AI Assistance for Court Review of Default Judgments

Theodora Worledge[1]*, Othman Bensouda Koraichi[2], Daniel Bernal[2], Aviv Caspi[2,3], Tatsunori Hashimoto[1], Carlos Guestrin[1], David Freeman Engstrom[2,4]

[1]Stanford Computer Science, Stanford University
[2]Deborah L. Rhode Center, Stanford Law School
[3]University of Chicago Law School
[4]Stanford Law School

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 100 debt collection cases granted default judgment by a California court, we find that 37 cases contained errors requiring amendment prior to judgment, four cases contained inconsistencies requiring reduced judgments, and four cases contained major defects that should have entirely prevented default judgment. To enhance the accuracy of court review, we built a Default Assistant that uses large language models with retrieval-augmented generation to flag case defects and provide recommendations for default judgment review. We equip human experts 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 are 5.6% more accurate and 25.9% faster in reviewing the average requirement than unaided reviewers, with both differences statistically significant ($p < 0.0003$). Statutory requirements demanding extensive document search realized the largest gains, with error reductions and time savings from AI assistance up to 53% and 33%, 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 while maintaining human oversight.

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 respond in less than 9% of debt collection cases filed in California courts (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. For instance, each year, the Los Angeles Superior Court routes as many as 30,000 debt collection default judgment requests to court staff for manual review. The high caseload creates severe time constraints. Court research attorneys have a few minutes per case to verify roughly a dozen statutory requirements. The failure of any requirement prohibits the entry of 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 a checklist of statutory and procedural requirements developed by legal and court professionals, we audited 100 debt-buyer plaintiff collections cases granted default judgment by a California Superior Court, hereafter referred to as the Court. We found that 37 cases contained errors that should have resulted in amended petitions prior to judgment. Four of the cases contained inconsistencies that should have resulted in a reduction in the amount of judgment requested by the plaintiff. Moreover, four cases 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. These estimates suggest there may be many hundreds of erroneous debt collection default judgments per year at this Court alone.

Courts across the United States need support in efficiently identifying unsatisfied statutory requirements in default judgment review. Although hiring additional court staff would help, the number of collections cases is rising across California (Johnson Raba 2023). Also, the number of collections filings will likely continue to rise as debt buyers leverage artificial intelligence (AI) to file at higher rates. Courts must themselves turn to automation 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

*Corresponding author: worledge@stanford.edu

outcomes. An assistive tool can help give courts the necessary capacity to rethink the policies they implement and enforce. Indeed, several judges at the Court are even now considering interpreting hearsay requirements more restrictively, despite higher costs in manual review.

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 5.6% more accurate and 25.9% faster than unassisted users. To support courts in efficiently issuing accurate default judgments, our work makes the following contributions:

1. **Dataset Generation:** We annotate 200 California debt collection cases requesting default judgment for case defects to produce a gold-labeled dataset for Default Assistant development and evaluation.
2. **Default Assistant Implementation:** We design and develop an LLM-based Default Assistant in accordance with California statutory and procedural requirements, working closely with court stakeholders.
3. **Human Evaluation:** We compare the performance of humans teamed up with the Default Assistant to unassisted humans and 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

Given the necessity of human oversight in high-stakes judicial settings, we prioritize designing the Default Assistant to fit the needs of a human user. Prior work proposes different forms of explainable AI to enhance human-AI teamwork. Local explainability methods, such as LIME (Ribeiro, Singh, and Guestrin 2016) and SHAP (Lundberg and Lee 2017) provide users with insight into which parts of an example were influential in its AI's prediction. However, such local explanations have been shown to increase over-reliance on incorrect model recommendations (Bansal et al. 2021). Prior works have hypothesized and demonstrated that explanations must enable users to verify model recommendations, rather than hint at the model's reasoning process to 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. Because the success of human-AI teams is influenced by task characteristics beyond AI system performance, such as task difficulty and user incentives (Vasconcelos et al. 2023), we evaluate the performance of human-Default Assistant teams in a simulated court setting.

