Evaluating Generative AI in Benefits Administration: A Demonstration Project

VARUN MAGESH*, OLIVIA H. MARTIN*, FAIZ SURANI*, AMY PEREZ, KIT RODOLFA, and DANIEL E. HO

Government benefits programs are a primary touchpoint between citizens and the state. Yet they form a core challenge for government modernization, with legacy systems that strain when demand is highest. Agencies are exploring artificial intelligence (AI) and machine learning (ML) tools for these systems while vendors eagerly market such solutions. The potential benefits and risks of these tools are profound when applied to benefits systems where timeliness and accuracy are essential to due process. We present a collaboration with the US Department of Labor (DOL) and the Colorado Department of Labor and Employment (CDLE) to develop and evaluate Generative AI tools to modernize a pillar of the social safety net: Unemployment Insurance (UI). We make four primary contributions. First, we established the first comprehensive sandbox environment for AI evaluation in benefits administration, enabling co-design of a GenAI system with agency staff and providing unique access to granular, individual-level adjudication data such as editing patterns and cross-adjudicator variation. Second, we developed a systematic methodology for eliciting and encoding expert quality assessment from adjudicators, contributing to the broader challenge of measuring adjudication quality and aligning AI systems with domain-expert values. Third, we conducted a randomized controlled trial evaluating our fact finding assistance system on real, historical cases, with outcome measures capturing both decision quality and fine-grained behavioral data. Fourth, our evaluation reveals a critical divergence: AI fact finding was a substantial improvement to historical (observational) baselines and examiners subjectively rated the system highly; but the system did not improve quality or efficiency in the sandbox control group, though it may reduce inter-adjudicator variance. This contrast demonstrates that rigorous, context-situated evaluation is essential to evaluate AI in legal contexts.

CCS Concepts: • **Applied computing** → **Law**; • **Human-centered computing** → **User studies**; • **Computing methodologies** → *Natural language generation.*

Additional Key Words and Phrases: generative AI, adjudication, randomized control trial

1 Introduction

Government agencies decide more cases each year than all federal courts combined, adjudicating the rights of immigrants, the disabled, veterans, and welfare claimants. Yet this system of "mass adjudication" is marked by significant delays and errors [5]. For instance, as of 2017, injured veterans waited an average of 5-7 years to have their appeals for disability benefits resolved, with some 7% estimated to pass away while waiting [36]. Error rates in the Department of Labor (DOL) worker's compensation system are estimated to be between 20-40% [24], while 44% of denials of SNAP (food stamps) benefits are incorrect [2]. During the pandemic – the worst unemployment crisis since the Great Depression – Unemployment Insurance (UI) systems buckled under a flood of claims, with timely payments cratering from 90% to 52% [65]. Caseload pressures fundamentally compromise accuracy, consistency, and due process. The result is a system of "benefits roulette," where outcomes may be driven more by the adjudicator assigned than the facts of the case [13, 34].

With high caseloads and insufficient resources, agencies have turned to technology modernization and innovation to try to bridge this gap. For instance, the US Patent & Trademark Office, which processes over a million applications each year, is using AI prototypes to reduce the time to search for prior art or marks that may conflict with claims of novelty [19]. The Social Security Administration is using AI to help flag potential errors in disability benefits decisions

*These authors contributed equally to this research.

Authors' Contact Information: Varun Magesh, mvarun@posteo.net; Olivia H. Martin, omartin@stanford.edu; Faiz Surani, faiz@law.stanford.edu; Amy Perez, amyperez747@gmail.com; Kit Rodolfa, krodolfa@law.stanford.edu; Daniel E. Ho, deho@law.stanford.edu.

and batch similar claims for more efficient review [25, 57, 58]. The Department of Labor has hosted a 'Vendor Day' for government technology vendors to share previews of new AI tools with government staff. But high-profile failures underscore the need for a responsible approach to — and rigorous evaluation of — these innovations. When Michigan's Department of Labor and Economic Opportunity deployed a model in 2013 to identify fraudulent UI claims, the system resulted in false accusations of more than 20,000 claimants, with some individuals facing fines as high as $100,000 [22]. Recent years have seen mounting concerns about disparities in algorithmic tools used in domains such as criminal justice [6, 23], healthcare [49], hiring [16, 55], and housing [1], while large language models (LLMs) have surfaced worries about their potential to leak private information [47, 48].

In few places is the promise for societal benefit and risk for significant harm for AI greater than in the high-stakes context of government benefits programs that shape millions of lives. Here we present a case study of co-design and real-world evaluation of a Generative AI system in the high-stakes setting of UI benefits adjudication. We focus on UI, one of the nation's largest social safety net programs that plays a particularly critical role at times of national crisis. During the first year of the COVID-19 pandemic, 46 million Americans (18% of the country's adult population) relied on the program's benefits [30]. Through a unique cross-sector partnership with US DOL and the Colorado Department of Labor and Employment (CDLE), which administers the state's UI program, we designed, developed, and evaluated state-of-the-art large language models (LLMs) to support the decision workflows of UI adjudicators in the state of Colorado. This collaboration provided extraordinary access to both individual-level historical claims data and the granular mechanics of adjudicator decision-making processes - a level of detail rarely available in studies of administrative adjudication, which typically rely on highly aggregated data such as grant rates per adjudicator.

We make four contributions. First, we establish the first comprehensive sandbox environment for AI evaluation in benefits administration, and leverage it to evaluate a novel, co-designed GenAI intervention. Our data allow us to observe not just final outcomes but fine-grained intermediate decision processes such as the number of questions generated, time in seconds for each decision, the extent and nature of edits to AI-generated content, and variation in usage patterns across adjudicator experience levels. Second, we develop a systematic methodology for eliciting and encoding quality assessment from government adjudicators, contributing to the broader challenging of developing contextual quality benchmarks for AI systems. Third, we conduct a randomized controlled trial evaluating our system on real, historical cases, measuring not only decision quality and time-to-completion but also detailed behavioral indicators of how adjudicators engage with AI-generated content. Fourth, our findings reveal a striking divergence between different measures of system performance: the AI system significantly outperformed historical baselines and received positive subjective ratings, meeting conventional criteria for adoption; at the same time, it showed no improvement in quality or efficiency relative to the sandbox control group receiving no AI assistance. The system may, however, help reduce inter-adjudicator variability. These findings have complex implications for the adoption of AI systems – showing the potential limitations of sandbox trials, user satisfaction, and historical comparisons in isolation – but also demonstrate the importance and feasibility of multifaceted evaluations of AI systems in consequential settings.

1.1 Institutional Context

Determining eligibility for UI benefits is a time-intensive and complicated task – so much so that that one adjudicator in CA was still referring to himself as "the new guy" after *17 years* of experience [51]. The process begins with a person submitting a claim explaining why and how they separated from their job. An adjudicator then decides their eligibility under state and federal law, generally evaluating the claimant on two criteria: whether they earned enough money during the applicable period and whether they left their job due to no fault of their own (e.g., laid off or needed to

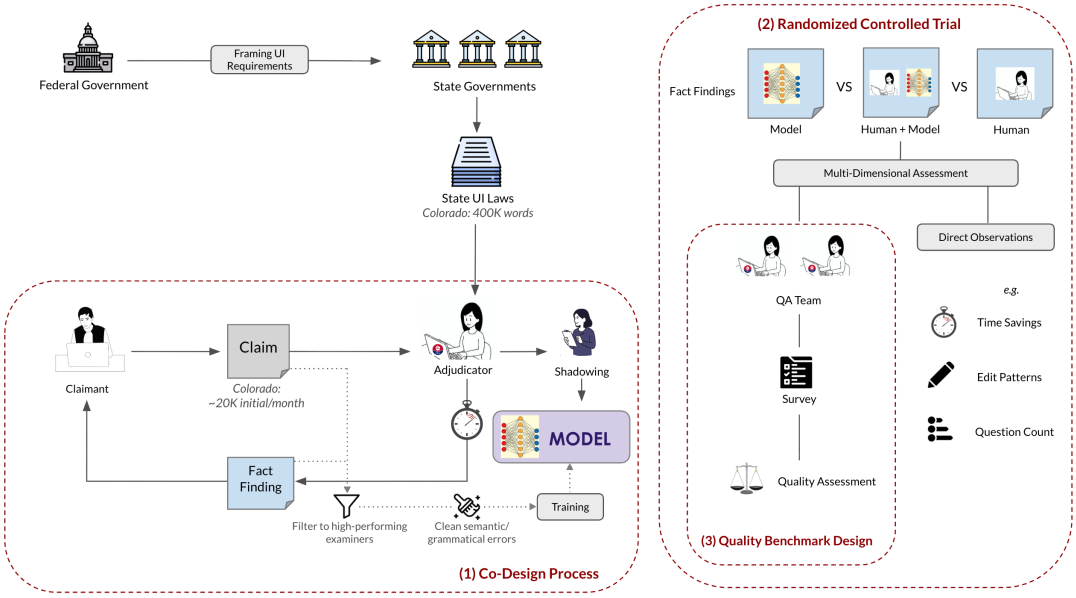

Fig. 1. Institutional and evaluation design for AI-assisted unemployment insurance (UI) adjudication. Federal law establishes broad UI eligibility requirements, which state governments implement through voluminous state laws and regulations (for example, Colorado's code section covering labor and employment is nearly 400,000 words long). Adjudicators apply these rules to decide eligibility on individual claims, conducting additional fact finding with claimants as necessary. By shadowing adjudicators, we identified the fact finding process as a significant pain point and co-designed our model, trained on historical claims, to assist in this process (left panel). We evaluated its effectiveness in a randomized controlled trial comparing fact finding generated by the model alone, by adjudicators working with the model, and by adjudicators working unaided (right panel). Outcomes were assessed by a QA team on quality and by measuring time savings, providing benchmarks for both accuracy and efficiency.

