

AI Assistance for Human Review of Default Judgments

Theodora Worledge¹, Othman Bensouda Koraichi², Daniel Bernal^{2,3}, Aviv Caspi^{2,3},
Tatsunori Hashimoto¹, Carlos Guestrin¹, David Freeman Engstrom^{2,4}

¹Stanford Computer Science, Stanford University

²Deborah L. Rhode Center, Stanford Law School

³University of Chicago Law School

⁴Stanford Law School

worledge@stanford.edu

Abstract

Overwhelmed courts in the United States review millions of default judgments each year. Unfortunately, such manual reviews are time-consuming and prone to error. In an audit of 188 debt collection cases granted default judgment by the Superior Court of Los Angeles, we find that 4% contained major defects that should have entirely prevented default judgment, 10% contained inconsistencies requiring reduced judgments, and 32% contained errors requiring amendment prior to judgment. To support courthouses in default judgment review, we collaborated with courthouse attorneys and judges in designing a Default Assistant. The Default Assistant employs large language models to evaluate a case with respect to predetermined legal requirements and provide cited recommendations for an expert user's review. We equip users to verify these recommendations by grounding the assistant's explanations in cited quotes and tables from the original case filings. We conduct a controlled study with 66 law students that conservatively simulates court review, with more time and resources than court staff. We nevertheless find users aided by the Default Assistant were 6.0% more accurate on the average requirement than unaided reviewers ($p < 1.0e-4$). Simultaneously, users were 25.9% faster in reviewing the average requirement than unaided reviewers ($p < 2.5e-10$). Statutory requirements demanding extensive document search realized the largest gains, with error reductions and time savings from AI assistance up to 62% and 34%, respectively, relative to unassisted user performance and with differences statistically significant ($p < 0.05$). Our work provides a proof-of-concept that AI assistants with citations have the potential to help resource-constrained courts conduct default judgment review more accurately and efficiently.

Introduction

The American civil justice system is failing to meet the needs of millions of litigants. Many of the 15 million civil cases filed in American state courts each year represent personal crises—a debt collection that results in wage garnishment, an eviction that leads to homelessness (Johnson Raba 2023; Garnham, Gershenson, and Desmond 2022)—that fuel cycles of unemployment, poverty, poor health, and family breakdown (Mullen 2019; Desmond and Kimbro 2015). Yet, despite these high stakes, roughly three-quarters of

these civil cases involve at least one person who cannot afford a lawyer (Agor, Graves, and Miller 2015). Many defendants do not take action to defend themselves in court; defendants only respond in 6% of debt collection cases filed at the Superior Court of Los Angeles County (SCLAC) (Johnson Raba 2023). With minimal adversarial process to surface evidence of defects in cases, courts, constrained by the impracticality of rigorous manual review, are more likely to issue erroneous default judgments (Jiménez 2015; Bookman 2024).

The court caseload for default judgment review is crushing. Each year, SCLAC—the largest trial court in the nation¹—routes as many as 30,000 debt collection default judgment requests to court staff for manual review. The high caseload creates severe time constraints. SCLAC research attorneys spend on average four minutes per case to verify over a dozen statutory requirements.² The failure of any requirement influences the judge's decision to grant default judgment in favor of debt buyers pursuant to CA Civil Code §§ 1788.58-60. Given such resource constraints, even the most diligent courts are prone to error.

Using annotation guidelines for statutory and procedural requirements developed by legal academics and court professionals, we audited 188 debt-buyer plaintiff collections cases granted default judgment by SCLAC. We found that 4% contained major defects that should have entirely prevented default judgment, such as a complaint that violated the statute of limitations or a case missing essential documentary evidence. 10% of the cases contained inconsistencies that should have resulted in a reduction in the amount of judgment requested by the plaintiff. Moreover, 32% contained errors that should have resulted in amended petitions prior to judgment. These estimates indicate there may be over a thousand erroneous debt collection default judgments per year in Los Angeles alone.

Courts across the United States need support in efficiently identifying unsatisfied statutory requirements in default judgment review. Although hiring more human reviewers would help, the number of collections cases is rising

¹<https://www.lacourt.ca.gov/pages/lp/court-communications/tp/about-the-court>

²Estimated from the number of cases requesting a default judgment per week at SCLAC and the number of dedicated attorney-hours per week.

across California (Johnson Raba 2023). Also, the number of collections filings may continue to rise as debt buyers leverage artificial intelligence (AI) to file at higher rates (Shah and Levy 2026). Courts should consider assistive AI to support expert human review. We believe that an assistive tool has the potential to significantly reduce the rate of unjust case outcomes and the life-altering consequences that follow.

Furthermore, courts are currently too capacity-constrained to reexamine their role in ensuring accurate outcomes. An assistive tool can help give courts the necessary capacity to rethink the policies they implement and enforce. Indeed, several judges at SCLAC are even now considering interpreting hearsay requirements more restrictively, despite higher costs in manual review.

Through an interdisciplinary collaboration with legal academics, SCLAC judges, SCLAC research attorneys, and computer scientists, we propose a “Default Assistant” that provides requirement-level recommendations to court staff and flags cases that are more likely than others to contain errors. The Default Assistant is built using large language models (LLMs) to efficiently parse pages of text and tables to identify evidence satisfying or failing statutory requirements. Importantly, we design the assistant to ground its recommendations in the case filings and cite the relevant quotes and tables in each accompanying explanation. The Default Assistant will support courts in upholding statutory requirements by providing recommendations, surfacing relevant evidence through precise citations, and leaving final decisions up to staff attorneys and judges.

In this work, we evaluate an initial version of the Default Assistant with law students on default judgment requests filed by debt-buyer plaintiffs. We find experimental evidence that the Default Assistant can improve case review: assisted users were 6.0% more accurate and 25.9% faster than unassisted users. Towards the end of supporting courts in issuing accurate and timely default judgments, our work makes the following contributions:

1. **Task Definition and Dataset Generation:** Our team of legal academics collaborated with SCLAC experts to define detailed case annotation guidelines for relevant statutory and procedural requirements. We work with research assistants to annotate defects of 188 California debt collection cases granted default judgment.
2. **Default Assistant Implementation:** We design and develop an LLM-based Default Assistant in accordance with the previously identified California statutory and procedural requirements, working closely with court stakeholders.
3. **Human Evaluation:** We compare the performance of humans using Default Assistant recommendations to that of unassisted humans. We find that the assisted humans are faster and more accurate at debt collection default judgment review.
4. **Fairness Evaluation:** Using augmented defendant names, we find no meaningful differences in standalone Default Assistant accuracy across defendants’ race and gender.

Related Work

Designing for Human-AI Collaboration. Given the necessity of human oversight in the judicial setting, a critical feature of the Default Assistant is to equip a human user to verify recommendations. Prior works propose different forms of explainable AI to enhance human-AI teamwork. Local explainability methods, such as LIME (Ribeiro, Singh, and Guestrin 2016), provide insight into which parts of an input were influential in the AI output. LLMs can provide natural language explanations of their answers.