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). In response to this proliferation of automated tooling, prior literature has raised concerns including automation bias, where users come to blindly over-rely on AI recommendations (Ruschemeier and Hondrich 2024), and opaque model reasoning that hinders fact-finding and verifiability (Kuźniacki et al. 2022; Koenecke et al. 2025). We tackle these concerns by citing AI recommendations to case materials and centering our evaluation around a human study, rather than a standalone system evaluation. Other works highlight the risk of allocational harms that amplify historical bias (Ajunwa 2019; Angwin et al. 2022). To mitigate the amplification of historical biases, we annotate debt collection court cases with new ground-truth labels, rather than using the Court's decisions, which may be unduly constrained by time and resource limitations. Crootof (2019) discusses the risk of technological lock-in, where courts are unable to adapt procedure due to technical overhead. We address this concern on the technical side by designing a modular Default Assistant where statutory requirements may be added, edited, and removed with minimal dependency on other requirements. While our approaches are no panacea for these multi-faceted challenges, these concerns have substantially shaped the design and evaluation of our proposed system.

Data

We collected a sample of 200 debt collection cases filed in 2023-2024 with debt buyer plaintiffs. We created this sample by randomly sampling 100 cases that were granted default judgments and another 100 cases that requested default judgments. All case files are pulled from the electronic filing and case management systems at the Court. 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.

Prior to annotation, our research team, led by a practicing lawyer and supervised by a law professor, collaborated with court experts to develop annotation guidelines grounded directly in procedural and statutory requirements (CA Civil Code §§1788.58–.60). Law student research assistants spent on average 14.0 minutes per case following 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

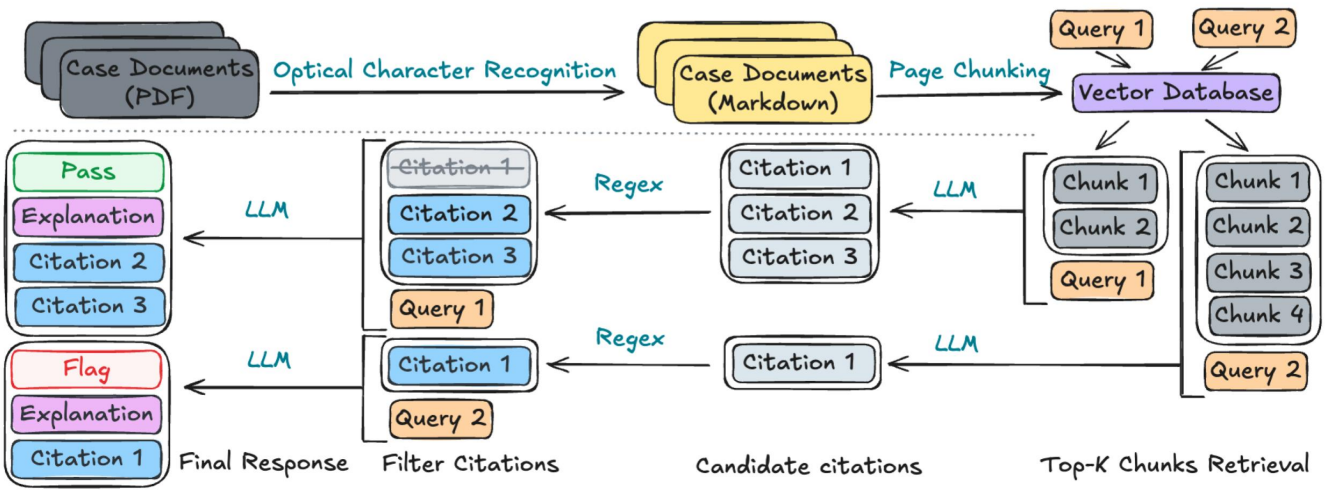

Figure 1: The Default Assistant processes case documents in PDF format to provide cited recommendations for case requirements.