care for a sick family member). The employer is also notified of the claim, asked to provide details about the worker's separation, and may dispute the claim by alleging that the worker was terminated for cause or quit without reason. If there are discrepancies between the claims, the adjudicator must conduct additional fact finding with the claimant, the employer, or both. For example, a claimant may offhandedly mention concerns about workplace staffing and safety as reasons for their quit, and it is the adjudicator's job to further understand these concerns to see if they meet legal eligibility requirements. Yet some claims require little to no fact finding, such as a layoff undisputed by an employer. Claimants and employers can appeal the initial determination, resulting in a hearing before an administrative law judge. Federal and state quality assurance (QA) teams also regularly review a random audit of eligibility determinations each quarter to ensure correct application of law and policy as well as adequate reasoning.

Overall, the fact finding process can take weeks, depending on the time taken to respond in the back-and-forth between the parties, requiring the adjudicator to identify which leads should be followed and carefully craft questions to efficiently uncover the underlying reason for the separation. We chose to focus our intervention on this aspect of the process given this institutional background and from a number of observations from our co-design process. First, during shadowing of adjudicators processing claims, it became clear that this was one of the more tedious and time consuming aspects of their workflow. In particular, we observed that many adjudicators have developed their own "question banks" (e.g. in Excel) that they have owned over their time on the job, from which they will search

to find an appropriate question, copy-and-paste it into CDLE's case management system, and then edit to reflect the details of the current claim. We posited that an effective language model might help reduce the burden of this manual process. Second, we also learned that writing effective questions involves a steep learning curve for less experienced adjudicators, particularly to ensure they elicited the necessary information to reach an accurate determination. A model that suggests potential topics for follow-up and related questions might help these adjudicators onboard more quickly and ensure their fact finding interactions were as comprehensive as possible. And, third, we hypothesized that such models might not only save adjudicators time in developing fact finding questionnaires, but also ensure better coverage of potential reasons for granting or denying benefits with a smaller number of back-and-forth rounds of fact finding, potentially reducing the total time required to reach a determination.

We therefore chose to use fine-tuned LLMs to facilitate this complex decision process, particularly in supporting high-quality follow-up questions for this back-and-forth "fact finding" process. We hypothesized that improving the quality of follow-up questions could improve decision accuracy (by ensuring adjudicators elicit the relevant information from both claimant and employer), timeliness (by reducing the number of rounds of back-and-forth follow-up), and consistency (by reducing variability in process and outcome across adjudicators). The system integrates two LLM components into the existing fact finding workflow:

- The first identifies topics for follow-up based on the initial claim materials, such as points of disagreement between the narratives of the claimant and employer, details of the claim that need elaboration, or aspects of a narrative that might be determinative of a decision to grant or deny benefits. Example topics might include "accusation of falling asleep on night shift" or "employer's response to COVID-19 concerns."
- The adjudicator can then select from among these topics as well as write custom instructions to prompt the second AI component, which drafts potential follow-up questions for streamlined fact finding to resolve key issues of the claim.

The output of this second component is a list of suggested questions the adjudicator may want to consider in developing their fact finding: they can edit the wording of the draft questions, remove those that do not seem useful, or add additional questions before finalizing the follow-up questionnaire. For ease of prototyping and evaluation, we developed a lightweight interface for adjudicators to interact with this AI system, depicted in Figure 2.

2 Related Work

Our work sits at the intersection of five discrete bodies of literature addressing the challenges of deploying AI systems in high-stakes public sector settings: administrative burden and state capacity, human-AI collaboration in practice, ML evaluation methodologies, legal frameworks for quality assessment, and responsible AI governance in the public sector.

First, a substantial body of work documents how administrative burdens affect both benefits claimants and the civil servants who serve them [33, 51]. Learning and compliance burdens create systematic inequalities in access to public programs [11]. The digitization of benefits systems, while promising efficiency gains, can exacerbate these burdens when poorly designed [20]. Yet at the same time, limited state capacity constrains agencies' ability to process claims in an accurate and timely fashion as demanded by due process [14]. Our work addresses these dual challenges by evaluating AI tools designed to reduce burdens on both claimants and agency staff while maintaining decision quality.

Second, despite the enthusiasm for AI assistance, empirical studies of human-AI interactions reveal a more complex reality [4, 54]. Recent evidence from healthcare shows that AI-assisted physician documentation can be nuanced. One study found that AI-generated notes were longer, often higher in quality, but also prone to factual errors and did not

clearly save physicians time [8]. Other work has found some efficiency gains, particularly in reducing documentation time, though effects on overall efficiency and burnout are mixed [18, 43, 60]. Moreover, human-alone or AI-alone performance can even exceed that of human-AI teams, particularly when humans struggle to calibrate trust in AI systems appropriately [9, 29, 61, 66]. These findings underscore the importance of studying AI deployment in specific organizational contexts rather than assuming universal benefits from human-AI collaboration.

Third, a growing critique within machine learning challenges conventional benchmarking practices that fail to capture real-world deployment complexity [40, 56]. Domain-independent metrics often miss critical contextual factors that determine system success or failure. Recent work calls for situated evaluation that considers the full deployment environment [67]. This literature directly motivates our co-designed quality benchmarking approach.

Fourth, our work pertains to a rich literature in administrative law on the challenges of mass adjudication in light of constitutional due process values. That literature highlights the difficult balance between efficiency and quality in benefits determination [35]. Mashaw's "managerial" conception of due process centered on continuous and systemic quality assurance measurement. A key challenge has been the availability of granular micro-data [52]. The Administrative Conference of the United States (ACUS) issued a formal recommendation that "Agencies, particularly those with large caseloads, should consider whether . . . artificial intelligence (AI) tools help quality assurance personnel identify potential errors or other quality issues" [37]. ACUS also recommended soliciting feedback from adjudicators about such systems. But the choice of quality measures for AI systems is not straightforward [45, 46]. Some scholarly work warns that AI systems may entrench a narrow efficiency focus at the expense of procedural fairness and substantive accuracy [12]. This tension is particularly acute in public benefits, where incorrect denials impose severe costs on vulnerable populations. Our evaluation framework explicitly addresses this by measuring both efficiency gains and quality impacts.

Finally, our work speaks to recent policy initiatives, such as the Biden Administration's efforts to establish requirements for responsible AI deployment in government [21, 50, 68]. However, significant gaps remain between high-level principles and operational implementation [44]. Recent work has demonstrated the important tensions that can arise between these goals, such as the trade-offs that can occur between methods for privacy-protection and bias mitigation [10, 42, 59]. Our work demonstrates the value of context-situated evaluation to make abstract responsible AI concepts more concrete and measurable.

Our paper contributes to these literatures through a concrete demonstration of responsible AI development in a public benefits context. Our results contribute evidence on effective human-AI collaboration patterns in bureaucratic settings, while our evaluation approach demonstrates the value of carefully designed quality benchmarks. By developing AI tools within existing legal and operational constraints, we provide a model for translating responsible AI principles into practice. However, our findings also reveal that even carefully designed sandboxes cannot fully anticipate the complexities of real-world deployment, suggesting the need for graduated implementation strategies rather than binary deployment decisions.

3 Methods

3.1 Data

The primary data for this project are historical UI claims adjudicated by CDLE. In particular, we focus on claims submitted starting on January 1, 2022 (to avoid the idiosyncrasies of claims filed at the height of the COVID pandemic), yielding a universe of more than 3.3 million job separation issues with over 486 million individual fact finding elements (e.g., question-answer pairs). Our models draw on several pieces of information about these claims: the structured and

unstructured fact finding responses from the claimant and their former employer(s); claimant demographic information; and data produced during the adjudication process (intermediate and final issue types and subtypes, structured and free-text reasons for a decision, and timeliness measures). To protect the privacy of UI claimants, all modeling and data analysis was performed within the sandbox established in CDLE's cloud-based secure compute environment.

3.2 Model Development and Prototype Interface

Based on our adjudicator shadowing, we identified fact finding as a time intensive task that could be improved with LLM assistance. We envisioned two models to reduce this friction and allow adjudicators to exercise their intent more efficiently: a topic suggestion model that would identify relevant follow-up topics and a fact finding draft model that would convert a series of topics and a free-text instruction into a set of questions grounded in the claims record.

