# Courts of Tomorrow:

## Evidence from a Nationwide Introduction of Generative AI

Sultan Mehmood ⓡ Christoph Goessmann ⓡ Elliott Ash*

February 7, 2026

**Abstract**

We present the first large-scale field experiment on integrating generative AI into a national judicial system. In partnership with Pakistan's judiciary, we enrolled about half of Pakistan's lower judiciary, or 1,559 trial-court judges, in a randomized controlled trial of a custom AI assistant, *JudgeGPT*. Judges were assigned to one of three arms: (i) Assistant + targeted training ($n = 1{,}197$), which bundled JudgeGPT access with assistant-specific instruction and risk mitigation; (ii) Assistant only ($n = 180$), which provided access plus a general placebo course on technology and AI in legal practice; and (iii) Control ($n = 182$), which received only the placebo course. Treated judges are more likely to use generative AI and to support broader AI adoption in the judiciary. Using an LLM-based measure of opinion quality, we show that Assistant + targeted training improves writing quality, while Assistant only reduces it relative to control. Administrative records further indicate productivity gains: when 20% of judges in a district receive Assistant + targeted training, case disposals increase by 0.2 standard deviations, about 4,000 additional cases per year. Overall, our results suggest that generative AI can increase judicial productivity and state capacity in resource-constrained settings, but that training is essential to realize these gains without degrading quality.

*Acknowledgments: We thank Justice Mansoor Ali Shah of the Supreme Court of Pakistan for his feedback, guidance, and support on the development and roll-out of JudgeGPT. We also thank the Pakistan's Judicial Academy for their feedback on the AI tool, and Georgiy Marinichev and Tim Ehrensperger for excellent research assistance. Finally, we thank W. Bentley MacLeod and participants at the American Economics Association for helpful feedback. **Ethics and approvals:** The study received approval from the ETH Zurich Ethics Commission (EK-2023-N-343-A) and the Institute of Business Administration (IBA) Karachi Institutional Review Board (IBA-Esub006/rec-24). **Preregistration:** The design and analysis were pre-specified in the AEA RCT Registry (Project ID AEARCTR-0012906). **Author contributions:** Authors contributed equally and order is randomized. **Contact:** Sultan Mehmood, The New Economic School, smehmood@nes.ru; Christoph Goessmann, Imperial College London, c.goessmann@imperial.ac.uk; Elliott Ash, ETH Zurich, ashe@ethz.ch.

# 1. INTRODUCTION

How does AI shape the productivity of the State? Rapid advances in generative AI have revived the idea that algorithmic tools can transform how governments reason, prioritize, and allocate, with consequences for economic development (Gans, 2025) and institutional design (Acemoglu, Ozdaglar, & Siderius, 2025). From medicine and law to education and public administration, large language models now offer real-time support in drafting, summarizing, and interpreting complex information. Yet we still lack credible evidence on whether AI actually improves state effectiveness. The judiciary is a central test case: court decisions determine liberty and property rights, set enforcement priorities, and anchor the rule of law.

This paper presents the first large-scale field experiment evaluating the integration of generative AI into judicial decision-making. In partnership with Pakistan's national judiciary, we randomly assigned over 1559 trial court judges in the country to receive access to JudgeGPT, custom generative AI tool attached to the country's own corpus of court rulings via Retrieval Augmented Generation (RAG). A randomly selected subset of judges also received targeted instruction on how to use the tool responsibly. The intervention was implemented through the Pakistan Judicial Academy's national "Technology and Law" program and closely follows a pre-registered research protocol and pre-analysis plan.

We document four main findings. First, providing judges access to JudgeGPT increases stated support for adopting generative AI in judicial work. On a 1–5 scale, access with training raises support by about 0.24 points (about 7 percent of the mean), with a similar increase for access without training, indicating that even limited exposure shifts attitudes toward adoption.

Second, perceived productivity improvements are concentrated among trained judges. Judges assigned to JudgeGPT and structured training become more likely to agree that AI can improve their productivity, with an effect of about 0.16 points (roughly 4 percent of the mean), while estimates for access without training are smaller and less precise.

Third, we look at the quality of opinions written before and after treatment, using a novel LLM annotation method based on pairwise comparisons. We find that AI and training is associated with increased opinion quality relative to the placebo group without tools or training. In the group that got AI tools but without training, opinion quality actually decreased.

Fourth, these shifts coincide with substantial increases in realized case disposal. Using district-year administrative statistics, districts with greater exposure to JudgeGPT and training close about 15 percent more cases relative to the mean, implying an increase of roughly 4,139 cases per district-year. Specifically, when 20% of judges in a

district receive Assistant + targeted training, case disposals increase by 0.2 standard deviations, about 4,000 additional cases per year.

To shed light on how judges use JudgeGPT, and whether the increase in case disposal comes at the expense of decision writing, we analyze usage logs of JudgeGPT and post-period judgment texts. Training changes the composition of requests made to the model. The share of messages devoted to legal research falls by about 7.8 percentage points, which is roughly a 25 percent decline relative to the mean share. At the same time, the share of prompts requesting writing support rises by about 2.8 percentage points, an 82 percent increase relative to the mean share. We find no meaningful change in non-legal use. This pattern suggests that training shifts judges away from treating the tool as a research engine and toward using it for structured drafting assistance, which is a possible channel for the large increase in cases closed.

These results add, first, to the growing body of work on the role of AI in justice systems, particularly debates around the capacity of algorithmic tools to promote fairness, consistency, and efficiency in legal decision-making (Kleinberg, Ludwig, Mullainathan, & Rambachan, 2018; Ludwig & Mullainathan, 2021). While much of the existing literature focuses on comparing algorithmic outputs with human judicial decisions to document bias and inconsistency (Kleinberg, Lakkaraju, Leskovec, Ludwig, & Mullainathan, 2018), this paper takes a different approach. We embed a large-scale field experiment in Pakistan to evaluate an AI assistant designed to support legal research and judgment writing. The intervention, delivered through the national judicial training academy program, includes targeted instruction to promote responsible use and mitigate concerns about bias and opacity in AI systems. Such concerns have been extensively documented both in general AI ethics literature (Acemoglu, Ozdaglar, & Siderius, 2023; Buolamwini & Gebru, 2018; Vicente & Matute, 2023) and in studies specifically addressing legal applications (Yang & Dobbie, 2020). By focusing on augmentation rather than substitution, our approach aspires to harness the benefits of AI while addressing the critical challenges that have limited its broader acceptance in judicial contexts.

Second, we contribute to the literature on how generative AI affects productivity in the public sector. While existing studies document substantial gains from AI in private-sector settings (Brynjolfsson, Li, & Raymond, 2025; Noy & Zhang, 2023), where market competition and profit motives drive adoption, the public sector presents a fundamentally different institutional environment (Agrawal, Gans, & Goldfarb, 2019; Muralidharan, Singh, & Ganimian, 2019). Frontline officials typically face weaker performance incentives, stronger tenure protections, and limited accountability. Adoption is rarely demand-driven and often stems from top-down mandates. These structural features raise the possibility that AI may fail to generate similar productivity improvements. Despite these concerns, empirical evidence from the public sector re-

mains scarce. This study offers one of the first large-scale field experiments evaluating the integration of generative AI in a core state institution. Our findings highlight that productivity gains of AI might be masked by general equilibrium effects (Acemoglu, Autor, Hazell, & Restrepo, 2022; Brynjolfsson, Rock, & Syverson, 2017).

Third, our paper contributes to the growing discourse on judicial effectiveness by examining a critical but relatively understudied dimension: the role of new technological tools and training in shaping judicial productivity. While earlier work has focused on institutional structures such as court design, appointment procedures, and administrative capacity (Ash & MacLeod, 2024; Landes & Posner, 2009; Liu, Lu, Peng, & Wang, 2022; Mehmood, 2022), as well as behavioral drivers of judicial performance (de Janvry, He, Sadoulet, Wang, & Zhang, 2023), empirical studies on public sector innovation have rarely examined the judiciary (Acemoglu, Akcigit, & Celik, 2022; Fujiwara, 2015; Muralidharan, Niehaus, & Sukhtankar, 2016). This absence is striking given that courts play a central role in enforcing constitutional rules, resolving disputes, and enabling reforms in areas like electoral governance and financial oversight. By evaluating the effects of generative AI tools and associated training on trial court judges, our study provides new evidence on how technological interventions can interact with institutional features in high-stakes public settings, especially given increasing pressure to adopt AI innovations (Pah et al., 2022).

The remainder of the paper proceeds as follows. Section 2 outlines the institutional background and the context. Section 3 introduces the data and empirical design. Section 4 presents the main results. Section 5 explores heterogeneity and reports robustness checks. Section 6 concludes.

## 2. Institutional Background

**Colonial Origins.** Pakistan's legal institutions trace their roots to the British colonial era, during which a common law system was established to serve the administrative needs of the Raj. This framework emphasized formal legal procedures, hierarchical authority, and the doctrine of precedent, laying the foundation for modern statutory interpretation and judicial reasoning. Following independence in 1947, Pakistan retained much of this legal infrastructure, including its multi-tiered court structure, procedural codes, and the use of English in legal proceedings. Although the post-colonial state introduced new legal statutes and Islamic jurisprudential elements over time, the underlying institutional architecture, particularly at the level of trial courts, continues to reflect colonial-era principles of adjudication, case management, and bureaucratic legalism. These historical legacies continue to shape the operation and challenges of the judiciary today (Siddique, 2013).

**Structure and Mandate of Trial Courts.** At the base of Pakistan's judiciary lie the district-level trial courts, which constitute the institutional core of legal service delivery. These courts include three key judicial ranks: Civil Judges, who primarily handle civil disputes such as contracts, land, and family law; Sessions Judges, who preside over criminal matters; and Additional Sessions Judges, who assist in adjudicating serious offenses such as murder, sexual violence, and public corruption. Trial courts are the primary forum through which ordinary citizens interact with the legal system. For most individuals, trial courts represent not only the first point of contact with the judiciary but also the only realistic opportunity to access justice. When these institutions fail to deliver timely decisions, the consequences can be severe (Butt, 2025).

**Institutional Constraints.** Trial court judges in Pakistan operate under formal rules and procedures, yet face severe institutional limitations. Most courts lack modern case management systems and legal research tools. Moreover, judges are often overburdened with high caseloads, and support staff is minimal. For instance, Pakistan employs less than 2 judges per 100,000 inhabitants (International Development Law Organization, 2018) compared to 22 per 100,000 in the European Union (Council of Europe, 2024). These frictions contribute to systemic delay: as of 2023, more than 2 million cases remained pending across the country's trial courts, with substantial variation across regions (Law and Justice Commission of Pakistan, 2023). Trial judges are recruited as career civil servants through competitive examinations administered by provincial public service commissions. Once appointed, they receive strong tenure protection and are rarely dismissed. However, judges operate within rigid bureaucratic hierarchies and face limited performance-based incentives. Unlike in private sector settings, promotions are slow and rarely contingent on case throughput or decision quality.

**Context for Technological Reform.** In recent years, Pakistan's judiciary has undertaken a series of institutional reforms to modernize its operations and address longstanding inefficiencies. A centerpiece of these reforms is the Technology and Law Initiative, a national program launched in 2023 to digitize case records, improve administrative processes, and integrate emerging technologies into judicial workflows. At the same time, global attention to advances in generative artificial intelligence, especially the adoption of tools like ChatGPT, sparked interest within Pakistan's judiciary about the feasibility of using similar systems to assist judges. In response, the judiciary began preparing for a phased introduction of AI-based legal tools aimed at improving case management, legal reasoning, and judgment writing. Our study was implemented within this broader context. In partnership with the Federal Judicial Academy of Pakistan (FJA), we designed a customized legal assistant based on generative AI, drawing on national case law and legal texts. Access to this system was randomly assigned to

a large group of trial court judges across Pakistan, alongside structured training sessions focused on responsible and effective use. The intervention reached roughly half of all sitting trial judges, providing a unique opportunity to assess how generative AI can shape decision-making and productivity within a public sector institution at scale.

To support this rollout and mitigate potential risks, such as misuse, over-reliance, or opacity, we collaborated with judicial authorities to develop a domain-specific large language model, JudgeGPT, trained on national legal texts. We designed and embedded a randomized controlled trial into the deployment process, assigning judges to receive structured access to the tool, with a subset also receiving training in responsible use. This setting provides a distinct opportunity to evaluate the effects of generative AI on frontline decision-making within a public institution, in a context where policy reform rather than market incentives drive adoption. In *Constitutional Petition. No. 1010-L*, the Supreme Court of Pakistan formally acknowledged the use of generative AI tools such as JudgeGPT in trial courts. The judgment recognized the potential of such co-developed technologies to support judicial functioning, while emphasizing the importance of regulatory caution and institutional oversight (Supreme Court of Pakistan, 2024).

## 3. Research Design

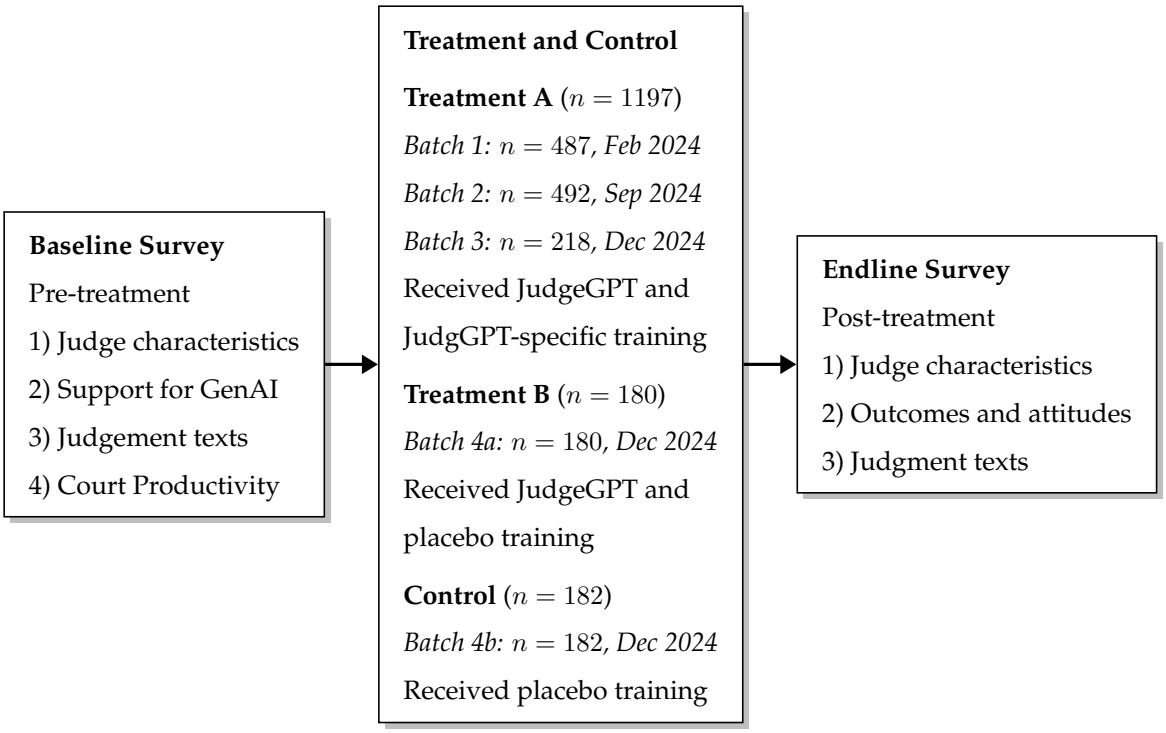
The randomized controlled trial comprises two treatment and one control arm, as illustrated in Figure I:

Treatment A ($n = 1197$) pairs access to the AI assistant JudgeGPT with targeted training on using the assistant effectively. Treatment B ($n = 180$) pairs access to JudgeGPT with a general placebo course on technology and AI use in legal practice, without any JudgeGPT-specific instruction. The control group ($n = 182$) does not receive access to JudgeGPT but only the placebo course.

This design, in which a subset of judges receive JudgeGPT-specific training and others a placebo course only, allows us to separate the effects of access to the AI assistant from those associated with structured guidance on the integration of JudgeGPT into legal workflows.