"Satisfied", "Not Satisfied", or "Unclear". Only 23 sub-requirement annotations out of 8,400 were marked "Unclear". When 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 at the Court; we instead provide lower bounds on the number of errors in the sample.

Table 1: Inter-annotator agreement and Cohen's κ for the annotations, prior to two-thirds majority aggregation into "gold labels". The measures Cohen's κ indicate near perfect agreement (0.81-1) in one requirement, substantial agreement (0.61–0.80) in three and moderate (0.41–0.60) in two (n=200). Annotators only classified cases as valid for the sworn declaration requirement. In this table, "Unclear" annotations are counted as "Not Satisfied".

Requirement	Agreement	Cohen's κ
Debtor Address	0.96	0.82
Chain of Title	0.97	0.68
Agreement to Debt	0.99	0.66
Request-Complaint Cons.	0.91	0.64
Charge-Off Balance	0.92	0.57
Last Payment Date & SoL	0.85	0.56
Sworn Declaration	1.00	-

The Cohen's kappa coefficients (Cohen 1960) in Table 1 show that our annotation guidelines and procedure achieve near-perfect to moderate inter-annotator agreement across all requirements (Hall and Wright 2008). To further improve the annotations, we take the two-thirds majority label for each requirement over three independent annotations from different research assistants for each case. We refer to these strengthened annotations as the "gold labels". We used 20 random cases for the development of the Default Assistant and held out 180 cases to preserve a robust evaluation set.

Default Assistant Implementation

We built an AI "Default Assistant" to flag case defects and provide recommendations for requirements evaluated in compliance with CA Civil Code §§1788.58–.60. 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 and Appendix Figure 5).

The Default Assistant processes each case file through a structured pipeline that converts PDFs into Markdown using optical character recognition (OCR), chunks and stores the files by page in a vector database, and performs retrieval-augmented generation (RAG) with additional citation steps to classify whether each statutory requirement is satisfied or not (Figure 1). The model grounds each recommendation in verified evidence from the original filings and provides an explanation citing specific quotes and tables.

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.