Due to the sensitive nature of UI claims data and the agencies' requirement that all data remain within CDLE's sandbox environment, we focused our model development on open models that could be fine-tuned and deployed entirely within this infrastructure [62]. To develop the question-drafting model, we curated a dataset with three components: claim details, follow-up questions, and related prompts to generate those questions (e.g., "Ask about the incident that occurred with the manager on Jan. 11"). The first two components are readily available in the historical adjudication records, but we lacked ground truth data for "LLM instructions that create good fact finding questions." We therefore employed a synthetic data approach by using a few-shot prompted model to generate instructions from the ground-truth questions we already had. Specifically, we prompted a model with the text of a claim and a set of real questions written by an adjudicator, then asked it to write an instruction that could generate those questions. To produce a diverse training set, we generated instructions using an ensemble of three different models (Meta's LLaMA-3 70B [28, 64], Alibaba's Qwen 32B [7], and Cohere's Command R [27]) at high temperature, conditioned on $n = 3$ random samples of few-shot examples from a handwritten set of 20.

Curating "ground truth" questions involves nuances, precisely because adjudicators operate under significant time pressure with variation in experience, leading to potentially vague or ungrammatical questions. We addressed this in two steps. First, we restricted the training examples to claims processed by adjudicators that CDLE managers identified as among the most experienced. In essence, this steers the AI system – based on institutional knowledge – toward simulating individuals known to produce higher quality fact finding. Second, we instructed a large model (LLaMA-3 70B) to revise the historical questionnaires to improve their grammar, specificity, and adherence to the synthetic instruction. The aim here was to simply to improve writing quality and accessibility without changing the underlying content of the fact finding. Using both the generated prompts and revised questions as training data, we fine-tuned a smaller LLaMA-3 8B model on the inverse fact finding drafting task: given a claim and a synthetic instruction, the model predicts the adjudicator-written (and synthetically revised) fact finding. We trained the LLaMA-3 8B model using Low-Rank Adaptation (LoRA) [39] on 46,000 training examples.[1]

We adopted a similar synthetic data strategy for the topic suggestion model. Given a list of revised fact finding questions, we few-shot prompted the ensemble of large LLMs (LLaMA-3 70B, Qwen 32B, Cohere Command R) to generate a list of topics that correspond to these ground truth questions. These generated topics were manually reviewed and revised to improve clarity and specificity. We then fine-tuned a smaller LLaMA-3 8B model using the same LoRA configuration to predict the list of topics given the text of the initial claim alone.

[1]We used LoRA with rank 16, zero dropout, training all attention and feed-forward network layers with FlashAttention-2 [15] and Unsloth kernels [31], 16-bit precision, and a learning rate of 3e-4 on an Nvidia A10G GPU.

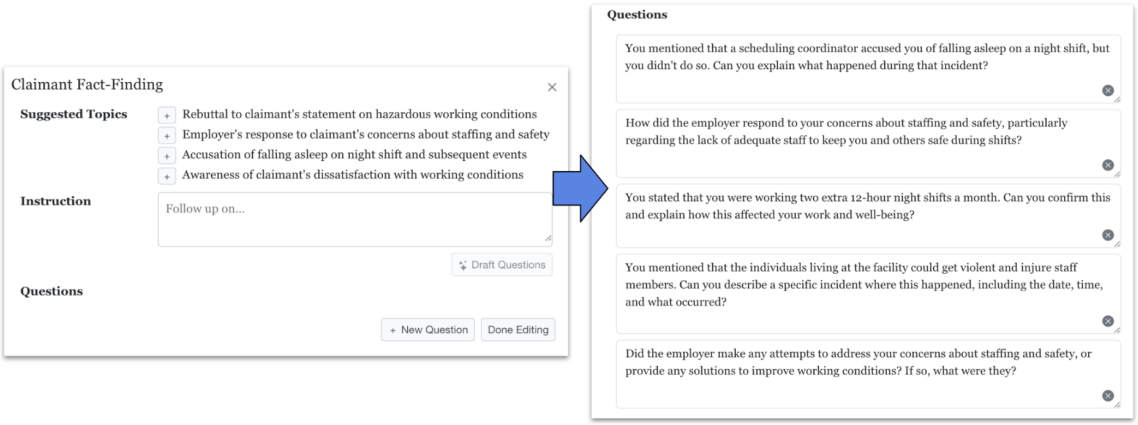

Fig. 2. Example of the prototype interface used for the trial. Adjudicators could choose among model-generated follow-up topics as well as specify custom instructions in the text box to specify additional topics for followup or desired properties of the output (e.g., number of questions). Upon clicking "Draft Questions", model-generated questions would be presented in editable text-boxes, along with the option to remove unwanted questions or add entirely new ones.

Finally, we developed a lightweight prototype interface for adjudicators to use and evaluate both of these models (Figure 2). For a given historical claim, the adjudicator is provided with the details of the initial application submitted by the claimant as well as their employer's response (collected via a standard form that is automatically sent when the claim is filed). Below these details, the adjudicator can decide whether to draft follow-up questions for the claimant, employer, or both parties (alternatively, they can indicate that enough information is already available to reach a decision and no additional follow-up is necessary). In each case, they can select from among AI-generated suggested topics, as well as write custom instructions (to provide either other topics for follow up or desired properties of the output, such as limiting the number of questions). Upon submission, the question drafting model generates a set of questions, each in its own editable text box, along with the option to remove unwanted questions or add new ones before finalizing the fact finding questionnaire.

3.3 Trial Design, Randomization, and Analysis

We evaluate the impact of our AI assistant intervention in a randomized, controlled crossover trial, employing 8 CDLE adjudicators to evaluate the system by addressing historical cases using the tool's assistance. Six members of CDLE's internal Quality Assurance (QA) team participated to assess the resulting fact finding questionnaires (see Figure 3 for a schematic overview of the trial design). To make comparisons holding either adjudicator or claim attributes constant, we randomized assignment to the model at the adjudicator-week level. That is, for all historical evaluation claims an adjudicator considers in a given week, they either work with or without the AI tool to develop follow-up fact finding questions. By exposing the same adjudicator to different conditions (e.g., with or without-model) for different batches of claims, we sought to be able to more directly assess heterogeneity at the within-adjudicator level in the model's effects. We opted for a within-subjects design due to the high level of inter-adjudicator disparities, which make achieving statistical power of a between-subjects design infeasible given operational constraints. Adjudicators participated in the trial based on a number of factors, including interest, availability, and manager selection. Adjudicators participating in the trial were generally pulled from the pool of adjudicators actively working on cases. High-quality evaluation of this

system came at a real cost for CDLE, and we were constrained by the availability of adjudicators the magnitude of CDLE's case backlog. CDLE controlled the process of recruiting adjudicators for the evaluation; the adjudicators were not randomly selected.

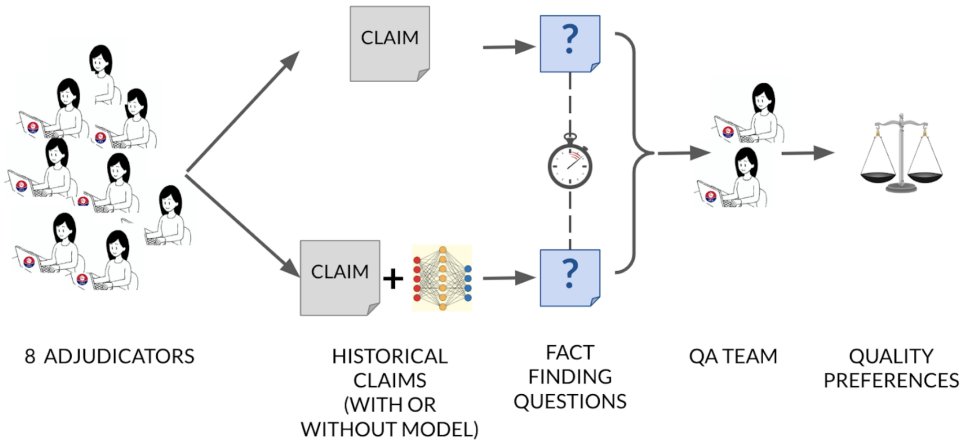

Fig. 3. Field trial design. Each week, participating adjudicators were randomized to either develop fact finding questions for historical claims either with or without the assistance of the model. Time to draft questionnaires was measured directly from interaction data with the prototype interface and draft quality was assessed by members of CDLE's QA team, who were shown head-to-head comparisons of questionnaires developed in each arm of the study.