However, local and free-form explanations have been shown to increase over-reliance on incorrect model recommendations (Bansal et al. 2021; Jacobs et al. 2021; Kim et al. 2025). Furthermore, Turpin et al. (2023) show that LLM-generated explanations are not necessarily faithful to the model’s underlying reasoning. A promising line of work shows that explanations that enable users to verify model recommendations with relevant evidence, rather than those that hint at the reasoning process, can reduce over-reliance and improve user performance (Fok and Weld 2024; Kim et al. 2025). Accordingly, we engineer a Default Assistant that provides recommendations with citations to case files to equip users in efficiently verifying the assistant and catching mistakes.

Methods for Cited Generation. Methods for generating cited outputs differ primarily in when citations are produced relative to the generated text. One line of work generates text and citations jointly, training or prompting models to produce inline citations alongside their answers (Nakano et al. 2021; Menick et al. 2022; Shao et al. 2024; Jiang et al. 2024; Zhang et al. 2025). Another generates first and attributes after the fact, revising the output with sources retrieved post hoc (Gao et al. 2022; Li et al. 2024).

Prior work provides empirical support for a third design choice: *attribute-first-then-generate*, in which supporting evidence is identified before generation rather than alongside or after it. Slobodkin et al. (2024) demonstrate that attributing evidence prior to generation yields fewer unsupported claims than the simultaneous generation of answers and citations. Worledge, Hashimoto, and Guestrin (2024) further show that generating model outputs from pre-identified source quotes improves citation precision and coverage over deployed systems, including post hoc citation. We follow this three-step process of identifying, verifying, and finally generating from source quotes for precise and comprehensive citations, enabling transparent tracing of each model generation to specific lines of the underlying case materials.

Evaluating Legal Human-AI Collaboration.

Because the success of human-AI teams is influenced by factors beyond AI system performance (Vasconcelos et al. 2023), we run a randomized controlled trial to evaluate the performance of humans using the Default Assistant for debt collection default judgment review. Prior work has studied AI-assisted users on other legal tasks. Choi, Monahan, and Schwarcz (2024) run a controlled study with 59 law students on legal drafting assignments, observing large efficiency gains from GPT-4 access with insignificant or small changes in work quality. Nielsen et al. (2024) compare AI

summaries and AI highlighting of legal complaints in a controlled experiment with 206 law students, finding that highlighting reduces task time by 30% without a measured quality drop, while summaries alone produce neither efficiency nor quality improvements. Tan et al. (2025) test a fine-tuned LLM for dataset license compliance with four software IP lawyers, finding substantial time savings without quantifying changes in accuracy. Similar to Nielsen et al. (2024), we consider AI outputs with explainability features (i.e., highlights and citations) and like Tan et al. (2025), we evaluate a custom AI system. Our work may be the first study of human-AI collaboration on a legal task to show significant and meaningful increases in both efficiency and accuracy.

Socio-Technical Discussion

Especially since the advent of LLMs, automated tooling has been considered and deployed for a litany of legal tasks, including legal research, case review, and contract analysis (Siino et al. 2025; Pasquale 2019).³ In response to the possibility of assistive AI in law and other high-stakes settings, prior literature has characterized the risks of such deployments.

In this section, we discuss how the design of our Default Assistant and controlled experiment are informed by our multidisciplinary team and SCLAC stakeholders in light of these risks. Our contributors include a law professor of legal ethics, a legal expert in debt collection, AI researchers, and SCLAC judges, leadership, and court research attorneys. We have identified three areas of risk: automation bias, historical bias, and technological lock-in.

A primary challenge of assistive AI is automation bias: the tendency for users to blindly rely on AI recommendations (Ruscheimer and Hondrich 2024), especially due to opaque model reasoning that hinders fact-finding and verifiability (Kuźniacki et al. 2022; Koencke et al. 2025). Automation bias is most damaging when it leads users to accept incorrect AI recommendations on tasks they would have completed correctly on their own. To mitigate this risk, we design the Default Assistant to cite each recommendation to specific lines in the case materials, aligning with broader legal norms that use citations to indicate source provenance. We also center our evaluation around a controlled user study, rather than a standalone system evaluation, and conduct an analysis on over-reliance and appropriate reliance.

Second, prior works highlight the risk of allocational disparities arising from deployed AI that amplify historical bias, along identity-based attributes (Ajunwa 2019; Angwin et al. 2022). In the development of our Default Assistant, it is possible that biases are learned from the outcomes in our debt collection default judgment dataset or from the training data of the underlying LLM itself. To mitigate the amplification of these biases, we annotate debt collection court cases with new ground-truth labels, rather than using the historical SCLAC decisions, which may be unduly constrained by time and resource limitations. To detect signs of bias in our

end-to-end system, we also conduct an evaluation of the stability of Default Assistant recommendations across different defendant names perceived to be of different race and gender identities.

Lastly, Crotoft (2019) explains that reliance on software with high technical overhead can add friction to legal evolution in courts. While this work takes substantive steps to address automation bias and historical bias, technological lock-in remains an open problem. We further discuss this concern in our limitations section. While the technical framing of our work is no panacea for these multifaceted challenges—which also require broader AI governance and institutional policies—these concerns have meaningfully shaped the design and evaluation of our system.

Data

We collected an initial sample of 200 debt collection cases filed in 2023-2024 by debt buyer plaintiffs. Specifically, we sampled 100 cases that requested default judgments and 100 cases that were granted default judgments. We removed two cases filed by original creditor plaintiffs (mistakenly labeled as debt-buyer cases in the court's filing system), two cases that were rejected prior to default judgment review due to administrative errors, and nine cases with fewer than three independent annotations from unique research assistants (methodology described below). Our final dataset is comprised of 96 cases that were sampled to have a request for default judgment and 92 cases that were granted default judgment. We note that the 92 cases sampled to have a request were all ultimately granted default judgment, resulting in a total of 188 cases that were all granted default judgment. We report historical court accuracy on these 188 cases. From this sample, we used 20 random cases for the development of the Default Assistant and held out 168 cases for a robust evaluation set.

All of the debt collection case files are public records and were accessed via the electronic filing and case management systems at SCLAC. The PDF files for each case include a complaint, request for default judgment, and often one or more declarations containing business record evidence. On average, there are about 55 pages of PDF content per case. These files form the evidentiary basis for default judgment review, yet vary widely in format—some are text-based, while others are scanned images.

Our research team, led by a practicing lawyer and supervised by a law professor, collaborated with court judges and research attorneys to develop annotation guidelines grounded directly in procedural and statutory requirements (CA Civil Code §§1788.58–.60). The requirements and instructions for conducting the review of each requirement on an individual case were subject to multiple rounds of revision between our team and our court collaborators. Once finalized, the lawyer on our team trained law student research assistants to follow these guidelines to annotate sub-requirements, which we combined into higher-level *requirements*, using the logical AND operator. For each annotation, assistants marked sub-requirements as “Satisfied”, “Not Satisfied”, or “Unclear”. Only 11 sub-requirement annotations

³The development of legal AI for law firms is moving a mile-a-minute; Harvey and Legora are two prominent companies in this space.