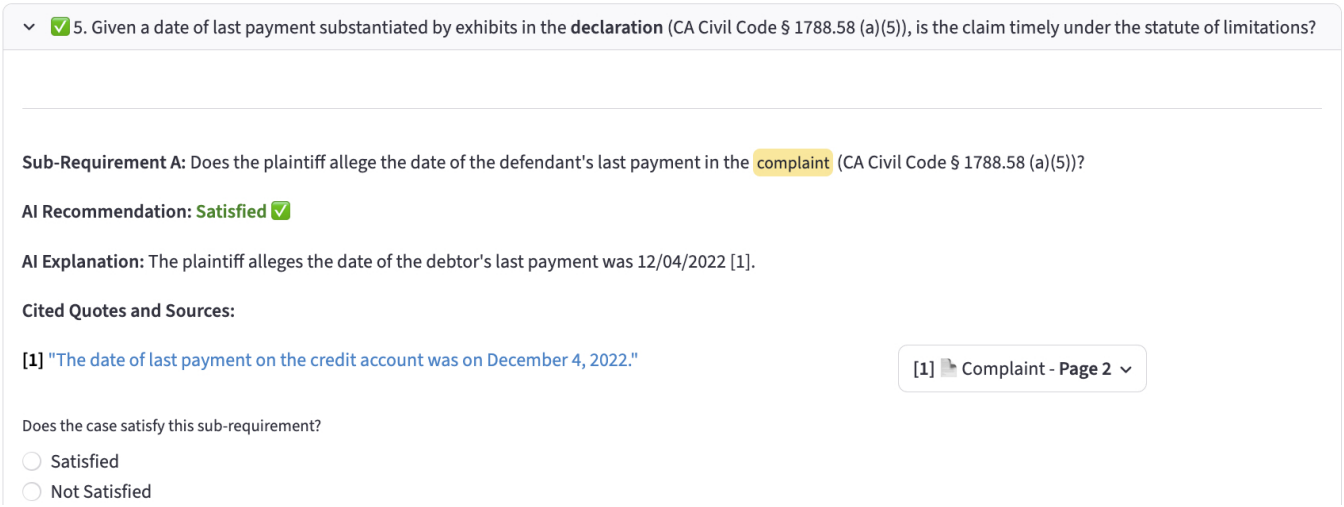

Figure 2: For each recommendation, the Default Assistant provides a binary recommendation and a free-form explanation cited to verified quotes from the case materials. Users may also view the original case file PDF page as shown in Appendix Figure 5.

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.

The three-step process of identifying, verifying, and finally generating from source quotes minimizes hallucinations, enforces citation accuracy, and enables transparent tracing of each model decision to the underlying case materials. Prior work provides empirical support for the *attribute-first-then-generate* design choice. Specifically, Slobodkin et al. (2024) demonstrate that attributing evidence prior to generation yields more stable grounding and fewer unsupported claims than the simultaneous generation of answer and citations. Worledge, Hashimoto, and Guestrin (2024) show that constraining model outputs to verifiable quoted or entailed spans under the *attribute-first-then-generate* paradigm improves citation coverage over that of deployed systems.

Supported Statutory Requirements

The Default Assistant flags case defects given procedural and statutory requirements from CA Civil Code §§ 1788.58-60. Working closely with court stakeholders, we identified the list of requirements that the court reviews for default judgment requests in debt-buyer plaintiff collections cases, as of September 2025.

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 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)).

The Default Assistant evaluates each requirement through a decomposition of sub-requirements, presented in Appendix Figure 3. These sub-requirements and requirements are shared with the gold labeling procedure. Requirement-level recommendations 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.

The Default Assistant uses a round of retrieval and cited generation to generate each sub-requirement recommendation. We define each round of retrieval and cited generation as a node in LangGraph[1]—a graph that asynchronously executes nodes in parallel and in order of dependency, permitting later nodes to use information identified in previous 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.

Human Study Methodology

We recruited 66 participants to review debt collection default judgment requests with and without the Default Assistant, under a 10-minute per-case time limit. Consistent with prior work studying AI in legal workflows (Nielsen et al. 2024), we worked with law student participants. The study was four hours long and included training and case review to simulate court staff workflows. Participants were randomly assigned to two conditions:

1. **Human:** No access to the Default Assistant
2. **Team:** Access to the Default Assistant

We co-designed an in-person, 80-minute training reflecting standard procedure with a research attorney from the Court 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. After training, participants were assigned to separate rooms 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 (Appendix Figure 4). 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 only a few minutes per case, we allowed more time to account for participant inexperience, setting the limit at 10 minutes based on pilot study averages. Participants were compensated with $100 gift cards and free dinners.

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 180 cases in the held-out test set were randomly selected to receive decisions from multiple participants within a branch, resulting in 217 evaluations per branch. As with the gold labels, 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 5.6%, while decreasing the average time spent reviewing each requirement for a 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.

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 the average case (n=217). Starred values (*) indicate statistically significant non-zero differences after Bonferroni correction ($p < 0.002$).

Metric	Absolute Gain	Relative Gain	p-value
Accuracy	4.9 pp	5.6%	2.7e-04*
Timing	-12.6s	-25.9%	1.7e-10*
Confidence	2.6	3.6%	1.1e-01

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 53% fewer errors in judging the *Charge-off Balance* requirement and 47% fewer for *Last Payment Date & SoL*, with differences significant ($p < 0.05$). The Default Assistant also expedites review time by 19% to 33%, 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*.