To provide a more consistent universe for comparisons, we focused our trial on the 94,279 issues in CDLE's data that began as quits, meaning that the separation was initiated by the claimant (the employee), drawing a stratified random sample of claims from this universe. The sampling was stratified by: Whether follow-up fact finding was sent; Whether the first follow-up fact finding was to employer or to claimant; Whether the employer responded to the initial automated questionnaire; Claimant demographic characteristics including race, ethnicity, gender, and language preference. Cases that were originally worked on by any of the eight participating adjudicators were excluded. A total of 200 claims was sampled, with each randomly assigned to 4 out of the 8 adjudicators (2 in the treatment condition and 2 in the control condition). We assigned each case to 4 out of 8 adjudicators so that we could cover more unique cases. Overlap in the cases enabled comparisons between adjudicators. We assigned each case to only 4 adjudicators so that we could sample more unique claims in our trial. Adjudicators developed fact finding questionnaires for 10 claims each week, initially randomized to either treatment or control with their assignment alternating each week. We structured the trial to last for at least 5 weeks based on a minimum commitment from DOL. In practice, we were able to extend the trial for a few weeks (and around 50 claims) longer, though adjudicators completed cases at their own pace after the initial five-week period. In total, 788 fact finding drafts were collected (one adjudicator was promoted midway through the trial and did not complete their assigned claims).

To evaluate the model's impact on the quality of the fact finding questionnaires, we recruited six members of CDLE's internal QA team and senior management to conduct a blind review of the questionnaires. For a given claim, reviewers were shown the initial claim details (in the same interface they were presented to the adjudicators in the trial) as well as a series of pairs of draft questionnaires and asked them to assess the comparative quality of the questionnaires. The choice of this head-to-head ranking approach reflects several considerations. First, our outcome of interest is inherently

relative - we are interested in whether AI-assisted fact findings are preferable or less preferable to fact findings written without the tool, which lends itself well to a head-to-head ranking approach. Second, fact findings can be complex, with many different axes on which they can be better or worse. Placing them side-by-side and asking the question 'which fact finding would be better to send to a real recipient?' can make the value trade-offs in different questionnaires clearer. It also reduces researcher degrees of freedom by avoiding defining any criteria on which to evaluate 'good' fact finding, consistent with the finding by Tamblyn and colleagues [63] that in value-laden quality evaluation contexts, ranking may produce more reliable results relative to criteria-based rating schemes. In total, we collected 1,598 comparisons spanning pairs of four types of questionnaires: 1) those written by adjudicators during the trial with assistance of the AI tool (treatment), 2) those written by adjudicators during the trial on their own (control), 3) the historical questionnaire that was actually sent when the claim was processed, and 4) questionnaires generated by the model alone, without adjudicator editing/review. In all cases, QA raters were blinded to the origin of the questionnaires they were comparing. Claims where an adjudicator chose not to follow up with either party (e.g., because they believed enough detail was already available to make a determination) were included in the head-to-head comparisons, and QA reviewers were asked to rank which follow-up action they felt was best given the details of the claim at hand.

For directly-measured interaction data (e.g., drafting time, follow-up rate, questionnaire length, etc.), differences in distributions were assessed with χ^2 tests between the treatment and control groups, while differences in means were assessed with t-tests with standard errors clustered at the adjudicator level. In analyses to test for differences in QA team preferences, the 5-point preference scale was mapped to -2 (strong preference for one version) to +2 (strong preference for the other version), with 0 indicating both versions were equally preferred, with mean ratings tested against the null hypothesis of no preference (t-test with standard errors clustered at the level of the pair of adjudicators being compared across). For heterogeneous treatment effects across subgroups, QA preference across the groups was regressed against an indicator of treatment with errors clustered at the adjudicator pair level.

3.4 Qualitative Feedback

One of the most common evaluation methods for AI systems lies in user satisfaction metrics [32, 38, 69]. ACUS similarly recommended soliciting feedback about quality assurance systems [37]. We hence also surveyed adjudicators about their experience using the AI tool. After working with the tool to develop fact finding questions in the course of the trial, we both held semi-structured feedback sessions as well as asked the adjudicators who participated to fill out a brief survey on several aspects of their experience with it:

- Their general impressions of the overall utility of the AI tool in the fact finding process
- The relevance and completeness of the topics surfaced by the tool for follow-up as well as how often they felt the tool was helping them find topics they might not have followed up on otherwise
- The quality, tone, and specificity of the draft questions suggested by the tool, and how often they needed to edit or add to the model-generated draft
- Their suggestions for improvements to the tool or similar tools that might be helpful in their workflows

The full text of the survey is available in Appendix Table A1.

4 Results

4.1 Adjudicators Found the AI Tool Useful

Through both follow-up surveys and feedback sessions, the adjudicators who participated in the trial expressed positive experiences with the tool. All 8 adjudicators rated the tool as very or somewhat useful and said they would use the tool regularly (75%) or sometimes (25%) if it were available in their workflows. Likewise, when asked about the overall quality of the draft questions suggested by the model, none of the adjudicators rated them poorly and 63% (5/8) rated them very good or excellent. One adjudicator, for instance, saw particular potential for the tool during times of high caseloads, noting, "All in all I feel like this tool works well and will be very useful especially during times of a backlog."

However, the adjudicators varied on specifically how they saw the tool as providing value in their workflows. Only half of the adjudicators believed use of the tool saved time in preparing follow-up questionnaires: while one adjudicator described the tool as "a great timesaver" another noted, "I think the tool is pretty cool overall, but I am not sure if it was really much of a timesaver." Four adjudicators indicated the model often or sometimes helped them identify topics for follow-up that they might have missed otherwise, but 4 said it rarely helped in this way. Additionally, while they reported finding the draft questions a great starting point that helped surface specific details of the claim for follow-up, all eight adjudicators indicated that they nonetheless found it important to edit and refine these initial drafts. Taken together, these results suggest that the adjudicators found the system useful, but that the nature of its benefits may vary across individuals. To explore this question more quantitatively, we turn to the results from our head-to-head field trial with real claims data.

4.2 Overall Effects of AI Assistance

Time. Because AI models can rapidly ingest and process the potentially complex details of a UI claim, we hypothesized that assistance of the models here might substantially reduce the time adjudicators need to spend drafting their follow-up questionnaires. Surprisingly, however, there was little overall difference in the average time adjudicators took to draft questionnaires prepared with the model's assistance versus those prepared without it, a statistically insignificant reduction of only 4 seconds (p=0.8). Nevertheless, in the distribution of drafting times (Figure 4(A)), we can see a modest increase in the fraction of claims that were processed very quickly (p=0.1). However this reflected a speeding up of claims that were already very fast to process (e.g., those requiring only 2-5 minutes), yielding little overall impact on the average drafting time. Feedback from adjudicators generally seemed to confirm these dynamics: for "easier" cases (such as those with little information provided or where no additional fact finding was necessary), adjudicators felt that the model helped quickly confirm that little had to be done allowing them to move onto other cases. However, for more complex cases, the time spent editing the AI-generated draft to ensure it focused on the most salient details of the case counterbalanced the speed with which it could generate an initial draft.

Comprehensiveness. At least as important as the drafting time is the comprehensiveness and quality of the questionnaire itself. A questionnaire that misses key issues could lead to additional rounds for fact finding (thereby increasing both the administrative burden on the claimant as well as the time spent waiting on needed benefits) or yield errors in the ultimate benefits determination. We therefore also sought to understand how questionnaires drafted with the assistance of the model compared to those drafted without it. Figure 4(B) shows that use of the model increased the length of questionnaires (p=0.002), marginally increasing the average number of questions asked when the adjudicator chose to follow up from 3.8 to 4.3. Notably, this increase in length did not correspond to a higher propensity for adjudicators to choose to follow-up in the first place. One concern with a generative system like this one might be that

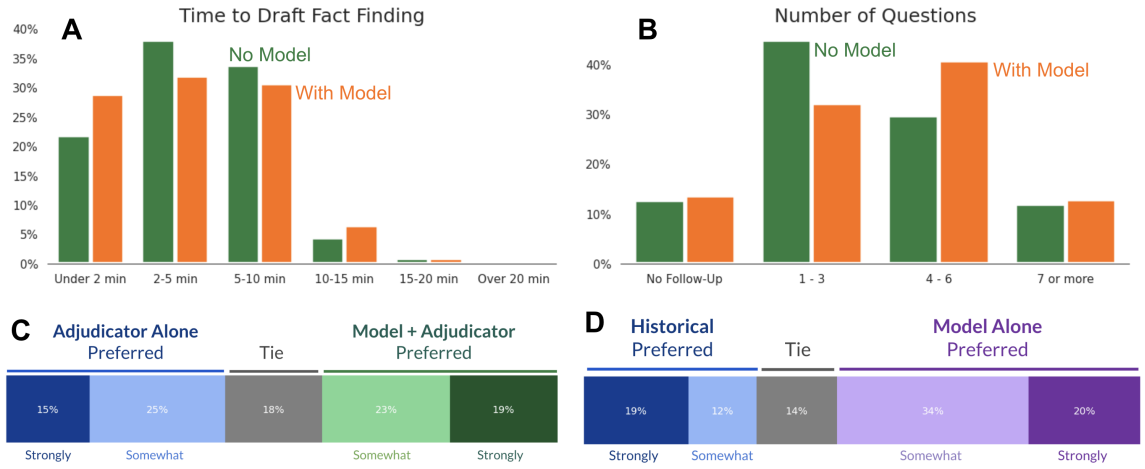

Fig. 4. Overall results of the trial. (A and B) Impact of AI assistance on drafting time and length of questionnaires, with results from drafts developed with the models in orange and those without in green (N=788). (A) The model moderately increased the fraction of drafts that were developed in under two minutes (p=0.1), but had little impact on the average drafting time (p=0.8). (B) While access to the model didn't change the rate at which adjudicators chose to follow-up, it did significantly increase the length of the questionnaires (p=0.002). (C and D) QA-team preferences in head-to-head comparisons of questionnaire drafts developed for the same historical claim under different conditions. (C) Overall, drafts developed by adjudicators participating in the trial with the model were preferred at similar rates to those developed by trial adjudicators alone (p=0.37; N=750). (D) However, strong preferences were observed for the model alone when compared to the historical fact finding questionnaire that was actually sent when the claim was processed (p=0.1; N=90).

it could increase the burden on claimants by making it "too easy" for an adjudicator to generate and send an additional questions to claimant, even when no follow-up is needed to come to a conclusion. However, we see no evidence of this dynamic occurring in practice: in both arms of the trial, adjudicators chose to follow-up in 87% of claims.

