Courts of Tomorrow: Evidence from a Nationwide Introduction of Generative AI

Sultan Mehmood ⓡ Christoph Goessmann ⓡ Elliott Ash*

March 27, 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 developed *JudgeGPT*, a custom-built generative AI assistant designed specifically for Pakistan's trial courts, and randomized 1,559 judges into one of three arms: (i) access to *JudgeGPT* with targeted training tailored to the tool; (ii) access to *JudgeGPT* with placebo training in technology and law; and (iii) a control group that received the placebo training but no access to the AI assistant. Treated judges are more likely to use generative AI and to support its broader adoption in the judiciary. Using an LLM-based, lawyer-validated measure of opinion quality, we find that *JudgeGPT* improves writing quality when paired with targeted training, but lowers it when provided without such training. Administrative records also point to sizable productivity gains: a one-standard-deviation increase in the share of judges assigned to *JudgeGPT* plus targeted training raises case resolutions by roughly 1,100 cases per district-year. Consistent with this pattern, usage data from *JudgeGPT* suggest that targeted training shifts judges away from open-ended legal search and toward more structured writing tasks, where the tool is likely more useful. We find no evidence that either treatment amplifies gender or ethnic bias in case outcomes or judicial language. Overall, the results suggest that generative AI can raise public-sector productivity, but that these gains may depend on targeted training that directs use toward tasks for which the tool is better suited.

*__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 generative AI affect state capacity? Recent theoretical work argues that AI may reshape production, the organization of expertise, and the direction of innovation, with potentially far-reaching consequences for institutions and long-run development (Aghion et al., 2019; Acemoglu, 2025). Early causal evidence from the private sector lends these arguments some empirical weight: access to large language models raises worker productivity substantially in information-intensive tasks (Noy and Zhang, 2023; Brynjolfsson et al., 2025). Yet whether these gains extend to the public sector remains an open question; governments operate under different incentive structures, face harder-to-measure outputs, and are bound by legal and constitutional constraints that private markets are not (Acemoglu and Lensman, 2024). The judiciary is an especially high-stakes setting to study these effects, because judges' decisions allocate liberty and property rights and shape the implementation of law and policy.

This paper presents evidence from 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 1,500 trial court judges, covering roughly half of the country's lower courts, to one of three groups. Judges in the first treatment arm received access to *JudgeGPT*, a custom generative AI legal assistant that is grounded in Pakistan's corpus of court rulings. This access was paired with targeted instruction on responsible and effective use. Judges in the second treatment arm received the same AI access paired with a general placebo course on technology and law (minimal material on AI). The control group received only the placebo course and no access to the AI assistant. The intervention was delivered through Pakistan's National Judicial Academy's "Technology and Law" program, enabling us to study how access to AI and complementary training shape judicial attitudes and productivity at scale.

The design produces rich findings. First, providing judges with access to *JudgeGPT* increases stated and actual adoption of generative AI in judicial work. Both access with targeted training and access with generic non-JudgeGPT training increase the frequency of *JudgeGPT* use by roughly 0.1–0.15 standard deviations (relative to the control group receiving no tool and a placebo course), indicating that exposure to the tool generates a first-stage shift into AI adoption. In contrast, we find no detectable effects on the use of other technologies, such as Microsoft Office, confirming that the interventions increase adoption of *JudgeGPT* specifically rather than digital tools use more broadly.

We also see effects on attitudes toward AI. Access to *JudgeGPT* increases judges' support for the use of generative AI in judicial decision-making. On a five-point scale, both treatment arms, including the arm without JudgeGPT-specific instruction, show agreement that AI can increase judicial productivity by about 0.25 points, a roughly seven percent increase over the sample mean. The similar increase in support for AI across both treatment arms is consistent with access to the tool itself shaping judges' attitudes toward AI in judicial work. Further, judges assigned access to *JudgeGPT* with training are likely to agree that AI can improve their productivity, with an increase of about 0.16 points on a five-point scale (roughly four percent of the mean). In contrast, judges who receive AI access with only the placebo technology and law training exhibit smaller changes in perceived productivity, with estimates that are indistinguishable from zero.

To shed light on how judges use *JudgeGPT* and what may account for the rise in case resolution, we analyze the *JudgeGPT* chat logs. These logs provide direct measures of engagement and the content of judges' requests. Because *JudgeGPT* is our OpenAI-based assistant integrated with a linked corpus of Pakistani rulings, we can observe how the tool is used in practice, not merely whether access is provided. We find usage patterns consistent with a shift toward structured drafting support. This channel is plausible given that large language models are well suited to writing assistance, and it aligns with evidence from the private sector showing sizable productivity gains concentrated in writing-intensive tasks (Noy and Zhang, 2023).

Using a sample of opinions from the judges written before and after the intervention, we can evaluate whether changes in the amount and character of AI usage affects judicial work product. We evaluate opinion quality before and after treatment using an LLM-based annotation approach that relies on pairwise comparisons and is validated against independent human coding. The results suggest that *JudgeGPT* with targeted training raises the quality of judicial opinions relative to the placebo group. In contrast, AI access without training is associated with a deterioration in opinion quality, suggesting that the effects of generative AI may depend on complementary guidance in how the tool is used.

Finally, we ask whether these changes in AI judicial support affect court efficiency. Using administrative statistics from the Justice Department of Pakistan, we find that districts with greater exposure to judges receiving *JudgeGPT* with targeted training resolve more cases per year. Overall, the results suggest that AI can improve judge performance along both the productivity and work quality margins.

These results speak to several strands of the literature. First, we contribute to the

emerging evidence on generative AI and workplace productivity. Recent studies document average productivity gains from LLM-based assistance in the private sector (Noy and Zhang, 2023), while also highlighting heterogeneity across workers, tasks, and educational backgrounds (Brynjolfsson et al., 2025; Dell'Acqua et al., 2026; Cruces et al., 2026). A natural next question is whether these gains extend to public-sector organizations, where production is governed by distinct rules and oversight (Finan et al., 2017). As emphasized by Agrawal et al. (2019), the policy and institutional environment can shape both the adoption and consequences of AI, raising a priori uncertainty about productivity effects in core public institutions. Public officials in such environments often face weaker performance incentives and stronger job protections, and technology adoption is frequently shaped by administrative directives rather than market demand (Muralidharan et al., 2021). These features make it unclear ex ante whether access to AI tools will deliver comparable productivity improvements in the public sector. Our study provides the first large-scale experimental evidence on the introduction of a generative AI tool in a central state institution. In the judiciary, decisions shape rights, liberties, and the rule of law, and AI may reconfigure task content and skill demands even when aggregate productivity effects are uncertain (Acemoglu et al., 2022; Brynjolfsson et al., 2017). Our results suggest that, in the judicial context, AI access with targeted training can translate into productivity gains.

Second, and more directly, the results add to the growing body of work on the role of AI in justice systems, particularly debates around the extent to which algorithmic tools can improve consistency, accuracy, and efficiency in legal decision-making (Ludwig and Mullainathan, 2021). While much of the existing literature compares algorithmic predictions with human judicial decisions to study misallocation, bias, or inconsistency (Kleinberg et al., 2018), related work also shows that GPT can serve as a scalable measurement tool for qualitative text, producing labels and attribute ratings that closely track human judgments across domains (Asirvatham et al., 2026). The intervention pairs access with targeted instruction, precisely intended to promote responsible use and to address concerns about opacity and biased reliance in AI systems, concerns that are prominent in both the broader AI ethics literature and legal applications (Buolamwini and Gebru, 2018; Vicente and Matute, 2023; Acemoglu et al., 2023; Yang and Dobbie, 2020). A complementary concern is whether courts can process cases quickly enough to deliver timely justice. Motivated by low case resolution rates and pervasive court congestion, we examine AI as a tool to accelerate judicial work. Court backlogs are widespread globally, in large part

because case closure rates remain low. As of 2024, India had over 51 million pending cases (The New York Times, 2024), US immigration courts held more than two million pending cases (U.S. Government Accountability Office, 2024), and the Crown Court of England and Wales had over 75,000 outstanding cases (Institute for Fiscal Studies, 2025). Our results suggest increased productivity and work quality.

Third, we contribute to work on state effectiveness by studying technological change in a core state institution. Prior research emphasizes how institutional design and personnel systems shape public-sector performance, including in the judiciary (Landes and Posner, 2009; Liu et al., 2022; Mehmood, 2022; Ash and MacLeod, 2024), and how incentives and monitoring affect bureaucratic effort (de Janvry et al., 2023). Yet there is limited causal evidence on how new technology alter the production of justice and the attitudes towards innovations inside courts, despite their central role in enforcing rules, resolving disputes, and facilitating reforms. We provide experimental evidence from frontline trial courts on the effects of a generative AI assistant and targeted training. In addition, to court productivity, we test whether access and training shift judges' beliefs about AI and willingness to adopt AI-enabled workflows (Pah et al., 2022), with results suggesting increases in support for AI and confidence in its use.

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, when a common law system was introduced to serve the administrative needs of the British *Raj*. The legal structure emphasized formal procedure, hierarchical authority, and the doctrine of precedent. After independence in 1947, Pakistan retained much of this infrastructure, including procedural codes and the use of English in legal proceedings. Over time, new statutes and Islamic jurisprudential elements were layered onto this inherited structure, but the organization of courts and case management, especially at the district level, continues to reflect colonial-era bureaucratic practice (Siddique, 2013). This legacy operates through a three-tier court hierarchy: district and sessions courts as courts of first instance (trial courts), provincial High Courts as the main appellate layer,

and the Supreme Court as the final court of appeal. Because cases begin in trial courts, where the core procedural and fact-finding stages of adjudication take place, including filing, hearings, witness examination, and evidence recording, much of the system's delay accumulates at the court of first instance (Law and Justice Commission of Pakistan, 2024). Figure A1 in Appendix A summarizes this structure in Panel A and describes the jurisdiction and composition of trial courts in Panel B.

Trial Courts. Courts of first instance, often referred to as the district judiciary in Pakistan, constitute the institutional core of legal service delivery. They adjudicate both civil and criminal cases and comprise three broad judicial ranks: Civil Judges (handling civil disputes such as land, contract, and family), Judicial Magistrates (lower-tier criminal cases), and Sessions and Additional Sessions Judges (serious, often violent criminal cases and civil cases). Because filing, hearings, witness examination, and evidence recording occur at this first instance stage, system-wide delay largely accumulates in these courts (Law and Justice Commission of Pakistan, 2024). For most citizens, they are the main point of contact with the formal justice system (Law and Justice Commission of Pakistan, 2023). The district judiciary faces acute capacity constraints, limited digital case-management infrastructure, and weak research and drafting support, while caseloads are high relative to staffing levels. For instance, Pakistan has fewer than 2 judges per 100,000 inhabitants (International Development Law Organization, 2018), compared with roughly 22 per 100,000 in the European Union (Council of Europe, 2024). Judges are recruited as career civil servants through provincial public service commissions, enjoy some tenure protections (mandatory retirements at age 60), and operate within a rigid hierarchy with limited performance-based incentives.

Context for Technological Reform. In recent years, Pakistan's judiciary has pursued institutional reforms to modernize court operations and address longstanding inefficiencies. A key component was the judiciary's "Technology and Law" program, which developed alongside broader judiciary-led technological upgrading, including the reconstitution of the National Judicial Automation Committee (NJAC) in 2023. Justice Syed Mansoor Ali Shah, a former Supreme Court Justice, played a prominent leadership role in these technology-focused judicial reform efforts during this period (Press Information Department, Government of Pakistan, 2023). At the same time, rapid advances in generative AI, especially the spread of tools such as ChatGPT, increased interest within Pak-

istan's judiciary in whether similar systems could assist judges, and some judges had already begun experimenting with these tools. In response, the judiciary began preparing for a phased introduction of an AI-based legal tool to facilitate accurate citation of case law and precedents. Our study was implemented within this broader reform context. In partnership with the Judicial Academy of Pakistan (FJA), we co-developed a customized generative AI legal assistant, *JudgeGPT*, using a retrieval-augmented generation (RAG) architecture built on national case law and legal texts, alongside structured training sessions focused on responsible and effective use. The intervention reached roughly half of all sitting trial judges (who consented to the use of this new technology), creating a unique opportunity to assess how generative AI affects attitudes, decision-making and productivity in a public-sector institution at scale. To support the rollout and mitigate potential risks, including misuse and over-reliance, we embedded a randomized controlled trial into the deployment process, assigning judges to structured access to the tool, with a subset also receiving training in responsible use. In *Constitutional Petition No. 1010-L*, the Supreme Court of Pakistan formally recognized the use of generative AI tools like *JudgeGPT* in trial courts. The ruling acknowledged the potential of such technologies co-developed by academia and state institutions to support judicial functions, while stressing the need for regulatory caution and institutional oversight (Supreme Court of Pakistan, 2024).

Ethics. Given this institutional setting and the potential risks of deploying generative AI in judicial work, ethical safeguards were built into the intervention from the outset, based on a collaborative understanding between the researchers and the judiciary. The intervention was embedded within ongoing national judicial training and technology initiatives administered by Pakistan's judiciary, and we worked in partnership with judicial authorities to randomize the timing and allocation of access in a manner consistent with their planned implementation strategy. All participating judges provided informed consent, and the study received ethical approval from two review bodies: 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). This dual approval process ensured a balanced approach, with input from both European and Pakistani institutions, and helped align the research with the ethical standards of both contexts. We also designed the study in line with recommendations for structured ethics reporting in social science research (Asiedu et al., 2021) and include an Ethics Appendix that discusses pol-

icy equipoise, the researcher's role, potential harms to participants and non-participants, conflicts of interest, and the possible misuse of the intervention and its findings. The intervention also included training programs across treatment arms, including a lecture on "legal ethics and technology." To judges assigned our *JudgeGPT*, we additionally 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. These trainings were designed as an additional safeguard to reduce uncritical reliance on generative AI and to ensure that judges had the framework needed to engage with the technology responsibly and ethically. While we cannot rule out misuse or misunderstanding, the aim was to make the tool's limitations more transparent and to encourage informed and cautious use in a sensitive institutional setting. Table A1 provides an overview of the treatment groups, detailing the training received and links to the lecture slides for each session.

3 Research Design and Data

3.1 Experimental Setup

Treatment Arms. We randomize judges into two treatment arms and one control group. Judges assigned to Treatment A ($n = 1,197$) receive access to *JudgeGPT* together with structured, practice-oriented training on its effective and responsible use. Judges assigned to Treatment B ($n = 180$) receive access to *JudgeGPT* together with a placebo course on AI and legal technology in judicial practice that does not provide instruction on use of the tool itself. Access to *JudgeGPT* was restricted to treatment-status judges through individualized, password-protected accounts. To limit contamination across arms, access was individualized and password-protected, login activity including IP addresses was monitored, and credential sharing was penalized by removal from the course and cancellation of platform access. The control group ($n = 182$) receives the same placebo course but does not receive access to *JudgeGPT*. For ethical and practical reasons, all participating judges, including those in the placebo arm, receive basic instruction on AI-related risks.[1] All judges completed a six-lecture program delivered on Zoom, with assignments, quizzes,

[1]We considered it important to provide basic guidance on the risks of generative AI even to judges in the placebo group, since they could still use tools such as ChatGPT or similar tools outside the study. In addition, because the intervention was delivered through the judiciary's broader Technology and Law initiative, we had agreed to provide training content viewed as valuable by the judiciary regardless of treatment status.

and a capstone project. Treatment A was implemented in three waves, whereas Treatment B and the control group were introduced in the later phase of the rollout. Figure 1 summarizes treatment assignment and training timing across batches. Assignment was randomized ex ante in Stata using judges' names and dates of birth. Individual treatment status was not disclosed prior to the start of training.