The Default Assistant enabled participants to achieve the error reductions reported in Table 3 by helping users increase their precision and, to a lesser extent, recall in rejecting defective cases (Table 4). While the small number of defects limits interpretation, we observe that the human-AI team generally increases precision over unassisted humans and raises recall for *Request-Complaint Consistency*. *Debtor Address* and *Charge-Off Balance* show apparent decreases in

[1] www.langchain.com/langgraph

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=217). Starred comparisons (**) indicate statistically significant non-zero differences ($p < 0.05$) after Benjamini-Hochberg correction. Calculated at full precision; rounded for display.

Requirement	Human Acc.	Team Acc.	Rel. Error Reduction	Rel. Time Savings
Last Payment Date & SoL	0.78	0.88	47%**	33%**
Charge-Off Balance	0.86	0.94	53%**	30%**
Debtor Address	0.86	0.91	39%	31%**
Chain of Title	0.87	0.92	41%	19%**
Agreement to Debt	0.95	0.98	64%	7%
Sworn Declaration	0.96	0.99	63%	30%**
Request-Complaint Consistency	0.90	0.90	5%	23%**

recall, but the team and unassisted humans differed on only two or fewer cases with defects for these requirements.

The Default Assistant consistently improves accuracy across major and minor error categories between the unassisted human and team branches (Table 5). *Major - Reject* errors are severe case defects that would preclude default judgment (i.e., time-barred cases, failure to include a valid declaration, or 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. Reductions in false rejections, i.e., the rejection of valid cases, drove improvements in *Major - Reduce* accuracy; the Default Assistant led to 64% reductions in false rejections. In contrast, reductions in false acceptances, i.e., the acceptance of invalid cases, drove improvements in *Minor - Amend* accuracy and *Major - Reject* accuracy improved due to both forms of error reduction.

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.

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. 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 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.

Discussion

Our human study demonstrates that the Default Assistant can help users review requirements more accurately and efficiently. While the Default Assistant reduces the time taken by participants to review requirements across the board, the strongest boosts in accuracy are for the *Charge-Off Balance* and *Last Payment Date & SoL* 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 100 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 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.

Table 4: Precision and recall for identifying invalid cases by requirement for human and team decisions, ordered by number of defects (n=217). Gains are the team metrics minus human metrics. AI assistance increases or maintains precision for all requirements and increases or maintains recall of 5 out of 7 requirements. However, the low number of defects limits interpretation. Calculated at full precision; rounded for display.

		Precision			Recall		
Requirement	Num. Defects	Human	Team	Gain	Human	Team	Gain
Last Payment Date & SoL	35	0.42	0.59	0.17	0.91	0.97	0.06
Debtor Address	24	0.42	0.59	0.17	0.75	0.67	-0.08
Request-Complaint Consistency	19	0.44	0.47	0.03	0.58	0.95	0.37
Charge-Off Balance	17	0.34	0.57	0.22	0.82	0.76	-0.06
Chain of Title	9	0.17	0.28	0.11	0.56	0.56	0.00
Agreement to Debt	1	0.08	0.20	0.12	1.00	1.00	0.00
Sworn Declaration	0	0.00	0.00	0.00	0.00	0.00	0.00

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=217). *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.

	Accuracy		True Rejection			True Acceptance		
Requirement	Human	Team	Human	Team	Rel. Change	Human	Team	Rel. Change
Major - Reject	0.81	0.90	0.79	0.86	9%	0.81	0.91	12%
Major - Reduce	0.90	0.91	0.58	0.95	64%	0.93	0.90	-3%
Minor - Amend	0.74	0.82	0.88	0.86	-2%	0.70	0.82	17%

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* and *Request-Complaint Consistency*) 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 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 generally achieve higher accuracy gains than the assistant alone, compared to the human baseline (Table 6). For example, humans using the Default Assistant gained 10 pp on *Last Payment Date & SoL* requirement accuracy, compared to an 8 pp gain for the Default Assistant alone, relative to the human baseline. This observation suggests that humans do not rely entirely on the Default Assistant. 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 either alone.