Quality. To understand how AI assistance affected the quality of the follow-up questionnaires, we enlisted members of CDLE's internal QA team to evaluate their preference across different drafts prepared for the same claim (blinded to whether adjudicators or the AI tool was involved in producing each draft). When we asked this team to make head-to-head comparisons of drafts written by adjudicators during the trial with vs without the AI tool (Figure 4(C)), we see little evidence to support the hypothesis that the model was helping these adjudicators improve the average quality of their follow-ups (p=0.37). To understand the performance of the model itself as well as the historical baseline, we also included drafts generated by the model alone (without edits from an adjudicator) as well as the historical questionnaires that were actually prepared when the claim was first assessed. Although we were able to collect relatively few such comparisons (N=90), the results in Figure 4(D) suggest a considerable preference by the QA team for the model outputs relative to what had actually been sent in practice: in 54% of comparisons the model-generated draft was preferred, and it was deemed just as good as the historical questionnaire an additional 14% of the time, compared to a preference for the historical draft just 31% of the time (p=0.1).

Taken together, these results indicate that the model alone could out-perform the status quo fact finding process, yet paradoxically provided no appreciable overall benefits to the adjudicators working in the context of the trial. We see three potential explanations for these results: First, because of the relatively small set of adjudicators taking part in the trial and their non-random selection, it may be the case that the adjudicators participating in the trial had deeper level of

expertise than the average historical adjudicator in the data. Second, design aspects of the trial (such as the lack of time pressure from a backlog of claims or consolidated presentation of relevant data in our prototype interface) may have created favorable conditions for the adjudicators to carefully consider the details of a claim and draft thorough follow-up questions. Or, third, this may be evidence for a "Hawthorne Effect," whereby participation in the trial itself induced adjudicators to take more care with preparing their questionnaires than the might have otherwise. Consistent with all three explanations, comparisons between questionnaires written by adjudicators during the trial (without AI assistance) against the historical questionnaires (Appendix Figure A1) indicated that the QA team preferred the draft written during the trial 56% of the time (p=0.001; N=105) vs just 26% of the time for the historical questionnaire. Regardless of the underlying mechanism, the evidence here suggests the model itself performs similarly to an experienced adjudicator working under ideal conditions, and therefore would likely provide quality benefits in practice, especially in times of high caseloads and systemic strain when large numbers of inexperienced adjudicators are hired to fill gaps.

4.3 Reducing Inter-Adjudicator Variability

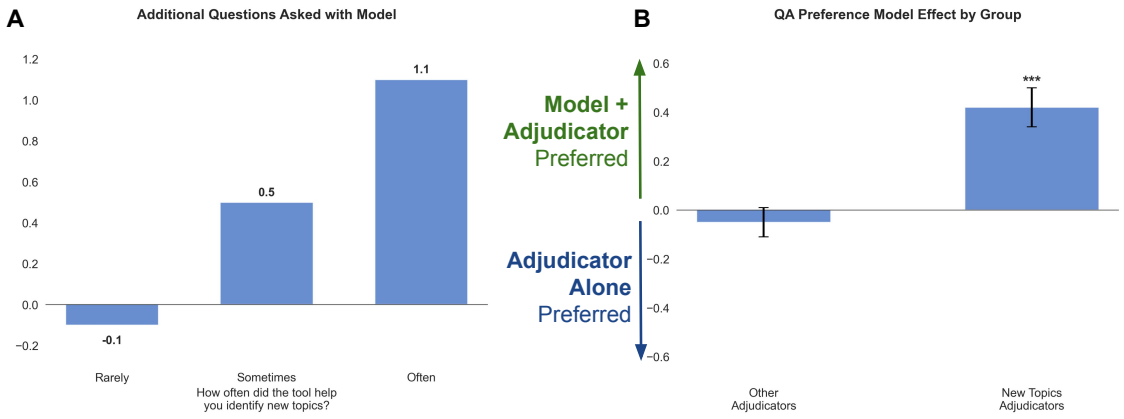

Fig. 5. AI Assistance May Help Reduce Quality Gaps Between Adjudicators. (A) Increases in questionnaire length when using the model were associated with identification of follow-up topics that might have been missed otherwise, with adjudicators who reported that they felt the model often helped them identify new topics asking 1.1 additional questions on average (p=0.01). (B) Among this group, the QA team strongly preferred questionnaires written with the assistance of the model to those written by the adjudicators alone (p<0.001).

Consistency in decision-making is a fundamental tenant of due process: the determination of whether you qualify for benefits should be a function only of the circumstances of your case and not the adjudicator who processed it. As such, we explored heterogeneity in the effects of our AI tool to understand whether it might help reduce variability across adjudicators.

We divided adjudicators into two groups based on their response to whether the model helped them identify new topics for follow-up: those who reported it "often" helped (N=2) versus those who reported "rarely" or "sometimes" (N=5). As shown in Figure 5(A), adjudicators who found the model helpful also asked more questions on average when using the model. However, increased fact finding length does not necessarily indicate improved quality. Therefore in Figure 5(B), we aggregate QA preferences on head-to-head comparisons of drafts written with and without AI assistance for each group, with the scale ranging from -2 (strongly preferred the draft written by an adjudicator alone) to +2

(strongly preferred the draft written with AI assistance). Consistent with the overall result (Figure 4(C)), there was no distinguishable difference in quality (as measured by QA preferences) for the adjudicators who reported that the the tool only rarely or sometimes helped them identify new topics. But for the adjudicators who reported that the tool "often" helped them identify new topics, and who asked additional questions as a result, we see see a strong preference by the QA team for AI-assisted drafts ($p<0.001$; $N=96$), indicating that the model meaningfully improved questionnaire quality among this group. Moreover, this improvement appears to close an existing quality gap: at baseline (that is, without assistance of the model), drafts written by other adjudicators were strongly favored over drafts written by this group, but this preference disappeared entirely when they were provided with the model (Appendix Figure A2). These results suggest that, although the tool may provide little benefit in terms of fact finding quality for more experienced adjudicators, it may nonetheless provide important benefits in reducing inter-adjudicator variability.

4.4 No Evidence for Over- or Under-Reliance on AI

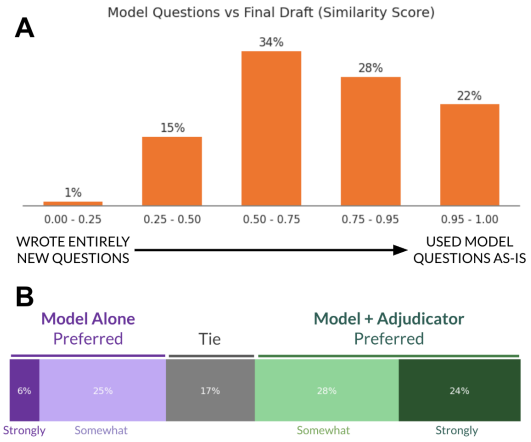

Fig. 6. No evidence for automation bias or algorithmic aversion. (A) Trigram similarity between initial AI-generated drafts and final drafts submitted by adjudicators indicates the collaborative nature of the human-AI system, with adjudicators infrequently using the model-generated draft without editing yet almost never disregarding this draft to develop an entirely new one. (B) QA team preferences across comparisons between questionnaires drafted by the model alone and those developed by the adjudicator with model assistance indicate the adjudicators improve upon these initial drafts ($p<0.001$; $N=202$).

One concern that is often voiced in the use of AI systems in high-stakes contexts like benefits determinations is whether humans might over- or under-rely on these tools. If, for instance, adjudicators were too trusting of model outputs — even when they may be wrong – the "humans in the loop" that are often posited as a safeguard against model errors might amount to little more than mere functionaries, a concern termed automation bias [3, 17, 26]. On the other hand, a general lack of trust in AI might lead adjudicators to ignore the system's useful outputs, even when there may be gains to be had by incorporating them, an issue referred to as algorithmic aversion [17]. Our detailed data allows us to examine examiner responses in detail, and we find little evidence that either mechanism is playing an appreciable role in our results.