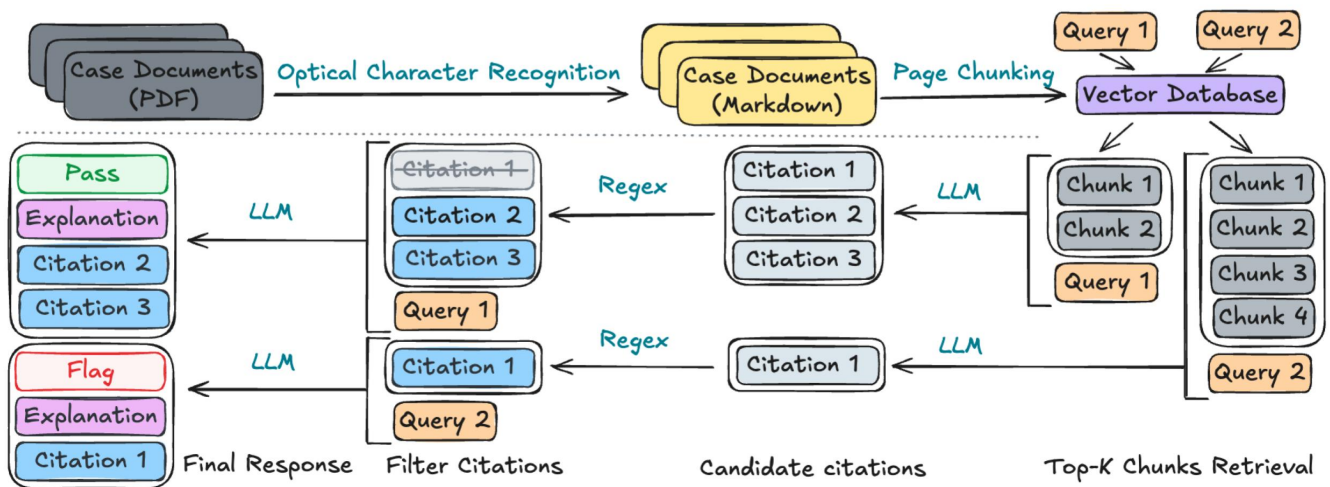


Figure 1: The Default Assistant processes PDF case documents to provide cited recommendations for case requirements.

out of 8,400 were marked “Unclear”.⁴ On average, the research assistants spent 14.1 minutes applying the guidelines to a case.

Requirement	Agreement	Cohen’s κ
Debtor Address	0.95	0.77
Chain of Title	0.97	0.65
Request-Complaint Cons.	0.90	0.61
Last Payment Date & SoL	0.88	0.60
Charge-Off Balance	0.92	0.50
Agreement to Debt	1.00	-
Sworn Declaration	1.00	-

Table 1: Inter-annotator agreement and Cohen’s κ for the annotations, prior to two-thirds majority aggregation into “gold labels”. The measures of Cohen’s κ indicate substantial agreement (0.61–0.80) in three requirements and moderate (0.41–0.60) in two ($n=188$). Annotators only classified cases as valid for the *Agreement to Debt* and *Sworn Declaration* requirements. In this table, “Unclear” annotations are counted as “Not Satisfied”.

The Cohen’s kappa coefficients (Cohen 1960) in Table 1 show that our annotation guidelines and procedure achieve substantial to moderate inter-annotator agreement across all requirements (Hall and Wright 2008). These agreement rates could be improved, so we strengthened our labels by collecting a third round of independent annotations and take the two-thirds majority label for each requirement over the three independent annotations from different research assis-

⁴When evaluating the Default Assistant, we convert “Unclear” annotations to “Not Satisfied” because the assistant should flag any case that requires further scrutiny. However, we do not count “Unclear” annotations toward major or minor error categories in our reported historical case outcomes; we instead provide lower bounds on the number of errors in the sample.

stants for each case. We refer to these strengthened annotations as the “gold labels”. These gold labels are used in the evaluation of both the historical court performance and the Default Assistant.

Statutory and Procedural Requirements

Working closely with court stakeholders, our legal experts identified the list of requirements that the court reviews for default judgment requests in debt-buyer plaintiff collections cases, as of September 2025. The gold coding methodology and the Default Assistant both evaluate each requirement listed below through a decomposition of sub-requirements that check for alleged facts and substantiating evidence.⁵ The seven legal requirements we study are as follows:

Request-Complaint Consistency: The Request for Default Judgment must not exceed the complaint’s prayer in damages or interest and may only request attorney fees or costs of suit if those were included in the complaint’s prayer.

Agreement to Debt: The complaint must include a copy of a contract or a monthly credit statement recording a purchase transaction, last payment, or balance transfer to show the defendant’s agreement to the debt (§ 1788.60 (b)).

Sworn Declaration: The case files must include at least one declaration signed under penalty of perjury in support of default judgment, signed by an individual with personal knowledge of the relevant business records (§ 1788.60 (a)).

Charge-Off Balance: The complaint must allege the charge-off balance (§ 1788.58 (a)(4)) and evidence in a sworn declaration must substantiate the alleged amount (§ 1788.60 (a)).

Last Payment Date & Statute of Limitations (SoL): The complaint must allege the date of last payment (§ 1788.58 (a)(5)), evidence in a sworn declaration must substantiate the

⁵Requirement-level recommendations from the assistant are obtained by the same procedure used to obtain requirement-level gold labels: if a case fails any sub-requirement, it fails the parent requirement.

alleged date (§ 1788.60 (a)), and the complaint filing date must fall within the four-year statute of limitations from a substantiated date of last payment.

Debtor Address: The complaint must allege the name and last-known address of the debtor from the charge-off creditor (§ 1788.58 (a)(7)) and evidence in a sworn declaration must substantiate the alleged address (§ 1788.60 (a)).

Chain of Title: The complaint must allege the names and addresses of all post-charge-off purchasers of the debt (§ 1788.58 (a)(8)) and evidence in a sworn declaration must substantiate the alleged chain of ownership (§ 1788.60 (a)).

We further categorize these requirements into major and minor error designations. *Major - Reject* errors are severe case defects that would preclude default judgment (i.e., time-barred cases, failure to include a valid declaration, and missing substantiation of the chain of title). *Major - Reduce* errors are inconsistencies—between the amounts requested in the complaint and default judgment—that would reduce the amount of the judgment. Finally, *Minor - Amend* errors are all of the remaining requirements—technical errors that would likely be remedied by an amended complaint.

Default Assistant Implementation

We built an AI “Default Assistant” to flag case defects and provide recommendations for compliance with CA Civil Code §§1788.58–.60 and other procedural requirements. If the Default Assistant finds inadequate evidence for a requirement, the assistant flags the requirement for court staff, who can verify the recommendation and make the ultimate decision. To facilitate verification, the Default Assistant includes citations to case materials for each recommendation, such as short quotes or tables that are highlighted within the original PDF (Figure 2).