Limitations & Future Work

Although we aim to simulate court review as closely as possible, our methodology differs in three key ways. First, collections case review was a new task for our participants, despite their broader 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 may affect how effectively users leverage the Default Assistant's recommendations.

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), and participants had 10 minutes per case, while real-world time limits are as short as four minutes.[2] A shorter time limit may have prevented unassisted users from thoroughly reviewing cases, potentially leading AI-assisted users to reject more case defects than users without AI.

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

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=217). 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.

Requirement	T	H	DA	$T-H$	$DA-H$	Amplification
Last Payment Date & SoL	**0.88**	0.78	0.86	0.10	0.08	1.29x
Charge-Off Balance	**0.94**	0.86	**0.94**	0.07	0.07	1.00x
Debtor Address	**0.91**	0.86	0.89	0.06	0.03	1.71x
Chain of Title	**0.92**	0.87	**0.92**	0.06	0.06	1.00x
Agreement to Debt	**0.98**	0.95	0.96	0.03	0.01	2.33x
Sworn Declaration	**0.99**	0.96	0.98	0.02	0.01	1.67x
Request-Complaint Consistency	0.90	0.90	**0.93**	0.00	0.03	0.17x

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.

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

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 four-hour study.

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. Before deploying the Default Assistant permanently, we are planning a year-long study at the Court to evaluate the long-term and downstream effects of incorporating the Default Assistant into live work streams. For our team, evidence that the Default Assistant's cited-recommendation approach can increase accuracy and efficiency in human case review was a prerequisite for real-world deployment.

Conclusion

Our audit of 100 California debt collection cases with default judgments found that 37 cases contained defects requiring amendment, four contained inconsistencies requiring reduced judgments, and four contained major defects precluding judgment, highlighting opportunities to reduce erroneous judgments. We built a Default Assistant to support accurate and efficient review by courthouse staff, complete with citations to quotes and tables within case files. In a 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 5.6% and reduced review time by over 25% in our simulated court 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 the Court, 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 Assistant may enable more thorough review of additional requirements, potentially shaping future protocols. More broadly, our findings point to AI's potential to assist with repetitive manual workloads in other courthouse settings.

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 complemen-

tary 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.

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.

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.

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.; Winogradska, 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.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Mullen, F. 2019. Fifty years after the Consumer Credit Protection Act: The high price of wage garnishments. *Mitchell Hamline L. Rev.*, 45: 191.

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*.

Ruschemeier, H.; and Hondrich, L. J. 2024. Automation bias in public administration–an interdisciplinary perspective from law and psychology. *Government Information Quarterly*, 41(3): 101953.

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.

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.