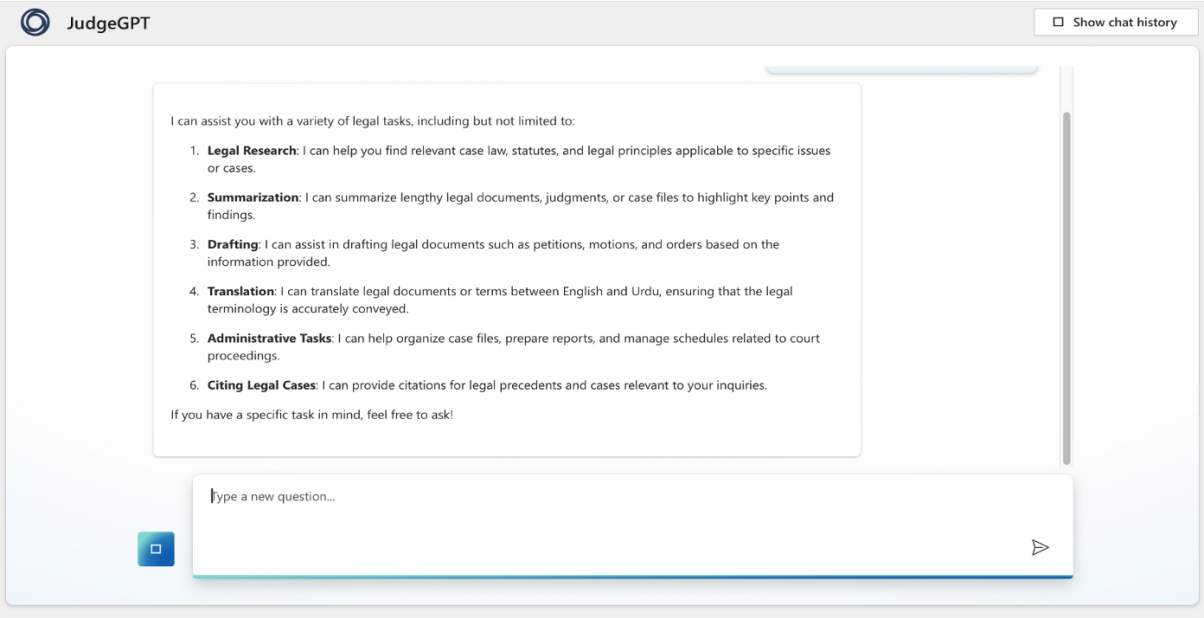
We first randomized, stratified by age and province, 979 judges into Treatment A via two batches. Batch 1 ($n = 487$) received JudgeGPT and the JudgeGPT-specific training from February 2024 onward, while Batch 2 ($n = 492$) was treated from September 2024 onward. Following that, by demand from the judiciary, we recruited an additional 580 judges and randomized them into Treatment A (Batch 3, $n = 218$), Treatment B (Batch 4a, $n = 180$), and Control (Batch 4b, $n = 182$) from January 2025 onward. We share more information regarding timing in Table I.

FIGURE I: Structure of Study




Note: Illustration of the study's treatment arms and timing combined with data types collected via baseline and endline surveys.

FIGURE II: JudgeGPT User Interface




Note: Depiction of JudgeGPT's user interface.

## 3.1. *JudgeGPT*

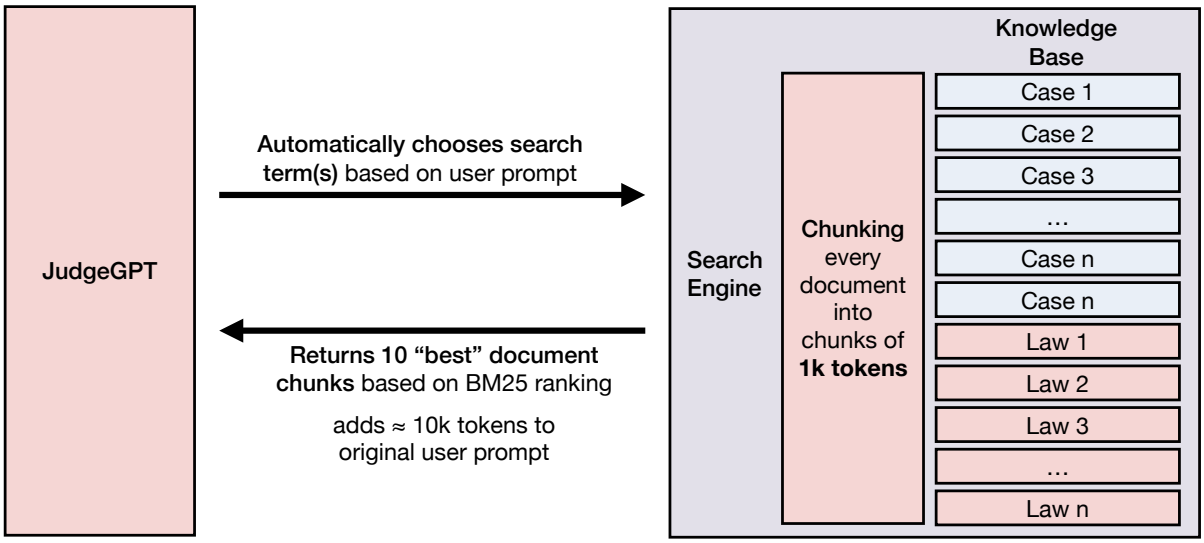
JudgeGPT is a chatbot based on OpenAI's GPT-4 family of models[1] that we customized to the Pakistani context by (1) adding a context-specific system prompt and (2) attaching a corpus of approximately 130,000 Pakistani judicial opinions and laws via Retrieval Augmented Generation (RAG). For illustration, we show JudgeGPT's user interface in Figure II.

The *system prompt*, not visible to users and protected from modification, primes the model for specific tasks, helps avoid repetition, sets standards, and acts as a guardrail against undesired and unsafe model behavior. In the case of JudgeGPT, we instructed the model in the following way:

> `You are JudgeGPT. You are a helpful legal assistant helping a Pakistan judge in all of his/her work, including for example legal research, summarization, drafting, translation, and any related administrative tasks. For that purpose, you are given Pakistani court cases with the full text of judgements.`
>
> `Answer questions based on the text and information I provide now, current Pakistan law, and the cases provided to you. Cite legal cases where appropriate and make sure to only cite cases given to you.`

[1] At the beginning of the study JudgeGPT used GPT-4-turbo 1106 followed by GPT-4o-mini from August 2024 on. Both models were set to temperature 0.

FIGURE III: JudgeGPT RAG




Note: Illustration of the Retrieval Augmented Generation (RAG) module that JudgeGPT relies on to respond to user requests.

```
Legal jargon is OK.
```

The system prompt's second paragraph instructs the model to primarily rely on information from our *RAG knowledge base* of Pakistani court cases and laws instead of any internalized model knowledge.

The underlying process is illustrated in Figure III. Whenever the user sends a request to JudgeGPT, the assistant decides if it needs to request data from the knowledge base. If that is the case, a search is conducted and the RAG system returns the 10 most relevant document chunks to be used by JudgeGPT to generate its response. In its response, JudgeGPT then cites the corresponding document chunks via footnotes in the appropriate places and lets the user click on those and verify any RAG content it cited from. We show an example of such a citation in Figure A1.

The knowledge base is made up of 129,235 documents in total. Those are 128292 judicial opinions from the Supreme Court of Pakistan ($n = 2670$), Lahore High Court ($n = 7539$), the High Court of Sindh ($n = 8627$), and easylaw.ai ($n = 109456$); plus an additional 943 Pakistani laws, acts, and ordinances.

In the RAG module, documents are split into chunks of 1000 tokens each—the approximate equivalent of two pages of text—with a 128 token overlap to preserve context across chunks. We decided to chunk documents so as to not overload the model's context window[2] and also to limit cost, as model usage is billed on a per-

[2]GPT-4-turbo 1106 and GPT-4o-mini, the models that JudgeGPT used, both have context windows of 128k tokens. That corresponds to approximately 250 pages of text. The chat conversation with the AI assistant, up to the last 10 messages, needs to fit into this context window.

token basis.

### 3.2. *Training Courses*

The Treatment A training course was designed to facilitate effective use of the tool in daily judicial practice. Judges in the treatment group followed a structured curriculum that addressed both the capabilities and limitations of generative AI in legal settings, combining lectures, ten practical home assignments, and a final project. As part of the final project, judges used *JudgeGPT* for tasks such as drafting judgments and conducting legal research, with several assignments specifically aimed at training participants to identify and manage hallucinated or fabricated outputs. To encourage engagement, the top one percent of participants were invited to an international conference on law and AI, and all successful participants received certificates jointly issued by ETH Zurich. Participation was high, with over 95 percent of judges submitting at least one assignment.

Details on the training content are available in Table A1. Panel A presents a sample of the instructional slides shown to judges exposed to Treatment A, while Panel B shows a sample from the placebo course included in Treatment B and the control. The notes accompanying the figure include additional information and direct links to the full set of slides used in both the treatment and placebo conditions.

### 3.3. *Data*

Our data are drawn from three primary sources that together offer a comprehensive view of judges' characteristics, experiences, and outputs. Summary statistics of our data can be found in Table A2.

**Data Sources.** The first source is the pre-RCT assessment, which collected baseline information on participating judges. This includes demographic variables such as age, years of judicial experience, court assignment, and province of service. The survey also incorporated several self-assessed measures, including confidence in public speaking, perceptions of work-life balance, and expectations regarding the training program.

The second data source is a post-course assessment conducted after the completion of the training program. This survey captures judges' responses regarding their engagement with the AI tool and their evaluations of the training. Respondents were asked whether they believed the course would improve judicial productivity and whether they supported the integration of AI into legal procedures. These responses offer insight into perceived effectiveness and broader institutional acceptance of the intervention.

The third and most detailed source is based on final project submissions. Each judge was asked to provide five judgments written prior to the intervention and five written after, regardless of treatment status. This yielded a dataset of approximately 3,000 judicial opinions—a rare resource given the lack of systematic digital records in Pakistan's lower courts. Using text analysis methods, we constructed several quantitative outcome measures from these documents. These include case delay (measured from filing to decision date when available), Gunning Fog indices to capture readability and counts of legal arguments. These measures constitute the primary outcomes used to evaluate the impact of *JudgeGPT* and the associated training on judicial performance and decision quality.

Fourth, through JudgeGPT we collect login timestamps, the browser's user agent[3], IP addresses, and chat histories. The chat histories include messages the judges wrote, the AI assistant's replies, and the documents that the AI assistant retrieved in the background. The IP addresses we collect to check against credential sharing within and across intervention arms.

Fifth, during the live sessions of each training course, we record attention and attendance via polls that test the content covered and ask judges for their name, birth year, and email address to aid linking attendance to our other datasets, such as the surveys and chat logs.

**Outcome Measures.** In evaluating outcomes, we rely on three complementary sets of measures. The first consists of self-reported data collected from judges regarding their use of AI tools, support for AI adoption in legal work, and views on the training program. Judges were asked how frequently they used AI in their work and responded to two key questions: "Do you support the use of generative AI assistants such as GPT in your role as a judge?" and Can the AI for Judges course improve your productivity?" Responses were recorded on a five-point scale to capture variation in attitudes and reduce measurement noise. While self-reported, these measures allow us to assess perceived usefulness and compare post-treatment responses with baseline beliefs and usage patterns.

The second set of outcomes is based on text analysis of judgments submitted as part of the final project. From these documents, we construct several quantitative indicators. Readability is proxied using the Gunning Fog Index, which incorporates sentence length and word complexity. Next, we measure opinion quality by comparing pairs of opinions, one from before treatment and one from after treatment. We make all pairwise comparisons based on the documents provided by the judges. We use a GPT-5-mini prompt to compare them on overall clarity, conciseness, adequate

[3]The browser's user agent is a self-identification string that contains information on the browser and operating system used. Example: 'Chrome Mobile 121.0.0 / Android 10.0.0'

presentation of the relevant facts, and the clarity/validity of the legal reasoning. For each judge, we then compute the share of comparisons where the post-treatment opinion was compared to the pre-treatment opinion. An improvement in opinion quality after treatment would thus be measured as a value higher than 0.5.

Third, we assemble the dataset based on the judicial statistics of courts in Pakistan. We create a panel dataset that allow us to asess the productivity impact of the introduction of the AI tool.

Taken together, these measures provide a basis for assessing the impact of *JudgeGPT* and related training on judicial performance, with attention to both efficiency and quality. The combination of survey responses and textual indicators offers a more complete view of how AI tools are integrated into legal reasoning.

### *3.4. Empirical Strategy*

To identify the effects of interest we use the experimental nature of our design. Specifically, we estimate the equation of the following form:

$$Y_i = \alpha_s + \beta_1 \text{JudgeGPT and Associated Training} + \beta_2 \text{JudgeGPT without Training} + \Gamma' \mathbf{X}_i + \varepsilon_i \tag{1}$$

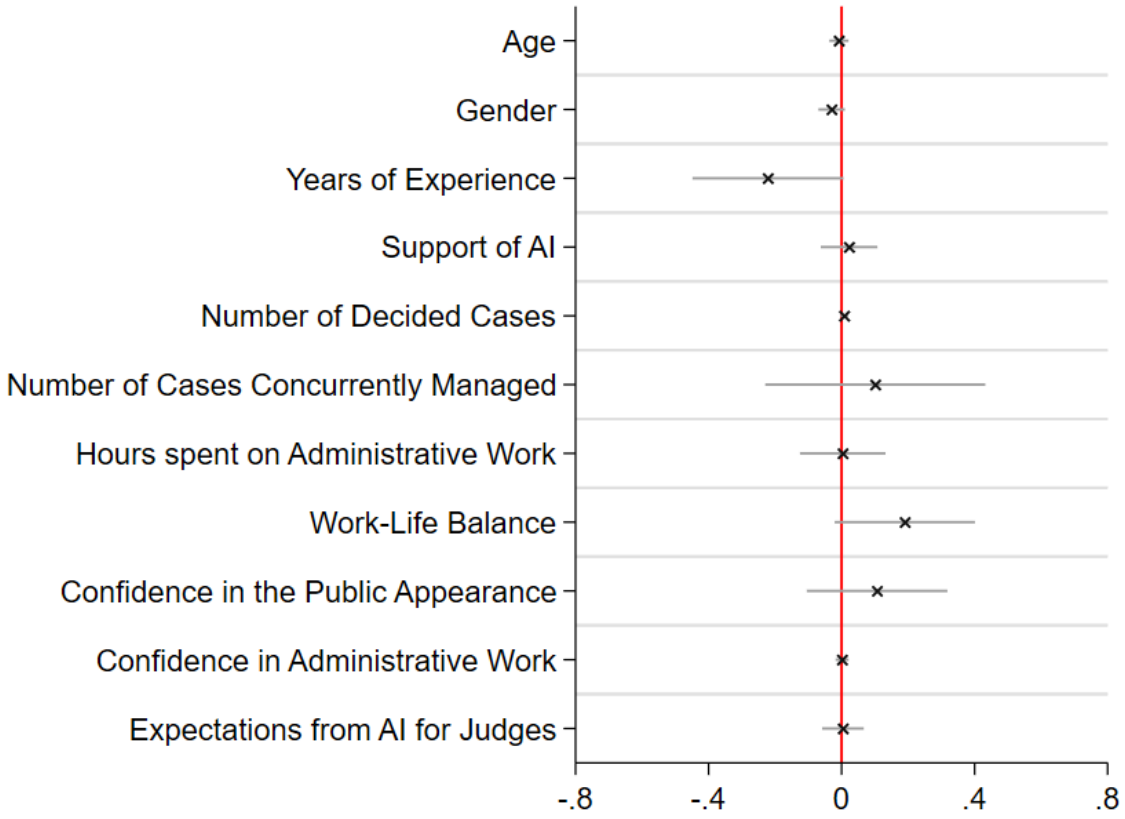
Where $\alpha_s$ is strata fixed effects, $\beta_1$ measures the effect of providing the JudgeGPT alongside with the course, $\beta_2$ measures the effect of providing JudgeGPT without the course. $\mathbf{X}$ is the vector of controls added to the regression to increase the precision of the estimates and $\Gamma$ is the vector of corresponding coefficients.

Since treatment was assigned randomly to the judges the OLS estimation of Equation 1 should yield consistent and unbiased estimates of the effect of interests under full compliance. However, since the majority but not all judges used JudgeGPT if they were assigned the treatment, the estimation might face the selection bias since those who used JudgeGPT are likely different from those who ignored the treatment because motivation, workload and some other observable and unobservable characteristics.