Experimental Structure. The staggered rollout of Treatment A (*JudgeGPT* and targeted training) and the larger size of that treatment arm were driven by institutional scheduling constraints under Pakistan's judiciary's centralized training calendar. Within those constraints, however, assignment remained randomized. In the robustness section, we re-estimate the main specifications using alternative batch-level comparisons, comparing each of Batches 1, 2, and 3, which received *JudgeGPT* with targeted training, to Batches 4a and 4b, which received either *JudgeGPT* with placebo training or placebo training alone. Across these comparisons, the estimates remain qualitatively similar, though they are somewhat less precisely estimated because of the smaller sample sizes in each comparison. All specifications also include strata fixed effects and follow the main analysis sample restrictions. A summary of treatment timing, assignment, and training timeline is illustrated in Figure 1.

3.2 JudgeGPT

JudgeGPT is a chatbot based on OpenAI's GPT-4 family of models, customized for the Pakistani context by (1) adding a context-specific system prompt and (2) attaching a corpus of approximately 130,000 Pakistani judicial opinions and 1,000 laws via retrieval-augmented generation (RAG, henceforth). For each query, the system retrieves the 10 most relevant document chunks and passes them to the model as context. The model then generates its response using both the retrieved materials and the underlying capabilities of the base model. Responses cite the retrieved passages in footnotes, which users can open to verify the underlying source material of laws and judicial opinions. Appendix Figure A2 shows the interface of the *JudgeGPT* (Panel A) and illustrates how case law and judicial opinions are cited within the platform (Panel B).

Specifically, the underlying RAG knowledge base contains 129,235 documents, comprising 128,292 judicial opinions and 943 Pakistani statutes from the Pakistan Code. The opinions span 1900–2024 and are drawn from the Supreme Court of Pakistan (n = 2,670), the Lahore High Court (n = 7,539), the High Court of Sindh (n = 8,627), and the easy-

law.ai database ($n = 109{,}456$), a source commonly used by Pakistani judges when drafting judgments. Because Pakistan's Supreme Court and provincial High Courts issue decisions that function as binding or highly persuasive precedent, these materials constitute the core body of legal authority used by trial judges in practice.

Figure 2 summarizes the retrieval workflow. Specifically, to support retrieval-augmented generation, each judgment and statute is segmented into chunks of roughly 1,000 tokens (a few thousand words), with a 128-token overlap across adjacent chunks to preserve local context. Given a user query, *JudgeGPT* maps the prompt into a retrieval query over the indexed legal corpus. A BM25-based retriever scores the candidate chunks and returns the 10 highest-ranked matches. These retrieved passages, totaling roughly 10,000 tokens, are then appended to the original prompt as additional context. The model then generates a response conditioned on this augmented input, with retrieved case law and statutory materials cited in footnotes where relevant. This design keeps the retrieved material within the model's context window while also limiting cost.[2]

The system prompt, which is not visible to users and is not directly user-editable, helps steer the model toward the intended tasks, reduces unnecessary repetition, sets response standards, and discourages some unsafe behavior. For *JudgeGPT*, we instructed the model as follows:

> `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 judgments.`
>
> `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. Legal jargon is OK.`

The first paragraph of the *JudgeGPT* prompt sets out the system's intended role and use cases, presenting it as a legal assistant for Pakistani judges across a broad set of tasks, including legal research, summarization, drafting, translation, and related administrative work. The second paragraph then instructs the model to base its responses on the materials provided to it, current Pakistani law, and retrieved case law, while citing only the authorities available within the platform.

[2]The GPT-4 family models used by *JudgeGPT* allow for a context window of 128,000 tokens, roughly 250 pages of text, and this capacity must also accommodate the recent conversation history.

In practice, however, adoption also depended on whether judges understood the interface, appropriate use cases, and key limitations. The intervention, therefore, combined access with training co-developed with the Federal Judicial Academy (FJA), Pakistan's primary institution for training judges and other judicial officers. The training introduced the AI tool for routine judicial work and emphasized careful use, including risks of bias and hallucination. The next subsection describes the two training arms and their differences.

3.3 Training Courses

The Treatment A training, delivered alongside access to *JudgeGPT*, was designed to support its effective use in routine judicial work. Judges in this group completed a structured curriculum on both the capabilities and limitations of generative AI in legal settings, combining lectures, ten practical take-home assignments, and a final project. The course comprised six 90-minute lectures delivered over three weeks, scheduled outside work hours from 7:30pm to 9:00pm Pakistan time, with participation via laptop or desktop on Zoom. In the final project, judges used *JudgeGPT* for core tasks such as legal research and drafting. The objective was also to reduce hallucination risks through repeated hands-on practice and explicit mitigation protocols, such as paying careful attention to source material and double-checking.

To protect research integrity and limit spillovers, judges were prohibited from sharing *JudgeGPT* account credentials. Violations resulted in disqualification from further access to the tool and the course. Compliance was monitored using platform logs, including anomalous IP address switching.

To encourage engagement, all successful participants received certificates jointly issued by ETH Zurich and the Federal Judicial Academy of Pakistan (FJA). In addition, we invited three judges from each batch, including the control group, to an AI and Judging Symposium in Zurich, which we announced in advance to encourage take-up. Participation was high among judges who consented to participate in the experiment: 1,559 enrolled, and over 95 percent submitted at least one assignment.

Appendix Table A1 summarizes the training conditions and the corresponding instructional materials. The table clarifies how access to *JudgeGPT* and course content differed across treatment arms and provides direct links to the lecture slides for all lectures in each condition, including the control group.

Targeted AI training versus placebo training. To distinguish tool-specific instruction from more general exposure to technology-in-law content, we implemented two separate training tracks within the Federal Judicial Academy's "Technology and Law" program. Judges in one group (Treatment A) received access to *JudgeGPT* and completed a targeted training on applied legal AI centered on our tool, with guided practice and assignments requiring its use for routine, writing-intensive tasks central to case processing, along with instruction on account protection and usage guidelines. In contrast, judges assigned to Treatment B (with *JudgeGPT*) and Control Group (without *JudgeGPT*) completed a placebo course covering broader court technology topics (e.g., case management systems, e-filing, and digital evidence, general AI risks on bias and hallucination) without *JudgeGPT*-specific instruction or guidance on deploying generative AI in judicial work. Because both tracks were embedded in the same program and delivered in a comparable format, differences in outcomes are less likely to reflect experimenter-demand effects or the salience of discussing AI and new digital tools per se. We designed the two training tracks as a compromise and extensive discussions with our judicial partners: to limit experimenter-demand concerns while also reducing risks associated with new AI tools, including bias and hallucinations.

3.4 Data and Descriptive Statistics

Our analysis draws on five complementary data sources covering judges' characteristics and attitudes, *JudgeGPT* use, written judgments, and court-level productivity. Appendix Table A2 defines all variables and describes their construction, while Appendix Table A3 reports summary statistics. Table A4 summarizes treatment timing, district coverage, and the maximum data span used. We briefly describe the data sources and main outcomes below, before turning to balance and attrition.

Data Sources. First, we use the pre-RCT assessment, which provides baseline information on participating judges, including demographics, experience, and self-reported pretreatment attitudes, such as their expectations about the AI-for-judges training. Second, we use a post-training survey administered roughly three months after course completion, which measures judges' engagement with the AI tool, their assessments of the training, and their views on whether the course could improve productivity and whether AI should be integrated into judicial work. Third, we assemble a new dataset of roughly 3,000 lower-court judicial opinions from final project submissions, including five judg-

ments per judge from before the experiment and five from after.[3] Such text data are rarely available in Pakistan's lower courts, making this a particularly valuable source for analysis. We use these texts to construct real outcome measures based on readability and on pairwise LLM evaluations of pre- and post-treatment judgments, which we validate against human assessments of decision quality. Fourth, from *JudgeGPT* itself, we collect login timestamps, IP addresses, and chat histories; we use the IP data to assess potential credential sharing and spillovers across intervention arms. Fifth, we digitize district-level judicial administrative data from the Law and Justice Commission of Pakistan (LJCP) yearbooks for 2021–2024, which report lower-court case closures overall and separately for civil and criminal cases. We use these publicly available administrative records to measure court productivity outside the experimental setting.

Outcome Measures. The treatment effects are evaluated using three broad sets of outcomes. The first consists of self-reported survey measures capturing support for generative AI, views on its judicial use, and beliefs about whether training programs can improve productivity. Judges were also asked how often they used AI in their work and were presented with questions such as: "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 four-point scale. Although self-reported, these measures help capture judges' attitudes toward AI, their reported use of it, and their views on the productivity effects of the training.

The second set of outcomes is based on text analysis of judgments. From these documents, we construct several quantitative indicators. Readability is proxied using the Gunning Fog Index, which incorporates sentence length and word complexity (Gunning, 1969). 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 actual judicial opinions. We use a GPT-5-mini prompt to compare them on overall clarity, conciseness, adequate presentation of the relevant facts, and the validity of the legal reasoning. See Appendix C1 for more detail on this measure and results from the human validation by licensed Pakistani lawyers. 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

[3]To our knowledge, lower-court judgments in Pakistan are generally not available in digitized form. Our data, though imperfect, constitute one of the first, and among the largest, collections of trial-court judgments assembled for Pakistan.

value higher than 0.5 since absent systematic improvement or deterioration, either opinion should be preferred with roughly equal probability.

The third set of outcomes is based on district-level judicial administrative data from the *Law and Justice Commission of Pakistan* (LJCP). The LJCP publishes annual statistical yearbooks reporting lower-court "case disposal", both in total and separately for civil and criminal cases. These records are aggregated at the district-court level and, to our knowledge, provide the only publicly available and consistently produced administrative source on lower-court performance in Pakistan. A useful feature of our setting is the broad geographic scope of the experimental rollout, which makes it possible to examine whether districts with a larger share of treated judges experience differential changes in cases resolved. We therefore construct district-level outcomes for total, civil, and criminal case closures and relate them to treatment exposure within the district court. These measures provide a basis for assessing the influence of *JudgeGPT* and training on judicial performance, with attention to both attitudes and actual output.

3.5 Estimation Approach

We estimate the causal effects of *JudgeGPT* access and training using the randomized assignment of judges to treatment arms. Because actual usage of the tool is not mechanically induced by assignment, we report both intent-to-treat (ITT) effects and treatment-on-the-treated effects using instrumental variables (2SLS).

Our baseline ITT specification is:

$$\begin{aligned} Y_i = {} & \beta_1 \text{ Assigned JudgeGPT Access and Targeted Training}_i \\ & + \beta_2 \text{ Assigned JudgeGPT Access and Placebo Training}_i \\ & + \alpha_s + \mathbf{X}_i'\Gamma + \varepsilon_i, \end{aligned} \tag{1}$$

where Y_i is an outcome for judge i, α_s are randomization-strata fixed effects (province $\times$ age-group strata), and $\mathbf{X}_i$ is a vector of pre-treatment covariates, including demographics, baseline workload and productivity, and baseline perceptions about AI and judicial work. The omitted category is the control group that received placebo training only and no access to *JudgeGPT*.

To estimate the effect of actual tool usage, we implement a two-stage least squares design that instruments realized access and use of *JudgeGPT* with randomized assignment.

The second-stage equation is:

$$\begin{aligned} Y_i = \pi_1 \, &\text{JudgeGPT Access and Targeted Training}_i \\ &+ \pi_2 \, \text{JudgeGPT Access and Placebo Training}_i \\ &+ \alpha_s + \mathbf{X}_i'\Gamma + u_i, \end{aligned} \tag{2}$$

where the endogenous regressors are instrumented with their corresponding assignment indicators. The first-stage relationships are reported in Section 4.1. Because judges in the control group did not receive access to *JudgeGPT*, noncompliance is plausibly one-sided, and the IV estimates can be interpreted as local average treatment effects for judges whose usage is induced by assignment.

Throughout, we report specifications with and without controls, although the inclusion of $\mathbf{X}_i$ primarily serves to improve precision given random assignment.

Balance and Attrition. Figure 3 assesses baseline balance between treated and control judges using pre-treatment characteristics collected at baseline. The figure plots standardized differences in means between all judges assigned to *JudgeGPT* treatment arms, including both Treatment A (with targeted training) and Treatment B (with generic training), and the control group, which received generic technology and law training without *JudgeGPT* access. The estimates are shown with 95% confidence intervals. Across demographics (age, gender), professional background (years of experience), baseline workload and productivity (cases decided, cases concurrently managed, and hours spent on administrative work), and pre-treatment attitudes toward generative AI (support for AI and expectations about AI for judges), the differences are small and the confidence intervals generally include zero. This pattern is consistent with successful random assignment and suggests that treatment status is not systematically related to observed baseline characteristics. Appendix Table A5 reports the corresponding pairwise mean comparisons between each treatment arm and the control group and includes the associated p-values. The conclusions are unchanged.

Attrition also appears limited and does not seem to undermine randomization. First, Appendix Table A6 repeats the balance exercise on the sample of survey respondents, that is, the non-attrited sample. The baseline covariates remain balanced across arms, suggesting that missing survey outcomes are unlikely to reflect differential selection on treatment assignment. Second, Appendix Table A8 reports Lee (2009) attrition bounds.

These bounds show that the main findings are not overturned even under relatively conservative assumptions. Our estimates remain generally positive for both treatment arms, indicating that differential survey response is unlikely to account for the results.[4]

Last, the results do not appear to be driven by any particular Treatment A batch. Because Treatment A (*JudgeGPT* with targeted training) was rolled out in three batches for logistical reasons dictated by the Pakistani judiciary, Appendix Table A11 compares each Treatment A batch separately to Batch 4a (*JudgeGPT* with generic training) and Batch 4b (control, no *JudgeGPT* but generic training). The estimates are broadly similar across the three treated Treatment A batches, suggesting that the results are not driven by a single batch.

4 Main Results

4.1 Usage of JudgeGPT

We begin by examining first-stage effects of randomized access to *JudgeGPT* on observed engagement with the tool. We compare take-up and continued use under targeted versus generic training, and trace how usage evolves over time. First-stage estimates for platform-recorded *JudgeGPT* usage are obtained from the following equations:

$$\begin{aligned}\mathbb{1}\{\text{Active JudgeGPT User}\}_i = {} & \pi_1 \,\text{JudgeGPT and Targeted Training}_i \\ & + \pi_2 \,\text{JudgeGPT and Placebo Training}_i \\ & + \alpha_s + \Gamma' X_i + \varepsilon_i\end{aligned} \tag{3}$$

$$\begin{aligned}\text{Number of JudgeGPT Logins}_i = {} & \kappa_1 \,\text{JudgeGPT and Targeted Training}_i \\ & + \kappa_2 \,\text{JudgeGPT and Placebo Training}_i \\ & + \alpha_s + \Gamma' X_i + \nu_i.\end{aligned} \tag{4}$$

$\mathbb{1}\{\text{Active JudgeGPT User}\}_i$ is an indicator equal to one if judge i logs into *JudgeGPT* at least twice, capturing adoption beyond a one-off visit. Number of JudgeGPT Logins$_i$ is the total number of logins over the post-assignment period and captures the intensity of engagement. α_s are randomization-strata fixed effects (province $\times$ age-group strata), and

[4] By construction, there is no attrition in the judicial administrative data on case disposal, which are observed at the district level for the full set of districts.