The Default Assistant processes each case file through a structured pipeline that converts PDFs into Markdown using optical character recognition (OCR), and chunks and stores the files by page in a vector database. For each sub-requirement, the assistant performs retrieval-augmented generation (RAG) with additional citation steps to generate a recommendation (Figure 1). The model grounds each recommendation in verified evidence from the original filings and provides an explanation citing specific quotes and tables. We define each round of retrieval and cited generation for a sub-requirement as a node in LangGraph⁶—a graph that asynchronously executes nodes in parallel and in order of dependency, permitting nodes to use information identified in any parent nodes. For each node, we wrote a retrieval query and a generation query. We iteratively improved these queries by evaluating performance on the development set. We provide additional implementation details in the following subsections.

Data Processing and Semantic Retrieval

After empirical testing and manual inspection of ten representative cases, we opted for Azure AI Document Intelligence as our OCR engine given multiple open and closed

source options. The Azure service preserves Markdown layout cues such as tables, headers, and key-value pairs, enabling the preservation of structured elements in the retrieval and LLM cited generation pipeline.

Following OCR, we chunk the Markdown text along the original page boundaries from the PDFs. Chunking along page boundaries ensures that tables, footers, and section headers remain coherent within the same chunk. Each page-level chunk is then embedded using OpenAI’s `text-embedding-3-small` and stored in a vector database for semantic retrieval. Top-*k* retrieval was set to return either 4 or 10 pages, depending on the requirement. Failures to retrieve relevant chunks were rare on the development set.

Cited Generation

To ensure that all Default Assistant recommendations are grounded in verifiable evidence, the assistant employs an *attribute-first-then-generate* citation framework (Slobodkin et al. 2024). For each step of case review, the Default Assistant first retrieves relevant documents from the vector database. Then, the LLM (`gpt-4.1`) is few-shot prompted to extract quotes and tables from the retrieved documents that are necessary to evaluate the current legal requirement. These extracted spans are then relocated in the original document; quotes are identified using fuzzy, white-space-agnostic matching, while tables are located using regex on the values. Each candidate citation text is replaced with its directly copied counterpart from the original case files. Only successfully located quotes and tables are retained and passed to the LLM with a few-shot generation query to compose the final response. The final response includes a binary recommendation that the legal requirement is “Satisfied” or “Not Satisfied” and a free-form explanation with citations to the verified source quotes.

Human Study Methodology

We recruited 66 participants to review debt collection default judgment requests. All participants were law students. There is reason to believe that law students are a reasonable study population; law students commonly graduate into clerkships and SCLAC is considering delegating debt collection default judgment review to clerks. Prior works studying AI assistance in legal workflows similarly work with law student participants for tasks relevant to law clerks and junior associates (Nielsen et al. 2024; Choi, Monahan, and Schwarcz 2024).

The study was four hours long and included training and case review to simulate court staff workflows. We co-designed an in-person, 80-minute training reflecting standard procedure with an SCLAC research attorney who specializes in default judgment review. The training covered relevant procedural and statutory requirements, illustrated by real case-file examples. The training also included practice reviews of acceptable and unacceptable debt-buyer cases, audience Q&A, and printed reference materials for use during the study. Participants were randomly assigned to two conditions:

⁶www.langchain.com/langgraph

5. Given a date of last payment substantiated by exhibits in the **declaration** (CA Civil Code § 1788.58 (a)(5)), is the claim timely under the statute of limitations?

Sub-Requirement A: Does the plaintiff allege the date of the defendant's last payment in the **complaint** (CA Civil Code § 1788.58 (a)(5))?

AI Recommendation: **Satisfied** ✓

AI Explanation: The plaintiff alleges the date of the debtor's last payment was 12/04/2022 [1].

Cited Quotes and Sources:

[1] "The date of last payment on the credit account was on December 4, 2022."

[1] Complaint - Page 2

Does the case satisfy this sub-requirement?

☐ Satisfied
☐ Not Satisfied

Figure 2: For each recommendation, the Default Assistant provides a binary recommendation and a free-form explanation cited to verified quotes from the case materials. Users may view the quotes highlighted in the original PDF in the citation pop-up.

1. **Human:** No access to the Default Assistant

2. **Team:** Access to the Default Assistant

After training, participants were directed to separate rooms by condition and given branch-specific instructions and websites, with participants blinded to other conditions. Both the human and team websites displayed the same requirements, case information, links to case files, and a 10-minute countdown timer. Branches differed only in whether Default Assistant recommendations and citations were shown on the website.

During the simulation, participants reviewed 12 randomly assigned cases, deciding whether each requirement was “Satisfied” or “Not Satisfied.” Responses were saved at a 10-minute per-case limit, with any incomplete requirements automatically marked as satisfied, though participants could submit earlier. While court staff typically spend about four minutes per case, we allowed for more time to account for participant inexperience, setting the limit at 10 minutes based on pilot study averages.

The websites collected sub-requirement decisions, time spent per requirement, and self-reported user confidence scores on a 0-100 scale for each requirement. Some of the 168 cases in the held-out test set were randomly selected to receive decisions from multiple participants within a branch, resulting in 204 evaluations per branch. Participants were compensated with \$100 gift cards and free dinners. We consulted with the IRB prior to recruiting participants; the board confirmed that formal review and approval were not required.

As with the gold labels and AI recommendations, we aggregate the sub-requirement decisions into requirement-level decisions. We refer to decisions from the human and team participants as *accepting* or *rejecting* and decisions from the Default Assistant as *flagging* or *not flagging* a case or requirement.

Human Study Results

The Default Assistant (DA) helped users improve their average requirement accuracy per case by 6.0%, while decreasing the average time spent reviewing each requirement per case by over 25% (Table 2). Differences in average requirement accuracy and review time per case between humans with and without the Default Assistant were found to be statistically significant ($p < 0.002$). Any increase in user confidence for those using the Default Assistant appears modest and tracks the increase in accuracy, however, the difference is not statistically significant.

Metric	Absolute Gain	Relative Gain	p-value
Accuracy	5.3 pp	6.0%	1.0e-04*
Timing	-12.8s	-25.9%	2.5e-10*
Confidence	2.2	3.0%	1.7e-01

Table 2: Users aided by the Default Assistant were both more accurate and faster than unaided users. This analysis considers the average paired accuracy, time, and user confidence (out of 100) over all requirements for a case ($n=204$). Starred values (*) indicate statistically significant non-zero differences after Bonferroni correction ($p < 0.002$).

Table 3 shows the rates of error reduction and time savings by requirement in the team setting, relative to the human baseline. The Default Assistant helped users achieve 62% fewer errors in judging the *Charge-off Balance* requirement, 47% fewer for *Last Payment Date & SoL*, and 48% fewer for *Debtor Address*, with differences significant ($p < 0.05$). The Default Assistant also expedites review time by 24% to 34%, with the differences in time between the human baseline and the assisted human found to be significant ($p < 0.05$) for all requirements but *Agreement to Debt* and *Chain of Title*.