To address this issue we employ intention to treat (ITT) identification strategy, instrumenting the compliance with the actual treatment assignment, with the following first stage equation:

$$\begin{aligned} \text{JudgeGPT and } \widehat{\text{Associated}} \text{ Training} = \alpha_s + \beta_1 \text{Assigned JudgeGPT and Training} \\ + \beta_2 \text{Assigned JudgeGPT without Training} + \Gamma' \mathbf{X}_i + \varepsilon_i \end{aligned} \tag{2a}$$

FIGURE IV: Balance Test




Note: The figure depicts the balance test comparing the treatment of getting the JudgeGPT and related training and control group with placebo training and no JudgeGPT on observable characteristics.

$$\widehat{\text{JudgeGPT without Training}} = \alpha_s + \beta_1 \text{Assigned JudgeGPT and Training} + \beta_2 \text{Assigned JudgeGPT without Training} + \Gamma' \mathbf{X}_i + \varepsilon_i \quad (2b)$$

Using treatment instrumented as in Equation 2 we can mitigate the compliance concerns and obtain the unbiased estimates of the effect of interest.

Figure IV presents evidence on the baseline balance between treated and control judges across a range of pre-treatment characteristics. We observe that the groups are similar in terms of age, gender, baseline support for AI, perceived productivity, and prior use of AI tools. These patterns suggest that the randomization was successful in generating comparable groups. In Appendix Table A3, we report pairwise comparisons between each treated group and the placebo group, along with associated

$p$-values. Across both aggregated and disaggregated comparisons, we find no systematic imbalances, supporting the validity of our experimental design.

**Court Level Analysis.** To analyze the productivity effects of introducing generative AI into the judicial process, we use district-level judicial statistics that we compiled for the study period. Because randomization occurred at the judge level rather than the district level, a naïve regression of the number of resolved cases on the number of treated judges at the district level may be biased. For example, judges in more developed districts—with better internet connectivity and greater access to computers—may have been more likely to participate in the experiment and may also be systematically more productive even absent the intervention.

To mitigate these concerns, we implement a difference-in-differences design. This approach does not require treated and comparison districts to be similar in levels; instead, it requires that they exhibit parallel pre-treatment trends. We estimate the following specification.

$$\begin{aligned}\text{\# of Cases Resolved} = \gamma_i + \gamma_t + \beta \times \mathbb{1}[\text{High Treatment Intensity}]_i \times \mathbb{1}[\text{Year = 2024}]_t + \\ + \Gamma' \mathbf{X}_{it} + \varepsilon_{it} \quad (3)\end{aligned}$$

We define treatment using only the first cohort of judges, since this group was the only one to receive access to the tool at the start of 2024. In all specifications, we control for the district's total number of judges, the composition of the caseload (the shares of pending criminal and civil cases), and other court characteristics, including the number of judicial vacancies, the number of cases carried over on the balance, and the numbers of civil and criminal cases appealed.

We define treatment intensity using only the first cohort of judges, the only group with tool access at the start of 2024. In all specifications, $\mathbf{X}_{it}$ includes the total number of judges in the district, the composition of the caseload (shares criminal and civil among pending cases), and court characteristics such as judicial vacancies, the stock of carryover cases, and the numbers of civil and criminal cases appealed. We classify a district as high-intensity if its exposure to the GPT-plus-training arm exceeds the median (equivalently, at least four such judges, about 20% of the treated judges).

## 4. Main Results

### *4.1. Support for AI and Perceived Gains*

We begin by examining whether access to JudgeGPT influenced judges' attitudes toward the use of generative AI tools in judicial work. Table V presents results based on

responses to the question, "Do you support the use of Generative AI assistants such as GPT in your job as a judge?", rated on a 1-to-5 scale. Judges who received access to the tool—either with structured training or without—are compared to those assigned to a placebo course without tool access. Columns 1 and 2 report OLS estimates, while Columns 3 and 4 present 2SLS estimates using the original random assignment as an instrument. Across both specifications, we find that access to JudgeGPT, particularly when accompanied by training, is associated with higher levels of support for AI use. The estimated effect is approximately 0.25 points, relative to a mean of 3.5, corresponding to an increase of about 7 percent.

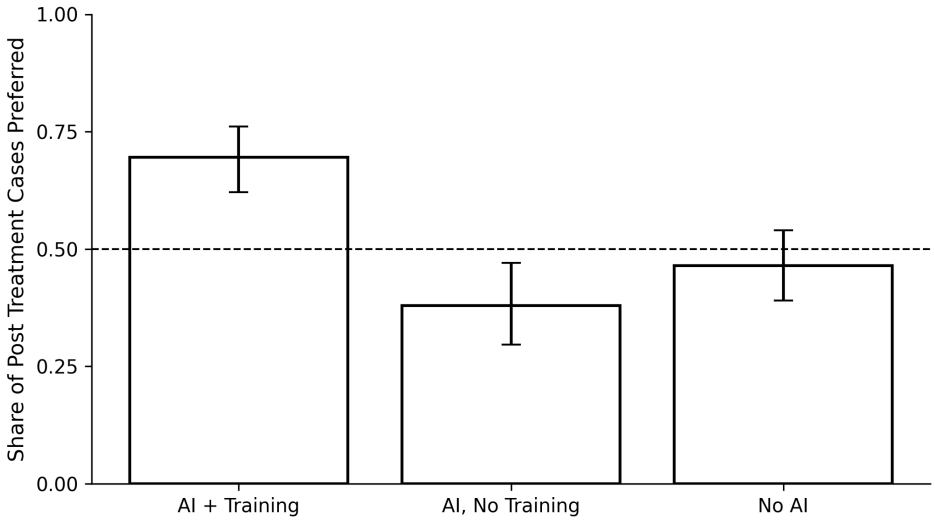
We also assess whether the intervention shaped judges' perceptions of the course itself. Table VI reports results based on responses to the question, "Can the AI for Judges Course improve your productivity?", also rated on a 1-to-5 scale. The treatment and control definitions mirror those used in the support outcome. Across both OLS and 2SLS specifications, judges in the treatment group reported higher perceived productivity, with an estimated effect of 0.157 points relative to a mean of 3.62. This represents an increase of approximately 5 percent. Taken together, the results suggest that access to JudgeGPT, particularly when paired with structured guidance, improves both support for AI integration and confidence in the productivity benefits of the training.

To assess whether these attitudinal shifts translated into actual behavioral engagement, we examine usage logs from the JudgeGPT system. Figure A2 hows a strong correspondence between self-reported frequency of generative AI use and actual tool activity. Judges who identified themselves as frequent users following the intervention sent, on average, over 200 more messages and logged into the system approximately 60 more times than those who did not. Panel A reports message frequency, and Panel B displays login counts, each stratified by post-treatment survey responses. These differences are both substantial and statistically significant. The close alignment between stated attitudes and revealed behavior suggests that the intervention not only shifted perceptions but also led to meaningful engagement with the tool.

### *4.2. Opinion Quality*

We next examine how the intervention may have affected decision quality, using available case-level data. We acknowledge that judicial quality is difficult to measure, and any empirical proxy remains inherently imperfect. Moreover, given that the average case in Pakistan takes over four years to resolve, it will take time before impacts on resolution rates or delays can be credibly assessed. In the shorter run, we focus on two indicative measures: the readability of judgments, and an LLM-annotated measure od case quality. These proxies, while limited, offer early insights into shifts in reasoning

FIGURE V: Treatment Effects on LLM-Annotated Opinion Quality




Note: The figure shows means and 95% confidence intervals (constructed from standard errors of the mean) in the LLM-annotated opinion quality measure, separately for Treatment A with tools and training (left), Treatment B with tools and without training (middle), and the active control group – placebo course and no AI.

depth and communicative clarity.

Table X reports the results for readability. Across both OLS and 2SLS specifications, we find no statistically significant effects of JudgeGPT access, with or without training, on the readability of judgments. We do not detect significant differences between treated and control groups. Judges with access to the tool, whether trained or not, did not produce opinions that were systematically clearer or more accessible based on this metric.

Next, Figure V shows the variation across the treatment groups in the LLM-annotated quality measure. This measure is the "win rate" of post-treatment opinions compared to pre-treatment opinions, where each pair is assessed. A value of 0.5 (horizontal dashed line) means that the win rate is 50/50 so there is no real difference between pre and post. As expected, in the active control group (right) is at 50/50, since those judges got no AI tools. In the large AI+Training group (left), meanwhile, post-treatment opinions tend to be preferred to pre-treatment opinions about 70% of the time, statistically higher than 50/50, and consistent with improved opinion quality. Meanwhile, in the middle group with AI but no training, opinion quality actually decreased, as post-treatment opinions are preferred only about 35% of the time to the pre-treatment opinions.

### *4.3. Court Productivity*

Finally, we assess whether individual-level experimental assignment translates into realized system throughput. We link treated judges to administrative court statistics and construct district-year exposure measures based on the number of treated judges. We then estimate district-level difference-in-differences specifications where the outcome is total cases disposed.

Two patterns emerge. First, Table VII shows that districts with greater JudgeGPT exposure resolve significantly more cases. The baseline magnitudes are about 15 percent of the mean, which corresponds to roughly 4,000 additional resolved cases per district-year. Second, Table VIII decomposes disposals by docket and shows increases in both civil and criminal cases. This broad-based response is difficult to reconcile with a simple reallocation story in which judges shift effort from one case category to another. Instead, the results are consistent with an expansion in effective judicial capacity, either through faster drafting and research, improved task management, or reduced time costs for routine procedural steps.

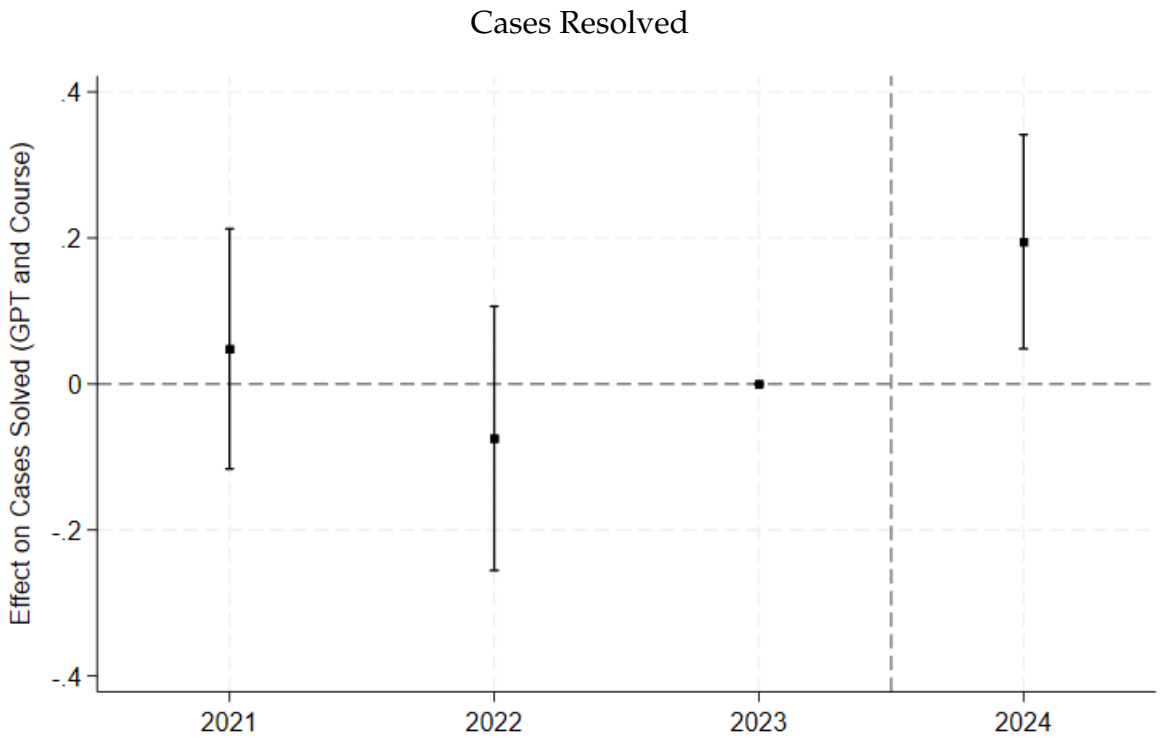
We define a district as high-intensity if it is above the median in exposure to GPT plus training (i.e., at least about 20% of judges in the district received GPT plus training). High-intensity districts experience a 0.2 standard deviation increase in case disposals, roughly 4,000 additional cases per year.

Figure VI complements the table-form evidence by showing the dynamic treatment effects from an event-study specification, normalized to 2023 as the omitted reference year. The 2021 and 2022 coefficients are close to zero and statistically indistinguishable from zero, which supports the absence of differential pre-trends prior to the treatment period. Relative to 2023, the estimated effect in 2024 is positive and precisely estimated, indicating a meaningful increase in cases resolved in the post period. Interpreted together with the event study evidence, the conclusion is that JudgeGPT exposure raises realized court throughput and that the productivity gains reflect higher disposal rather than a reshuffling across civil and criminal dockets.

## 5. Supporting Results

We complement our primary analysis with a set of supporting results. These include evaluation of heterogeneity, tests for experimenter demand effects, potential spillovers across treatment groups, adjustments for multiple inference, and randomization inference. We also consider external validity and assess sensitivity to alternative specifications. These analyses help contextualize the main findings and assess whether they are likely to arise from design artifacts or sampling variation.

FIGURE VI: JudgeGPT and Court Productivity




Note: This figure presents a Difference-in-Differences (DiD) event study evaluating the impact of introduction of JudgeGPT on court efficiency. The figure tracks changes in case disposal. The regression controls for district and year fixed effects. The total number of judges in the district, the share of criminal cases in disposal and institution, the number of civil and criminal appeals, the number of vacant judges' positions, and the total cases in balance are added as controls. 2023 is the reference year. Robust standard errors clustered at the district level.

**Heterogeneous treatment effects.** We follow our pre-analysis plan to examine heterogeneity in treatment effects by judge age and province of assignment. These dimensions were specified in advance, based on the possibility that responses to the intervention may vary with baseline familiarity or openness to technology. In particular, younger judges may be more receptive to generative AI tools, while regional differences in institutional practices or resource constraints may also shape adoption and use. We view this analysis as exploratory, aimed at understanding how context and background characteristics may relate to variation in treatment effects.

Panel A of Table A4 explores heterogeneity in self-reported support for artificial intelligence across the provinces of Pakistan and the federal capital territory. The estimated effects are broadly similar across regions, with consistent shifts in attitudes following the intervention. This suggests that geographic variation in attitudinal responses is limited. Panel B of Table A4 presents analogous results for self-reported increases in productivity.[4] Here, the estimates vary more across provinces, and standard errors are relatively large, limiting the scope for strong statistical inference. Collectively, the results point to only modest provincial heterogeneity. Despite considerable differences in linguistic and ethnic composition across regions, judges across district courts appear to exhibit similar shifts in attitudes in response to the intervention.

We also explore variation by baseline familiarity with technology, using information reported in the baseline survey. Table A5 presents these results. Judges who indicated familiarity with basic tools such as email, laptops, or messaging applications exhibited smaller changes in attitudes. In contrast, those who reported using software such as Microsoft Office or web search tools experienced somewhat larger shifts in support for the use of generative AI. This pattern is evident in both treatment groups—those who received the AI tool and those who also received the accompanying usage course. These findings are consistent with the age-related patterns discussed above and suggest that prior exposure to digital technologies may be one factor influencing responsiveness to the intervention.

**Experimenter Demand.** We consider whether the results may have been influenced by experimenter demand effects. Several design features suggest this is unlikely. First, while judges were informed that they were participating in a research study, they were not made aware of the specific metrics that would be used to evaluate outcomes, such as judgment clarity or citation of precedent. Second, the intervention was embedded within an official judicial training program, with our involvement presented as part of a broader professional development initiative. In particular, the group that received only access to JudgeGPT but no training had minimal contact with the research team

[4]Judges in Islamabad appear to show the largest shifts in support for generative AI and perceived productivity gains. However, we interpret these results with caution, as the number of courts in Islamabad is limited and the sample includes only 39 judges.

beyond receiving their login credentials. Third, judges were instructed not to share their credentials, and we explicitly stated that sharing would result in removal from the course. Platform usage data confirms that nearly all activity came from unique, judge-specific IP addresses, indicating individual engagement. Logs also show no evidence of unauthorized access by judges in the control or placebo groups. Finally, our ability to record the full content of JudgeGPT sessions, including timestamps, message content, and frequency, provides further evidence that observed usage reflects genuine engagement rather than performative behavior. Taken together, these safeguards suggest that the treatment effects observed are unlikely to be driven by experimenter demand.