X_i is a vector of baseline covariates, including demographics (age, gender, experience), prior attitudes toward AI, baseline workload and productivity (cases decided, cases concurrently managed, and administrative hours), and baseline perceptions (work-life balance, confidence in public speaking and administrative work, and expectations about AI for judges). Both assignment indicators are equal to zero for judges in the placebo-training-only control group, which did not receive access to *JudgeGPT*.

Table 1 reports the first-stage estimates from Equation 3 and Equation 4. Assignment to *JudgeGPT* induces take-up, with higher adoption under targeted, tool-specific training than under placebo training. In columns 1–2, about 50 percent of judges in the targeted-training arm become active users, compared with about 20 percent under placebo training; the difference is statistically significant ($p < 0.01$). Columns 3–4 show a similar contrast on the intensive margin. Judges assigned targeted training logged in about 55 times on average over the post-assignment period, versus about 10 logins under placebo training. This gap is again statistically significant ($p < 0.01$) and qualitatively large (roughly sixfold), consistent with targeted, tool-specific instruction generating higher adoption than generic training.[5]

Figure 4 illustrates the relationship between the extensive and intensive margins by showing how continued *JudgeGPT* use evolves over time. The targeted-training group remains active at substantially higher rates throughout the post-assignment period, with only a gradual decline in logins. Even after about a year, 55% of users continue to log in *JudgeGPT* regularly. In contrast, the placebo-training group exhibits a more steep drop in usage: activity falls much more steeply from about 55% to 20% in about a year. Consistent with this visual pattern, the fitted linear trends indicate a much flatter decay in the targeted-training group than in the placebo group, and the difference in slopes is statistically significant ($p < 0.01$). These dynamics suggest that access to *JudgeGPT* alone can induce initial engagement, but tool-specific instruction is associated with more persistent use, consistent with training lowering the cost of integrating the tool into routine judicial workflow.

Table 2 reports estimates for Stated Frequent User of Generative AI, a binary indicator based on the survey item asking how often the judge uses generative AI models. Columns (1) and (2) present ITT estimates without and with controls, while Columns (3)

[5]Because the control group did not receive access to *JudgeGPT*, the first stage measures take-up among judges offered the tool, rather than changes from a pre-existing baseline. Take-up is meaningful, though incomplete, motivating the treatment-on-the-treated estimates, which we will discuss in the next subsection.

and (4) report the corresponding 2SLS estimates. Across specifications, assignment increases the probability that judges describe themselves as frequent users by about 0.10 in the ITT specifications. In the 2SLS specifications, JudgeGPT Access and Targeted Training increase frequent-use reporting by about 0.15. Relative to the sample mean, these effects correspond to roughly a 15 percent increase in the likelihood of reporting frequent generative AI use. Appendix Figure A3 shows that this self-reported measure is strongly correlated with actual usage, as captured by platform login activity. Judges reporting frequent post-treatment use send more than 200 additional messages and log in roughly 60 more times than judges reporting infrequent use. Panel A reports messages, and Panel B reports logins. This correspondence suggests that the survey measure captures meaningful variation in actual engagement with generative AI.

Microsoft Office provides a useful benchmark for whether the intervention shifted broad technological confidence rather than attitudes specific to generative AI. If the estimates above simply reflected experimenter-demand effects or a generalized enthusiasm for technology, we would expect to see similar improvements for a widely used tool such as Microsoft Office. We do not. Table 3 shows that the coefficients are small and statistically indistinguishable from zero in both the ITT and 2SLS specifications. This contrast supports the interpretation that the attitudinal effects are specific to generative AI rather than to software use more generally.[6]

4.2 How Judges Used *JudgeGPT*

Because we developed and deployed the *JudgeGPT* platform ourselves, we observe judges' interactions directly through the chat logs. We classify conversations by task using LLM-based coding.

Figure 5 compares judges assigned to targeted training with those assigned to placebo training, excluding the pure control group that did not have *JudgeGPT* access. Panel A aggregates conversations into three broad categories. Relative to the placebo trained *JudgeGPT*, judges who received targeted training devote a substantially larger share of their conversations to writing support and a smaller share to legal analysis, with little difference in other or non-core uses. The estimated shift toward writing support is sizable, about 8.8 percentage points, while the corresponding decline in legal research and analy-

[6]We find a similar null effect for WhatsApp, another widely used technology among judges. We do not report these results because baseline use is already near universal, about 95 percent, leaving little scope for detectable increases.

sis is about 8.4 percentage points. Panel B reports the same exercise using the full set of disaggregated task categories. The increase in writing-related use is driven primarily by style and language improvement, grammar and spelling correction, and translation.

These results suggest that targeted training changed the composition of use. It shifted judges toward editorial and language-support tasks rather than broader legal research. This pattern may also help interpret some of the changes we observe in opinion quality and court productivity, to which we turn in the next subsections.

4.3 Effect on Attitudes

Having shown that assignment to *JudgeGPT* and training influenced take-up and usage, we next examine effects on judges' attitudes. We use the experimental design to estimate intent-to-treat (ITT) and instrumental-variables (2SLS) effects on judge attitudes.

Table 4 reports estimates for *Support for Using GPT*, a 1–4 Likert-scale measure constructed from the survey item: "Do you support the use of generative AI assistants such as GPT in your job as a judge?" Columns (1) and (2) present ITT estimates without and with controls, while Columns (3) and (4) report the corresponding 2SLS estimates. Across specifications, assignment to *JudgeGPT* increases stated support by about 0.20 points, and the 2SLS estimates are similar in magnitude. Relative to the sample mean of 3.545, this implies a roughly 6–7 percent increase in judges' stated support for using generative AI tools in their own judicial work.

This increase in frequent use is relevant for interpreting perceived productivity effects, since any efficiency gains from generative AI are likely to depend on adoption in judicial work. Consistent with this mechanism, Table 5 reports estimates for GPT Course Can Improve Productivity, a 1–4 Likert-scale measure based on the survey item asking whether the AI-for-judges course can improve judges' productivity. In the ITT specifications, assignment to *JudgeGPT* access with targeted training increases agreement by about 0.15 points, whereas assignment to *JudgeGPT* access with placebo training yields smaller, statistically insignificant estimates. The 2SLS estimates in Columns (3) and (4) are similar in magnitude. Relative to the mean outcome of 3.620, this corresponds to roughly a 5 percent increase in perceived productivity gains from an AI-for-judges course. Although point estimates are somewhat larger under targeted training, we cannot reject equality between the targeted and placebo training coefficients.

Appendix Figure Figure A4 provides a visual summary of the main attitude outcomes

across treatment arms. Panel A shows that judges assigned to *JudgeGPT*, especially those receiving the targeted course, are more likely to agree that GPT can improve productivity. Panel B shows a similar pattern for support for using GPT in judicial work. Panel C shows that treated judges are also more likely to describe themselves as frequent users of generative AI. Although descriptive, the figure closely mirrors the regression results by showing consistently higher mean outcomes in the treatment arms than in the control group. Finally, because the control group had no direct access to *JudgeGPT*, noncompliance is plausibly one-sided. The IV estimates therefore capture the effect for judges whose use is induced by assignment, which in this setting coincides with the average treatment effect on treated users.

4.4 Effect on Opinion Writing

We next examine whether the intervention affected the opinion writing quality. Because the average lower-court case in Pakistan takes roughly four years to resolve, it will take time before downstream effects on case duration, appeals, or reversals in the High Courts can be meaningfully assessed. We therefore begin with text-based measures of opinion quality, described in greater detail in Appendix C. These measures offer an earlier indication of whether the intervention changed aspects of written decisions such as clarity, conciseness, and legal reasoning.

Our main measure relies on blinded pairwise comparisons of pre- and post-treatment opinions using LLM evaluators, an approach that is well suited to relative document assessment rather than the construction of stand-alone absolute quality scores (Ash and Hansen, 2023). To assess whether this proxy tracks professional legal judgment, we also conducted a human validation exercise with a licensed Pakistani Supreme Court lawyer with more than ten years of experience. The lawyer rated 90 pre/post opinion pairs, comparing anonymized opinion texts on the same -2 to $+2$ scale used in the LLM evaluation along dimensions of clarity, conciseness, factual presentation, and legal reasoning. As shown in Table A9, GPT agrees with the lawyer on the direction of quality differences in 70.1 percent of non-tie pairs, and agreement rises to 83.9 percent in the subset of cases where GPT and Gemini independently agree with one another. Taken together, these results suggest that, while necessarily noisy, the LLM-based measure captures meaningful variation in opinion quality and is informative about changes in judicial writing quality.

The statistical results are shown in Figure 6. The figure reports the share of within-

judge pre- versus post-treatment comparisons in which the post-treatment judgment is preferred by the model, with 95 percent confidence intervals. The horizontal dashed line at 50 percent corresponds to no systematic change in quality, such that pre- and post-treatment judgments are equally likely to be preferred.

Two patterns stand out. First, in the *JudgeGPT* plus targeted training group, post-treatment judgments are preferred in about 70 percent of comparisons. This share statistically exceeds 50 percent and is consistent with improved opinion quality. Second, in the *JudgeGPT* plus placebo training group, post-treatment judgments are preferred in about 35 percent of comparisons (below the 50 percent threshold), suggesting worse relative performance in the absence of targeted training. The active control group, which received placebo training but no *JudgeGPT*, remains close to 50%, as expected.

4.5 Effects on In-Group Bias

One concern is that AI assistance could affect in-group bias, either in case outcomes or in the language judges use in written opinions. We examine this possibility in two ways. First, we test whether access to *JudgeGPT* differentially affects case outcomes by litigant identity. Columns 1 and 2 of Table A9 ask whether treatment effects on defendant success vary by whether the defendant is female. The coefficient on the female-defendant indicator is negative, pointing to a baseline gender gap in case outcomes. The interaction terms are small in magnitude and imprecisely estimated. This pattern provides no clear evidence that the intervention systematically widened the gender gap against female defendants. Columns 3 and 4 of Table A9 repeat this exercise for religious identity and yield a similar pattern, with no clear evidence that treatment effects differed systematically between Muslim defendants and others under either targeted or generic training.

Second, Figure A5 asks whether AI assistance changes the expression of bias in judicial opinions. Using the same LLM-based approach as in recent work, we score each opinion on a 0–3 scale for anti-women and pro-Islam bias based on the language of the opinion itself (Asher et al., 2024; Ash et al., 2025). This follows earlier work and is described in greater detail in Appendix C2. The prompt instructs the model to distinguish potentially biased reasoning from the routine application of law. These dimensions are relevant in a meaningful share of cases, with gender flagged in about 39 percent of opinions and Islamic or Muslim identity in about 32 percent. We then aggregate these opinion-level measures to the judge level and estimate pre-post treatment differences across experi-

mental groups, considering both average bias scores and the prevalence with which each dimension appears in post-treatment opinions. Across both exercises, the estimates are small and imprecise, and the confidence intervals generally include zero. These results provide little evidence that AI assistance altered in-group bias in this setting.

4.6 Effect on Court productivity

A central question in our setting is whether *JudgeGPT* improved court productivity. Answering it is challenging even with judge-level random assignment, because productivity is observed only in aggregate. We therefore use district-level administrative statistics compiled by the Justice Department (LJCP), which record the number of cases closed in each district. These data measure the number of cases closed at the district-court level, allowing us to examine whether districts with greater exposure to the intervention experienced changes in case-processing volume.

To link judge-level randomization to this court-level outcome, we construct a district-level measure of treatment intensity. We define treatment intensity as the share of Batch 1 judges in a district assigned to *JudgeGPT* and targeted training in March 2024. These are the earliest treated judges for whom we can conduct a before-and-after analysis. This choice is guided by data availability: the most recent LJCP statistics we observe are reported at year-end 2024. We relate treatment intensity to district-level outcomes measured in December 2024, capturing effects roughly nine months after treatment, relative to the previous year's case closures. We estimate the following difference-in-differences specification:

$$\begin{aligned}\text{Cases Resolved} = \beta \times \text{Share Judges Treated}_i \times \mathbb{1}[\text{Year} = 2024]_t + \\ \gamma_i + \theta_t + \Gamma' \mathbf{X}_{it} + \varepsilon_{it}\end{aligned} \tag{5}$$

$\text{Cases Resolved}_{it}$ is the number of cases closed in district i in year t. ShareTreated_i is the share of Batch 1 judges in district i assigned to *JudgeGPT* and targeted training in March 2024. $\mathbf{1}\{t = 2024\}$ is a post indicator that equals one for year-end 2024 outcomes and zero in earlier years in our district-year panel. We include district fixed effects γ_i, year fixed effects θ_t, and time-varying controls X_{it}, including the share of criminal cases disposed, the number of judges, and the balance of civil and criminal cases. Standard errors are clustered at the district level.

Table 6 reports estimates of Equation 5 and shows that districts with greater *JudgeGPT* exposure resolve significantly more cases. The results are robust across the inclusion of year fixed effects, controlling for court size, and adding additional controls. Quantitatively, moving from zero to full exposure (a one-unit increase in the share measure) would be associated with a 0.164 standard-deviation increase in cases resolved. Using the LJCP distribution of district-year cases closed, this corresponds to roughly 10,000 additional resolved cases per district-year, which is about a 30% increase relative to the sample mean (approximately 30000 cases per district-year).

Appendix Figure A6 provides a distributional counterpart to the Table 6 estimates. The figure plots kernel densities of cases resolved after partialling out district and year fixed effects, where zero means a district resolved exactly as many cases as its own historical average, and negative values reflect below-average years. High- and low-treatment districts are defined as above and below the median share of Batch 1 judges assigned JudgeGPT with targeted training. Low-treatment districts cluster tightly around -10,000, with almost all mass between -20,000 and 0; high-treatment districts spread broadly rightward with substantial mass between 0 and +20,000 and a long tail beyond. The gap between the two distributions is consistent in direction with our point estimates, as high-treatment districts systematically resolve more cases relative to their own historical baseline than low-treatment districts, and the spread of high-treatment mass across the full right tail suggests the effect is broad-based rather than driven by a small number of outliers.

Because trial courts adjudicate both civil and criminal matters, we also examine heterogeneity by case category using the district-level aggregates, which separately report the number of civil and criminal cases closed. Table 7 suggests an increase in both types of cases, with a larger increase in criminal dispositions. One possible explanation is that criminal matters tend to turn over faster than civil disputes, so short-run changes may be more readily reflected in criminal cases. Importantly, the estimated effects are not concentrated in a single category. While this does not rule out reallocation across dockets, the broad-based pattern is more naturally interpreted as consistent with an increase in effective adjudicatory capacity, for example, through faster drafting and legal research, improved task management, or reduced time costs for routine procedural steps.

Appendix Figure A7 provides a graphical complement to the table estimates by plotting coefficients from an event-study specification normalized to 2023 as the omitted year. We view this figure as intended to assess whether the estimated productivity effects are

visible under alternative implementations of the same design. In particular, we report both a standard difference-in-differences event-study and an event-study produced by a synthetic difference-in differences estimator, following Arkhangelsky et al. (2021). Across both approaches, pre-period estimates are small, while the 2024 coefficient is positive and precisely estimated, consistent with higher post-period case dispositions in more exposed districts.