Both the interaction data and QA preferences indicate that the adjudicators are taking an active role in refining and improving on the initial model drafts. By comparing the textual similarity between model-generated drafts and the

final versions submitted by adjudicators (Figure 6(A)), we see that only infrequently (22% of the time) do adjudicators use this initial model draft as-is. In most cases, they edited this initial draft to ensure it focused on the questions that would be most relevant to their determination. Likewise, we see that these edits improved the quality of the resultant questionnaires: the QA team showed a strong preference for drafts produced by adjudicators with AI assistance over those generated by the model alone (Figure 6(B); $p<0.001$). At the same time, we do not find evidence for algorithmic aversion: although adjudicators edited the AI-generated drafts, they almost never (only 1% of the time) disregarded them entirely, writing a new draft with little trigram similarity to the what the model had produced. Likewise, adjudicators consistently used topics suggested by the first model in prompting the second one (Appendix Figure A3) and QA preferences between adjudicators alone and the model alone favored the adjudicators. Although these interaction dynamics may vary with time in a longer-term deployment and would need to be continuously monitored, at least in the context of this field trial, the human adjudicators and AI models appeared to be working in a collaborative manner.

5 Discussion

Our demonstration explored whether a generative AI tool might help improve the efficiency and quality of decision-making in UI, a core component of the social safety net. Through iterative co-design and rigorous evaluation, we found a complex picture: while adjudicators overwhelmingly favored the system and found it subjectively helpful, the tool produced no measurable improvements in either time savings or average quality when adjudicators used it compared to working alone. However, two important patterns emerged from this seemingly null result. First, heterogeneous treatment effects suggest that the model may help less experienced adjudicators improve their fact finding quality, potentially reducing inter-adjudicator variability. Second, when evaluated in isolation, the model itself significantly outperformed historical questionnaires.

These results illustrate both the promise and complexity of AI tools in government service delivery. The disconnect between subjective satisfaction and lack of objective efficiency gains speaks to broader questions about human-AI collaboration that the field is only beginning to understand. For example, one recent randomized controlled trial examining AI assistance for software engineers found that the assistance simultaneously increased the number of vulnerabilities in the code *and* engineers' confidence that their code had no vulnerabilities [53]. In our setting, adjudicators spent considerable time editing model outputs, offsetting anticipated time savings — a pattern that challenges assumptions about AI as a simple productivity multiplier. That dynamic may evolve over time as adjudicators gain trust in the model and as the system itself learns from accumulated edits, potentially reducing the need for manual corrections in later phases. The heterogeneous effects we observed, where AI assistance may have helped close quality gaps between adjudicators, suggests the tool's greatest value may lie in establishing consistency rather than improving peak performance. This addresses fundamental due process concerns about benefits determinations varying based on which adjudicator handles a case rather than its merits.

Our findings still point to specific deployment scenarios where such tools could provide value despite modest overall effects. During system strain — when agencies face surging caseloads and must rapidly onboard inexperienced staff — the model could provide a quality baseline while helping experienced adjudicators manage overwhelming workloads. Additionally, given the model's strong performance against historical questionnaires, in-line integration when claimants are first applying could spot missing information in real-time, addressing the primary source of administrative delay: iterative back-and-forth between claimants and adjudicators. While these applications require further validation, they offer promising directions grounded in our empirical findings.

This work ultimately demonstrates why rigorous, context-situated evaluation is essential for responsible AI deployment in high-stakes settings. It confirms the ACUS recommendation that assessments of quality assurance must be multifaceted [37]. Our initial hypothesis centered on efficiency gains — a seemingly reasonable expectation that mirrors how vendors typically market these systems. Yet the trial revealed a more complex reality. This disconnect reflects a broader challenge in AI evaluation, where conventional benchmarks and isolated performance metrics fail to predict real-world effectiveness. During our collaboration, vendor demonstrations focused almost exclusively on processing speed without corresponding attention to decision quality: a narrow framing that risks accelerating incorrect determinations. Our evaluation framework deliberately balanced efficiency and accuracy metrics, revealing benefits like reduced inter-adjudicator variability that pure speed measures would miss.

The study's limitations also offer important lessons. Our small sample of relatively experienced adjudicators, recruited non-randomly by CDLE leadership, constrains generalizability.[2] Yet these constraints reflect the realities of evaluating AI tools in operational government settings; for example, synthetic evaluations using crowd workers would likely have missed the nuanced patterns we observed. The participating adjudicators' experience level, combined with quality improvements for those who found the tool most helpful, suggests our results may underestimate benefits. Moreover, our sandbox environment, while enabling responsible development, revealed inherent limitations of pre-deployment testing. The potential Hawthorne effect — where trial participants outperformed historical baselines without AI assistance — underscores that sandboxes, while essential for responsible pre-deployment testing, must be complemented by rigorous evaluation during actual deployment to capture real-world performance dynamics.

These evaluation challenges have direct implications for procurement and policy. Agencies face pressure to adopt AI solutions based on vendor promises and idealized demonstrations that may not reflect actual deployment conditions. While recent policy initiatives have called for innovation sandboxes and responsible AI frameworks, our experience shows that sandboxes alone are insufficient without evaluation methodologies that test contextual effectiveness. Nore are subjective user satisfaction metrics sufficient. The gap between our model's strong isolated performance and its modest impact when integrated into human workflows illustrates why agencies need graduated implementation strategies with continuous evaluation rather than binary deployment decisions.

In light of growing excitement around generative AI, governments face a critical juncture. These technologies offer genuine potential to address longstanding challenges in service delivery. However, realizing these benefits requires the following: targeted deployment for specific use cases where evidence supports value, recognition that human-AI collaboration involves complex trade-offs that vary across users and contexts, and commitment to field testing before procurement decisions where possible. Most critically, the disconnect between what seems reasonable in theory and what emerges in practice underscores why agencies must build capacity for and insist upon careful evaluation. Only through such assessment can governments distinguish vendor hype from genuine innovation, ensuring that AI adoption advances rather than undermines the fundamental purposes of public benefits administration.

Acknowledgments

We thank the teams at US DOL and CDLE for their partnership and willingness to allocate adjudicator and QA time to this project. We especially thank Brandon McClure, Phil Spesshardt, Amy Perez, and Nicole Zeichner for their work in setting up this project. We thank participants from the Office of Management Budget and the Council of Economic

[2] Our experiment may not be considered small for the context of within-government evaluations in which recruitment is particularly challenging; for example, one influential experiment on algorithm-assisted decisionmaking was conducted on only one judge [41].

Advisors for their valuable feedback during our White House briefing. We are grateful to Stanford Impact Labs and Arnold Ventures for supporting this research.

References

[1] 2019. HUD v Facebook. https://www.hud.gov/sites/dfiles/Main/documents/HUD_v_Facebook.pdf

[2] 2023. SNAP Case and Procedural Error Rates. https://www.fns.usda.gov/snap/qc/caper#1

[3] Saar Alon-Barkat and Madalina Busuioc. 2023. Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice. *Journal of Public Administration Research and Theory* 33, 1 (Jan. 2023), 153–169. doi:10.1093/jopart/muac007

[4] Kasun Amarasinghe, Kit T. Rodolfa, Hemank Lamba, and Rayid Ghani. 2023. Explainable machine learning for public policy: Use cases, gaps, and research directions. *Data & Policy* 5 (2023), e5. https://www.cambridge.org/core/journals/data-and-policy/article/explainable-machine-learning-for-public-policy-use-cases-gaps-and-research-directions/B5B66B3C3B16196482984E878D795161 Publisher: Cambridge University Press.

[5] David Ames, Cassandra Handan-Nader, Daniel E Ho, and David Marcus. 2020. Due Process and Mass Adjudication. *Stanford Law Review* 72, 1 (Jan. 2020). https://www.stanfordlawreview.org/print/article/due-process-and-mass-adjudication/

[6] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica, May* 23 (2016), 2016.

[7] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen Technical Report. doi:10.48550/arXiv.2309.16609 arXiv:2309.16609 [cs].

[8] Hayden P. Baker, Emma Dwyer, Senthooran Kalidoss, Kelly Hynes, Jennifer Wolf, and Jason A. Strelzow. 2024. ChatGPT's Ability to Assist with Clinical Documentation: A Randomized Controlled Trial. *Journal of the American Academy of Orthopaedic Surgeons* 32, 3 (February 2024), 123–129. doi:10.5435/JAAOS-D-23-00474

[9] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 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.

[10] Miranda Christ, Sarah Radway, and Steven M. Bellovin. 2022. Differential Privacy and Swapping: Examining De-Identification's Impact on Minority Representation and Privacy Preservation in the U.S. Census. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, San Francisco, CA, USA, 457–472. doi:10.1109/SP46214.2022.9833668

[11] Julian Christensen, Lene Aarøe, Martin Baekgaard, Pamela Herd, and Donald P. Moynihan. 2020. Human Capital and Administrative Burden: The Role of Cognitive Resources in Citizen-State Interactions. *Public Administration Review* 80, 1 (2020), 127–136.