**External Validity.** The setting of our study—trial courts in Pakistan—shares institutional features with many lower-tier judicial systems in other South Asian and developing country contexts. These include common law foundations, bureaucratic judicial hierarchies, and limited technological penetration in courtrooms. While our findings may not extend directly to high-income judiciaries, they provide relevant insights for jurisdictions seeking to introduce AI tools within constrained administrative systems. We follow the SANS framework proposed by List (2020) to reflect on external validity. Regarding selection, our sample covers nearly half of all sitting trial court judges in Pakistan, reducing concerns about selective participation. In terms of naturalness, the intervention was delivered through official judicial training channels and integrated into day-to-day work processes. The AI assistant was deployed to support real judgment writing rather than artificial exercises. Finally, in terms of scalability, both the software and training were implemented using standard public sector resources, making replication in similar settings feasible.

**Ethics.** This study received ethical clearance from two institutional review boards: the local IRB at the Institute of Business Administration (IBA) Karachi (IBA-Esub006/rec-24) and the international ethics committee at ETH Zurich (EK-2023-N-343-A). The interventions were embedded within ongoing national judicial training and technology initiatives administered by Pakistan's judiciary. We worked in partnership with judicial authorities to randomize the timing and allocation of access, consistent with their planned implementation strategy. Beyond procedural approval, we took several steps to ensure that the study was ethically responsible in substance. Recognizing that generative AI tools may carry both benefits and risks, particularly around hallucination, over-reliance, or opaque reasoning, bias, we designed the training component to inform judges of these limitations and to encourage responsible use. For those assigned to training, we explained the nature of large language models, provided examples of both helpful and problematic responses, and emphasized the need to verify outputs against authoritative legal sources. While we cannot rule out misuse or misunder-

standing, we aimed to make the tool's limitations more transparent and to open the black box of generative AI to its intended users. These efforts reflect our commitment to promoting not just adoption, but informed and cautious integration of new technologies in sensitive institutional settings.

**Sample Size and Randomization Inference.** With over 1,500 judges, this study is among the largest randomized evaluations in judicial settings. However, effect sizes on quality outcomes are modest and precision varies across subsamples. To strengthen inference, we conduct randomization inference exercises that compare our estimated coefficients to distributions obtained by reshuffling treatment assignments 1,000 times. The resulting p-values are reported in Figure A3 and indicate that the main findings are unlikely to result from idiosyncratic assignment. For each of our result we find that estimated coefficients are extreme comparing to generated counterfactuals, for 3000 permutation not a single one appeared to be of a greater magnitude that the actual coefficient, which suggest that effect is present in the data and unlikely to be driven by chance.

**Statistical Power.** In our study, we report two null results. Figure A4 presents the corresponding power analysis, which serves to reinforce our conclusions. For a range of regression coefficients (slopes) estimated in our models, we plot the statistical power of the tests given the sample size, the standard deviation of the treatment variable (estimated from the data), and the standard deviation of the regression residuals obtained from specifications including the full set of fixed effects, as in our main tables. Our sample size observations enables us to detect even very small effects with high power, except within a narrow interval around zero—approximately one standard deviation of the treatment variable—where power is reduced.

**Robustness of Alternative Difference-in-Difference Estimators..** We conduct several exercises to assess the robustness and stability of our estimated productivity effect. First, we estimate a synthetic difference-in-differences model following Arkhangelsky, Athey, Hirshberg, Imbens, and Wager (2021). As shown in Figure A6, the results are very similar to our baseline estimates.

Second, we re-estimate the baseline specification while excluding one district at a time to verify that the findings are not driven by any single court. Figure A7 reports these leave-one-out estimates; in every case, the coefficients remain positive and statistically significant at conventional levels.

Finally, we evaluate whether our model is correctly specified using the framework developed by Callaway, Goodman-Bacon, and Sant'Anna (2024). Specifically, we estimate a non-parametric specification designed to be robust to biases that can arise with continuous treatments. We discretize treatment intensity into quantiles and estimate a

first-difference model.[5] Figure A5 reports the result, showing that statistically significant treatment effects are concentrated in districts in the fourth and fifth quantiles of the number of judges.

## 6. Conclusion

This paper provides experimental evidence on whether generative AI can raise the productivity of a core state institution. In partnership with Pakistan's judiciary, we randomized access to a domain-adapted assistant, JudgeGPT, across 1,559 trial court judges, and layered in variation in structured, assistant-specific training. We have three main findings. First, judges exposed to JudgeGPT report higher support for adopting generative AI in judicial work, and these stated views align closely with revealed engagement in system logs. Second, perceived productivity gains are concentrated among judges who receive targeted training, consistent with the idea that effective use is not automatic and depends on workflow integration and technological knowledge. Third, AI and training improves opinion quality while AI without training harms opinion quality. Fourth, districts with greater exposure to trained JudgeGPT judges dispose substantially more cases. The increase is economically large, broad-based across civil and criminal dockets, and supported by event-study evidence showing no differential pre-trends.

The combination of the usage patterns and the district-level output results points to a concrete productivity channel. Training shifts how judges deploy the tool away from open-ended legal research and toward structured drafting support. In a setting where judges face high caseloads and limited staff, small reductions in the time required for recurring tasks such as summarizing facts, organizing issues, producing first-draft reasoning, and translating standard language can scale into meaningful increases in completed cases. The fact that disposals rise in both civil and criminal matters also argues against a pure reallocation story in which judges substitute effort across dockets. Instead, the evidence is consistent with a shift in the effective production function of adjudication, in which generative AI complements judicial labor by lowering time costs on routine components of decision production.

A central concern with productivity interventions in courts is that higher output might come at the expense of decision quality. Using the judgment texts we observe, we do not detect deterioration in basic writing and content proxies, including readability and counts of legal arguments. These measures are imperfect and cannot speak to all dimensions of correctness or fairness. They nonetheless provide reassuring evidence that the productivity gains we document do not mechanically reflect shorter or less structured decisions in the short run. More generally, the design of JudgeGPT,

[5] the exact equation is $\Delta Y_i = \alpha + \sum_{q \in Q} \beta_q \mathbb{1}[\text{ of judges} \in q] + \varepsilon_i$.

including retrieval augmentation tied to Pakistani legal materials and training that emphasizes verification and risk mitigation, highlights an important lesson for policy. Realizing productivity gains requires more than tool access. It requires institutional scaffolding that makes correct use easier than incorrect use.

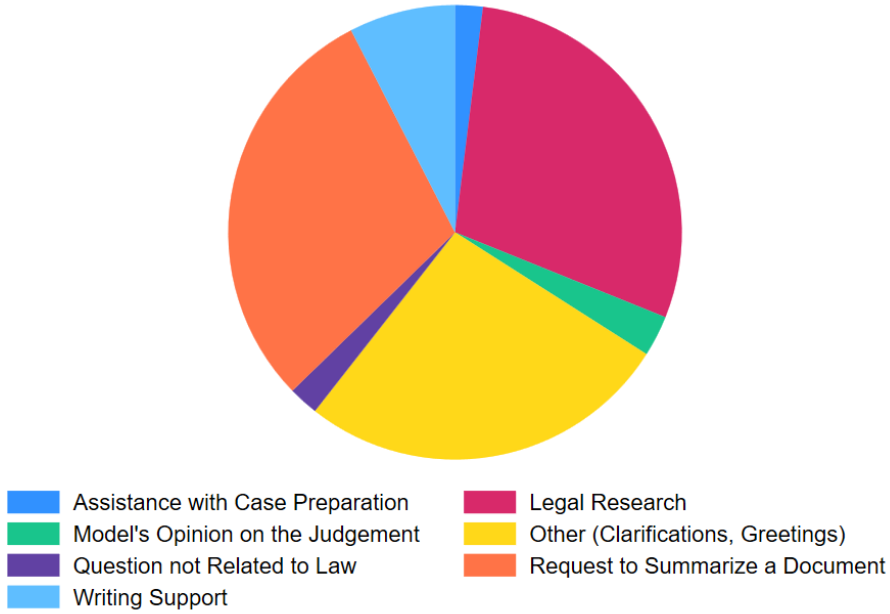
Our findings also speak to the broader debate on AI and state capacity. Public-sector organizations often lack the high-powered incentives and rapid experimentation that facilitate technology adoption in firms. This makes it an open question whether the productivity gains documented for generative AI in private settings translate to government. The judiciary is a demanding environment in which mistakes are costly, procedures are rigid, and outputs are scrutinized. Yet we still observe meaningful improvements in realized cases resolved when the tool is embedded in a training framework. This suggests that, even in bureaucratic settings, generative AI can deliver measurable efficiency gains when deployment is paired with human capital investments and clear norms of use.

More broadly, this experiment illustrates a template for evaluating frontier technologies inside the state. Rather than treating AI as an opaque replacement for bureaucratic discretion, we study AI as an augmenting input whose effects depend on training, institutional constraints, and accountability. The evidence from Pakistan's trial courts indicates that generative AI can raise the productivity of adjudication at scale, and that the gains are largest when deployment is accompanied by structured instruction that channels use toward tasks where the tool is most reliable. The practical implication is straightforward. For resource-constrained judiciaries facing persistent backlogs, generative AI is not a panacea, but it is a one potential lever for increasing cases resolved when paired with training that emphasizes verification, responsible use, and continuous evaluation.

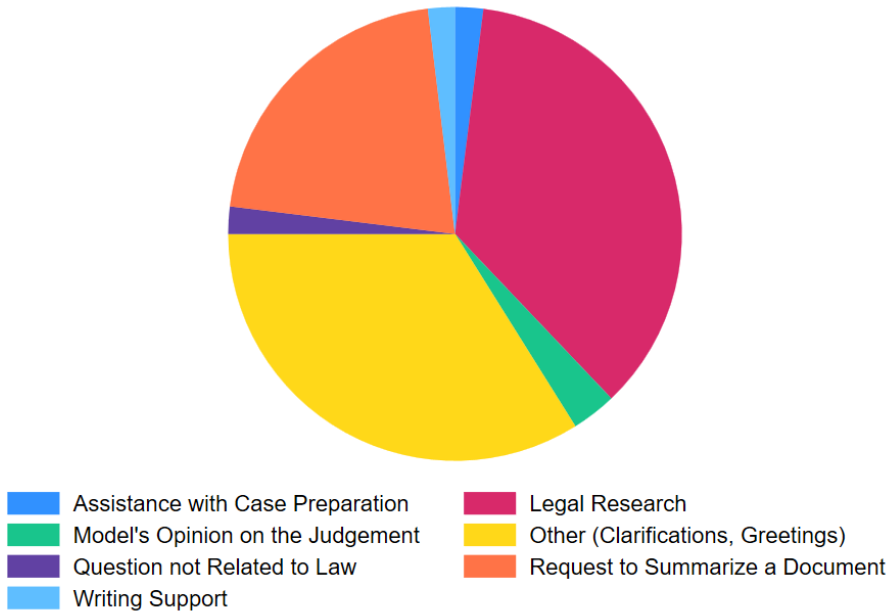
## FIGURES AND TABLES

FIGURE VII: Topics of Judges' Conversations with the Model

(A) Panel A. Distribution of Messages Topics for Judges with Training




(B) Panel B. Distribution of Messages Topics for Judges without Training




Note: The figure depicts the distribution of topics of judges' conversations with the model extracted via DeBerta zero-shot classifier. The shares of most frequent ways of model utilization is presented. Panel A reports the results for judges with assigned training while panel B reports the results for the control group that received JudgeGPT but did not received the course. The topics are Assistance with Case Preparation, i.e. a request to prepare questions to plaintiff or defendant or outlining the main argument for each position, legal research, e.g. searching for the relevant cases or laws, Model's Opinion on the Judgment, i.e.

TABLE I
TREATMENT ROLL-OUT SUMMARY

| | JudgeGPT subscription & GPT instruction course | | Judge GPT Subscription & Generic Technology and Law Course | | Generic Technology & Law Course | |
|---|---|---|---|---|---|---|
| | Judges | Districts | Judges | Districts | Judges | Districts |
| Jan 2022 - Jan 2024 | 0 | 0 | 0 | 0 | 0 | 0 |
| Feb 2024 | 487 | 94 | 0 | 0 | 0 | 0 |
| Sep 2024 | 979 | 108 | 0 | 0 | 0 | 0 |
| Dec 2024 | 1197 | 110 | 180 | 62 | 182 | 66 |
| End of Treatment | Jan 2025 | | | | | |

Note: Batches represent different experimental conditions

TABLE II
JUDGEGPT ACCESS AND GENERATIVE AI USAGE

| | OLS | | 2SLS | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | Stated Frequent User of Generative AI | | | |
| JudgeGPT Access and Associated Training | 0.175*** | 0.175*** | 0.122** | 0.121** |
| | (0.037) | (0.035) | (0.049) | (0.047) |
| JudgeGPT Access without Training | 0.194*** | 0.199*** | 0.155** | 0.162** |
| | (0.052) | (0.050) | (0.068) | (0.065) |
| Strata Fixed Effects | Yes | Yes | Yes | Yes |
| Controls | No | Yes | No | Yes |
| Adj. $R^2$ | .039 | .071 | .036 | .068 |
| Kleibergen-Paap F-statistic | | | 2734.089 | 2763.592 |
| P-value (Access with Training = Access without Training) | .643 | .559 | .471 | .361 |
| Mean Dep. Var. | .82 | .82 | .82 | .82 |
| Observations | 1081 | 1081 | 1081 | 1081 |

Note: Newey-West standard errors in parentheses. The dependent variable is a binary indicator of a judge reporting usage of GPT at least few times a month as answer to the question 'How often do you use generative AI models?'. Used JudgeGPT and Received a Course is a binary indicator of the judge with the course assigned logged into JudgeGPT at least once, while Used JudgeGPT without a Course is a binary indicator of a judge logging into GPT while not having the course as batch4a. Control group did not received GPT access while getting a placebo course. Controls include pre-treatment characteristics reported in the balance test and strata fixed effects. In columns (3)-(4) treatment and control A are instrumented with intial assignment of treatment and control A. $R^2$ and Kleibergen-Paap F-statistic are reported. $^{*}$ $p < 0.10$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

TABLE III
JUDGEGPT ACCESS AND FAMILIARITY WITH LLMS

| | OLS | | 2SLS | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | Familiar with LLMs | | | |
| JudgeGPT Access and Associated Training | 0.218*** | 0.130*** | 0.134** | 0.025 |
| | (0.040) | (0.041) | (0.055) | (0.058) |
| JudgeGPT Access without Training | 0.144** | 0.149** | 0.057 | 0.040 |
| | (0.066) | (0.066) | (0.083) | (0.083) |
| Pretreatment Familiarity with LLMs | | 0.259*** | | 0.282*** |
| | | (0.031) | | (0.032) |
| Strata Fixed Effects | Yes | Yes | Yes | Yes |
| Controls | No | Yes | No | Yes |
| Adj. $R^2$ | .031 | .115 | .027 | .108 |
| Kleibergen-Paap F-statistic | | | 3010.196 | 2269.758 |
| Mean Dep. Var. | .625 | .625 | .625 | .625 |
| P-value (Access with Training = Access without Training) | .198 | .744 | .222 | .809 |
| Observations | 1078 | 1078 | 1078 | 1078 |

Note: Newey-West standard errors in parentheses. The dependent variable is a an indicator of judge including GPT / Gemini or other Large Language Models into technologies the judge is familiar with. JudgeGPT Access and Associated Training is a binary indicator of the judge with the training assigned logged into JudgeGPT at least once, while JudgeGPT Access without Training is a binary indicator of a judge logging into GPT while not having the training as first Control Group (batch4a). Control group did not received GPT access while getting a placebo course. Controls include pre-treatment characteristics reported in the balance test and strata fixed effects. In columns (3)-(4) treatment and control A are instrumented with intial assignment of treatment and control A. $R^2$ and Kleibergen-Paap F-statistic are reported. $^{*}$ $p < 0.10$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

TABLE IV
JUDGEGPT ACCESS AND FAMILLIARITY WITH WHATSAPP

| | OLS | | 2SLS | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | Familliar with WhatsApp | | | |
| JudgeGPT Access and Associated Training | -0.006 | -0.021 | 0.001 | -0.017 |
| | (0.016) | (0.018) | (0.021) | (0.025) |
| JudgeGPT Access without Training | 0.011 | 0.016 | 0.014 | 0.014 |
| | (0.022) | (0.023) | (0.030) | (0.030) |
| Pretreatment Familliarity with WhatsApp | | 0.036*** | | 0.034** |
| | | (0.013) | | (0.014) |
| Strata Fixed Effects | Yes | Yes | Yes | Yes |
| Controls | No | Yes | No | Yes |
| Adj. $R^2$ | .019 | .052 | .019 | .052 |
| Kleibergen-Paap F-statistic | | | 3010.196 | 1640.078 |
| Mean Dep. Var. | .961 | .961 | .961 | .961 |
| P-value (Access with Training = Access without Training) | .361 | .103 | .567 | .213 |
| Observations | 1078 | 1078 | 1078 | 1078 |