Figure A8 complements this analysis by making the role of treatment exposure intensity explicit. Rather than imposing a linear dose-response relationship, we ask how case dispositions change across districts that receive progressively higher shares of treated judges. Following the recent estimator of Callaway et al. (2024) for continuous treatments, we compare each exposure bin to districts with zero exposure.[7] We discretize treatment intensity into quantiles defined by the treated judge share and estimate the corresponding differences in cases resolved relative to districts with zero treated judges.

In the lower quantiles, where treated judges comprise between zero and sixteen percent of the bench, point estimates are positive but confidence intervals are wide, spanning both sides of zero: a small number of treated judges is simply not enough to shift district-level case resolution in a detectable way. Only once the treated share exceeds sixteen percent does the estimate become positive and precisely estimated, with confidence intervals that exclude zero. This pattern suggests that productivity gains from AI assistance materialize once a sufficient share of the bench is exposed.

5 Robustness Analysis

We complement the main analysis with a range of sensitivity checks, including tests for spillovers, additional balance and attrition tests, multiple hypothesis testing, and randomization inference. We also discuss external validity. These exercises suggest that the findings are robust to standard concerns. Ethical considerations are discussed in Appendix B1. Appendix B2 reports the full pre-analysis plan and documents the deviations from it, which arose when additional data became available or when institutional constraints imposed by the judiciary required adjustments. In line with (Banerjee et al.,

[7] The estimating equation follows Callaway et al. (2024):

$$\Delta Y_i = \alpha + \sum_{q \in Q} \beta_q \, \mathbf{1}[J_i \in q] + \Delta X_i + \varepsilon_i,$$

where J_i denotes the share of treated judges in district i.

2020), we treat the PAP as a guide rather than a rigid constraint, especially in a high-stakes experiment that required practical judgment in response to stakeholder demands and judicial scheduling constraints.

Spillovers. To support stable treatment assignment, we implemented safeguards to limit unauthorized access. Each account was protected by an automatically generated password, and credential sharing was explicitly prohibited under threat of disqualification. We then directly probe spillovers among control judges using two complementary district-level exposure measures. Table A7 reports the results. Panel A tests for spillovers measured by *observed usage*. We construct a district-level intensity proxy from platform logs, the number of unique API calls per judge in the district, and interact this measure with an indicator for being assigned no access. If control judges benefit indirectly through nearby use (e.g., shared outputs or informal diffusion), their outcomes should be higher in districts with greater logged activity. We find no evidence of such spillovers: the interaction estimates are close to zero across outcomes. Panel B tests for spillovers measured by *treatment density*. We interact the no-access indicator with the district share of judges assigned JudgeGPT access. If spillovers operate primarily through the presence of treated colleagues, control judges should look different in districts with higher treated shares. The corresponding estimates are imprecise and do not suggest systematic differences in control outcomes by treated share. These tests are consistent with spillovers being limited in our setting; to the extent any spillovers remain, they would attenuate treatment-control differences.

Additional Balance and Attrition Tests. We complement the covariate-by-covariate balance table with a joint test of orthogonality. Following McKenzie (2015) and Bruhn and McKenzie (2009), we regress treatment assignment on the full set of baseline covariates (and randomization strata, where applicable) and test the joint null that these covariates do not predict assignment. This joint test, reported in Table A10, provides a compact summary of balance across many characteristics and is informative in cases where individual differences may be small but systematically aligned. We also extend our attrition analysis beyond the conventional (Lee, 2009) bounds reported earlier. We implement generalized Lee bounds that relax unconditional monotonicity by allowing selection responses to vary with pre-treatment covariates (Semenova, 2025). Table A8 reports these results and shows that the targeted-training results remain positive even under this demanding

exercise: for all three outcomes, the generalized bounds continue to lie on the positive side, albeit with wider intervals than under standard Lee bounds. The placebo-training comparison is noticeably weaker. Under standard Lee bounds, the effects remain positive for some outcomes, but under generalized Lee bounds the intervals widen substantially and no longer rule out zero. These results suggest that the positive attitudinal effects are not merely an artifact of selective attrition, and that they are substantively more robust for the targeted training arm than for the placebo-training arm.

Multiple Hypothesis Testing Adjustments. Because we examine a number of outcomes, we also assess whether the main findings are sensitive to adjustments for multiple inference. Following standard practice, we group conceptually related outcomes into families and report adjusted p-values alongside the conventional ones. In particular, we implement procedures that control the family-wise error rate and the false discovery rate, allowing us to distinguish results that remain robust under more demanding correction methods from those that are less precisely estimated. As emphasized in practical overviews such as McKenzie (2021), no single adjustment is universally preferred, so reporting several approaches can be informative when outcomes are correlated. The adjusted results, reported in Table A9, show that broadly the results remain statistically meaningful under these corrections, whereas weaker comparisons become less precise. This pattern suggests that the main conclusions are not driven by selective emphasis on individual significant coefficients.

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 A9 and indicate that the main findings are unlikely to result from idiosyncratic assignment. For each of our results, we find that estimated coefficients are extreme compared to generated counterfactuals, for 3000 permutation not a single one appeared to be of a greater magnitude than the actual coefficient, which suggests that the effect is present in the data and unlikely to be driven by chance.

Bootstrap Standard Errors. We also complement conventional inference with wild bootstrap (MacKinnon and Webb, 2017), which are often useful when treatment is assigned finite-sample uncertainty may not be well captured by asymptotic approximations. In our setting, this provides a more demanding check on whether the main results are sensitive to the way standard errors and p-values are computed. Figure A10 reports these results, where we Wild bootstrap using 999 replications. The resulting estimates remain similar in magnitude to the baseline results, although the p-values are modestly larger. Even under this more conservative procedure, however, the main effects continue to clear conventional significance thresholds. These results suggest that the core findings are not an artifact of overly optimistic conventional inference but instead remain present under a stricter bootstrap-based approach.

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 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.

Excluding Regions. To assess whether the results are driven by any region, we re-estimate the baseline specification repeatedly, omitting each region in turn. Figure A11 reports the resulting leave-one-out estimates for all four main outcomes. Across these exercises, the point estimates remain tightly clustered around the baseline and preserve the same qualitative pattern as in the main results. For the three survey-based outcomes in Panels A to C, the coefficients are broadly positive and remain statistically distinguishable from zero throughout. For case resolution in Panel D, the estimates are likewise stable and remain positive, with only modest movement in magnitude across exclusions. Overall, the exercise provides some reassurance that the findings reflect a pattern common across Pakistan's administrative geography rather than the outsized influence of any particular region.

Heterogeneity by Province. Table PAP 4 reports treatment effects separately by province. The pattern is broadly consistent across Punjab, Sindh, Khyber Pakhtunkhwa, and Balochistan: judges assigned to JudgeGPT with targeted training show positive shifts in support for generative AI and higher login counts across all four provinces, with coefficients on logins ranging from 67 in Punjab to 132 in Balochistan. Effects on self-reported frequent use and course completion are more variable and estimated with less precision, reflecting smaller within-panel samples and greater outcome variance. The placebo training arm yields systematically smaller attitudinal shifts and lower login counts, consistent with the main results. The residual "Other Provinces" panel is an exception, where the targeted training coefficient on AI support is the largest in the table but the placebo arm produces an anomalous negative estimate on course completion; given the small and het-

erogeneous set of courts in this category, we treat these estimates with caution. Overall, the provincial results are broadly consistent with the main findings. Adoption and attitudinal change are not concentrated in any single region, and the core pattern—targeted training amplifying the effect of tool access—is evident across Pakistan's major provinces.

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

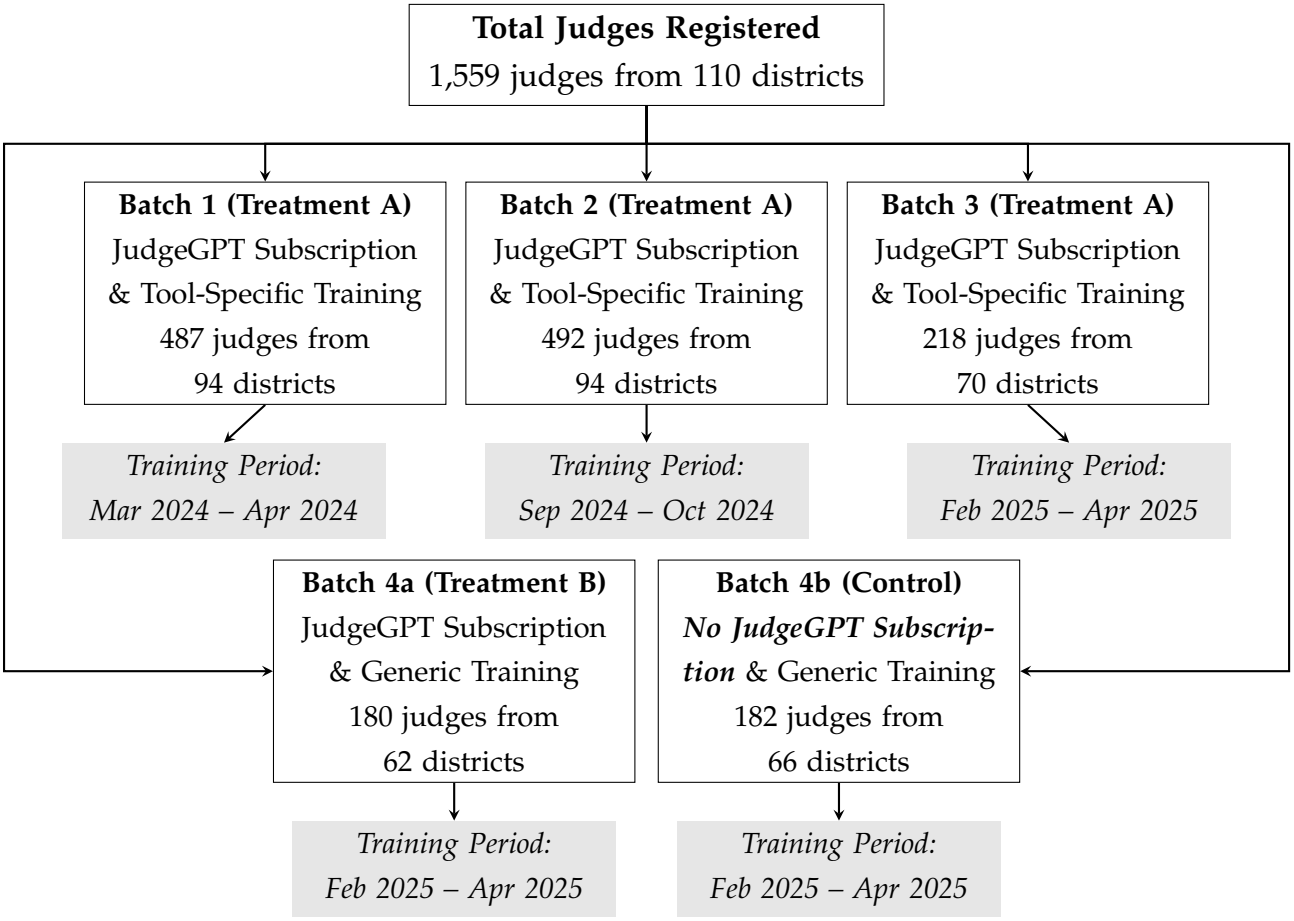
of legal arguments. These measures are imperfect and cannot speak to all dimensions of correctness or fairness. They nonetheless provide reassuring evidence that the productivity gains we document do not mechanically reflect shorter or less structured decisions in the short run. More generally, the design of JudgeGPT, 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.