Appendix

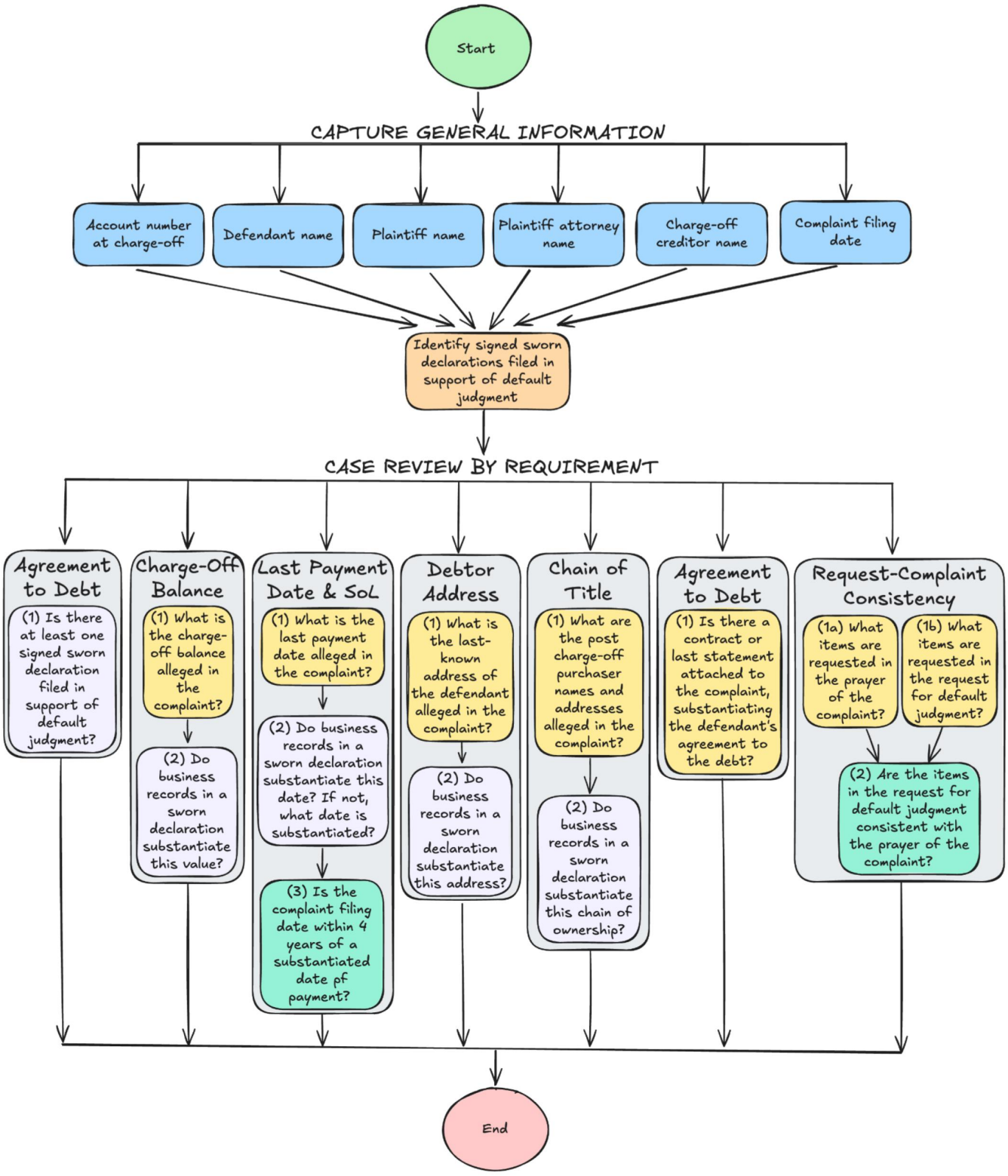

Figure 3: The Default Assistant gathers general case information before reviewing the case for each of the seven requirements. Each box in this graph, aside from "Start" and "End", retrieve relevant case documents, use information identified by previous nodes in the graph, and generate recommendations with citations for each requirement.