Note: Newey-West standard errors in parentheses. The dependent variable is a an indicator of judge incuding WhatsApp into technologies the judge is familiar with. JudgeGPT Access and Associated Training is a binary indicator of the judge with the training assigned logged into JudgeGPT at least once, while JudgeGPT Access without Training is a binary indicator of a judge logging into GPT while not having the training as first Control Group (batch4a). Control group did not received GPT access while getting a placebo course. Controls include pre-treatment characteristics reported in the balance test and strata fixed effects. In columns (3)-(4) treatment and control A are instrumented with intial assignment of treatment and control A. $R^2$ and Kleibergen-Paap F-statistic are reported. $^{*}$ $p < 0.10$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

TABLE V
JUDGEGPT ACCESS AND SUPPORT OF GENERATIVE AI USAGE

| | OLS | | 2SLS | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | Support the Usage of GPT | | | |
| JudgeGPT Access and Associated Training | 0.220*** | 0.224*** | 0.233*** | 0.243*** |
| | (0.056) | (0.054) | (0.076) | (0.072) |
| JudgeGPT Access without Training | 0.221*** | 0.215*** | 0.260** | 0.263*** |
| | (0.080) | (0.078) | (0.105) | (0.100) |
| Pretreatment Support of Generative AI Usage | | 0.175*** | | 0.175*** |
| | | (0.033) | | (0.032) |
| Strata Fixed Effects | Yes | Yes | Yes | Yes |
| Controls | No | Yes | No | Yes |
| Adj. $R^2$ | .021 | .079 | .021 | .079 |
| Kleibergen-Paap F-statistic | | | 2783.507 | 2798.802 |
| Mean Dep. Var. | 3.545 | 3.545 | 3.545 | 3.545 |
| P-value (Access with Training = Access without Training) | .982 | .89 | .711 | .781 |
| Observations | 1083 | 1083 | 1083 | 1083 |

Note: Newey-West standard errors in parentheses. The dependent variable is a support of GPT usage rated on the scale 1-5, specifically the answer to the question Do you support the use of Generative AI assistants such as GPT in your job as a judge? . JudgeGPT Access and Associated Training is a binary indicator of the judge with the training assigned logged into JudgeGPT at least once, while JudgeGPT Access without Training is a binary indicator of a judge logging into GPT while not having the training as first Control Group (batch4a). Control group did not received GPT access while getting a placebo course. Controls include pre-treatment characteristics reported in the balance test and strata fixed effects. In columns (3)-(4) treatment and control A are instrumented with intial assignment of treatment and control A. $R^2$ and Kleibergen-Paap F-statistic are reported. $^{*}$ $p < 0.10$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

TABLE VI
EFFECT OF JUDGEGPT ACCESS AND ASSOCIATED TRAINING ON SUPPORT OF AI FOR JUDGES COURSE

| | OLS | | 2SLS | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | GPT Course Can Improve Productivity | | | |
| JudgeGPT Access and Associated Training | 0.164*** | 0.163*** | 0.149** | 0.157** |
| | (0.050) | (0.048) | (0.068) | (0.065) |
| JudgeGPT Access without Training | 0.107 | 0.123 | 0.068 | 0.095 |
| | (0.078) | (0.078) | (0.100) | (0.097) |
| Pretreatment Support of the Course | | 0.045 | | 0.045 |
| | | (0.037) | | (0.036) |
| Strata Fixed Effects | Yes | Yes | Yes | Yes |
| Controls | No | Yes | No | Yes |
| Adj. $R^2$ | .013 | .068 | .013 | .068 |
| Kleibergen-Paap F-statistic | | | 2783.507 | 2798.802 |
| P-value (Access with Training = Access without Training) | .387 | .555 | .272 | .396 |
| Mean Dep. Var. | 3.62 | 3.62 | 3.62 | 3.62 |
| Observations | 1083 | 1083 | 1083 | 1083 |

Note: Newey-West standard errors in parentheses. The dependent variable is a an answer to the question 'Can AI for Judges Course improve your productivity?' rated on the scale 1-5. JudgeGPT Access and Associated Training is a binary indicator of the judge with the training assigned logged into JudgeGPT at least once, while JudgeGPT Access without Training is a binary indicator of a judge logging into GPT while not having the course as control A (batch4a). Control group did not received GPT access while getting a placebo course. Controls include pre-treatment characteristics reported in the balance test and strata fixed effects. In columns (3)-(4) treatment and control A are instrumented with intial assignment of treatment and control A. $R^2$ and Kleibergen-Paap F-statistic are reported. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE VII
JUDGEGPT ACCESS AND COURT EFFICIENCY

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | Cases Resolved | | | |
| # Judges With JudgeGPT and Associated Training × Post | 1.203*** | 0.530*** | 0.532*** | 0.188*** |
| | (0.280) | (0.161) | (0.163) | (0.069) |
| Year Fixed Effects | Yes | Yes | Yes | Yes |
| District Fixed Effects | No | Yes | Yes | Yes |
| # of Judges in the District | No | No | Yes | Yes |
| Controls | No | No | No | Yes |
| Adj. $R^2$ | .113 | .748 | .748 | .923 |
| Mean Dep. Var. | 0 | 0 | 0 | 0 |
| # Clusters | 122 | 122 | 122 | 122 |
| Observations | 483 | 483 | 483 | 483 |

Note: Robust standard errors in parentheses are clustered by District. The unit of observation is district-year. The dependent variable is the standardized number of cases disposed in district year. # Judges With JudgeGPT and Associated Training × Post is the number of district judges with access to JudgeGPT and course interacted with 2024 indicator. The total number of judges in the district, share of criminal cases in disposal and institution, number of civil and criminal appeals, number of vacant judges positions and total cases in balance are added as controls. District Fixed Effects and year fixed effects are included. Only judges that were treated at the beggining of the 2024 are included Adjusted $R^2$ is reported. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE VIII
JUDGEGPT ACCESS AND COURT EFFICIENCY BY CASE TYPE.

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | Civil Cases Resolved | | Criminal Cases Resolved | |
| # Judges With JudgeGPT and Associated Training × Post | 0.261*** | 0.109* | 0.688*** | 0.252*** |
| | (0.077) | (0.056) | (0.210) | (0.086) |
| Year Fixed Effects | Yes | Yes | Yes | Yes |
| District Fixed Effects | Yes | Yes | Yes | Yes |
| # of Judges in the District | Yes | Yes | Yes | Yes |
| Controls | No | Yes | No | Yes |
| Adj. $R^2$ | .891 | .934 | .619 | .898 |
| Mean Dep. Var. | 0 | 0 | 0 | 0 |
| # Clusters | 122 | 122 | 122 | 122 |
| Observations | 483 | 483 | 483 | 483 |

Note: Robust standard errors in parentheses are clustered by District. The unit of observation is district-year. The dependent variables are the standardized number of criminal and civil cases disposed in district year. # Judges With JudgeGPT and Associated Training × Post is the number of district judges with access to JudgeGPT and course interacted with 2024 indicator. The total number of judges in the district, share of criminal cases in disposal and institution, number of civil and criminal appeals, number of vacant judges positions and total cases in balance are added as controls. District Fixed Effects and year fixed effects are included. Only judges that were treated at the beggining of the 2024 are included Adjusted $R^2$ is reported. $^{*}\ p < 0.10$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$

TABLE IX
TRAINING OF JUDGES AND PURPOSES OF GENERATIVE AI USAGE

| | 2SLS | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | Share of Messages Using AI for Legal Research | | Share of Messages Asking for Writing Support | |
| JudgeGPT Access and Associated Training | -0.077*** | -0.078*** | 0.025*** | 0.028*** |
| | (0.029) | (0.028) | (0.004) | (0.005) |
| Adj. $R^2$ | .017 | .036 | .013 | .018 |
| Kleibergen-Paap F-statistic | 726934.826 | 135859.516 | 726934.826 | 135859.516 |
| Mean Dep. Var. | .307 | .307 | .034 | .034 |
| Observations | 954 | 954 | 954 | 954 |
| | Share of Messages Unrelated to Law | | Share of Messages Asking Model's Opinion | |
| JudgeGPT Access and Associated Training | -0.005 | -0.004 | 0.009* | 0.009* |
| | (0.011) | (0.010) | (0.005) | (0.005) |
| Controls | No | Yes | No | Yes |
| Strata Fixed Effects | No | Yes | No | Yes |
| Adj. $R^2$ | 0 | .023 | .002 | -.004 |
| Kleibergen-Paap F-statistic | 726934.826 | 135859.516 | 726934.826 | 135859.516 |
| Mean Dep. Var. | .021 | .021 | .03 | .03 |
| Observations | 954 | 954 | 954 | 954 |

Note: Newey-West standard errors in parentheses. The dependent variables are a number of messages classified as pursuing legal research goals, asking to provide a writing support, unrelated to law and the number of messages classified as asking model's opinion about the decision divided by the total number of messages. Used JudgeGPT and Received a Course is a binary indicator of the judge with the course assigned logged into JudgeGPT at least once. The control group received GPT access while getting a placebo course. Controls include pre-treatment characteristics reported in the balance test and strata fixed effects. In columns (3)-(4) treatment and control A are instrumented with intial assignment of treatment and control A. $R^2$ and Kleibergen-Paap F-statistic are reported. $^{*}$ $p < 0.10$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

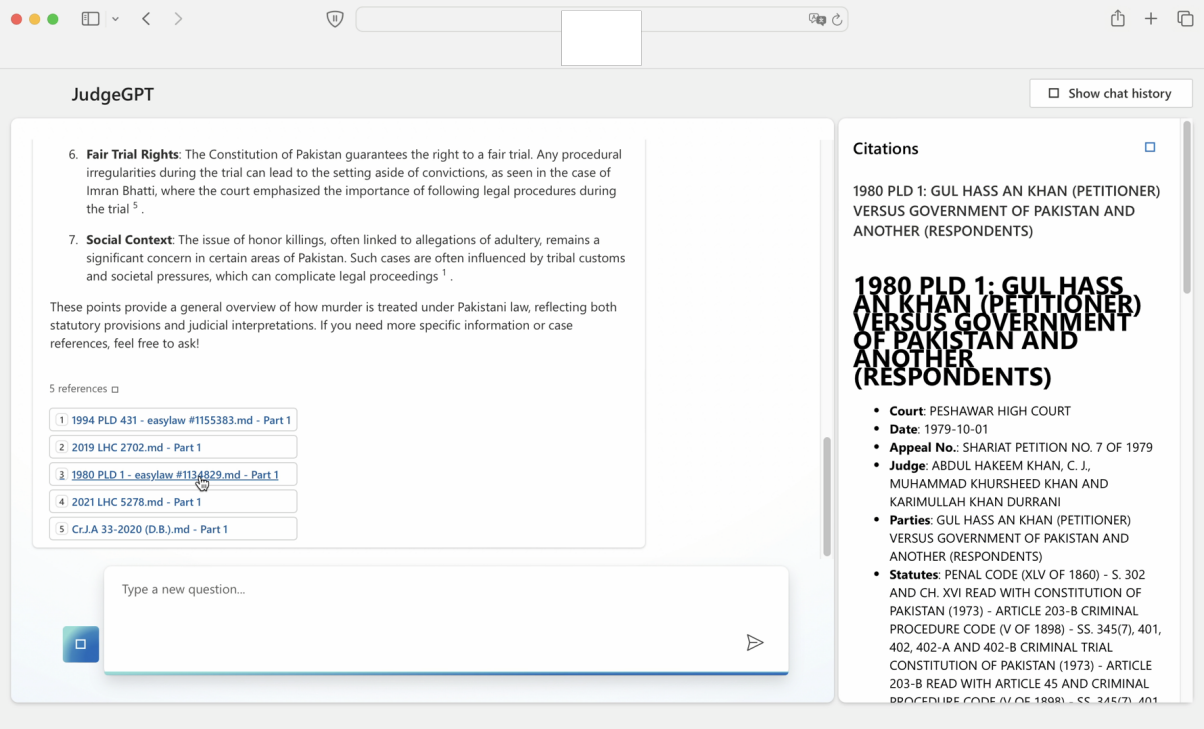
TABLE X
JUDGEGPT ACCESS AND TEXT COMPLEXITY

| | OLS | | 2SLS | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | Gunning Fog Readability | | | |
| JudgeGPT Access and Associated Training | 1.003 | 2.078 | -0.467 | 0.269 |
| | (1.591) | (2.281) | (0.565) | (1.780) |
| JudgeGPT Access without Training | 1.368 | 3.772 | -0.709 | 2.349 |
| | (1.604) | (4.157) | (1.131) | (4.059) |
| Pretreatment Text Readability | | 0.337 | | 0.320 |
| | | (0.724) | | (0.689) |
| Adj. $R^2$ | -.001 | -.013 | -.001 | -.013 |
| Kleibergen-Paap F-statistic | | | 8.617 | 7.887 |
| Anderson-Rubin P-val. | | | .715 | .836 |
| Mean Dep. Var. | 17.003 | 17.137 | 17.003 | 17.137 |
| P-value (Access with Training = Access without Training) | .551 | .591 | .778 | .52 |

Note: Newey-West standard errors in parentheses. The dependent variable is a gunning fod readability metrics, calculated using mean number of words in the sentences and number of complex words in Panel A and the number of legal arguments calculated as the number of references to statutes and precedents in the case. The dependent variable is taken in mean for pre and post treatment cases for each judge. JudgeGPT Access and Associated Training is a binary indicator of the judge with the training assigned logged into JudgeGPT at least once, while JudgeGPT Access without Training is a binary indicator of a judge logging into GPT while not having the training as first Control Group (batch4a). Control group did not received GPT access while getting a placebo course. Controls include pre-treatment characteristics reported in the balance test and strata fixed effects. In columns (3)-(4) treatment and control A are instrumented with intial assignment of treatment and control A. $R^2$ and Kleibergen-Paap F-statistic are reported. $^{*}$ $p < 0.10$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

# APPENDIX A. ADDITIONAL FIGURES AND TABLES

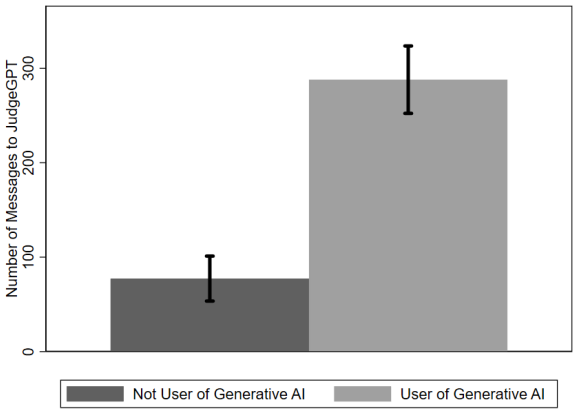
FIGURE A1: JudgeGPT RAG Citation



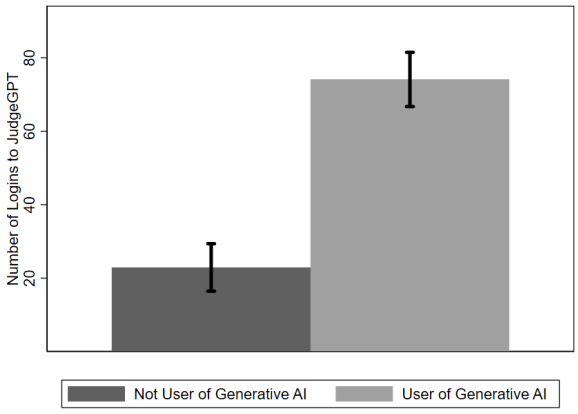

Note: Example of an interactive RAG citation inside JudgeGPT's user interface.