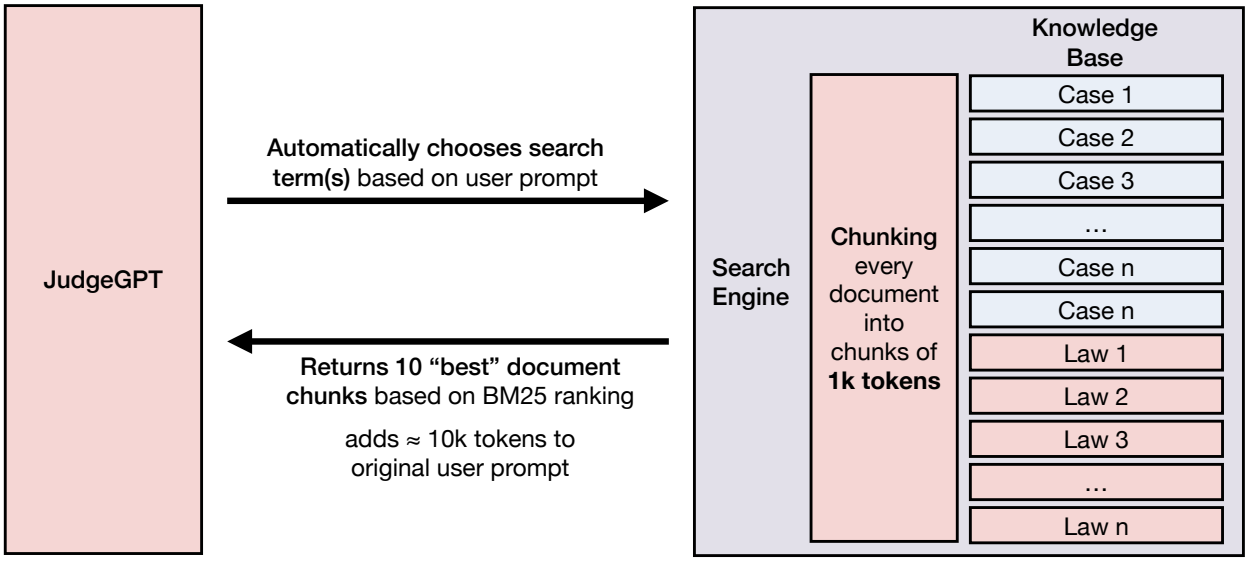
TABLES AND FIGURES

Figure 1: Treatment Assignment and Timing



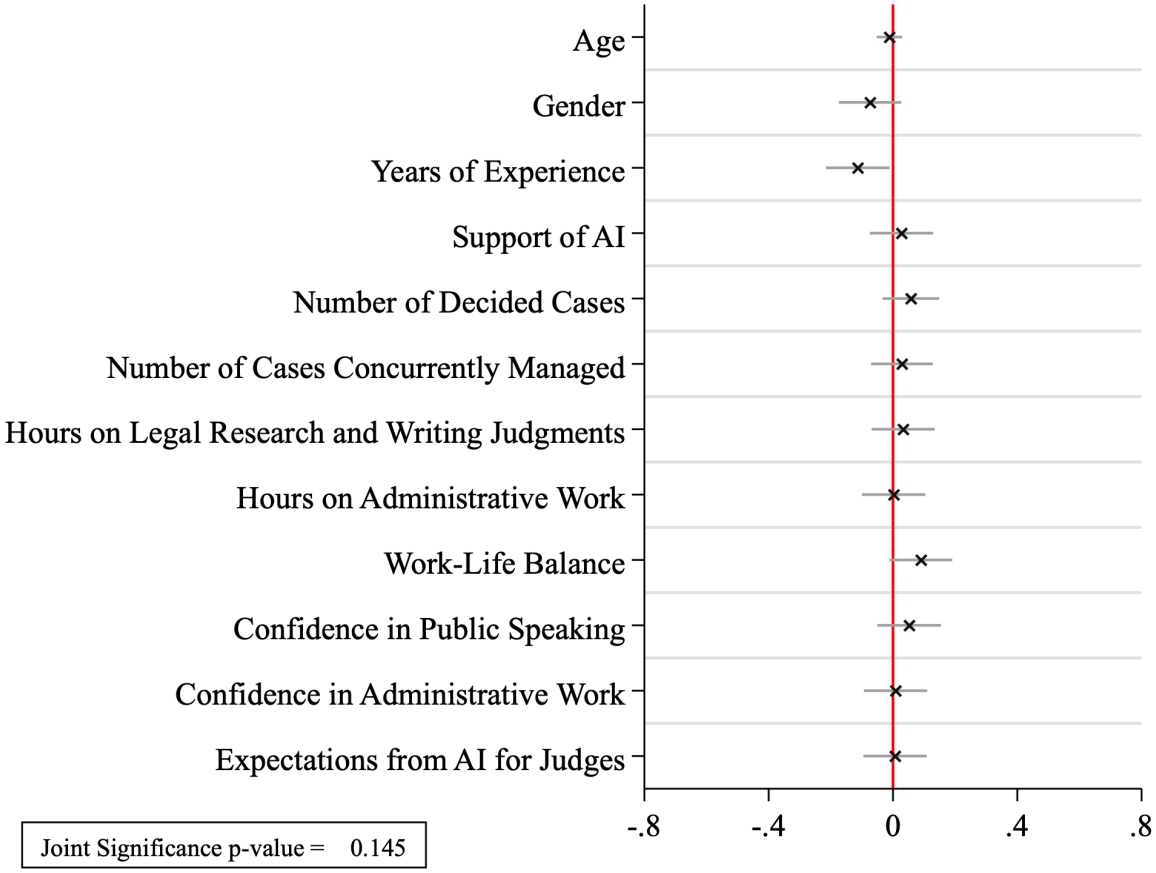
Note: This figure illustrates participant assignment and training timing in the randomized controlled trial embedded within Pakistan's Federal Judicial Academy "Technology and Law" program. A total of 1,559 trial-court judges from 110 districts were enrolled and assigned across treatment and control batches. The first treatment arm (Batches 1, 2, and 3; 1,197 judges in total) received access to the custom generative AI assistant, *JudgeGPT*, together with targeted, practice-oriented training on its effective and responsible use. The second treatment arm (Batch 4a; 180 judges) received access to *JudgeGPT* alongside a generic technology-and-law course that did not provide instruction on use of the tool itself. The control group (Batch 4b; 182 judges) received the same generic course without access to *JudgeGPT*.

Figure 2: JudgeGPT Case Retrieval System



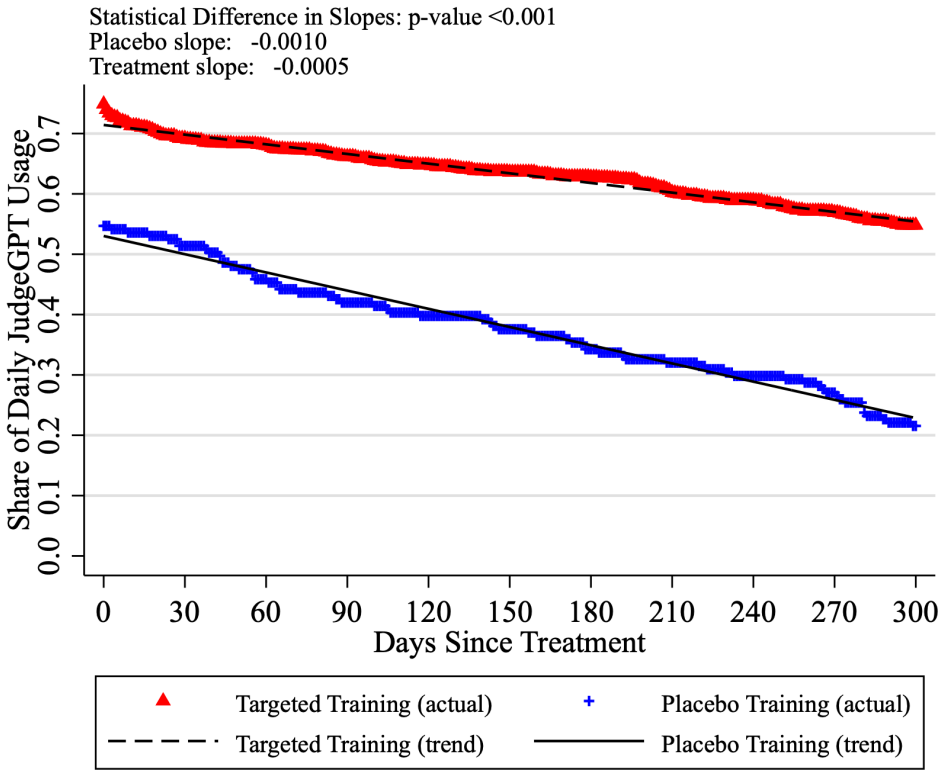
Note: Illustration of the Retrieval Augmented Generation (RAG) module that JudgeGPT relies on to respond to user requests. The knowledge base consists of Pakistani judicial opinions and statutes that are pre-chunked into segments of roughly 1,000 tokens and indexed by a search engine. For each user prompt, *JudgeGPT* automatically formulates search terms, queries the indexed corpus, and retrieves the 10 highest-ranked chunks using BM25. These retrieved chunks are appended to the prompt (about 10,000 tokens of additional context) and serve as the evidentiary basis for the generated answer. Footnote citations in the interface link each claim to the corresponding retrieved chunk, enabling users to inspect the source text.

Figure 3: Treatment Balance by Judge Characteristics



Note: This figure assesses baseline balance between judges assigned to the *JudgeGPT* treatment arms and those assigned to the control group. The treatment group combines Treatment A (*JudgeGPT* with targeted training) and Treatment B (*JudgeGPT* with generic training); the control group received generic technology and law training without *JudgeGPT* access. The points plot standardized mean differences in pre-treatment characteristics (treatment minus control), with 95% confidence intervals.

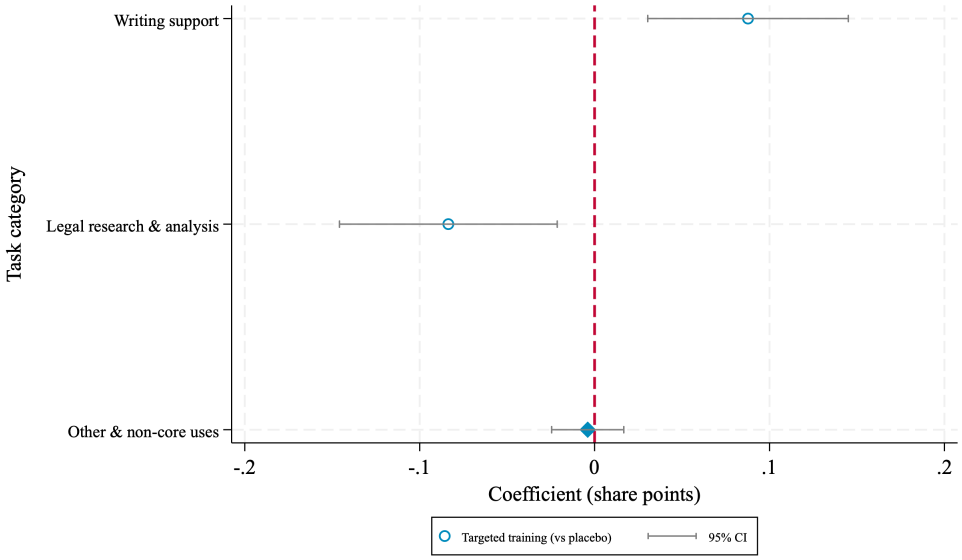
Figure 4: Dynamics of JudgeGPT Usage, by Training Type



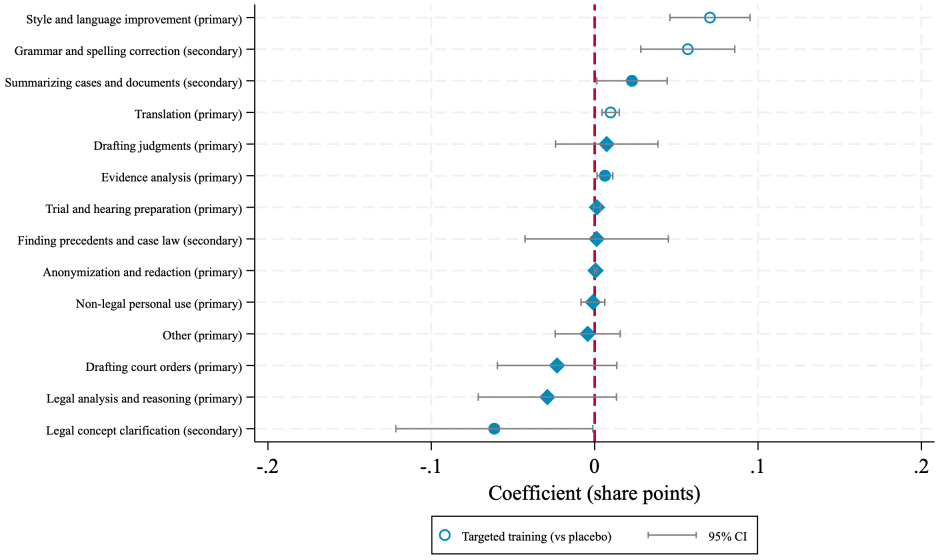
Note: The figure plots the share of judges who remain active on *JudgeGPT* over time, comparing the randomly assigned targeted-training and placebo-training groups (both received access to *JudgeGPT*). For each judge, we define the engagement window as the number of days between the first and last recorded login. A judge is coded as active at day t if this window is at least t days; judges with a single login are active only at $t = 0$. We compute group-level activity shares by averaging these indicators within assignment group for $t = 0, \ldots, 365$.

Figure 5: Effects of Training Type on JudgeGPT Usage Categories

Panel A: Unified Categories

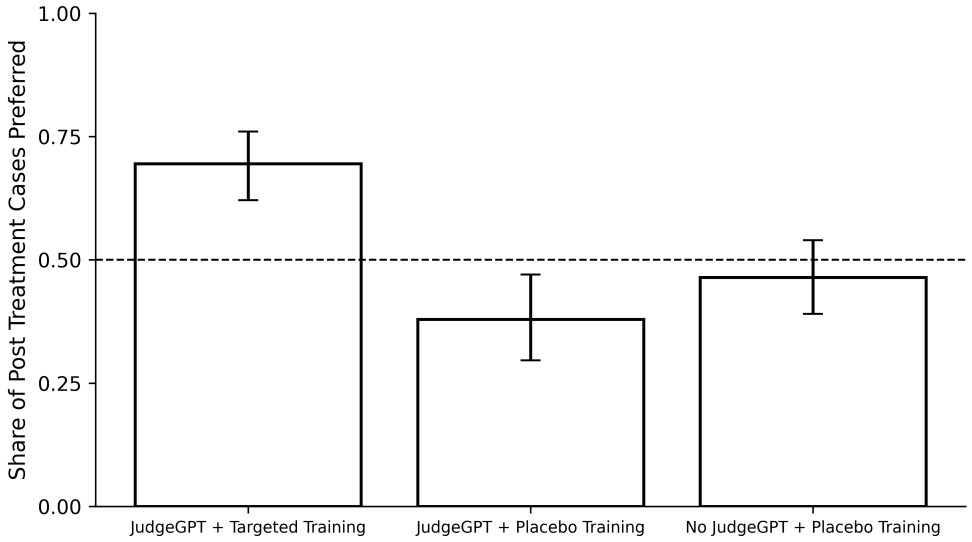


Panel B: Specific Categories



Note: The figure shows means and 95% confidence intervals estimated from regressions on count of different themes of JudgeGPT Chatlog. All specifications include strata fixed effects and baseline controls.

Figure 6: Treatment Effects on Opinion Quality



Note: The figure shows means and 95% confidence intervals (constructed from standard errors of the mean) in the pre-post change 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. The LLM made pairwise comparisons of all pairs of pre-treatment opinions and post-treatment opinions within judge. The bars give the share of comparisons where the post-treatment opinion was preferred, meaning that above 50% indicates an increase in quality, and below 50% indicates a decrease in quality.

TABLE 1: JudgeGPT Access and Platform Usage

	Extensive Margin		Intensive Margin	
	(1)	(2)	(3)	(4)
	Use of JudgeGPT			
Assigned JudgeGPT Access and Targeted Training	0.545***	0.535***	58.500***	56.453***
	(0.015)	(0.016)	(2.814)	(2.799)
Assigned JudgeGPT Access and Placebo Training	0.217***	0.222***	9.508***	9.519***
	(0.031)	(0.032)	(1.679)	(2.047)
Strata Fixed Effects	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes
Adj. R^2	.153	.158	.085	.094
p-val. (Targeted Training = Placebo Training)	< 0.001	< 0.001	< 0.001	< 0.001
Mean Dep. Var.	.445	.445	46.535	46.535
Observations	1559	1559	1559	1559

Note: Robust standard errors are in parentheses. The table reports intent-to-treat effects of randomized assignment to *JudgeGPT* and training on *JudgeGPT* usage. *JudgeGPT* Access and Targeted Training equals one for judges assigned access with tool-specific training. *JudgeGPT* Access and Placebo Training equals one for judges assigned access with generic technology-in-law training. The control group received placebo training only and no access to *JudgeGPT*. Extensive margin (cols. 1–2): indicator equal to one if the judge logs into *JudgeGPT* at least twice within 300 days of the first login (a window chosen to align with the persistence patterns in data). Intensive margin (cols. 3–4): number of logins per judge-month over the post-assignment period. All specifications include randomization-strata fixed effects; columns with controls additionally include pre-treatment judge characteristics. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE 2: JudgeGPT Access and Generative AI Usage

	ITT		2SLS	
	(1)	(2)	(3)	(4)
	Stated Frequent User of Generative AI			
Assigned JudgeGPT Access and Targeted Training	0.113**	0.118***		
	(0.046)	(0.045)		
Assigned JudgeGPT Access and Placebo Training	0.131**	0.142**		
	(0.058)	(0.056)		
JudgeGPT Access and Targeted Training			0.122**	0.127***
			(0.049)	(0.048)
JudgeGPT Access and Placebo Training			0.155**	0.167**
			(0.068)	(0.065)
Strata Fixed Effects	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes
Adj. R^2	.016	.044	.036	.064
Kleibergen-Paap F-statistic			2734.089	2701.6 3
p-val. (Targeted Training = Placebo Training)	0.658	0.534	0.471	0.364
Mean Dep. Var.	0.820	0.820	0.820	0.820
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. Controlsinclude age, gender, years of experience, support for AI, baseline workload and productivity (cases decided, cases concurrently managed, and hours spent on administrative work), and baseline perceptions (work-life balance, confidence in public speaking, administrative work, and expectations about AI for judges). 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 3: JudgeGPT Access and Familliarity with Microsoft Office