The Default Assistant enabled participants to achieve error reduction through increases in precision and, to a lesser

Requirement	Human Acc.	Team Acc.	Rel. Error Reduction	Rel. Time Savings
Last Payment Date & SoL	0.78	0.88	47%**	34%**
Charge-Off Balance	0.86	0.95	62%**	30%**
Debtor Address	0.85	0.92	48%**	31%**
Chain of Title	0.88	0.93	44%	18%
Agreement to Debt	0.95	0.98	60%	7%
Sworn Declaration	0.97	0.99	57%	30%**
Request-Complaint Consistency	0.90	0.91	5%	24%**

Table 3: Paired human and citation-assisted team performances by requirement, ordered by absolute accuracy gain. We report accuracy, relative error reduction, and relative time savings, taken with respect to the human baseline (n=204). Starred comparisons (**) indicate statistically significant non-zero differences ($p < 0.05$) after Benjamini-Hochberg correction. Calculated at full precision; rounded for display.)

extent, recall in rejecting defective cases (Table 4). The small number of defects limits interpretation; for instance, the apparent changes in recall for *Last Payment Date & SoL*, *Debtor Address*, and *Request-Complaint Consistency* correspond to differences over only one or two cases. However, we observe that human-AI teams generally increase precision over unassisted humans without evidence of reducing recall, and even appear to increase recall for *Request-Complaint Consistency*.

The Default Assistant improves accuracy across major and minor error categories between the unassisted human and team branches (Table 5). *Major - Reject* accuracy increased due to improvement in true rejections (+9 pp), i.e., the rejection of invalid cases, and improvement in true acceptances (+9 pp), i.e., the acceptance of valid cases. Gains in true rejections (+33 pp) drove improvements in *Major - Reduce* accuracy, while gains in true acceptances (+13 pp) drove improvements in *Minor - Amend* accuracy.

Complementary Performance

The team setting generally achieves stronger performance than the unassisted humans and Default Assistant individually (Table 6). We observe that unassisted human performance falls behind the Default Assistant on every requirement; we present an amplification multiplier that indicates how much of the DA-human performance gap is actually realized—or counteracted—when humans collaborate with the Default Assistant. This multiplier indicates that the team setting compensates for and even exceeds the DA-Human performance gap for most requirements.

Reliance

We observe that users may rely on some incorrect AI recommendations, but do catch the majority of AI mistakes. Given the set of all incorrect AI recommendations for a requirement, we quantify over-reliance by examining the number of cases where assisted humans relied on the incorrect recommendations and unassisted humans otherwise reviewed the requirement correctly (Figure 3). Our analysis is limited by sample size constraints, but we observe that over-reliance affects few cases per requirement. However, its presence across requirements suggests it remains difficult to fully mitigate.

When it comes to users appropriately relying on correct AI recommendations, we find that assisted users largely, although not perfectly, leverage accurate recommendations. Out of all correct AI recommendations for a requirement, we measure appropriate reliance as the number of cases where assisted humans agree with the correct recommendation and unassisted humans otherwise reviewed the requirement incorrectly (Figure 4). By generally adopting correct AI recommendations, assisted users achieve accuracy gains that unassisted users do not. Importantly, the accuracy gains from appropriate reliance counteract the smaller losses from over-reliance.

Fairness Evaluation

We examine differences in Default Assistant accuracy across the race and gender of defendants and find that they are minimal. Using regex, we find and replace original defendant names with common, racially distinct names selected by Yin, Alba, and Nicoletti (2024) across all pages of files for a case. We focus on defendants because, with rare exceptions, the debt buyer plaintiffs tend to be corporations and not individuals. The augmented defendant names fall into eight categories across perceived race (Asian, Black, Hispanic, and White) and gender (Female and Male). We evaluate each of 155 cases eight times with names randomly selected from the eight categories.

Table 7 shows that observed paired differences are minimal and do not exceed 0.0076. Moreover, per-case paired differences in average requirement accuracy across demographics are bounded within $[-0.02, 0.02]$ with $p < 0.05$ under two one-sided tests. The fact-oriented nature of the requirements evaluated by the Default Assistant—especially given the grounding of recommendations in cited quotes and tables—likely reduces room for bias, in contrast to settings that solicit subjective, open-ended opinions from LLMs. Nonetheless, exploring whether attributes beyond names—such as addresses and transactional patterns—introduce bias into Default Assistant recommendations remains important future work.

Discussion

Our human study demonstrates that the Default Assistant can help users review requirements more accurately and ef-

Requirement	Num. Defects	Precision			Recall		
		Human	Team	Gain	Human	Team	Gain
Last Payment Date & SoL	32	0.41	0.57	0.17	0.91	0.97	0.06
Debtor Address	19	0.34	0.57	0.23	0.68	0.63	-0.05
Request-Complaint Consistency	18	0.46	0.49	0.03	0.61	0.94	0.33
Charge-Off Balance	14	0.30	0.57	0.27	0.79	0.86	0.07
Chain of Title	6	0.12	0.21	0.09	0.50	0.50	0.00
Agreement to Debt	0	0.00	0.00	0.00	0.00	0.00	0.00
Sworn Declaration	0	0.00	0.00	0.00	0.00	0.00	0.00

Table 4: Precision and recall for identifying invalid cases by requirement for human and team decisions, ordered by number of defects (n=204). Gains are the team metrics minus human metrics. AI assistance increases or maintains precision for all requirements and increases or maintains recall of 6 out of 7 requirements. However, the low number of defects limits the interpretation of these measurements. Calculated at full precision; rounded for display.

Requirement	Num. Defects	Accuracy		True Rejection			True Acceptance		
		Human	Team	Human	Team	Rel. Change	Human	Team	Rel. Change
Major - Reject	11	0.81	0.90	0.73	0.82	13%	0.82	0.91	11%
Major - Reduce	18	0.90	0.91	0.61	0.94	55%	0.93	0.91	-2%
Minor - Amend	44	0.73	0.83	0.86	0.89	3%	0.69	0.82	19%

Table 5: The accuracy, rate at which invalid cases are rejected (True Rejection), rate at which valid cases are accepted (True Acceptance), and relative changes in true rejection/acceptance with AI assistance, taken with respect to the human baselines (n=204). *Major - Reject*: case defects unlikely to be remedied by an amended complaint. *Major - Reduce*: case defects that won't preclude judgment, but would reduce the amount of the judgment. *Minor - Amend*: case defects that can be amended by the plaintiff without additional evidence. Calculated at full precision; rounded for display.

ficiently. While the Default Assistant reduces the time taken by participants to review requirements across the board, the strongest boosts in accuracy are for the *Last Payment Date & SoL*, *Charge-Off Balance*, and *Debtor Address* requirements where the Default Assistant baseline strongly out-performs the human baseline (Table 6).

The unassisted participants in our study have a high true rejection rate (Table 5), compared to the court. The court has historically rejected cases very sparingly. Out of the random sample of 92 cases that requested default judgment, the court did not reject any cases for the requirements reviewed in our audit, despite our finding 6 cases with *Major - Reject* defects. It is possible that the inclination of our unassisted participants to reject cases limited observable improvement in true rejection rate that may otherwise occur with the Default Assistant in the court setting. Nonetheless, we observe that participants aided by the Default Assistant correctly rejected more instances of *Major - Reject* and *Major - Reduce* defects than unassisted participants. Notably, these gains came without substantially lowering true acceptance rates, suggesting the assistant reined in bias rather than inducing over-rejection. In fact, Default Assistant use increased both the true rejection and true acceptance rates for the *Major - Reject* category.

