total, and you will be compensated with a $20 Amazon gift card. The experiment involves reading a few short legal complaint documents and answering multiple choice questions about the complaints.

To sign up for the experiment, go to this link: [link here]

Please feel free to share this invitation with any [law school name here] law school peers who would be interested in participating. A law school email address at [law school name here] is needed to sign up for the experiment, and any student currently studying at [law school name here] law is eligible to participate.

Thank you for your interest in our experiment!

Sincerely,
The ETH Zurich LawTech team

Full Question Text Per Document

Document 1

In which state was Arising Industries, Inc. qualified and licensed to do business?
- Michigan
- Georgia (answer)
- Kentucky
- California

According to the complaint, what work role did Mr. Smalls perform at Arising Industries, Inc.?
- Axle and tire installation (answer)
- Painter
- Automotive mechanic
- Warehouse clerk

According to the complaint, what did Mr. Smalls begin to experience at the Hazlehurst, Georgia plant in March 2016?
- Discrimination on the basis of sexual orientation
- Discrimination on the basis of religion
- Discrimination on the basis of gender
- Discrimination on the basis of race (answer)

According to the complaint, what was Mr. Edward Justice's role at the Hazlehurst, Georgia plant?
- Customer
- Co-worker
- Manager (answer)
- Applicant

According to the complaint, what happened to Mr. Smalls in or about May 2016?

- He was promoted (answer)
- He was dismissed
- He was given leave of absence
- He received a gratuity

According to the complaint, what did Mr. Smalls request from Mr. Justice on December 1, 2016?
- An employee reference
- A paycheck
- A separation notice (answer)
- Wage increase

Based on your best assessment of the case given the information provided in the document, should the plaintiff prevail on the merits?
- Definitely yes
- Probably yes
- Neither yes nor no
- Probably no
- Definitely no

Please explain the reasons for your assessment of the merits of plaintiff's argument.

[freeform text input, 20 character minimum, no pasting permitted]

Document 2

For what purposes was 88 Street formed?
- For the purposes of selling a parcel of real estate in Elmhurst, Queens
- For the purposes of renovating a building in Elmhurst, Queens
- For the purposes of engaging in real estate development in Queens (answer)
- For the purposes of limiting landlord liability

Who among the plaintiffs agreed to appoint Mr. Dai the 'managing member' of 88 Street?
- All
- None (answer)
- None, except Mr. Aung Myint Soe

- All, except Mr. Bin Bin Zhou

How did Mr. Dai's role at 88 Street appear to change over time?
- He was increasingly excluded from internal decision-making processes
- He increasingly assumed the duties and responsibilities of the managing member, and likely held himself out as such (answer)
- He increasingly sold his capital contributions and thereby lost his right to vote
- He increasingly urged greater transparency in the business accounting methods of 88 Street

What was the purpose of the purchase of the Elmhurst properties by 88 Street?
- To convert cash and cash equivalents into assets
- To develop these properties and re-sell them (answer)
- To set up a headquarters for 88 Street within these properties
- To have apartments built into these properties and to rent them out eventually

Why was Mr. Dai's purported transfer of a membership interest to Khan invalid on its face?
- It violated the clear letter of the operating agreement (answer)
- It lacked a sufficiently clear description of the property
- It violated involved illegal activities
- It was severely one-sided

What was the plaintiffs' understanding from 2011 through July 2015?
- That Somerset Financial Group held the majority of shares in 88 Street
- That each owned a portion of 88 Street proportional to their initial capital investments (answer)
- That they collectively were managing members of 88 Street
- That the Elmhurst properties were losing value

What was the alleged role of Astoria Blvd. in the transfer of property?
- It's a subsidiary of Somerset Financial fully owned by it
- 88 Street took out a mortgage from Astoria Blvd
- Somerset Financial assigned the loan it had made to 88 Street to Astoria Blvd (answer)
- Astoria Blvd was used as collateral for the loan taken from Somerset Financial

What was the last action Mr. Khan allegedly undertook with respect to the parcels of land originally purchased by 88 Street?
- The parcels were transferred to Avenue Astoria without any consideration (answer)
- The parcels were used as collateral for the loan taken out from Somerset Financial
- The parcels were transferred to Dai for a price well below the market
- The parcels remained the property of 88 Street

Based on your best assessment of the case given the information provided in the document, should the plaintiff prevail on the merits?
- Definitely yes
- Probably yes
- Neither yes nor no
- Probably no
- Definitely no

Please explain the reasons for your assessment of the merits of plaintiff's argument.

[freeform text input, 20 character minimum, no pasting permitted]

Document 3

In which state is Portfolio Recovery Associates organized?
- Delaware (answer)
- New York
- Michigan
- California

In which state does Portfolio Recovery Associates have an address for service?
- Delaware
- New York (answer)
- Michigan
- California

Which actions of Portfolio can be characterized as unfair practices under 15 U.S. Code § 1692f?
- Portfolio called itself Portfolio LLC A/P/O GE Capital Retail Bank in the collection claim
- Portfolio acted as a debt collector despite not being registered as a debt collector

- Portfolio tried to collect credit card debt that it didn't own (answer)
- Portfolio used false representations to obtain information concerning a consumer

The action of the plaintiff against the defendant Portfolio Recovery Associates aims to ?
- Secure an injunction against debt collection
- Secure a declaratory judgment that Portfolio Recovery Associates engaged in unlawful conduct
- Secure a declaratory judgment about the lack of standing for the debt collection
- Obtain compensatory damages for previous court proceeding and legal fees (answer)

Which of the following is not included in the compensation plaintiff is seeking?
- Actual damages in the amount of $640
- Compensation for plaintiff's lost time, estimated at a reasonable hourly rate (answer)
- Statutory damages of $1,000
- Costs and reasonable attorney fees of this action

What is the statutory basis for statutory damages demanded by the Plaintiff?
- 15 U.S.C. § 1692c of the FDCPA
- 15 U.S.C. § 1692e of the FDCPA
- 15 U.S.C. § 1692f of the FDCPA
- 15 U.S.C. § 1692k of the FDCPA (answer)

Based on your best assessment of the case given the information provided in the document, should the plaintiff prevail on the merits?
- Definitely yes
- Probably yes
- Neither yes nor no
- Probably no
- Definitely no

Please explain the reasons for your assessment of the merits of plaintiff's argument.

[freeform text input, 20 character minimum, no pasting permitted]