	ITT		2SLS	
	(1)	(2)	(3)	(4)
	Familliar with Microsoft Office			
Assigned JudgeGPT Access and Targeted Training	-0.012	-0.008		
	(0.045)	(0.045)		
Assigned JudgeGPT Access and Placebo Training	-0.094	-0.083		
	(0.066)	(0.065)		
JudgeGPT Access and Targeted Training			-0.013	-0.009
			(0.049)	(0.047)
JudgeGPT Access and Placebo Training			-0.111	-0.099
			(0.078)	(0.076)
Strata Fixed Effects	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes
Adj. R^2	.031	.048	.024	.042
Kleibergen-Paap F-statistic			2960.705	2903.574
Mean Dep. Var.	0.690	0.690	0.690	0.690
p-val. (Targeted Training = Placebo Training)	0.118	0.149	0.115	0.145
Observations	1076	1076	1076	1076

Note: Robust standard errors are in parentheses. The dependent variable is *GPT Course Can Improve Productivity*, measured on a 1–5 scale. Columns (1) and (2) report intent-to-treat estimates. *Assigned JudgeGPT Access and Targeted Training* equals one for judges assigned *JudgeGPT* access with tool-specific training, and *Assigned JudgeGPT Access and Placebo Training* equals one for judges assigned *JudgeGPT* access with the generic technology-in-law training. Columns (3) and (4) report 2SLS estimates in which *JudgeGPT Access and Targeted Training* and *JudgeGPT Access and Placebo Training* are instrumented with their corresponding assignment indicators. All specifications include strata fixed effects; columns with controls additionally include age, gender, years of experience, support for AI, baseline workload and productivity (cases decided, cases concurrently managed, and hours spent on administrative work), and baseline perceptions (work-life balance, confidence in public speaking, administrative work, and expectations about AI for judges). The table reports adjusted R^2 and the Kleibergen-Paap first-stage F-statistic. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE 4: JudgeGPT Access and Support for Generative AI Usage

	ITT		2SLS	
	(1)	(2)	(3)	(4)
	Support the Usage of GPT			
Assigned JudgeGPT Access and Targeted Training	0.210***	0.212***		
	(0.072)	(0.070)		
Assigned JudgeGPT Access and Placebo Training	0.223**	0.218**		
	(0.090)	(0.088)		
JudgeGPT Access and Targeted Training			0.226***	0.228***
			(0.077)	(0.074)
JudgeGPT Access and Placebo Training			0.263**	0.255**
			(0.105)	(0.102)
Strata Fixed Effects	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes
Adj. R^2	.013	.077	.018	.083
Kleibergen-Paap F-statistic			2960.705	2862.091
Mean Dep. Var.	3.547	3.547	3.547	3.547
p-val. (Targeted Training = Placebo Training)	0.824	0.919	0.606	0.700
Observations	1076	1076	1076	1076

Note: Robust standard errors are in parentheses. The dependent variable is *Support the Usage of GPT*, measured on a 1–4 scale. Columns (1) and (2) report intent-to-treat estimates. *Assigned JudgeGPT Access and Targeted Training* equals one for judges assigned *JudgeGPT* access with tool-specific training, and *Assigned JudgeGPT Access and Placebo Training* equals one for judges assigned *JudgeGPT* access with the generic technology-in-law training. Columns (3) and (4) report 2SLS estimates in which *JudgeGPT Access and Targeted Training* and *JudgeGPT Access and Placebo Training* are instrumented with their corresponding assignment indicators. All specifications include strata fixed effects; columns with controls additionally include age, gender, years of experience, support for AI, baseline workload and productivity (cases decided, cases concurrently managed, and hours spent on administrative work), and baseline perceptions (work-life balance, confidence in public speaking, administrative work, and expectations about AI for judges). The table reports adjusted ^{2} and the Kleibergen-Paap first-stage F-statistic. $^*p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE 5: JudgeGPT Access and Support for Judicial AI Course

	ITT		2SLS	
	(1)	(2)	(3)	(4)
	GPT Course Can Improve Productivity			
Assigned JudgeGPT Access and Targeted Training	0.140**	0.149**		
	(0.064)	(0.062)		
Assigned JudgeGPT Access and Placebo Training	0.069	0.079		
	(0.087)	(0.085)		
JudgeGPT Access and Targeted Training			0.151**	0.160**
			(0.069)	(0.066)
JudgeGPT Access and Placebo Training			0.081	0.092
			(0.101)	(0.099)
Strata Fixed Effects	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes
Adj. R^2	.007	.061	.012	.066
Kleibergen-Paap F-statistic			2960.705	2862.091
p-val. (Targeted Training = Placebo Training)	0.268	0.277	0.347	0.354
Mean Dep. Var.	3.622	3.622	3.622	3.622
Observations	1076	1076	1076	1076

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 1-4. 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 age, gender, years of experience, support for AI, baseline workload and productivity (cases decided, cases concurrently managed, and hours spent on administrative work), and baseline perceptions (work-life balance, confidence in public speaking, administrative work, and expectations about AI for judges). 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 6: JudgeGPT Access and Court Efficiency

	(1)	(2)	(3)	(4)
	Cases Resolved			
Share JudgeGPT and Targeted Training × Post	0.157***	0.154***	0.174***	0.164***
	(0.009)	(0.007)	(0.009)	(0.008)
District Fixed Effects	Yes	Yes	Yes	Yes
Year Fixed Effects	No	Yes	Yes	Yes
# of Judges in the District	No	No	Yes	Yes
Controls	No	No	No	Yes
Adj. R^2	.857	.874	.894	.917
Mean Dep. Var.	0	0	0	0
# Clusters	118	118	118	118
Observations	472	472	472	472

Note: Robust standard errors, clustered by district, are in parentheses. The unit of observation is district-year. The dependent variable is the number of cases resolved in a district-year, standardized to have a mean of zero and a standard deviation of one. *Share JudgeGPT and Targeted Training × Post* equals the number of judges assigned *JudgeGPT* with targeted training devided on district courts working strenght, interacted with a post indicator for 2024. All specifications include year fixed effects; columns (2)–(4) additionally include district fixed effects. Column (3) controls for the number of judges in the district, and column (4) adds the full set of district-year controls that include the share of criminal cases in institution, the share of criminal cases disposed, the working strenght, the balance of civil cases, the balance of criminal cases. The table also reports the number of clusters and the number of observations. Only districts with treated judges at the beginning of 2024 (first batch) are included.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE 7: JudgeGPT Access and Court Efficiency, by Case Type

	(1)	(2)	(3)	(4)
	Civil Cases Resolved		Criminal Cases Resolved	
Share JudgeGPT and TargetedTraining × Post	0.064***	0.054***	0.229***	0.219***
	(0.008)	(0.007)	(0.010)	(0.010)
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	.909	.928	.874	.899
Mean Dep. Var.	0	0	0	0
# Clusters	118	118	118	118
Observations	472	472	472	472

Note: Robust standard errors in parentheses are clustered by District. The unit of observation is district-year. The dependent variables are the standardized (to have a mean of zero and a standard deviation of one) number of criminal and civil cases disposed in district year. *Share JudgeGPT and Targeted Training × Post* equals the number of judges assigned *JudgeGPT* with targeted training devided on district courts working strenght, interacted with a post indicator for 2024. The share of criminal cases in institution, the share of criminal cases disposed, the working strenght, the balance of civil cases, the balance of criminal cases 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

REFERENCES

Acemoglu, Daron, "The simple macroeconomics of AI," *Economic Policy*, 2025, *40* (121), 13–58.

_ **and Todd Lensman**, "Regulating Transformative Technologies," *American Economic Review: Insights*, 2024, *6* (3), 359–376.

_ **, Asuman Ozdaglar, and James Siderius**, "A Model of Online Misinformation," *The Review of Economic Studies*, 12 2023, *91* (6), 3117–3150.

_ **, David Autor, Jonathon Hazell, and Pascual Restrepo**, "Artificial Intelligence and Jobs: Evidence from Online Vacancies," *Journal of Labor Economics*, 2022, *40* (S1), S293–S340.

Aghion, Philippe, Céline Antonin, and Simon Bunel, "Artificial intelligence, growth and employment: The role of policy," *Economie et Statistique/Economics and Statistics*, 2019, (510-511-512), 150–164.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb, "Economic Policy for Artificial Intelligence," *Innovation Policy and the Economy*, 2019, *19* (1), 139 – 159.

Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager, "Synthetic Difference-in-Differences," *American Economic Review*, 2021, *111* (12), 4088–4118.

Ash, Elliott and Stephen Hansen, "Text algorithms in economics," *Annual Review of Economics*, 2023, *15*, 659–688.

_ **and W. Bentley MacLeod**, "Mandatory Retirement for Judges Improved the Performance of US State Supreme Courts," *American Economic Journal: Economic Policy*, February 2024, *16* (1), 518–48.

_ **, Daniel L Chen, and Arianna Ornaghi**, "Gender attitudes in the judiciary: Evidence from US circuit courts," *American Economic Journal: Applied Economics*, 2024, *16* (1), 314–350.

_ **, Dominik Stammbach, and Kevin Tobia**, "What is (and was) a person? Evidence on historical mind perceptions from natural language," *Cognition*, 2023, *239*, 105501.

_ , **Stephen Hansen, Yabra Muvdi, and Claudia Marangon**, "Large language models in economics," in "The Palgrave Handbook of Economics and Language," Springer, 2025, pp. 191–210.

Asher, Sam, Paul Novosad, and Charlie Rafkin, "Intergenerational Mobility in India: New Measures and Estimates across Time and Social Groups," *American Economic Journal: Applied Economics*, April 2024, *16* (2), 66–98.

Asiedu, Edward, Dean Karlan, Monica Lambon-Quayefio, and Christopher Udry, "A call for structured ethics appendices in social science papers," *Proceedings of the National Academy of Sciences*, 2021, *118* (29), e2024570118.

Asirvatham, Hemanth, Elliott Mokski, and Andrei Shleifer, "GPT as a Measurement Tool," Technical Report, National Bureau of Economic Research 2026.

Banerjee, Abhijit, Esther Duflo, Amy Finkelstein, Lawrence F Katz, Benjamin A Olken, and Anja Sautmann, "In praise of moderation: Suggestions for the scope and use of pre-analysis plans for rcts in economics," Technical Report, National Bureau of Economic Research 2020.

Bruhn, Miriam and David McKenzie, "In pursuit of balance: Randomization in practice in development field experiments," *American economic journal: applied economics*, 2009, *1* (4), 200–232.

Brynjolfsson, Erik, Daniel Rock, and Chad Syverson, "Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics," Working Paper 24001, National Bureau of Economic Research November 2017.

_ , **Danielle Li, and Lindsey Raymond**, "Generative AI at Work*," *The Quarterly Journal of Economics*, 02 2025, *140* (2), 889–942.

Buolamwini, Joy and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in Sorelle A. Friedler and Christo Wilson, eds., *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, Vol. 81 of *Proceedings of Machine Learning Research* PMLR 2018, pp. 77–91.

Callaway, Brantly, Andrew Goodman-Bacon, and Pedro H. C. Sant'Anna, "Event Studies with a Continuous Treatment," *AEA Papers and Proceedings*, May 2024, *114*, 601–05.

Clarke, Damian, Daniel Pailañir, Susan Athey, and Guido W. Imbens, "Synthetic difference-in-differences estimation," 2023.

Council of Europe, *European Judicial Systems - CEPEJ Evaluation Report 2024 - Part 1: General Analyses*, Council of Europe, 2024.

Cruces, Guillermo, Diego Fernández Meijide, Sebastian Galiani, Ramiro H Gálvez, and María Lombardi, "Does Generative AI Narrow Education-Based Productivity Gaps? Evidence from a Randomized Experiment," Technical Report, National Bureau of Economic Research 2026.

de Janvry, Alain, Guojun He, Elisabeth Sadoulet, Shaoda Wang, and Qiong Zhang, "Subjective Performance Evaluation, Influence Activities, and Bureaucratic Work Behavior: Evidence from China," *American Economic Review*, March 2023, *113* (3), 766–99.

Dell'Acqua, Fabrizio, Edward McFowland III, Ethan Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani, "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality," *Organization Science*, 2026. Forthcoming.

Finan, Frederico, Benjamin A Olken, and Rohini Pande, "The personnel economics of the developing state," *Handbook of economic field experiments*, 2017, *2*, 467–514.

Gunning, Robert, "The fog index after twenty years," *Journal of Business Communication*, 1969, *6* (2), 3–13.

Institute for Fiscal Studies, "Crown Court backlog exacerbated by post-pandemic productivity slump," June 2025. Accessed 2025-09-24.

International Development Law Organization, "IDLO Discusses Legal Reform With Pakistan Supreme Court Judges," 2018.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human Decisions and Machine Predictions," *The Quarterly Journal of Economics*, 2018, *133* (1), 237–293.

Landes, William M. and Richard A. Posner, "Rational Judicial Behavior: A Statistical Study," *Journal of Legal Analysis*, 07 2009, *1* (2), 775–831.

Law and Justice Commission of Pakistan, "National Judicial Policy Implementation Committee: Judicial Statistics of Pakistan 2023," 2023. Accessed June 2025.

_ , "Judicial Statistics of Pakistan 2024," 2024.

Lee, David S., "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *The Review of Economic Studies*, July 2009, *76* (3), 1071–1102.

Liu, Ernest, Yi Lu, Wenwei Peng, and Shaoda Wang, "Judicial independence, local protectionism, and economic integration: Evidence from China," Technical Report, National Bureau of Economic Research 2022.

Ludwig, Jens and Sendhil Mullainathan, "Fragile Algorithms and Fallible Decision-Makers: Lessons from the Justice System," *Journal of Economic Perspectives*, November 2021, *35* (4), 71–96.

MacKinnon, James G and Matthew D Webb, "Wild bootstrap inference for wildly different cluster sizes," *Journal of Applied Econometrics*, 2017, *32* (2), 233–254.

McKenzie, David, "Tools of the Trade: a joint test of orthogonality when testing for balance," World Bank Blogs (Development Impact) February 2015. Accessed 2026-03-07.

_ , "An Updated Overview of Multiple Hypothesis Testing Commands in Stata," July 2021. Accessed March 22, 2026.

Mehmood, Sultan, "The Impact of Presidential Appointment of Judges: Montesquieu or the Federalists?," *American Economic Journal: Applied Economics*, 2022, *14* (4), 411–45.

Muralidharan, Karthik, Paul Niehaus, Sandip Sukhtankar, and Jeffrey Weaver, "Improving last-mile service delivery using phone-based monitoring," *American Economic Journal: Applied Economics*, 2021, *13* (2), 52–82.

Noy, Shakked and Whitney Zhang, "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence," *Science*, 2023, *381* (6654), 187–192.

Pah, Adam, David Schwartz, Sarath Sanga, Charlotte Alexander, Kristian Hammond, Luis Amaral, and SCALES OKN Consortium, "The Promise of AI in an Open Justice System," *AI Magazine*, 2022, *43* (1), 69–74.

Press Information Department, Government of Pakistan, "PR No. 128: Digitalization of the Judiciary in Pakistan," Press release November 2023. Accessed 2026-02-23.

Semenova, Vira, "Generalized lee bounds," *Journal of Econometrics*, 2025, *251*, 106055.

Siddique, Osama, *Pakistan's Experience with Formal Law: Alienation, Orthodoxy and the Search for Legal Empowerment*, Cambridge: Cambridge University Press, 2013.

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

The New York Times, "A Lifelong Nightmare: Seeking Justice in India's Overwhelmed Courts," January 2024. Accessed 2024-11-22.

U.S. Government Accountability Office, "U.S. Immigration Courts See Significant and Growing Backlog," 2024. Accessed 2024-11-29.