The *Last Payment Date & SoL*, *Debtor Address*, *Charge-Off Balance*, and *Sworn Declaration* requirements see the greatest improvements in timing (Figure 3). We note that these four requirements all require a user to comb through multiple pages of evidence for specific facts, unlike other

requirements which either require referencing a narrowly-scoped location in the case materials (*Agreement to Debt*) or substantial additional reasoning (*Chain of Title*). Future work seeking performance gains from AI assistants may benefit from considering tasks where humans can verify provided answers faster than providing an answer in the first place.

Our finding that the Default Assistant benefited users in tasks requiring extensive document review might also signal the potential for impact in other contexts. For example, court staff reviewing eviction default judgments must examine leases, rent ledgers, and other exhibits to verify alleged damages and rent owed—a time-consuming process. We hypothesize that an AI assistant that provides recommendations with precise citations might offer significant advantages to diligent court staff.

Interestingly, humans working with the Default Assistant consistently achieve higher accuracy gains than the assistant alone, compared to the human baseline (Table 6). While not shown to be a statistically significant effect, this observation suggests that the users do not rely entirely on the Default Assistant. Indeed, while over-reliance is present, users still catch most AI mistakes. Users and the assistant may make different errors, allowing each to correct the other. Understanding these complementary strengths can guide the design of systems where human-AI teams outperform their individual counterparts.

Requirement	T	H	DA	$T - H$	$DA - H$	Amplification
Last Payment Date & SoL	0.88	0.78	0.87	0.10	0.09	1.17x
Charge-Off Balance	0.95	0.86	0.94	0.09	0.08	1.06x
Debtor Address	0.92	0.85	0.89	0.07	0.04	1.88x
Chain of Title	0.93	0.88	0.92	0.05	0.04	1.38x
Agreement to Debt	0.98	0.95	0.96	0.03	0.01	3.00x
Sworn Declaration	0.99	0.97	0.98	0.02	0.01	2.00x
Request-Complaint Consistency	0.91	0.90	0.93	0.00	0.03	0.17x

Table 6: Comparison of team (T), unassisted human (H), and Default Assistant (DA) accuracy across requirements, with highest (tied) accuracy per row in bold ($n=204$). Absolute gains and amplification ($\frac{T-H}{|DA-H|}$) quantify how collaboration often exceeds the DA-human gap. Ordered by team gain over DA. Calculated at full precision; rounded for display.

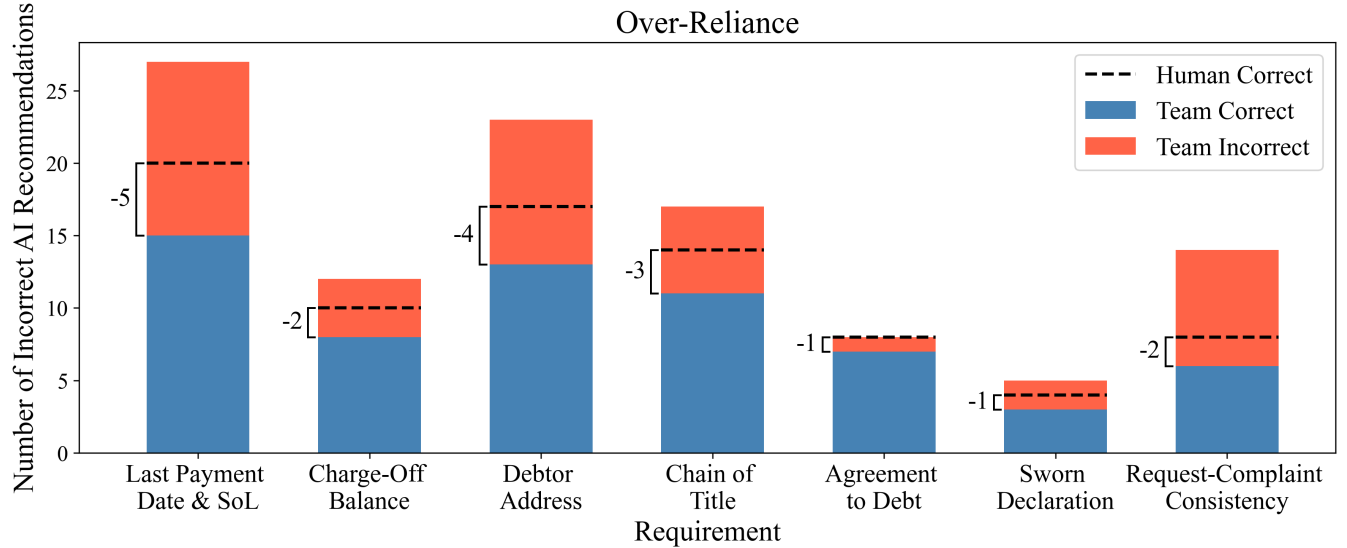


Figure 3: Comparison of the number of incorrect AI recommendations that the assisted humans corrected (blue) to the number of these cases that were correctly reviewed by unassisted humans (dashed line). While the differences are small, there is a general trend of users over-relying on incorrect AI recommendations.

Group 1	Group 2	Avg. Paired Diff.	p-value
Female	Male	7.6e-03	1.5e-03
Hispanic	Black	-9.2e-04	1.0e-02
Asian	White	1.4e-03	1.5e-02
White	Black	3.7e-03	2.4e-02
Hispanic	White	-4.6e-03	3.2e-02
Asian	Black	5.1e-03	3.9e-02
Asian	Hispanic	6.0e-03	4.7e-02

Table 7: Paired differences in average requirement accuracy, aggregated over all permutations of demographic pairs per case ($n=155$) and tested for equivalence using two one-sided tests (TOST). All comparisons are equivalent within $[-0.02, 0.02]$ with $p < 0.05$ under TOST, after Benjamini-Hochberg correction.

Limitations & Future Work

Although we aim to simulate court review as closely as possible, our methodology differs in three key ways. First, our

participants were law students who were new to the task of collections case review, despite their general familiarity with legal terminology and case review. In contrast, court staff often have months of experience handling hundreds of cases and receiving judicial feedback. Expertise on the task may impact how effectively users leverage the Default Assistant’s recommendations. Nonetheless, our law student study provides an empirical proof-of-concept that justifies the future investment of SCLAC in a more ecologically valid evaluation.