[12] Danielle Keats Citron and Frank Pasquale. 2014. The scored society: Due process for automated predictions. *Wash. L. Rev.* 89 (2014), 1. Publisher: HeinOnline.

[13] Mallory E Compton, Matthew M Young, Justin B Bullock, and Robert Greer. 2023. Administrative Errors and Race: Can Technology Mitigate Inequitable Administrative Outcomes? *Journal of Public Administration Research and Theory* 33, 3 (July 2023), 512–528. doi:10.1093/jopart/muac036

[14] Amy Perez Daniel E. Ho, Olivia Martin and Kit Rodolfa. Forthcoming, 2025. Evaluation as Due Process: Civil Service in an Automated Age. *Admin. Law Review* (March Forthcoming, 2025).

[15] Tri Dao. 2023. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning. *arXiv preprint arXiv:2307.08691* (2023).

[16] Jeffrey Dastin. 2022. Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women *. In *Ethics of Data and Analytics*. Auerbach Publications. Num Pages: 4.

[17] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–12.

[18] Matthew J. Duggan, Julietta Gervase, Anna Schoenbaum, et al. 2025. Clinician Experiences With Ambient Scribe Technology to Assist With Documentation Burden and Efficiency. *JAMA Network Open* 8, 2 (2025), e2460637. doi:10.1001/jamanetworkopen.2024.60637 Published Online: February 19, 2025.

[19] David Freeman Engstrom and Daniel E. Ho. 2021. Artificially intelligent government: a review and agenda. In *Research Handbook on Big Data Law*. 57–86.

[20] Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Publishing Group. Google-Books-ID: pn4pDwAAQBAJ.

[21] Executive Office of the President. 2020. *Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government*. Technical Report E.O. 13960. https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government

[22] Ryan Felton. 2016. Michigan unemployment agency made 20,000 false fraud accusations – report. *The Guardian* (Dec. 2016). https://www.theguardian.com/us-news/2016/dec/18/michigan-unemployment-agency-fraud-accusations

[23] Riccardo Fogliato, Alice Xiang, Zachary Lipton, Daniel Nagin, and Alexandra Chouldechova. 2021. On the Validity of Arrest as a Proxy for Offense: Race and the Likelihood of Arrest for Violent Crimes. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. Association for Computing Machinery, New York, NY, USA, 100–111. doi:10.1145/3461702.3462538

[24] GAO. 2008. *Federal Workers' Compensation: Better Data and Management Strategies Would Strengthen Efforts to Prevent and Address Improper Payments*. Technical Report GAO-01-1016. Government Accountability Office.

[25] Kurt Glaze, Daniel E. Ho, Gerald K. Ray, and Christine Tsang. 2022. Artificial Intelligence for Adjudication: The Social Security Administration and AI Governance. In *The Oxford Handbook of AI Governance*. Oxford University Press, Handbook on AI Governance (forthcoming).

[26] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association* 19, 1 (Jan. 2012), 121–127. doi:10.1136/amiajnl-2011-000089

[27] Aidan Gomez. 2024. Command R: Retrieval-Augmented Generation at Production Scale. Cohere Blog. https://cohere.com/blog/command-r

[28] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal,

Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The Llama 3 Herd of Models. doi:10.48550/arXiv.2407.21783 arXiv:2407.21783 [cs].

[29] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24. Publisher: ACM New York, NY, USA.

[30] Nick Gwyn. 2022. *Historic Unemployment Programs Provided Vital Support to Workers and the Economy During Pandemic, Offer Roadmap for Future Reform.* Technical Report. Center on Budget and Policy Priorities, Washington, D.C. https://www.cbpp.org/research/economy/historic-unemployment-programs-provided-vital-support-to-workers-and-the-economy

[31] Daniel Han and Michael Han. 2024. Unsloth: Faster LLM Fine-tuning. GitHub repository. https://github.com/unslothai/unsloth

[32] Nouzha Harrati, Imed Bouchrika, Abdelkamel Tari, and Ammar Ladjailia. 2016. Exploring user satisfaction for e-learning systems via usage-based metrics and system usability scale analysis. *Computers in Human Behavior* 61 (2016), 463–471. doi:10.1016/j.chb.2016.03.051

[33] Pamela Herd and Donald Moynihan. 2019. Administrative Burdens in the Social Safety Net. *Journal of Economic Perspectives* (2019).

[34] Daniel E. Ho. 2017. Does Peer Review Work? An Experiment of Experimentalism. *Stanford Law Review* 69, 1 (Jan. 2017). https://www.stanfordlawreview.org/print/article/does-peer-review-work-an-experiment-of-experimentalism/

[35] Daniel E. Ho, Cassandra Handan-Nader, David Ames, and David Marcus. 2019. Quality Review of Mass Adjudication: A Randomized Natural Experiment at the Board of Veterans Appeals, 2003–16. *Journal of Law, Economics, and Organization* 35, 2 (2019), 239–294. doi:10.1093/jleo/ewz001 Advance Access published March 29, 2019.

[36] Daniel E Ho, Cassandra Handan-Nader, David Ames, and David Marcus. 2019. Quality Review of Mass Adjudication: A Randomized Natural Experiment at the Board of Veterans Appeals, 2003–16. *The Journal of Law, Economics, and Organization* 35, 2 (July 2019), 239–288. doi:10.1093/jleo/ewz001

[37] Daniel E. Ho, David Marcus, and Gerald K. Ray. 2021. *Quality Assurance Systems in Agency Adjudication: Emerging Practices and Insights.* Technical Report. Administrative Conference of the United States. https://www.acus.gov Report prepared for the Administrative Conference of the United States (ACUS).

[38] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2023. Measures for Explainable AI: Explanation Goodness, User Satisfaction, Mental Models, Curiosity, Trust, and Human-AI Performance. *Frontiers in Computer Science* 5 (2023), 1096257. doi:10.3389/fcomp.2023.1096257 Published February 5, 2023.

[39] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685* (2021). https://doi.org/10.48550/arXiv.2106.09685

[40] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2022. Evaluation Gaps in Machine Learning Practice. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (2022), 1859–1876.

[41] Kosuke Imai, Zhichao Jiang, D James Greiner, Ryan Halen, and Sooahn Shin. 2023. Experimental evaluation of algorithm-assisted human decision-making: application to pretrial public safety assessment*. *Journal of the Royal Statistical Society Series A: Statistics in Society* 186, 2 (May 2023), 167–189. doi:10.1093/jrsssa/qnad010

[42] Jennifer King, Daniel Ho, Arushi Gupta, Victor Wu, and Helen Webley-Brown. 2023. The Privacy-Bias Tradeoff: Data Minimization and Racial Disparity Assessments in U.S. Government. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 492–505. doi:10.1145/3593013.3594015

[43] Tsai-Ling Liu, Timothy C. Hetherington, Ajay Dharod, Tracey Carroll, Richa Bundy, Hieu Nguyen, Henry E. Bundy, McKenzie Isreal, Andrew McWilliams, and Jeffrey A. Cleveland. 2024. Does AI-Powered Clinical Documentation Enhance Clinician Efficiency? A Longitudinal Study. *NEJM AI* 1, 12 (2024), –. doi:10.1056/Aloa2400659 Published November 22, 2024.

[44] Olivia Martin, Faiz Surani, Kit Rodolfa, Amy Perez, and Daniel E. Ho. 2024. The Spectrum of AI Integration: The Case of Benefits AdjudicationOlivia Martin, Faiz Surani, Kit Rodolfa, Amy Perez, and Daniel E. Ho). In *AI: Legal Issues, Policy, and Practical Strategies*, Cynthia H. Cwik, Christopher A. Suarez, and Lucy L. Thomson (Eds.). American Bar Association.

[45] Jerry Mashaw. 1974. The Management Side of Due Process: Some Theoretical and Litigation Notes on the Assurance of Accuracy, Fairness and Timeliness in the Adjudication of Social Welfare Claims. *Faculty Scholarship Series* (Jan. 1974). https://openyls.law.yale.edu/handle/20.500.13051/353

[46] Jerry L. Mashaw. 1983. *Bureaucratic Justice: Managing Social Security Disability Claims.* Yale University Press. https://www.jstor.org/stable/j.ctt1dt009d

[47] Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. 2023. Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity Theory. doi:10.48550/arXiv.2310.17884 arXiv:2310.17884 [cs].

[48] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable Extraction of Training Data from (Production) Language Models. doi:10.48550/arXiv.2311.17035 arXiv:2311.17035 [cs].

[49] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453. https://doi.org/10.1126/science.aax2342

[50] White House Office of Science and Technology Policy. 2022. *Blueprint for an AI Bill of Rights.* Whitepaper. https://www.whitehouse.gov/ostp/ai-bill-of-rights/

[51] Jennifer Pahlka. 2023. *Recoding America: Why Government Is Failing in the Digital Age and How We Can Do Better.* Macmillan.