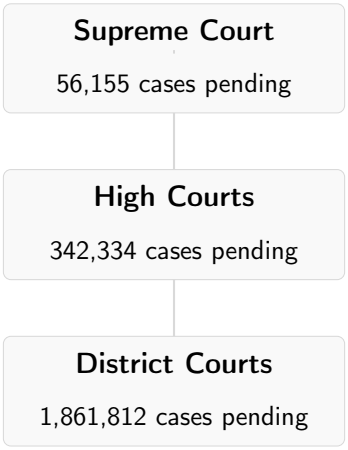
Vicente, Lucía and Helena Matute, "Humans inherit artificial intelligence biases," *Scientific Reports*, October 2023, *13* (1), 15737.

Yang, Crystal S. and Will Dobbie, "Equal Protection Under Algorithms: A New Statistical and Legal Framework," *Michigan Law Review*, 2020, *119* (2), 291–358.

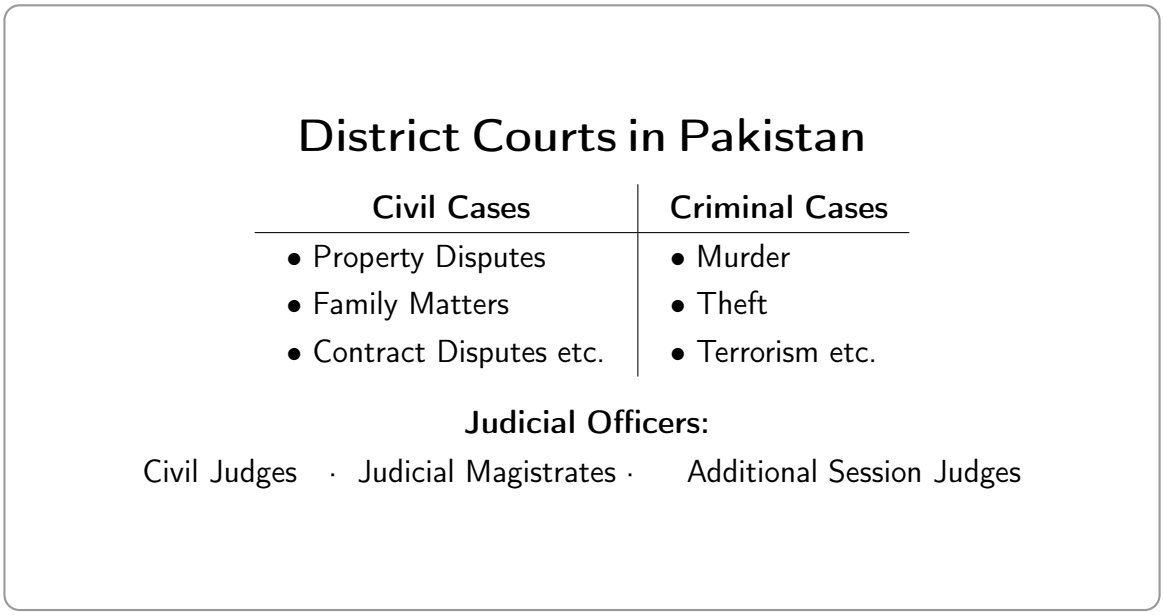
Appendix A. Additional Figures and Tables

Figure A1: Institutional background

Panel A: Courts Tiers in Pakistan



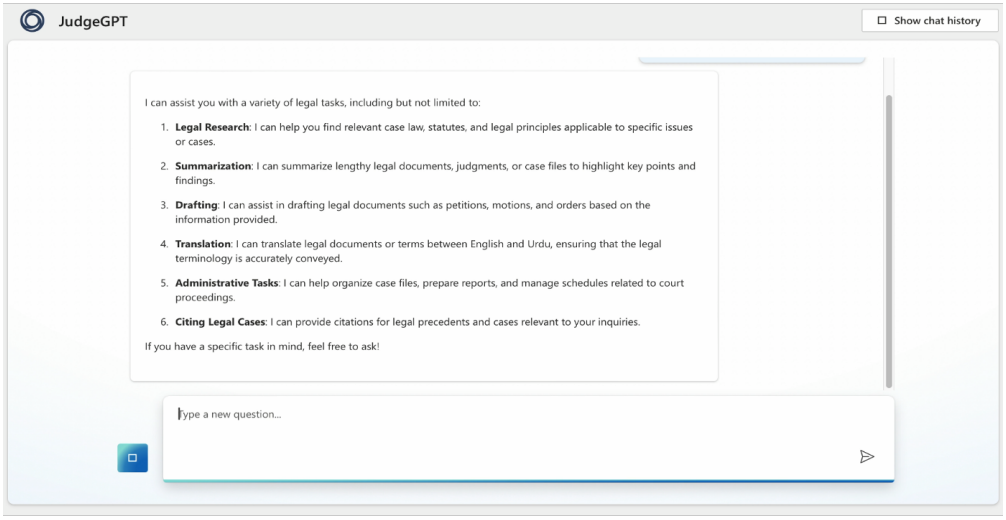
Panel B: District Courts in Pakistan



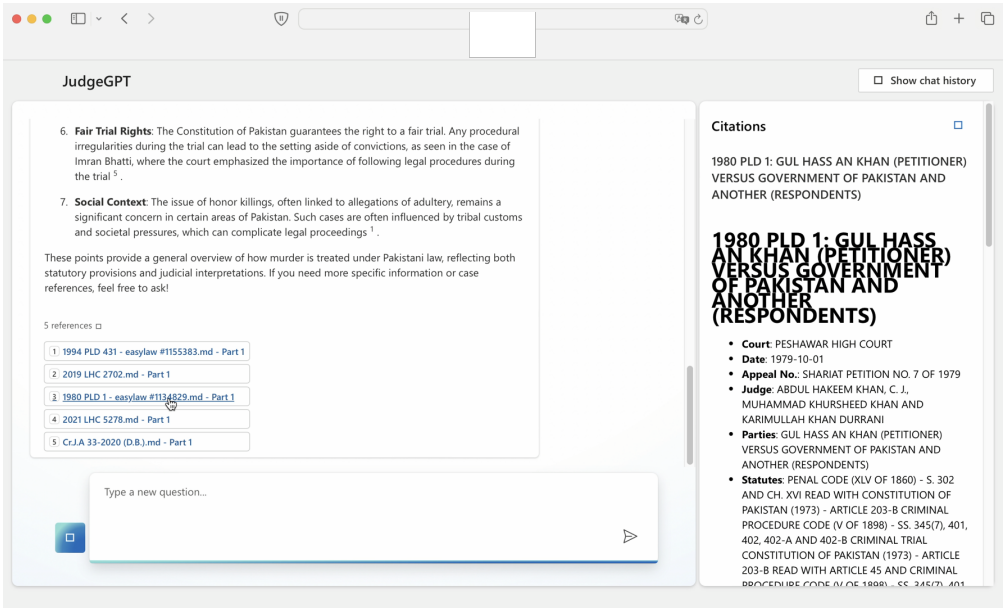
Note: Panel A illustrates Pakistan's three-tier court hierarchy: District Courts (courts of first instance), High Courts (provincial appellate courts), and the Supreme Court (final court of appeal). Case counts report pending matters at each level in 2024 (Law and Justice Commission of Pakistan, 2024). Cases originate in District Courts and may proceed upward on appeal. Panel B summarizes the jurisdiction and composition of District Courts. Trial courts adjudicate both civil (e.g., property, family, contract) and criminal (e.g., murder, theft, terrorism) matters. The district judiciary consists of Civil Judges, Judicial Magistrates, and Sessions and Additional Sessions Judges, who conduct the core fact-finding and adjudicatory functions of the system.

Figure A2: JudgeGPT AI Assistant

Panel A: Generic JudgeGPT Interface



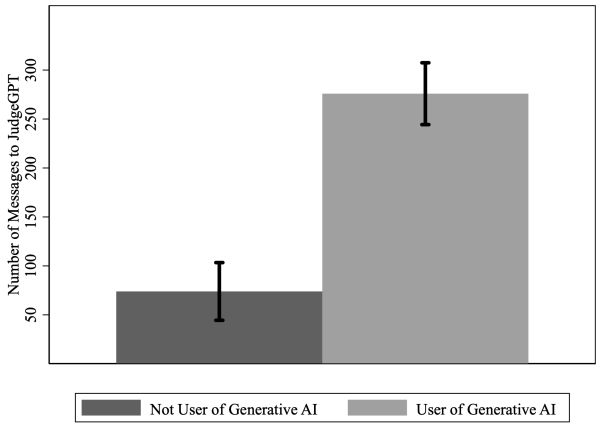
Panel B: JudgeGPT RAG Citation



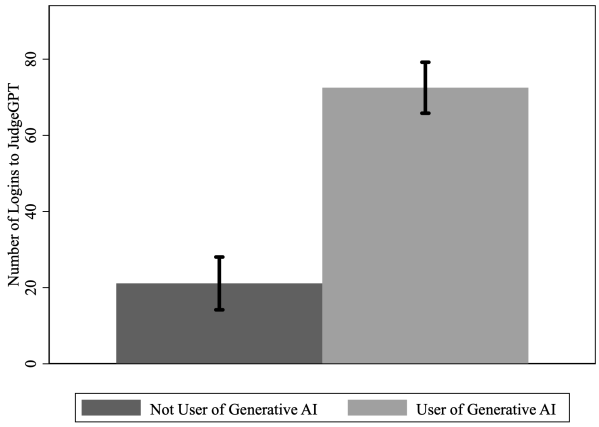
Note: Panel A depicts the user interface of *JudgeGPT*, an AI-powered legal assistant designed to support legal professionals across a range of tasks. Panel B shows *JudgeGPT*'s retrieval-augmented generation (RAG) interface. The model generates a draft answer (left panel) and attaches citations to specific claims. The references box at the bottom lists the retrieved documents (Pakistani rulings) from the linked corpus. Clicking a reference opens the source in a citation panel (right), which displays the case title, court, date, and the extracted passage supporting the cited claim. Judges can navigate across references to verify legal propositions and copy verified citations into draft opinions.

Figure A3: First-Stage: Stated and Actual Generative AI Usage

(a) Panel A. Judge GPT Message Frequency



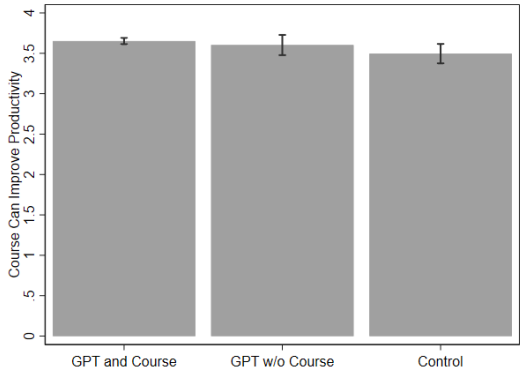
(b) Panel B. JudgeGPT Actual Logins



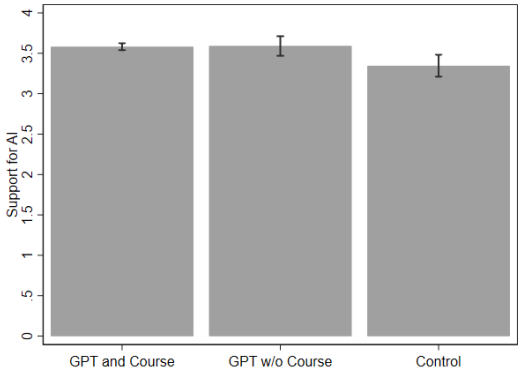
Note: The figure depicts means and standard deviations for usage of JudgeGPT, specifically, number logins and messages to JudgeGPT among people who described themselves as frequent generative AI users after the treatment and not frequent generative AI users. Specifically, answered that they are using 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 A4: Distribution of Main Outcome across Treatment Arms

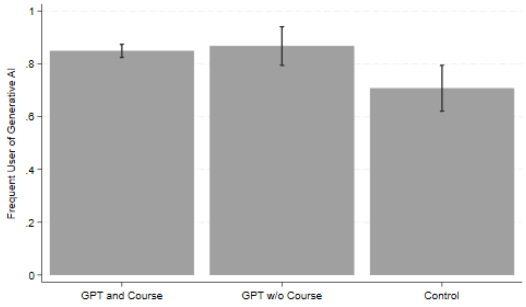
Panel A: GPT Can Improve Productivity



Panel B: Support the Usage of GPT



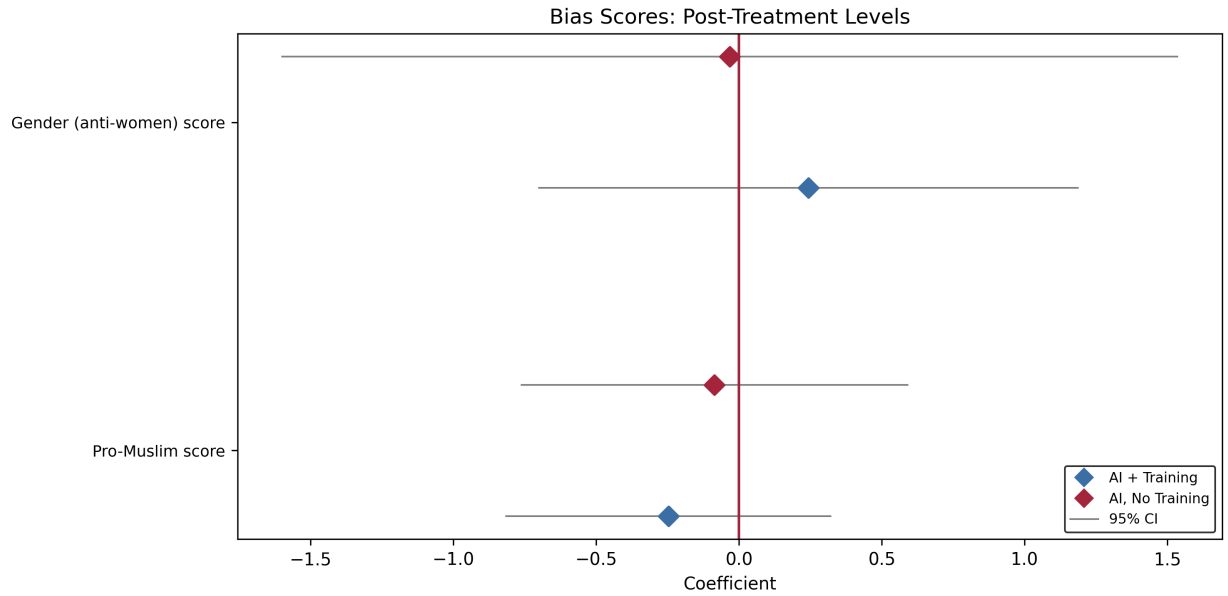
Panel C: Stated Frequent User of Generative AI



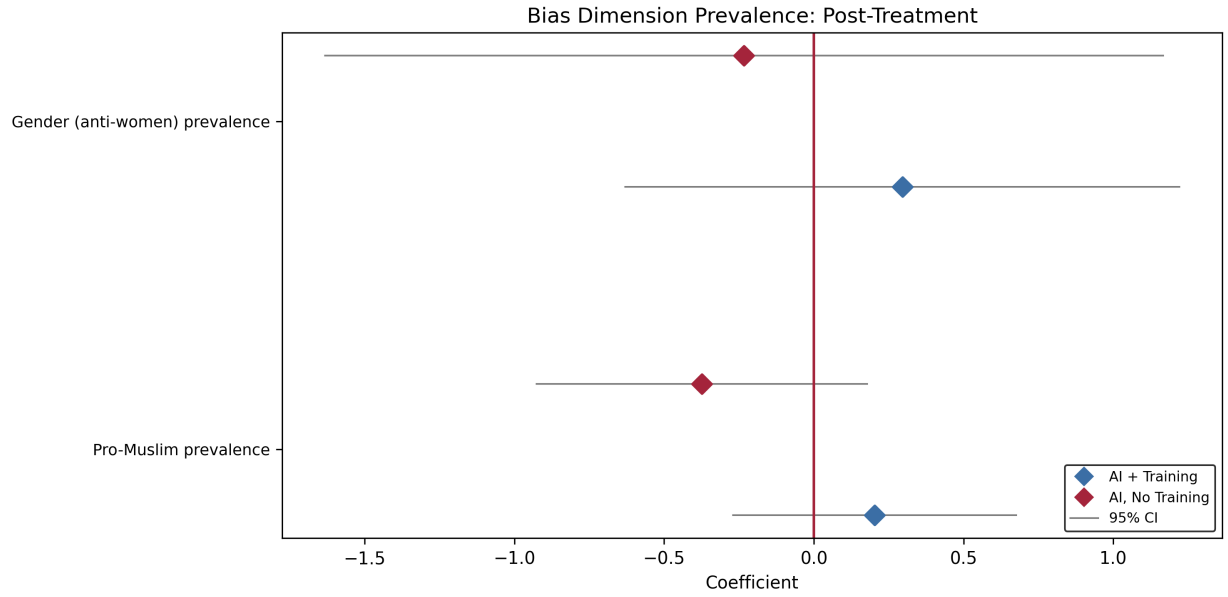
Note: Panel A: Displays the mean response to whether the provided course can improve productivity. Panel B: Measures the level of institutional support for AI integration. Results are based on responses to the question, "Do you support the use of Generative AI assistants such as GPT in your job as a judge?", rated on a 1-to-5 scale. Panel C: Shows the proportion of Frequent Users of Generative AI. We define frequent users ar Judges, who 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?'. Treatment significantly increases the likelihood of becoming a frequent user compared to the control group.