Second, our unassisted participants presented a strong baseline, rejecting more meritless cases than historical Court judgments. Their performance likely benefited from extra resources: the study solicited decisions through an explicit checklist of requirements—unlike standardized Court protocol—which may itself boost accuracy (Haynes et al. 2009). Also, participants were given 10 minutes per case, whereas real-world time constraints clock in at an average of four minutes per case. A shorter time limit may prevent unassisted users from thoroughly reviewing cases, potentially

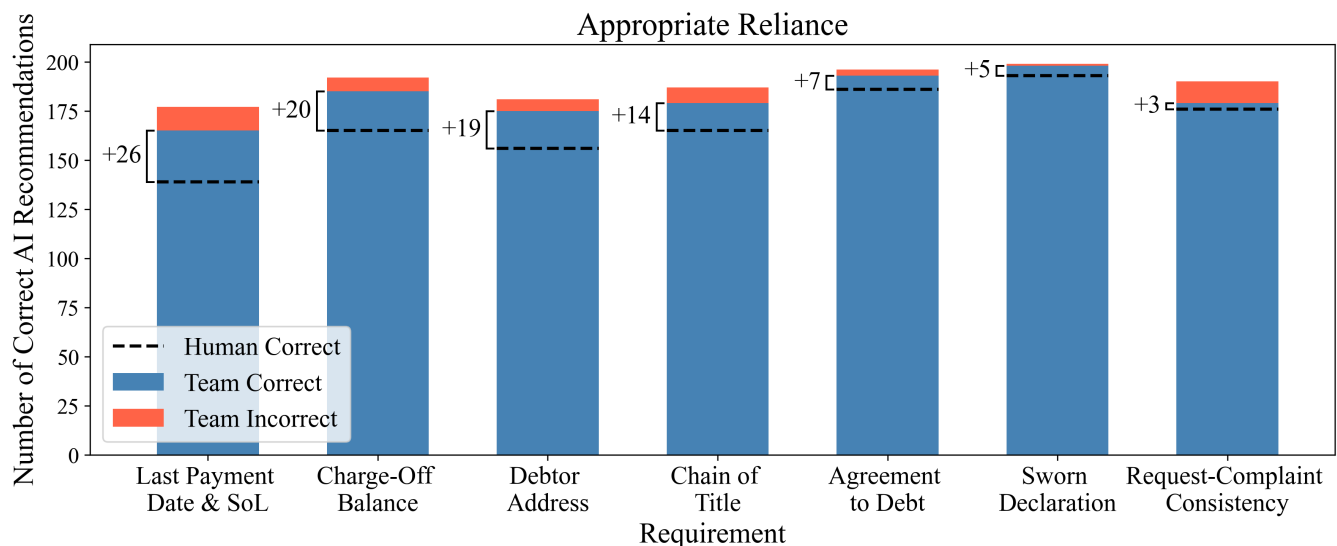


Figure 4: Comparison of the number of correct AI recommendations that the assisted humans agreed with (blue) to the number of these cases that were incorrectly reviewed by unassisted humans (dashed line). Assisted users generally align with correct AI recommendations and through doing so, achieve accuracy gains unrealized by their unassisted counterparts.

leading AI-assisted users to reject more case defects than users without AI. However, such a reduction in the allotted time per case may also reduce observed relative changes in time efficiency. On the extreme, if users with and without AI assistance run up against the time limit, then there would be no relative efficiency gain.

Third, participants used the Default Assistant over the course of a limited data collection period. Long-term usage of the Default Assistant over weeks and months may lead to failure modes such as over-reliance or under-reliance. User behaviors and downstream performance may also change as users calibrate their judgment of the strengths and weaknesses of the Default Assistant over longer time periods than a point-in-time study.

Another limitation of our work is that of technological lock-in, where software entrenches institutional practices and can become a barrier to change. We design a modular Default Assistant where statutory requirements may be added, edited, and removed with minimal dependency on other requirements. However, by virtue of encoding institutional knowledge into a software system, we are inevitably increasing the work and skill to evolve that knowledge. This limitation remains an open problem and will require institutional planning and resource allocation to comprehensively address it.

The Court is continuing to refine its internal review criteria, motivating further development on the Default Assistant to accommodate the new requirements and improve overall performance. As the next step in our staged evaluation, we are planning a year-long study with SCLAC to evaluate the long-term and downstream effects of incorporating the Default Assistant into debt collection default judgment review.

Conclusion

Our audit of 188 California debt collection cases with default judgments found that 32% contained defects requiring amendment, 10% contained inconsistencies requiring reduced judgments, and 4% contained major defects precluding judgment, highlighting opportunities to reduce erroneous judgments. We built a Default Assistant to support accurate and timely review by courthouse staff, complete with citations to quotes and tables within case files. In a controlled study simulating the court, we studied human-Default Assistant collaboration in terms of accuracy and efficiency. We found that providing reviewers with Default Assistant recommendations citing case materials increased average requirement accuracy by 6.0% and reduced review time by over 25% in our study setting. The assistant's recommendations showed similar accuracy across cases, regardless of defendant names signaling different racial or gender identities.

While the results from our simulated court setting may not precisely translate to real court settings, our findings establish a proof of concept: the Default Assistant can help reviewers identify case defects more accurately and efficiently. By providing AI recommendations grounded in citations to original case files, we equip reviewers to leverage AI while remaining informed decision-makers. This evidence licenses the next step beyond a simulation to a field study at SCLAC, where the assistant may ultimately help the court reduce unjust wage garnishments, suggest amendments, and more consistently sanction plaintiffs acting in bad faith. Judges are already debating debt collection review procedures, and tools like the Default Assistant may enable more thorough review, potentially shaping future protocols. More broadly, our findings point to AI's potential to assist with repetitive manual workloads in other courthouse settings.

Acknowledgments

We are very grateful to the Superior Court of Los Angeles County judges, research attorneys, staff, and leadership for their continued collaboration, feedback, and openness to innovation. We are deeply grateful to Parker Grove, Victor Wu, Brian Xu, and many others for their invaluable data annotation work. We also thank Margaret Hagan for her indispensable help during the study, our participants, and all members of the Rhode Center for their support. Thank you to the participants of JURIX AIDA2J 2025, attendees of ACM CS & Law 2026, Guestrin and Hashimoto lab members, Myra Cheng, Vishakh Padmakumar, and the SLS law faculty for impactful feedback.

CG was supported by a grant from Stanford HAI. TH was supported by a grant from Stanford HAI, DSO labs, gifts from Open Philanthropy, Amazon, Schmidt Sciences, and the Tianqiao and Chrissy Chen Foundation and a grant under the NSF CAREER IIS-2338866. This work does not necessarily reflect the position or policy of the government and no official endorsement should be inferred.