FIGURE A2: First-State: Stated and Actual Generative AI Usage

(A) Panel A. Judge GPT Message Frequency



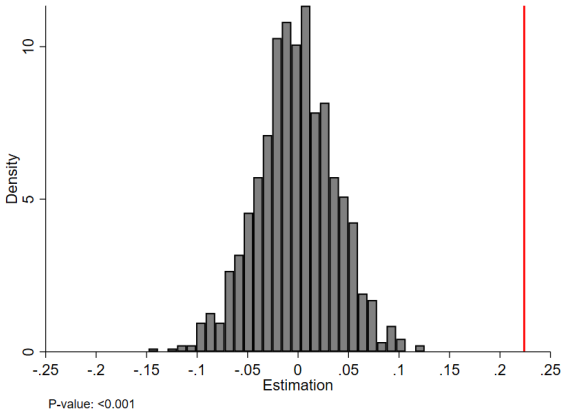

(B) Panel B. JudgeGPT Actual Logins



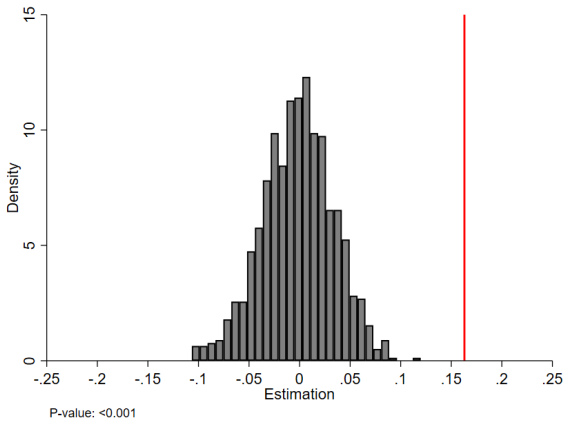

Note: The figure depicts means and standard deviations for usage of JudgeGPT, specifically, number logins and messages to JudgeGPT among people who described themselves as frequent generative AI users after the treatment and not frequent generative AI users. Specifically, answered that they are using generative AI at least few times a month to the question 'Which of the following best describes how often you use GPT or other AI systems in your judicial work?' The figure shows that active users of JudgeGPT tend to describe themselves as frequent generative AI users in general after the treatment.

FIGURE A3: Permutation Inference Test

(A) Panel A. Support of Generative AI Usage



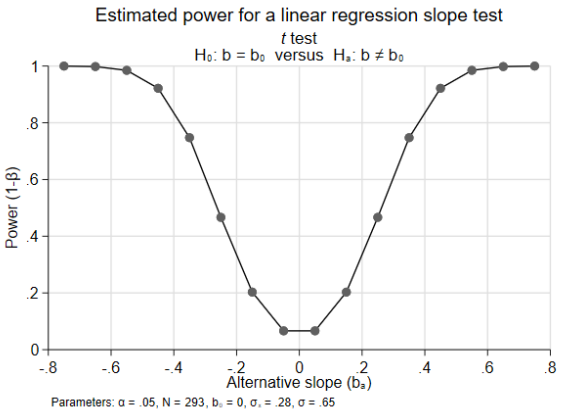

(B) Panel B. Support of AI for Judges Course



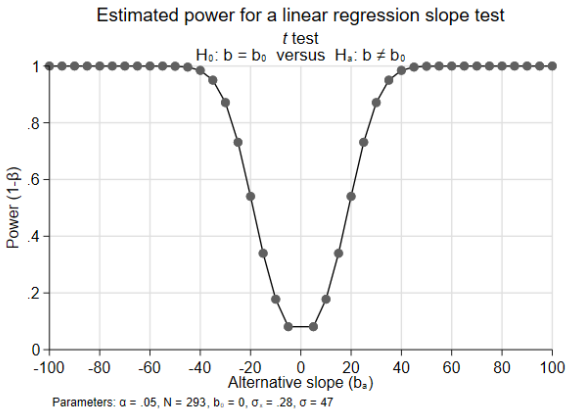

Note: The figure depicts permutation based inference results. Graphs present estimated density of permutations of treatment for our main outcomes. For each outcome 1000 permutation of treatment were analyzed, for supreme court references permutation was stratified by judge to preserve the treatment structure. Read line marks the actual OLS estimate of the coefficient of interest.

FIGURE A4: Power Analysis For Null Results

(A) Panel A. Judgment Readability

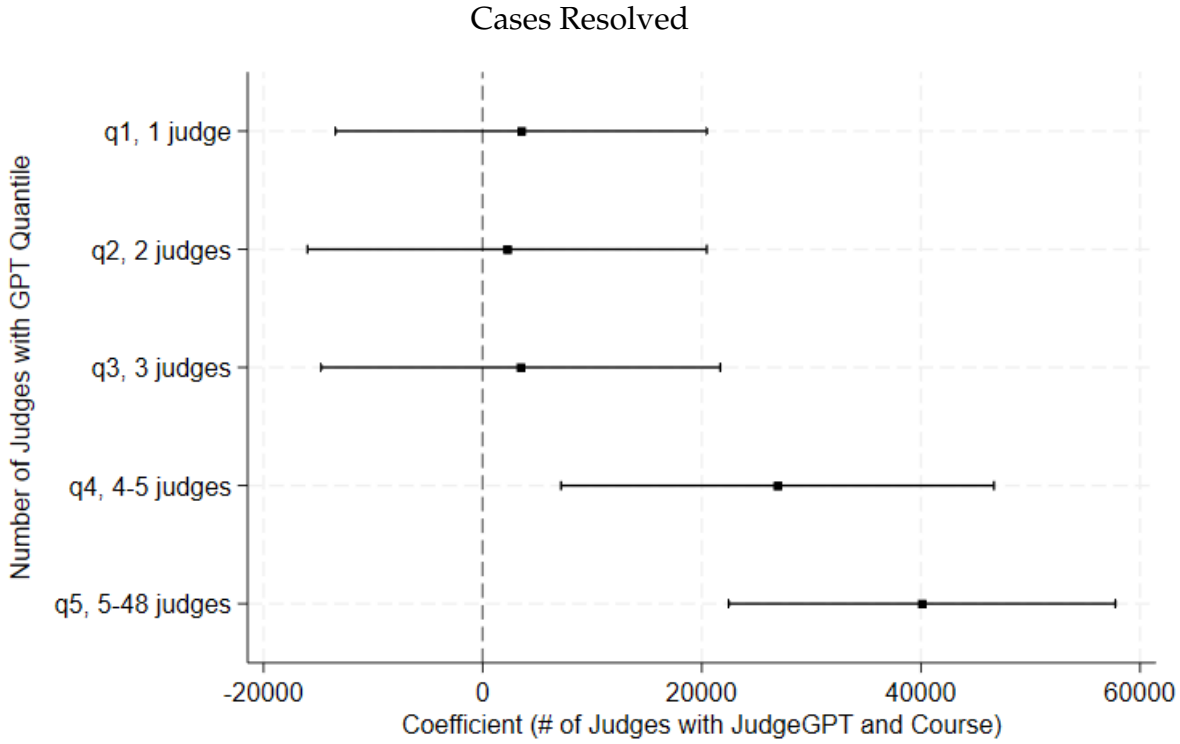



(B) Panel B. # of Legal References



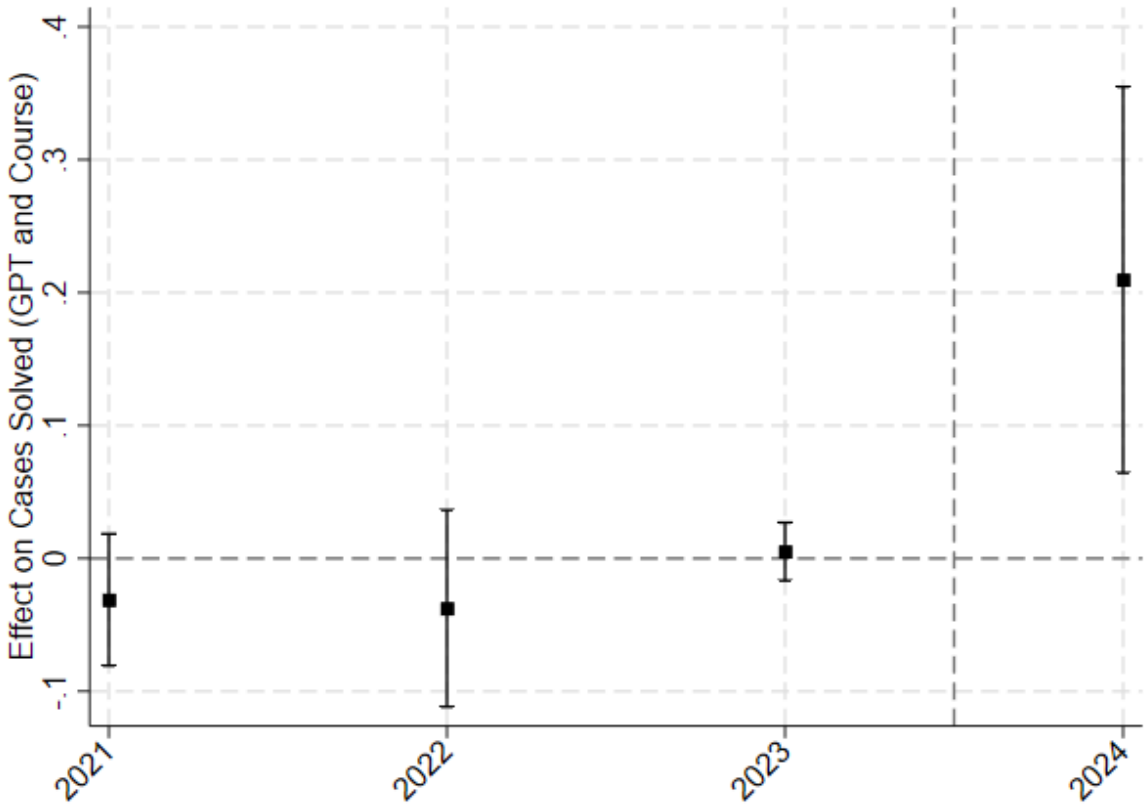

Note: The figure depicts the power analysis of regression coefficients for case quality null results. Panel A reports the power analysis for gunning fog readability, while panel B focuses on the number of legal references. For regression coefficients on x-axis the power of our test is depicted on y-axis, given sample size, estimated standard deviation of error term and regressor. The power calculations are provided for the main treatment, which includes JudgeGPT and Associated training; residuals are calculated using regression with fixed effects.

FIGURE A5: Robustness of the Diff-in-Diff Result to Callaway et al. (2024) Approach



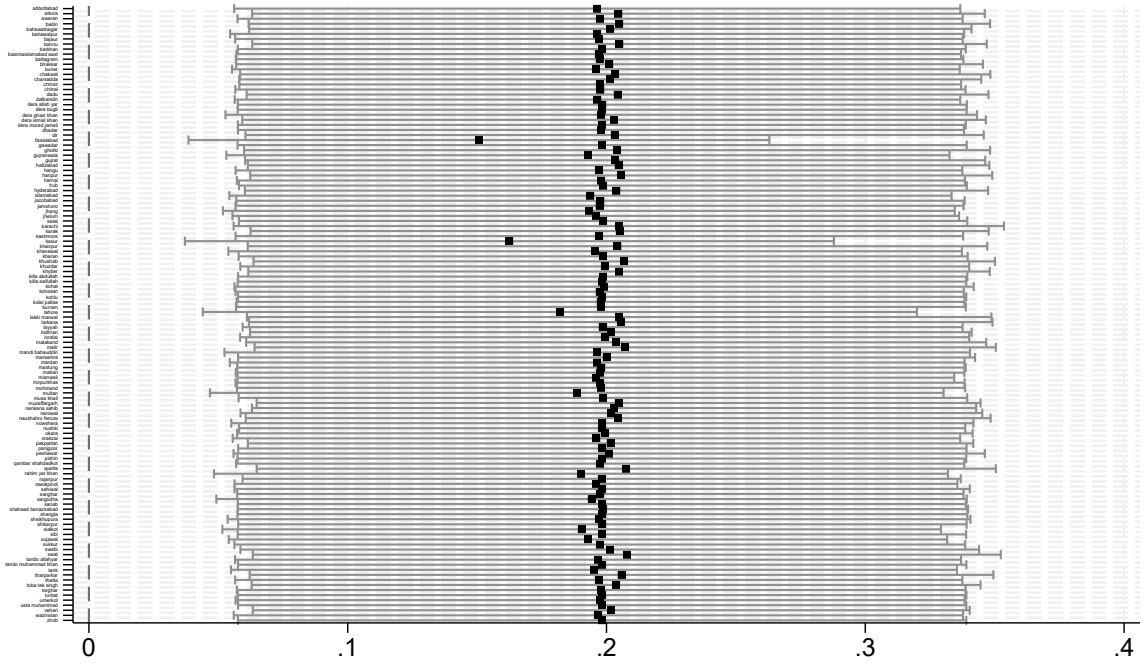

Note: This figure presents estimates of a non-parametric specification of Callaway et al. (2024). The continuous treatment was discretized by quantiles and an indicator for each quantile is included into regression with the number of cases resolved in first differences as a dependent variable. The coefficient for each quantile is plotted with districts with 0 judges treated as a reference group.

FIGURE A6: Synthetic Difference-in-Differences




Note: his figure presents a Synthetic Difference-in-Differences event study evaluating the impact of introduction of JudgeGPT on court efficiency. The figure tracks changes in case disposal. The total number of judges in the district, share of criminal cases in disposal and institution, number of civil and criminal appeals, number of vacant judges positions and total cases in balance are added as controls. The SDID estimates are computed using the cluster bootstrap procedure described in Section 4.4 of Clarke, Pailañir, Athey, and Imbens (2023), following Algorithm 4 in Arkhangelsky et al. (2021).

FIGURE A7: Robustness to Excluding One District at a Time




Note: This figure presents the results of estimation of the baseline difference in difference specification with one district excluded at a time. 95% confidence intervals along with the point estimates are reported for the district on y axis excluded.

TABLE A1
TREATMENT DESCRIPTION AND SLIDES

| Treatment Group | Treatment Description | Treatment Details | Link to the Slides |
|---|---|---|---|
| Judge GPT and Related Training | This group received our full treatment, containing of access to JudgeGPT and AI course. Lectures were given by professor Elliott Ash from ETH Zurich and covered the topics of access to the tool, capabilities of JudgeGPT, limitations and strengths of AI, samples of possible applications in judicial process. | • Fine-tuned on the corpus of judgments AI tool<br>• The course on Law and AI consisting of 6 lectures<br>• 5 home assignments requiring the use of AI and quizzes in class • Final project on AI application in legal process<br>• Technical support throughout the course<br>• Certificates and invitation to Law and Technology Workshop in Zurich for successful participants | Link |
| Judge GPT without Related Training | This treatment group received JudgeGPT and placebo course in Law and Technology. Lectures were given by professor Elliott Ash from ETH Zurich. The course covered topics in technological applications in legal process, such as cites for legal research, topics in machine learning, such as general overview of machine learning tools and biases of AI. | • Fine-tuned on the corpus of judgments AI tool<br>• The course on Law and Technology consisting of 6 lectures<br>• 5 home assignments and quizzes in class<br>• Final project on legal research<br>• Certificates and invitation to Law and Technology Workshop in Zurich for successful participants | Link |
| Placebo Course | This treatment group did not receive an access to Judge GPT while receiving a placebo course in Law and Technology. Lectures were given by professor Elliott Ash from ETH Zurich. The course covered topics in technological applications in legal process, such as cites for legal research, topics in machine learning, such as general overview of machine learning tools and biases of AI. | • The course on Law and Technology consisting of 6 lectures<br>• 5 home assignments and quizzes in class<br>• Final project on legal research<br>• Certificates and invitation to Law and Technology Workshop in Zurich for successful participants | Link |

Note: The table describes levels of our treatment and provides link to slides of lectures that were

TABLE A2
SUMMARY STATISTICS OF MAIN VARIABLES

**Panel A: Pretreatment Characteristics**

| | Mean | Median | SD | Min | Max | N obs |
|---|---|---|---|---|---|---|
| Age | 42.70 | 42 | 7.041 | 25 | 60 | 1559 |
| Years of Experience | 11.54 | 9.200 | 7.203 | 0 | 70 | 1559 |
| Support of AI | 3.428 | 4 | 0.824 | 1 | 4 | 1559 |
| Number of Cases Pending | 533.2 | 330 | 804.5 | 0 | 24000 | 1559 |
| Number of Cases Decided per Month | 115.0 | 74 | 141.9 | 0 | 2500 | 1559 |
| Number of Cases Comfortable Managing at the Same Time | 142.6 | 40 | 349.8 | 0 | 6007 | 1559 |
| Hours per Week Spent on Legal Research | 16.95 | 14 | 21.42 | 0 | 500 | 1559 |
| Hours per Week Spent on Administrative Work | 11.82 | 8 | 13.92 | 0 | 300 | 1559 |
| Confidence in Public Speaking | 6.506 | 7 | 2.115 | 1 | 10 | 1559 |
| Workload | 6.320 | 6 | 2.183 | 1 | 10 | 1559 |
| Worklife Balance | 5.746 | 5 | 2.111 | 1 | 10 | 1559 |
| Self-Assessed Public Speaking | 7.528 | 8 | 2.047 | 1 | 10 | 1559 |
| Hours Spent on Administrative Needs | 7.502 | 8 | 1.891 | 1 | 10 | 1559 |
| Expectations from the Course | 3.724 | 4 | 0.612 | 1 | 4 | 1559 |
| Familiar with AI | 0.263 | 0 | 0.440 | 0 | 1 | 1078 |
| Familiar with WhatsApp | 0.429 | 0 | 0.495 | 0 | 1 | 1078 |