[52] Nicholas R. Parrillo (Ed.). 2017. *Administrative Law from the Inside Out: Essays on Themes in the Work of Jerry L. Mashaw.* Cambridge University Press, Cambridge. doi:10.1017/9781107159518 Publication date: March 23, 2017.

[53] Neil Perry, Megha Srivastava, Deepak Kumar, and Dan Boneh. 2023. Do Users Write More Insecure Code with AI Assistants?. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23).* ACM, New York, NY, USA, 16. doi:10.1145/3576915.3623157

[54] Deborah Plana, Dennis L. Shung, Alyssa A. Grimshaw, Anurag Saraf, Joseph J. Y. Sung, and Benjamin H. Kann. 2022. Randomized Clinical Trials of Machine Learning Interventions in Health Care: A Systematic Review. *JAMA Network Open* 5, 9 (Sept. 2022), e2233946. doi:10.1001/jamanetworkopen.2022.33946

[55] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. 469–481.

[56] Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The Fallacy of AI Functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22).* Association for Computing Machinery, New York, NY, USA, 959–972. doi:10.1145/3531146.3533158

[57] Gerald Ray and Glenn Sklar. 2019. *An Operational Approach to Eliminating Backlogs in the Social Security Disability Program.* Technical Report. McCrery-Pomeroy SSDI Solutions Initiative. 40 pages. http://www.crfb.org/project/ssdi/operational-approach-eliminating-backlogs-social-securitydisability-program

[58] Gerald K. Ray and Jeffrey S. Lubbers. 2014. Government Success Story: How Data Analysis by the Social Security Appeals Council (with a Push from the Administrative Conference of the United States) Is Transforming Social Security Disability Adjudication. *Geo. Wash. L. Rev.* 83 (2014), 1575.

[59] Alexis R. Santos-Lozada, Jeffrey T. Howard, and Ashton M. Verdery. 2020. How differential privacy will affect our understanding of health disparities in the United States. *Proceedings of the National Academy of Sciences* 117, 24 (June 2020), 13405–13412. doi:10.1073/pnas.2003714117 Company: National Academy of Sciences Distributor: National Academy of Sciences Institution: National Academy of Sciences Label: National Academy of Sciences Publisher: Proceedings of the National Academy of Sciences.

[60] Maxime Sasseville, Farzaneh Yousefi, Steven Ouellet, Florian Naye, Théo Stefan, Valérie Carnovale, Frédéric Bergeron, Linda Ling, Bobby Gheorghiu, Simon Hagens, Samuel Gareau-Lajoie, and Annie LeBlanc. 2025. The Impact of AI Scribes on Streamlining Clinical Documentation: A Systematic Review. *Healthcare (Basel)* 13, 12 (2025), 1447. doi:10.3390/healthcare13121447 Published June 16, 2025, Editor: Daniele Giansanti, PMID: 40565474, PMCID: PMC12193156.

[61] Daniela Sele and Marina Chugunova. 2024. Putting a human in the loop: Increasing uptake, but decreasing accuracy of automated decision-making. *PLOS ONE* 19, 2 (2024), e0298037. doi:10.1371/journal.pone.0298037

[62] Yashothara Shanmugarasa, Ming Ding, Chamikara Mahawaga Arachchige, and Thierry Rakotoarivelo. 2025. SoK: The Privacy Paradox of Large Language Models: Advancements, Privacy Risks, and Mitigation. In *Proceedings of the 20th ACM Asia Conference on Computer and Communications Security (ASIA CCS '25).* Association for Computing Machinery, New York, NY, USA, 425–441. doi:10.1145/3708821.3733888

[63] Robyn Tamblyn, Nadyne Girard, James Hanley, Bettina Habib, Adrian Mota, Karim M. Khan, and Clare L. Ardern. 2023. Ranking versus rating in peer review of research grant applications. *PLOS ONE* 18, 10 (Oct. 2023), e0292306. doi:10.1371/journal.pone.0292306 Publisher: Public Library of Science.

[64] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[65] U.S. Department of Labor, Employment & Training Administration (ETA). 2024. Benefits: Timeliness and Quality Reports. https://oui.doleta.gov/unemploy/btq/btqrpt.asp

[66] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. 2024. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour* 8, 12 (2024), 2293–2303. doi:10.1038/s41562-024-02024-1

[67] Laura Weidinger, Inioluwa Deborah Raji, Hanna Wallach, Margaret Mitchell, Angelina Wang, Olawale Salaudeen, Rishi Bommasani, Deep Ganguli, Sanmi Koyejo, and William Isaac. 2025. Toward an Evaluation Science for Generative AI Systems. *arXiv preprint arXiv:2503.05336* (2025). First two authors contributed equally.

[68] Shalanda Young. 2024. *Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence.* Memorandum for the Heads of Executive Departments and Agencies M-24-10. Office of Management and Budget, Washington, D.C. 34 pages. https://www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Use-of-Artificial-Intelligence.pdf

[69] Fan Zhang, Jiaxin Mao, Yiqun Liu, Xiaohui Xie, Weizhi Ma, Min Zhang, and Shaoping Ma. 2020. Models Versus Satisfaction: Towards a Better Understanding of Evaluation Metrics. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20).* Association for Computing Machinery, New York, NY, USA, 379–388. doi:10.1145/3397271.3401162

A Appendix

Section	Questions (Response Scale)
General Impressions	(1) Overall, how useful did you find the AI tool? (1 = Not at all useful, 5 = Very useful) (2) If the AI tool were available in myUI+, how likely would you be to use it on a regular basis? (1 = Would not use on any issues, 5 = Would use on most issues)
Follow-up Topics	(1) How often were the suggested topics identified by the model relevant to the issue? (1 = Never relevant, 5 = Always relevant) (2) Did the model help you identify topics for follow-up that you might not have identified on an initial read of the issue? (1 = Didn't help on any issue, 5 = Helped on every issue) (3) How often did you feel like you needed to add follow-up topics that the model missed? (1 = Never, 5 = Always)
Draft Questions	(1) Overall, did you feel like the AI tool saved you time in preparing follow-up fact findings? (1 = Strongly Disagree, 5 = Strongly Agree) (2) How would you rate the overall quality of the draft questions generated by the model for eliciting the needed follow-up information? (1 = Very Poor, 5 = Excellent) (3) Did the tone and style of the draft questions feel appropriate for corresponding with claimants or employers? (1 = Never Appropriate, 5 = Always Appropriate) (4) In general, how specific to the details of the issue were the draft questions? (1 = Much Too Vague, 5 = Highly Specific) (5) How often did you feel like you needed to edit the draft follow-up questions or add new questions? (1 = Never, 5 = Always)
General Feedback	(1) Are there improvements to the tool that would make it more useful? What would be your ideal tool to help make your job easier? (2) Any other thoughts or suggestions you'd like to share?

Appendix Table A1. Adjudicator Feedback Survey Instrument

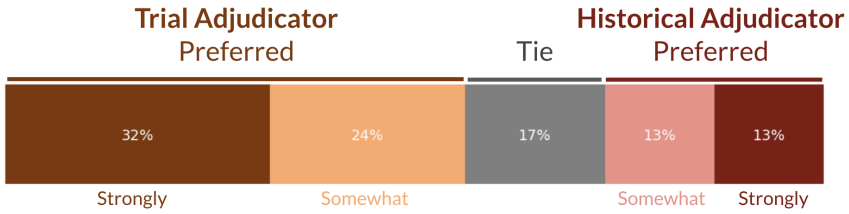

Fig. A1. Questionnaires written by adjudicators participating in the trial (without assistance of the AI model) were strongly preferred over the historical questionnaires actually sent when the claim was adjudicated in practice (p=0.001; N=105).

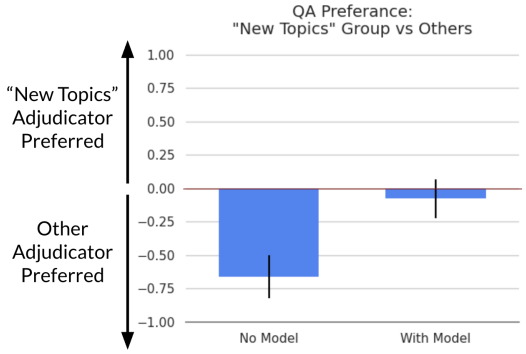

Fig. A2. At baseline (without AI assistance), drafts developed by the adjudicators who reported the model was particularly helpful in identifying new follow-up topics significantly under-performed relative to other adjudicators. However, when provided with the model, this quality gap closes entirely (p=0.001; N=138).

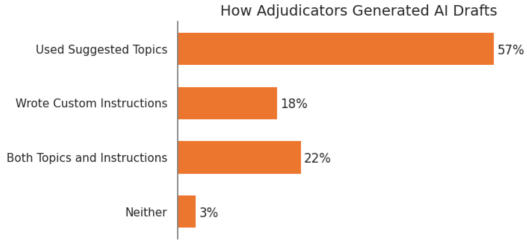

Fig. A3. Adjudicators made use of both model-suggested topics and custom instructions in prompting the question-drafting model.