References

- Agor, P. H.; Graves, S. E.; and Miller, S. 2015. The Landscape of Civil Litigation in State Courts. *Available at SSRN 2700745*.
- Ajunwa, I. 2019. The paradox of automation as anti-bias intervention. *Cardozo L. Rev.*, 41: 1671.
- Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2022. Machine bias. In *Ethics of data and analytics*, 254–264. Auerbach Publications.
- Bansal, G.; Wu, T.; Zhou, J.; Fok, R.; Nushi, B.; Kamar, E.; Ribeiro, M. T.; and Weld, D. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–16.
- Bookman, P. K. 2024. Default Procedures. *U. Pa. L. Rev.*, 173: 1419.
- Choi, J. H.; Monahan, A. B.; and Schwarcz, D. 2024. Lawyering in the age of artificial intelligence. *Minn. L. Rev.*, 109: 147.
- Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1): 37–46.
- Crootof, R. 2019. “Cyborg Justice” and the Risk of Technological–Legal Lock-In. *Columbia law review*, 119(7): 233–251.
- Desmond, M.; and Kimbro, R. T. 2015. Eviction’s fallout: housing, hardship, and health. *Social forces*, 94(1): 295–324.
- Fok, R.; and Weld, D. S. 2024. In search of verifiability: Explanations rarely enable complementary performance in AI-advised decision making. *AI Magazine*, 45(3): 317–332.
- Gao, L.; Dai, Z.; Pasupat, P.; Chen, A.; Chaganty, A. T.; Fan, Y.; Zhao, V. Y.; Lao, N.; Lee, H.; Juan, D.-C.; et al. 2022. Rarr: Researching and revising what language models say, using language models. *arXiv preprint arXiv:2210.08726*.
- Garnham, J. P.; Gershenson, C.; and Desmond, M. 2022. New data release shows that 3.6 million eviction cases were filed in the United States in 2018. *Eviction Lab*, 11.
- Hall, M. A.; and Wright, R. F. 2008. Systematic content analysis of judicial opinions. *Calif. L. Rev.*, 96: 63.
- Haynes, A. B.; Weiser, T. G.; Berry, W. R.; Lipsitz, S. R.; Breizat, A.-H. S.; Dellinger, E. P.; Herbosa, T.; Joseph, S.; Kibatala, P. L.; Lapitan, M. C. M.; et al. 2009. A surgical safety checklist to reduce morbidity and mortality in a global population. *New England journal of medicine*, 360(5): 491–499.
- Jacobs, M.; Pradier, M. F.; McCoy Jr, T. H.; Perlis, R. H.; Doshi-Velez, F.; and Gajos, K. Z. 2021. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational psychiatry*, 11(1): 108.
- Jiang, Y.; Shao, Y.; Ma, D.; Semnani, S.; and Lam, M. 2024. Into the Unknown Unknowns: Engaged Human Learning through Participation in Language Model Agent Conversations. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 9917–9955. Miami, Florida, USA: Association for Computational Linguistics.
- Jiménez, D. 2015. Dirty Debts Sold Dirt Cheap. *Harv. J. on Legis.*, 52: 41.
- Johnson Raba, C. 2023. One-Sided Litigation: Lessons from Civil Docket Data in California Debt Collection Lawsuits. *University of Illinois Chicago School of Law July*.
- Kim, S. S. Y.; Vaughan, J. W.; Liao, Q. V.; Lombrozo, T.; and Russakovsky, O. 2025. Fostering Appropriate Reliance on Large Language Models: The Role of Explanations, Sources, and Inconsistencies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI ’25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713941.
- Koenecke, A.; Stiglitz, J.; Mimno, D.; and Wilkens, M. 2025. Tasks and Roles in Legal AI: Data Curation, Annotation, and Verification. *arXiv preprint arXiv:2504.01349*.
- Kuźniacki, B.; Almada, M.; Tyliński, K.; Górski, Ł.; Wino-gradska, B.; and Zeldenrust, R. 2022. Towards eXplainable Artificial Intelligence (XAI) in tax law: the need for a minimum legal standard. *World tax journal*, 14: 573–616.
- Li, W.; Li, J.; Ma, W.; and Liu, Y. 2024. Citation-Enhanced Generation for LLM-based Chatbots. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1451–1466. Bangkok, Thailand: Association for Computational Linguistics.
- Menick, J.; Trebacz, M.; Mikulik, V.; Aslanides, J.; Song, F.; Chadwick, M.; Glaese, M.; Young, S.; Campbell-Gillingham, L.; Irving, G.; et al. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.
- Mullen, F. 2019. Fifty years after the Consumer Credit Protection Act: The high price of wage garnishments. *Mitchell Hamline L. Rev.*, 45: 191.

- Nakano, R.; Hilton, J.; Balaji, S.; Wu, J.; Ouyang, L.; Kim, C.; Hesse, C.; Jain, S.; Kosaraju, V.; Saunders, W.; et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.
- Nielsen, A.; Skylaki, S.; Norkute, M.; and Stremitzer, A. 2024. Building a better lawyer: Experimental evidence that artificial intelligence can increase legal work efficiency. *Journal of Empirical Legal Studies*, 21(4): 979–1022.
- Pasquale, F. 2019. A rule of persons, not machines: the limits of legal automation. *Geo. Wash. L. Rev.*, 87: 1.
- Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.
- Ruscheimer, H.; and Hondrich, L. J. 2024. Automation bias in public administration—an interdisciplinary perspective from law and psychology. *Government Information Quarterly*, 41(3): 101953.
- Shah, A. V.; and Levy, J. Y. 2026. Access to Justice in the Age of AI: Evidence from US Federal Courts. Technical report, Working Paper, Massachusetts Institute of Technology and University of
- Shao, Y.; Jiang, Y.; Kanell, T.; Xu, P.; Khattab, O.; and Lam, M. 2024. Assisting in Writing Wikipedia-like Articles From Scratch with Large Language Models. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 6252–6278. Mexico City, Mexico: Association for Computational Linguistics.
- Siino, M.; Falco, M.; Croce, D.; and Rosso, P. 2025. Exploring llms applications in law: A literature review on current legal nlp approaches. *IEEE Access*.
- Slobodkin, A.; Chen, Y.; Zhang, Y.; Khashabi, D.; and Durrett, G. 2024. Attribute First, then Generate: Improving Faithfulness and Attribution in Retrieval-Augmented Generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*. Bangkok, Thailand: Association for Computational Linguistics.
- Tan, J.; Rajbahadur, G. K.; Li, Z.; Song, X.; Lin, J.; Li, D.; Zheng, Z.; and Hassan, A. E. 2025. LicenseGPT: A fine-tuned foundation model for publicly available dataset license compliance. In *Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering*, 39–50.
- Turpin, M.; Michael, J.; Perez, E.; and Bowman, S. 2023. Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36: 74952–74965.
- Vasconcelos, H.; Jörke, M.; Grunde-McLaughlin, M.; Gerstenberg, T.; Bernstein, M. S.; and Krishna, R. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1): 1–38.
- Worledge, T.; Hashimoto, T.; and Guestrin, C. 2024. The Extractive-Abstractive Spectrum: Uncovering Verifiability Trade-offs in LLM Generations. *arXiv:2411.17375*.
- Yin, L.; Alba, D.; and Nicoletti, L. 2024. OpenAI’s GPT is a Recruiter’s Dream Tool. Tests Show There’s Racial Bias. *Bloomberg*. Accessed: 2026-01-26.
- Zhang, J.; Bai, Y.; Lv, X.; Gu, W.; Liu, D.; Zou, M.; Cao, S.; Hou, L.; Dong, Y.; Feng, L.; and Li, J. 2025. LongCite: Enabling LLMs to Generate Fine-grained Citations in Long-Context QA. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 5098–5122. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.