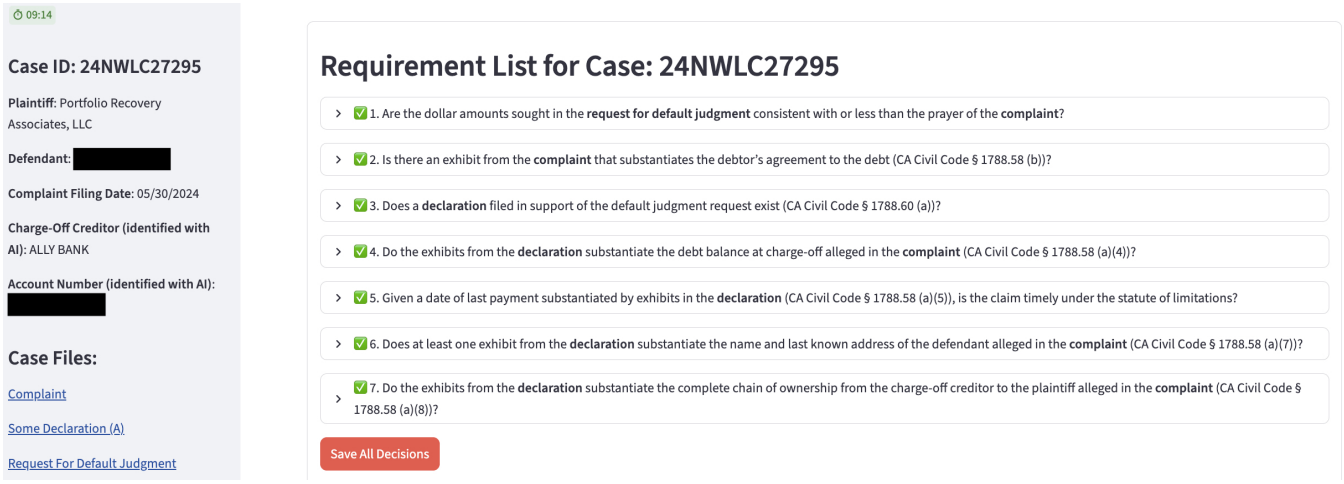

Figure 4: Participants review each case using general case information, case files, and the sub-requirements under the drop-down tabs for each requirement. Only participants assigned to a branch with the Default Assistant are shown "Charge-Off Creditor (Identified with AI)" and "Account Number (Identified with AI)", in the sidebar. All other case information is pulled from the case management system. Redactions by the authors.

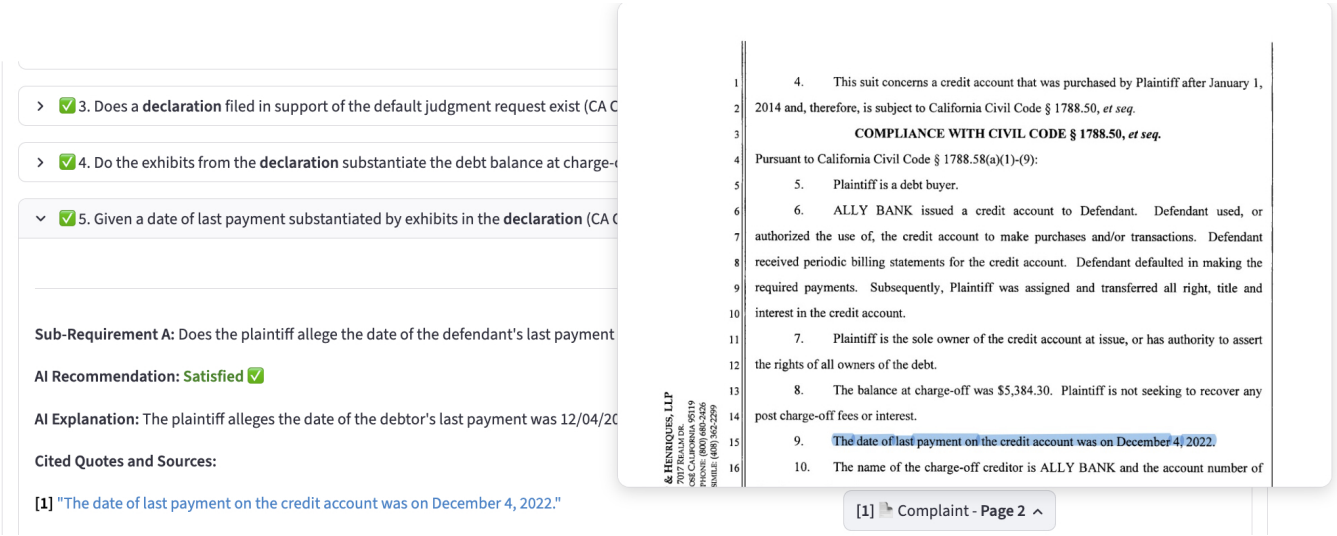

Figure 5: Each citation in the "Cited Quotes and Sources" section includes a pop-over that links to the original case-file PDF page with the cited information, highlighted in blue when available. The screenshot shows the page for citation [1] of "Sub-Requirement A." Redactions by the authors.