**Panel B: Judge Level Outcomes**

| | Mean | Median | SD | Min | Max | N obs |
|---|---|---|---|---|---|---|
| JudgeGPT and Related Training | 0.610 | 1 | 0.488 | 0 | 1 | 1559 |
| JudgeGPT without Training | 0.0712 | 0 | 0.257 | 0 | 1 | 1559 |
| Frequent AI User | 0.820 | 1 | 0.385 | 0 | 1 | 1081 |
| Support of Generative AI Usage | 3.428 | 4 | 0.824 | 1 | 4 | 1559 |
| Support of AI for Judges Course | 3.620 | 4 | 0.575 | 1 | 4 | 1083 |
| Familiar with AI | 0.625 | 1 | 0.484 | 0 | 1 | 1078 |
| Familiar with WhatsApp | 0.961 | 1 | 0.194 | 0 | 1 | 1078 |

**Panel C: Case Analysis Outcomes**

| | Mean | Median | SD | Min | Max | N obs |
|---|---|---|---|---|---|---|
| Judgments Readability, post | 0.00766 | -0.0488 | 0.668 | -0.178 | 9.485 | 293 |
| # of Legal Arguments, post | 124.5 | 123 | 51.81 | 3.714 | 270 | 293 |

**Panel D: District Level Panel**

| | Mean | Median | SD | Min | Max | N obs |
|---|---|---|---|---|---|---|
| # of Cases Resolved | 26509.1 | 10862 | 49275.0 | 0 | 682933 | 601 |
| # of Civil Cases Resolved | 11218.5 | 3050 | 19875.2 | 0 | 184715 | 601 |
| # of Criminal Cases Resolved | 15005.5 | 6435 | 31077.4 | 0 | 503817 | 601 |
| # of Judges in Districts | 20.99 | 13 | 30.45 | 0 | 272 | 601 |
| # of Vacant Judge Positions | 7.295 | 3 | 13.21 | 0 | 113 | 601 |
| # of Cases on Balance | 14509.9 | 3745 | 28358.7 | 0 | 351174 | 601 |
| # of Cases in Civil Appeal | 433.8 | 198 | 666.2 | 0 | 5826 | 601 |
| # of Cases in Criminal Appeal | 10.60 | 4 | 22.32 | 0 | 348 | 601 |
| Share of Criminal Cases Disposed | 0.570 | 0.639 | 0.261 | 0 | 0.992 | 601 |
| Share of Criminal Cases Instituted | 0.582 | 0.636 | 0.246 | 0 | 0.980 | 601 |



Note: This table reports mean, median, standard deviation for main outcome variables in our analysis. The unit of analysis is judge for panels A, B and C and district-year for panel D.

# TABLE A3
# BALANCE TEST OVER JUDGE CHARACTERISTICS AND PRE-TREATMENT OUTCOMES

| | (1) | | (2) | | (3) | | (1)-(2) | | (1)-(3) | | (2)-(3) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | JudgeGPT And Course | | Only Judge GPT | | General Course | | | | Pairwise t-test | | | |
| Variable | N | Mean/(SE) | N | Mean/(SE) | N | Mean/(SE) | N | P-value | N | P-value | N | P-value |
| Age | 1196 | 42.729 | 180 | 42.800 | 183 | 42.432 | 1376 | 0.451 | 1379 | 0.911 | 363 | 0.467 |
| | | (0.205) | | (0.510) | | (0.505) | | | | | | |
| Gender | 1196 | 1.788 | 180 | 1.822 | 183 | 1.820 | 1376 | 0.329 | 1379 | 0.348 | 363 | 0.851 |
| | | (0.012) | | (0.029) | | (0.028) | | | | | | |
| Years of Experience | 1196 | 26.558 | 180 | 55.928 | 183 | 33.666 | 1376 | 0.210 | 1379 | 0.638 | 363 | 0.412 |
| | | (5.006) | | (22.022) | | (15.441) | | | | | | |
| AI Support | 1196 | 3.431 | 180 | 3.406 | 183 | 3.432 | 1376 | 0.644 | 1379 | 0.976 | 363 | 0.687 |
| | | (0.024) | | (0.063) | | (0.061) | | | | | | |
| Number of Cases on the Desk | 1196 | 523.565 | 180 | 654.539 | 183 | 476.929 | 1376 | 0.340 | 1379 | 0.136 | 363 | 0.192 |
| | | (15.508) | | (138.142) | | (37.349) | | | | | | |
| Number of Decided Cases | 1196 | 117.902 | 180 | 106.933 | 183 | 103.716 | 1376 | 0.133 | 1379 | 0.095* | 363 | 0.751 |
| | | (4.372) | | (7.367) | | (8.275) | | | | | | |
| Number of Cases Concurrently Managed | 1196 | 140.920 | 180 | 174.989 | 183 | 121.530 | 1376 | 0.430 | 1379 | 0.231 | 363 | 0.147 |
| | | (9.596) | | (38.760) | | (17.581) | | | | | | |
| Hours spent on Legal Research and Writing Judgments | 1196 | 17.081 | 180 | 15.711 | 183 | 17.328 | 1376 | 0.336 | 1379 | 0.654 | 363 | 0.243 |
| | | (0.677) | | (0.864) | | (1.025) | | | | | | |
| Hours spent on Administrative Work | 1196 | 11.755 | 180 | 12.883 | 183 | 11.230 | 1376 | 0.301 | 1379 | 0.561 | 363 | 0.225 |
| | | (0.412) | | (1.082) | | (0.794) | | | | | | |
| Workload | 1196 | 6.375 | 180 | 6.078 | 183 | 6.202 | 1376 | 0.110 | 1379 | 0.345 | 363 | 0.678 |
| | | (0.063) | | (0.161) | | (0.156) | | | | | | |
| Work/Life Balance | 1196 | 5.759 | 180 | 5.683 | 183 | 5.721 | 1376 | 0.736 | 1379 | 0.800 | 363 | 0.926 |
| | | (0.061) | | (0.157) | | (0.150) | | | | | | |
| Confidence in the Public Appearance | 1196 | 7.564 | 180 | 7.311 | 183 | 7.508 | 1376 | 0.174 | 1379 | 0.786 | 363 | 0.392 |
| | | (0.058) | | (0.163) | | (0.154) | | | | | | |
| Confidence in Administrative Work | 1196 | 7.518 | 180 | 7.400 | 183 | 7.497 | 1376 | 0.535 | 1379 | 0.997 | 363 | 0.676 |
| | | (0.054) | | (0.142) | | (0.142) | | | | | | |
| Expectations from AI for Judges | 1196 | 3.722 | 180 | 3.739 | 183 | 3.716 | 1376 | 0.697 | 1379 | 0.971 | 363 | 0.799 |
| | | (0.018) | | (0.045) | | (0.046) | | | | | | |

Note: The table presents the balance test. The treatment group was assigend a JudgeGPT subscription and course explaining the use of JudgeGPT (batches 1-3), the control A group got only JudgeGPT access (batch 4a) while the control B group got general course in law and technology. Table reports mean and standard deviations for pretreatment covariates as well as 3 pairwise t-tests p-value. Heteroskedacity-robust standard errors are in paranthesis. 18 strata fixed effects are included. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

# TABLE A4
## EFFECT HETEROGENEITY BY PROVINCE

| | 2SLS | | | | |
|---|---|---|---|---|---|
| **Panel A. Effect on Support for AI** | | | | | |
| Province: | Balochistan | Islamabad | KPK | Punjab | Sindh |
| | (1) | (2) | (3) | (4) | (5) |
| | Support the Usage of GPT | | | | |
| JudgeGPT Access and Associated Training | 0.509** | 0.890*** | 0.242* | 0.179* | 0.247 |
| | (0.211) | (0.191) | (0.132) | (0.093) | (0.210) |
| JudgeGPT Access without Training | 0.434 | 0.008 | 0.207 | 0.169 | 0.433* |
| | (0.270) | (0.429) | (0.189) | (0.140) | (0.261) |
| Pretreatment Support of the Course | -0.131 | 0.087 | 0.138* | 0.008 | 0.070 |
| | (0.156) | (0.159) | (0.072) | (0.053) | (0.103) |
| P-value (Access with Training = Access without Training) | .705 | .006 | .827 | .921 | .165 |
| Mean Dep. Var. | 3.586 | 3.487 | 3.449 | 3.584 | 3.564 |
| **Panel B. Self Reported Productivity** | | | | | |
| | GPT Course Can Improve Productivity | | | | |
| JudgeGPT Access and Associated Training | 0.177 | 0.642*** | 0.106 | 0.124 | 0.179 |
| | (0.350) | (0.232) | (0.143) | (0.082) | (0.179) |
| JudgeGPT Access without Training | 0.363 | -1.684*** | -0.070 | 0.084 | 0.255 |
| | (0.386) | (0.505) | (0.196) | (0.132) | (0.230) |
| Pretreatment Support of the Course | -0.064 | 0.243 | 0.236*** | -0.004 | -0.019 |
| | (0.121) | (0.238) | (0.070) | (0.053) | (0.083) |
| Controls | Yes | Yes | Yes | Yes | Yes |
| Strata Fixed Effects | Yes | Yes | Yes | Yes | Yes |
| Kleibergen-Paap F-statistic | 30.531 | 34.048 | 1222.835 | 1072.143 | 474.536 |
| P-value (Access with Training = Access without Training) | .329 | $< 0.001$ | .261 | .704 | .569 |
| Mean Dep. Var. | 3.657 | 3.59 | 3.538 | 3.65 | 3.636 |
| Observations | 70 | 39 | 225 | 497 | 220 |

Note: Newey-West standard errors in parentheses. The figure depicts ITT estimates of the effect of AI for Judges by 5 provinces.The dependent variable is a support of GPT usage rated on the scale 1-5, specifically the answer to the question Do you support the use of Generative AI assistants such as GPT in your job as a judge? in Panel A and an answer to the question 'Can AI for Judges Course improve your productivity?' rated on the scale 1-5. JudgeGPT Access and Associated Training is a binary indicator of the judge with the training assigned logged into JudgeGPT at least once, while JudgeGPT Access without Training is a binary indicator of a judge logging into GPT while not having the course as control A (batch4a). Control group did not received GPT access while getting a placebo course. Controls include pre-treatment characteristics reported in the balance test and strata fixed effects. In columns (3)-(4) treatment and control A are instrumented with intial assignment of treatment and control A. Kleibergen-Paap F-statistic is reported. $^{*}$ $p < 0.10$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

# TABLE A5
# EFFECT HETEROGENEITY BY TECHNOLOGY FAMILIARITY

| | 2SLS | | | | |
|---|---|---|---|---|---|
| **Panel A. Effect on Support for AI** | | | | | |
| Familiar with: | Email | Microsoft Office | Search Engine | Laptop | WhatsApp |
| | (1) | (2) | (3) | (4) | (5) |
| | Support the Usage of GPT | | | | |
| JudgeGPT Access and Associated Training | 0.735 | 1.309** | 1.271** | 0.737 | 0.755 |
| | (0.508) | (0.533) | (0.509) | (0.508) | (0.522) |
| JudgeGPT Access without Training | 0.734 | 1.797*** | 1.759*** | 0.589 | 0.752 |
| | (0.569) | (0.549) | (0.515) | (0.602) | (0.570) |
| Pretreatment Support of the Course | 0.094 | 0.097 | 0.102 | 0.089 | 0.071 |
| | (0.067) | (0.078) | (0.071) | (0.063) | (0.059) |
| P-value (Access with Training = Access without Training) | .998 | .001 | $< 0.001$ | .663 | .991 |
| Mean Dep. Var. | 3.598 | 3.609 | 3.609 | 3.602 | 3.586 |
| **Panel B. Self Reported Productivity** | | | | | |
| | GPT Course Can Improve Productivity | | | | |
| JudgeGPT Access and Associated Training | -0.094 | 0.156 | 0.075 | -0.088 | -0.066 |
| | (0.363) | (0.440) | (0.500) | (0.375) | (0.382) |
| JudgeGPT Access without Training | -0.046 | 0.582 | 0.538 | -0.185 | -0.010 |
| | (0.437) | (0.452) | (0.505) | (0.475) | (0.443) |
| Pretreatment Support of the Course | 0.035 | 0.072 | 0.049 | 0.011 | 0.024 |
| | (0.057) | (0.061) | (0.060) | (0.055) | (0.055) |
| Controls | Yes | Yes | Yes | Yes | Yes |
| Strata Fixed Effects | Yes | Yes | Yes | Yes | Yes |
| Kleibergen-Paap F-statistic | 1715.028 | 3136.124 | 1716.3 | 1727.592 | 1945.839 |
| P-value (Access with Training = Access without Training) | .852 | $< 0.001$ | $< 0.001$ | .749 | .817 |
| Mean Dep. Var. | 3.686 | 3.715 | 3.706 | 3.685 | 3.666 |
| Observations | 420 | 340 | 371 | 425 | 463 |

Note: Newey-West standard errors in parentheses. The figure depicts ITT estimates of the effect of AI for Judges by familiarity with technology.The dependent variable is a support of GPT usage rated on the scale 1-5, specifically the answer to the question Do you support the use of Generative AI assistants such as GPT in your job as a judge? in Panel A and an answer to the question 'Can AI for Judges Course improve your productivity?' rated on the scale 1-5. JudgeGPT Access and Associated Training is a binary indicator of the judge with the training assigned logged into JudgeGPT at least once, while JudgeGPT Access without Training is a binary indicator of a judge logging into GPT while not having the course as control A (batch4a). Control group did not received GPT access while getting a placebo course. Controls include pre-treatment characteristics reported in the balance test and strata fixed effects. In columns (3)-(4) treatment and control A are instrumented with intial assignment of treatment and control A. Kleibergen-Paap F-statistic is reported. $^{*}\ p < 0.10$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$

TABLE A6
BALANCE TEST OVER JUDGE CHARACTERISTICS AND PRE-TREATMENT OUTCOMES - MALE VS FEMALE

| | (1) Male Plaintiff | | (2) Female Plaintiff | | (1)-(2) Pairwise t-test | |
|---|---|---|---|---|---|---|
| Variable | N | Mean/(SE) | N | Mean/(SE) | N | P-value |
| Age | 1221 | 42.775 | 338 | 42.414 | 1559 | 0.401 |
| | | (0.202) | | (0.378) | | |
| Gender | 1221 | 1.805 | 338 | 1.760 | 1559 | 0.084* |
| | | (0.011) | | (0.023) | | |
| Years of Experience | 1221 | 31.283 | 338 | 28.977 | 1559 | 0.844 |
| | | (5.650) | | (10.250) | | |
| AI Support | 1221 | 3.419 | 338 | 3.462 | 1559 | 0.392 |
| | | (0.024) | | (0.044) | | |
| Number of Cases on the Desk | 1221 | 516.113 | 338 | 594.985 | 1559 | 0.299 |
| | | (15.919) | | (74.324) | | |
| Number of Decided Cases | 1221 | 114.875 | 338 | 115.317 | 1559 | 0.953 |
| | | (4.263) | | (6.155) | | |
| Number of Cases Concurrently Managed | 1221 | 148.962 | 338 | 119.515 | 1559 | 0.058* |
| | | (10.878) | | (11.121) | | |
| Hours spent on Legal Research and Writing Judgments | 1221 | 17.321 | 338 | 15.618 | 1559 | 0.071* |
| | | (0.667) | | (0.666) | | |
| Hours spent on Administrative Work | 1221 | 11.974 | 338 | 11.281 | 1559 | 0.388 |
| | | (0.407) | | (0.692) | | |
| Workload | 1221 | 6.279 | 338 | 6.467 | 1559 | 0.139 |
| | | (0.064) | | (0.110) | | |
| Work/Life Balance | 1221 | 5.722 | 338 | 5.831 | 1559 | 0.401 |
| | | (0.060) | | (0.115) | | |
| Confidence in the Public Appearance | 1221 | 7.505 | 338 | 7.612 | 1559 | 0.381 |
| | | (0.059) | | (0.108) | | |
| Confidence in Administrative Work | 1221 | 7.500 | 338 | 7.512 | 1559 | 0.914 |
| | | (0.055) | | (0.100) | | |
| Expectations from AI for Judges | 1221 | 3.737 | 338 | 3.675 | 1559 | 0.117 |
| | | (0.017) | | (0.036) | | |

The table presents the balance test over assignment of gender in cases shown to the judges. The treatment group was assigned a female plaintiff while the control group the male plaintiff. Table reports mean and standard deviations for pretreatment covariates as well as pairwise t-tests p-value. Heteroskedacity-robust standard errors are in paranthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE A7
BALANCE TEST OVER JUDGE CHARACTERISTICS AND PRE-TREATMENT OUTCOMES - NON-MUSLIM VS MUSLIM

| | (1) | | (2) | | (1)-(2) | |
|---|---|---|---|---|---|---|
| | Non-Muslim Plaintiff | | Muslim Plaintiff | | Pairwise t-test | |
| Variable | N | Mean/(SE) | N | Mean/(SE) | N | P-value |
| Age | 1209 | 42.629 | 350 | 42.929 | 1559 | 0.488 |
| | | (0.202) | | (0.381) | | |
| Gender | 1209 | 1.803 | 350 | 1.769 | 1559 | 0.177 |
| | | (0.011) | | (0.023) | | |
| Years of Experience | 1209 | 31.432 | 350 | 28.543 | 1559 | 0.799 |
| | | (5.716) | | (9.831) | | |
| AI Support | 1209 | 3.423 | 350 | 3.446 | 1559 | 0.642 |
| | | (0.024) | | (0.043) | | |
| Number of Cases on the Desk | 1209 | 525.982 | 350 | 558.191 | 1559 | 0.377 |
| | | (25.139) | | (26.403) | | |
| Number of Decided Cases | 1209 | 111.698 | 350 | 126.274 | 1559 | 0.155 |
| | | (3.720) | | (9.542) | | |
| Number of Cases Concurrently Managed | 1209 | 139.687 | 350 | 152.560 | 1559 | 0.571 |
| | | (9.763) | | (20.498) | | |
| Hours spent on Legal Research and Writing Judgments | 1209 | 17.334 | 350 | 15.631 | 1559 | 0.069* |
| | | (0.673) | | (0.652) | | |
| Hours spent on Administrative Work | 1209 | 11.868 | 350 | 11.669 | 1559 | 0.790 |
| | | (0.417) | | (0.623) | | |
| Workload | 1209 | 6.285 | 350 | 6.440 | 1559 | 0.214 |
| | | (0.064) | | (0.107) | | |
| Work/Life Balance | 1209 | 5.705 | 350 | 5.889 | 1559 | 0.140 |
| | | (0.061) | | (0.108) | | |
| Confidence in the Public Appearance | 1209 | 7.498 | 350 | 7.631 | 1559 | 0.273 |
| | | (0.059) | | (0.106) | | |
| Confidence in Administrative Work | 1209 | 7.495 | 350 | 7.526 | 1559 | 0.786 |
| | | (0.055) | | (0.097) | | |
| Expectations from AI for Judges | 1209 | 3.722 | 350 | 3.729 | 1559 | 0.854 |
| | | (0.018) | | (0.030) | | |

The table presents the balance test over assignment of gender in cases shown to the judges. The treatment group was assigned a female plaintiff while the control group the male plaintiff. Table reports mean and standard deviations for pretreatment covariates as well as pairwise t-tests p-value. Heteroskedacity-robust standard errors are in paranthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE A8
JUDGEGPT ACCESS AND BIAS

| | OLS | | 2SLS | |
|---|---|---|---|---|
| **Panel A: Gender Bias** | | | | |
| | (1) | (2) | (3) | (4) |
| | Plaintiff is Likely to Win | | | |
| JudgeGPT and Training × Female Plaintiff | -0.360*** | -0.355*** | -0.196** | -0.185** |
| | (0.059) | (0.059) | (0.079) | (0.078) |
| JudgeGPT without Training × Female Plaintiff | -0.416*** | -0.428*** | -0.158 | -0.175 |
| | (0.113) | (0.112) | (0.136) | (0.134) |
| JudgeGPT Access and Associated Training | 0.188*** | 0.179*** | -0.008 | -0.023 |
| | (0.023) | (0.023) | (0.051) | (0.051) |
| JudgeGPT Access without Training | 0.250*** | 0.238*** | -0.028 | -0.035 |
| | (0.055) | (0.055) | (0.089) | (0.089) |
| Female Plaintiff | 0.771*** | 0.769*** | 0.666*** | 0.661*** |
| | (0.049) | (0.049) | (0.062) | (0.061) |
| Mean Dep. Var. | .336 | .336 | .336 | .336 |
| **Panel B: Religious Bias** | | | | |
| | Plaintiff is Likely to Win | | | |
| JudgeGPT and Training × Non-Muslim Plaintiff | 0.245*** | 0.238*** | 0.194** | 0.185** |
| | (0.078) | (0.080) | (0.093) | (0.094) |
| JudgeGPT without Training × Non-Muslim Plaintiff | 0.478*** | 0.478*** | 0.394*** | 0.398*** |
| | (0.122) | (0.122) | (0.146) | (0.145) |
| JudgeGPT Access and Associated Training | -0.058 | -0.059 | -0.081 | -0.083 |
| | (0.076) | (0.077) | (0.085) | (0.086) |
| JudgeGPT Access without Training | -0.316*** | -0.317*** | -0.314** | -0.316** |
| | (0.113) | (0.113) | (0.129) | (0.129) |
| Non-Muslim Plaintiff | -0.612*** | -0.604*** | -0.582*** | -0.574*** |
| | (0.071) | (0.072) | (0.082) | (0.083) |
| Strata Fixed Effects | Yes | Yes | Yes | Yes |
| Controls | No | Yes | No | Yes |
| Mean Dep. Var. | .26 | .26 | .26 | .26 |
| Observations | 1559 | 1559 | 1559 | 1559 |

Note: Robust standard errors in parentheses are clustered by Judge. The unit of observation is judge. The dependent variable is a an indicator of judge stating that the plintiff is more likely to win the case. JudgeGPT Access and Associated Training is a binary indicator of the judge with the training assigned logged into JudgeGPT at least once, while JudgeGPT Access without Training is a binary indicator of a judge logging into GPT while not having the training as first Control Group (batch4a). Female (Muslim) plaintiff is an indicator of plaintiff having female (muslim) name while all other case characteristics kept constant. Control group did not received GPT access while getting a placebo course. Strata (age by province) Fixed Effects are included. Controls include all the variables presented in the balance test such as demographic characteristics, support of AI and attitudes towards different aspects of judicial work. In columns (3)-(4) treatment A and treatment B are instrumented with intial assignment of treatment A and treatment B and their interaction with indicator of assigned female (muslim) plaintiff. Adjusted $R^2$ is reported. Panel B presents the religious bias panel A reports gender bias. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

## APPENDIX B. FURTHER DATA AND PRE-ANALYSIS PLAN

*Preanalysis Plan, ID AEARCTR-0012906)*

**Introduction**

This study evaluates the impact of integrating generative AI technology in the Pakistani judicial system. Approximately 800 trial court judges, or one-third of the national judiciary, will be equipped with AI tools as part of the Pakistan Judicial Academy's "Artificial Intelligence Training Initiative." The project aims to assess AI's role in reducing case backlogs, improving disposal rates, and enhancing the quality of legal analysis. This initiative could set a precedent for the integration of AI in judicial systems in developing countries, offering insights into leveraging technology to improve state capabilities and judicial productivity.

**Study Design**

A randomized controlled trial will be implemented, with trial court judges randomly assigned to receive generative AI assistance through OpenAI's ChatGPT and related training. The judges will be divided into two groups: the treatment group, receiving AI tools and training, and the control group, continuing with traditional judicial processes until a subsequent round of the course at the end of the year. This phased roll-out ensures a comprehensive evaluation of AI's impact on judicial efficiency across varied court settings in Pakistan.

**Outcome Variables**

The primary outcome variables include:

- Attendance, participation tracking
- AI tool utilization
- Judgment text analysis on readability, argument diversity, legal arguments
- AI utility perceptions
- Productivity, AI expectations

**Regression Specification**

We will estimate OLS regressions with the following specifications:

- no covariates
- FE for court
- FE for stratification unit

**Heterogeneity Analysis**

The study will explore heterogeneity in treatment effects by examining:

- Heterogeneity by Age of Judges
- Heterogeneity by education level of Province
- Differences in outcomes based on the type of cases handled by the judges (civil vs. criminal).
- Variation in AI tool usage patterns among judges and their correlation with efficiency improvements.
- The impact of judges' prior experience with technology on the effectiveness of AI assistance.

**Hypotheses**

H1: Participation in the course will increase AI assistant usage.

H2: AI assistance will improve the quality of legal reasoning in judgments.

H3: The perceived usefulness of AI tools by judges will increase.

## REFERENCES


Acemoglu, D., Akcigit, U., & Celik, M. A. (2022, July). Radical and incremental innovation: The roles of firms, managers, and innovators. *American Economic Journal: Macroeconomics*, *14*(3), 199–249. Retrieved from https://www.aeaweb.org/articles?id=10.1257/mac.20170410 doi: 10.1257/mac.20170410

Acemoglu, D., Autor, D., Hazell, J., & Restrepo, P. (2022). Artificial intelligence and jobs: Evidence from online vacancies. *Journal of Labor Economics*, *40*(S1), S293-S340. doi: 10.1086/718327

Acemoglu, D., Ozdaglar, A., & Siderius, J. (2023, 12). A model of online misinformation. *The Review of Economic Studies*, *91*(6), 3117-3150. Retrieved from https://doi.org/10.1093/restud/rdad111 doi: 10.1093/restud/rdad111

Acemoglu, D., Ozdaglar, A., & Siderius, J. (2025, June). *Ai and social media: A political economy perspective* (Working Paper No. 33892). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w33892 doi: 10.3386/w33892

Agrawal, A., Gans, J., & Goldfarb, A. (2019). Economic policy for artificial intelligence. *Innovation Policy and the Economy*, *19*(1), 139 - 159. Retrieved from https://EconPapers.repec.org/RePEc:ucp:ipolec:doi:10.1086/699935

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., & Wager, S. (2021). Synthetic Difference-in-Differences. *American Economic Review*, *111*(12), 4088–4118.

Ash, E., & MacLeod, W. B. (2024, February). Mandatory retirement for judges improved the performance of us state supreme courts. *American Economic Journal: Economic Policy*, *16*(1), 518–48. Retrieved from https://www.aeaweb.org/articles?id=10.1257/pol.20210667 doi: 10.1257/pol.20210667

Brynjolfsson, E., Li, D., & Raymond, L. (2025, 02). Generative ai at work*. *The Quarterly Journal of Economics*, *140*(2), 889-942. Retrieved from https://doi.org/10.1093/qje/qjae044 doi: 10.1093/qje/qjae044

Brynjolfsson, E., Rock, D., & Syverson, C. (2017, November). *Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics* (Working Paper No. 24001). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w24001 doi: 10.3386/w24001

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. A. Friedler & C. Wilson (Eds.), *Conference on fairness, accountability and transparency, fat 2018, 23-24 february 2018, new york, ny, usa* (Vol. 81, p. 77-91). PMLR. Retrieved from http://proceedings.mlr.press/v81/buolamwini18a.html

Butt, J. S. (2025, March). *Justice delayed: The tragic consequences of pakistan's legal system crisis. The Friday Times*. Retrieved from https://thefridaytimes.com/04-Mar-2025/justice-delayed-the-tragic-consequences-of-pakistan-s-legal-system-crisis (Citizens' Voice, Main Slider)

Callaway, B., Goodman-Bacon, A., & Sant'Anna, P. H. C. (2024, May). Event studies with a continuous treatment. *AEA Papers and Proceedings*, *114*, 601–05. Retrieved from https://www.aeaweb.org/articles?id=10.1257/pandp.20241047 doi: 10.1257/pandp.20241047

Clarke, D., Pailañir, D., Athey, S., & Imbens, G. W. (2023). Synthetic difference-in-differences estimation.

Council of Europe. (2024). *European judicial systems - cepej evaluation report 2024 - part 1: General analyses*. Council of Europe.

de Janvry, A., He, G., Sadoulet, E., Wang, S., & Zhang, Q. (2023, March). Subjective performance evaluation, influence activities, and bureaucratic work behavior: Evidence from china. *American Economic Review*, *113*(3), 766–99. Retrieved from https://www.aeaweb.org/articles?id=10.1257/aer.20211207 doi: 10.1257/aer.20211207

Fujiwara, T. (2015, March). Voting Technology, Political Responsiveness, and Infant Health: Evidence From Brazil. *Econometrica*, *83*, 423-464.

Gans, J. S. (2025, June). *Growth in ai knowledge* (Working Paper No. 33907). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w33907 doi: 10.3386/w33907

International Development Law Organization. (2018). *Idlo discusses le-*

*gal reform with pakistan supreme court judges.* Retrieved 2025-11-18, from https://www.idlo.int/news/highlights/idlo-legal-reform-pakistan-supreme-court-judges

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, *133*(1), 237-293. Retrieved from https://ideas.repec.org/a/oup/qjecon/v133y2018i1p237-293..html

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018, May). Algorithmic fairness. *AEA Papers and Proceedings*, *108*, 22–27. Retrieved from https://www.aeaweb.org/articles?id=10.1257/pandp.20181018 doi: 10.1257/pandp.20181018

Landes, W. M., & Posner, R. A. (2009, 07). Rational judicial behavior: A statistical study. *Journal of Legal Analysis*, *1*(2), 775-831. Retrieved from https://doi.org/10.1093/jla/1.2.775 doi: 10.1093/jla/1.2.775

Law and Justice Commission of Pakistan. (2023). *National judicial policy implementation committee: Judicial statistics of pakistan 2023.* Retrieved from https://ljcp.gov.pk (Accessed June 2025)

Liu, E., Lu, Y., Peng, W., & Wang, S. (2022). *Judicial independence, local protectionism, and economic integration: Evidence from china* (Tech. Rep.). National Bureau of Economic Research.

Ludwig, J., & Mullainathan, S. (2021, November). Fragile algorithms and fallible decision-makers: Lessons from the justice system. *Journal of Economic Perspectives*, *35*(4), 71–96. Retrieved from https://www.aeaweb.org/articles?id=10.1257/jep.35.4.71 doi: 10.1257/jep.35.4.71

Mehmood, S. (2022). The impact of presidential appointment of judges: Montesquieu or the federalists? *American Economic Journal: Applied Economics*, *14*(4), 411–45. Retrieved from https://www.aeaweb.org/articles?id=10.1257/app.20210176 doi: 10.1257/app.20210176

Muralidharan, K., Niehaus, P., & Sukhtankar, S. (2016). Building state capacity: Evidence from biometric smartcards in india. *American Economic Review*, *106*(10), 2895–2929.

Muralidharan, K., Singh, A., & Ganimian, A. J. (2019). Disrupting education? experimental evidence on technology-aided instruction in india. *American Economic Review*, *109*(4), 1426–60. Retrieved from https://www.aeaweb.org/articles?id=10.1257/aer.20171112 doi: 10.1257/aer.20171112

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, *381*(6654), 187–192. Retrieved from https://www.science.org/doi/10.1126/science.adh2586 doi: 10.1126/science.adh2586

Pah, A., Schwartz, D., Sanga, S., Alexander, C., Hammond, K., Amaral, L., & OKN Consortium, S. (2022). The promise of ai in an open justice system. *AI Magazine*, *43*(1), 69-74. doi: 10.1002/aaai.12039

Siddique, O. (2013). *Pakistan's experience with formal law: Alienation, orthodoxy and the search for legal empowerment*. Cambridge: Cambridge University Press.

Supreme Court of Pakistan. (2024). *Judgment in Suo Moto Case C.P. No. 1010-L of 2022*. (Available at https://www.supremecourt.gov.pk/downloads_judgements/c.p._1010_l_2022.pdf)

Vicente, L., & Matute, H. (2023, October 3). Humans inherit artificial intelligence biases. *Scientific Reports*, *13*(1), 15737. Retrieved from https://doi.org/10.1038/s41598-023-42384-8 doi: 10.1038/s41598-023-42384-8

Yang, C. S., & Dobbie, W. (2020). Equal protection under algorithms: A new statistical and legal framework. *Michigan Law Review*, *119*(2), 291–358. Retrieved from https://repository.law.umich.edu/mlr/vol119/iss2/3