Figure A5: Treatment Effects on Biases

Panel A: Effect on Bias Scores

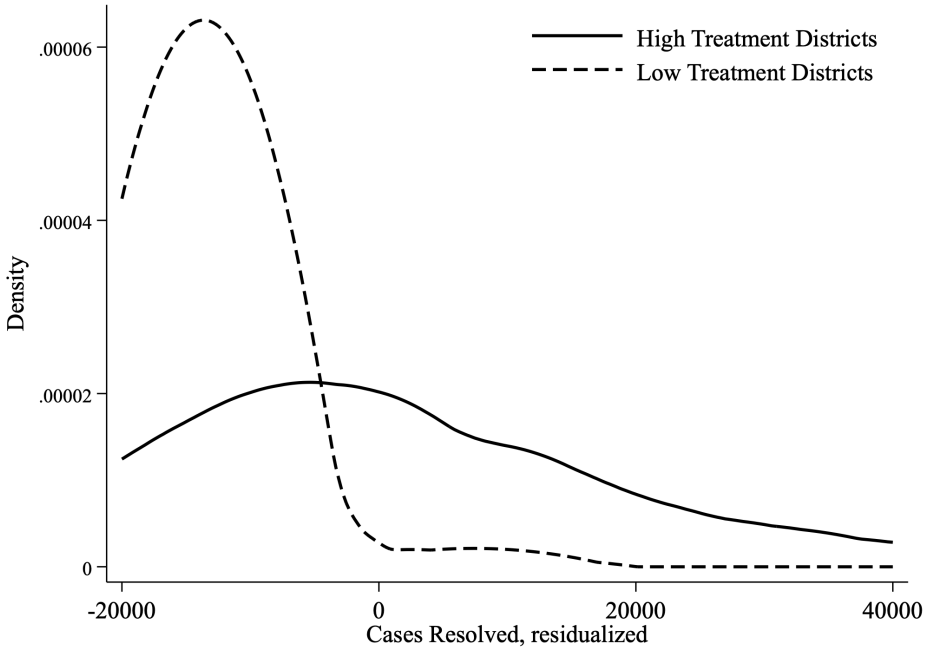


Panel B: Effect on Bias Dimension



Note: Panel A presents OLS coefficients from judge-level regressions of mean post-treatment bias score (0–3 scale) on treatment arm indicators, with strata fixed effects, survey controls, and HC1 robust standard errors. The omitted category is No AI. Whiskers show 95% confidence intervals. Marker shapes indicate significance: diamond ($p \geq 0.10$), square ($p < 0.10$), circle ($p < 0.05$), hollow circle ($p < 0.01$). $N = 41$ judges in batches 3, 4a, and 4b. Panel B presents OLS coefficients from judge-level regressions of the fraction of post-treatment opinions where the bias dimension is relevant, on treatment arm indicators, with strata fixed effects, survey controls, and HC1 robust standard errors. The omitted category is No AI. Whiskers show 95% confidence intervals. $N = 41$ judges in batches 3, 4a, and 4b.

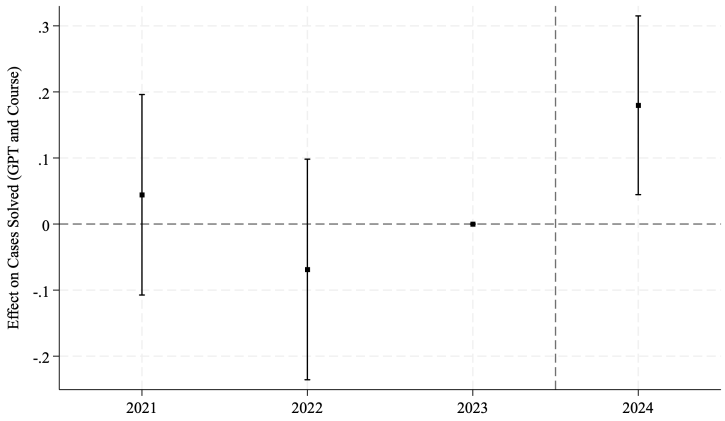
Figure A6: Distribution of Cases Resolved



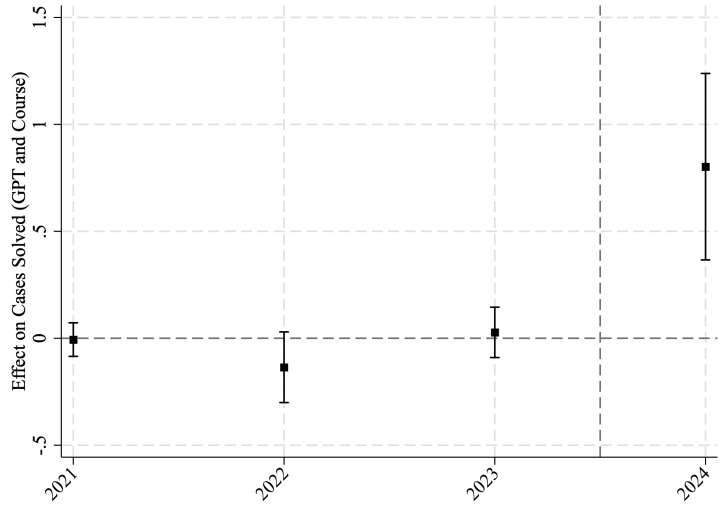
Note: A kernel density plot residualized on district and year fixed effects. That is, it reports the distribution of cases resolved after netting out each district's average across years and the national average for each year. A value of zero means a district resolved exactly as many cases as its own historical average in that year. A positive value means it resolved more than its average, and a negative value means it resolved fewer. The plots are shown separately for high- and low-treatment districts. High-treatment districts are those above the median share of Batch 1 judges assigned to *JudgeGPT* with targeted training, while low-treatment districts are those below the median.

Figure A7: Cases Resolved by Share of Treated Judges in District

Panel A: Standard Difference-in-Differences

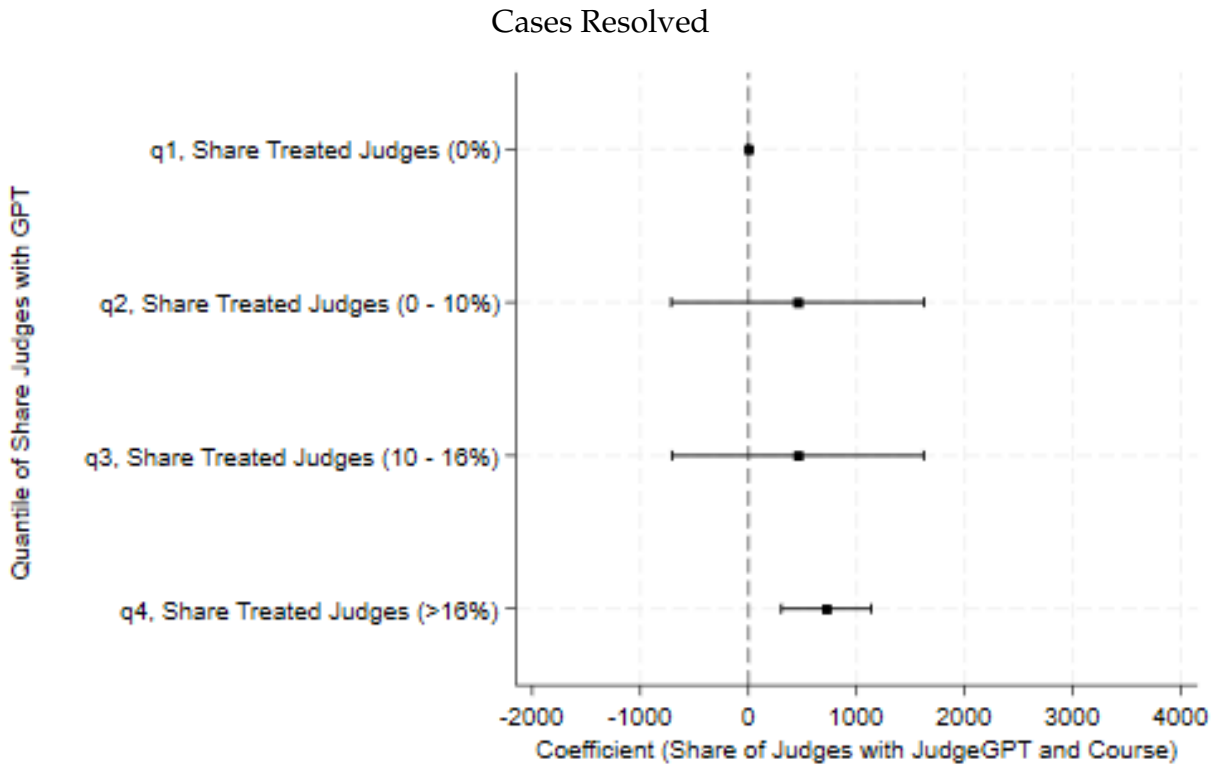


Panel B: Synthetic Difference-in-Differences



Note: Panel A presents 95% confidence intervals from regression of normalized number of cases resolved on dummy judges treated above median interacted with year FE, the share of criminal cases in institution, the share of criminal cases disposed, the number of vacant posts, the balance of cases, the balance of civil cases, the balance of criminal cases are added as controls. Panel B presents synthetic difference-in-difference estimation of normalized number of cases resolved on dummy share judges treated above median interacted with year FE, the share of criminal cases in institution, the share of criminal cases disposed, the working strenght, the balance of civil cases, the balance of criminal cases are added as controls. All specifications include year and district FE. The SDID estimates are computed using the cluster bootstrap procedure described in Section 4.4 of Clarke et al. (2023), following Algorithm 4 in Arkhangelsky et al. (2021).

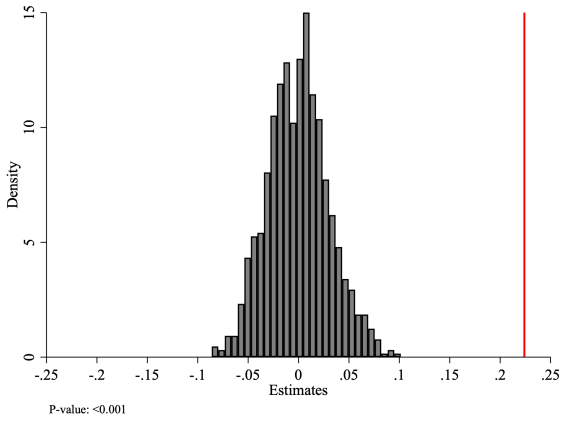
Figure A8: Continuous Treatment Analysis due to Callaway et al. (2024)



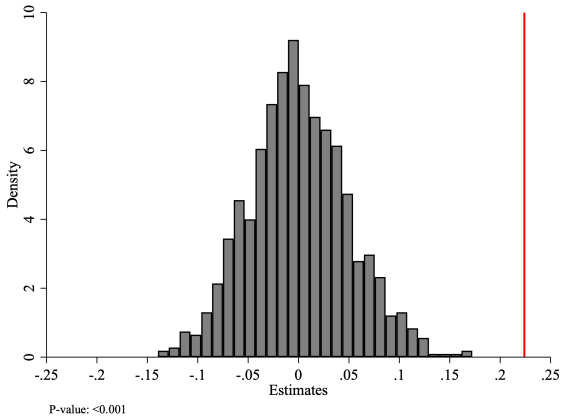
Note: This figure presents estimates of a non-parametric specification of Callaway et al. (2024). The continuous treatment was discretized by quantiles and an indicator for each quantile is included into regression with the number of cases resolved in first differences as a dependent variable. The coefficient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

Figure A9: Permutation Inference Test

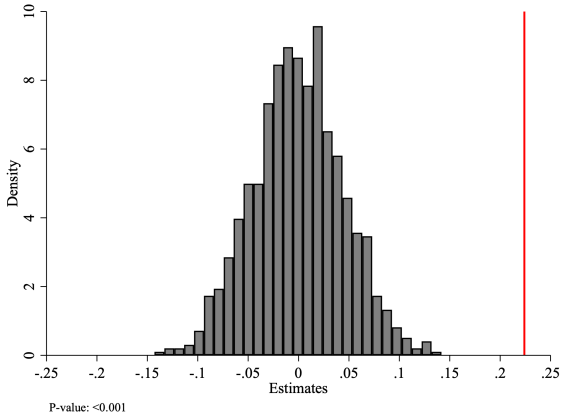
Panel A: Support of Generative AI Usage



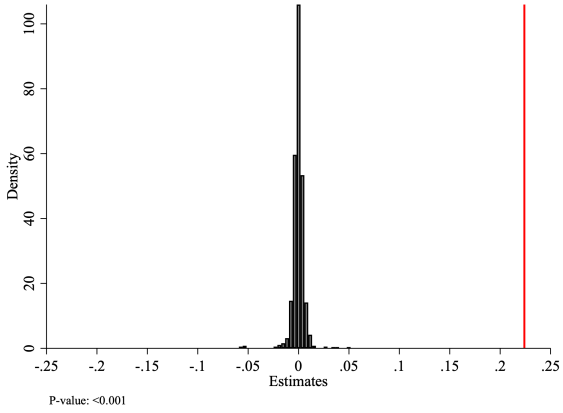
Panel B: Stated Frequent User of Generative AI



Panel C. Support of AI for Judges Course

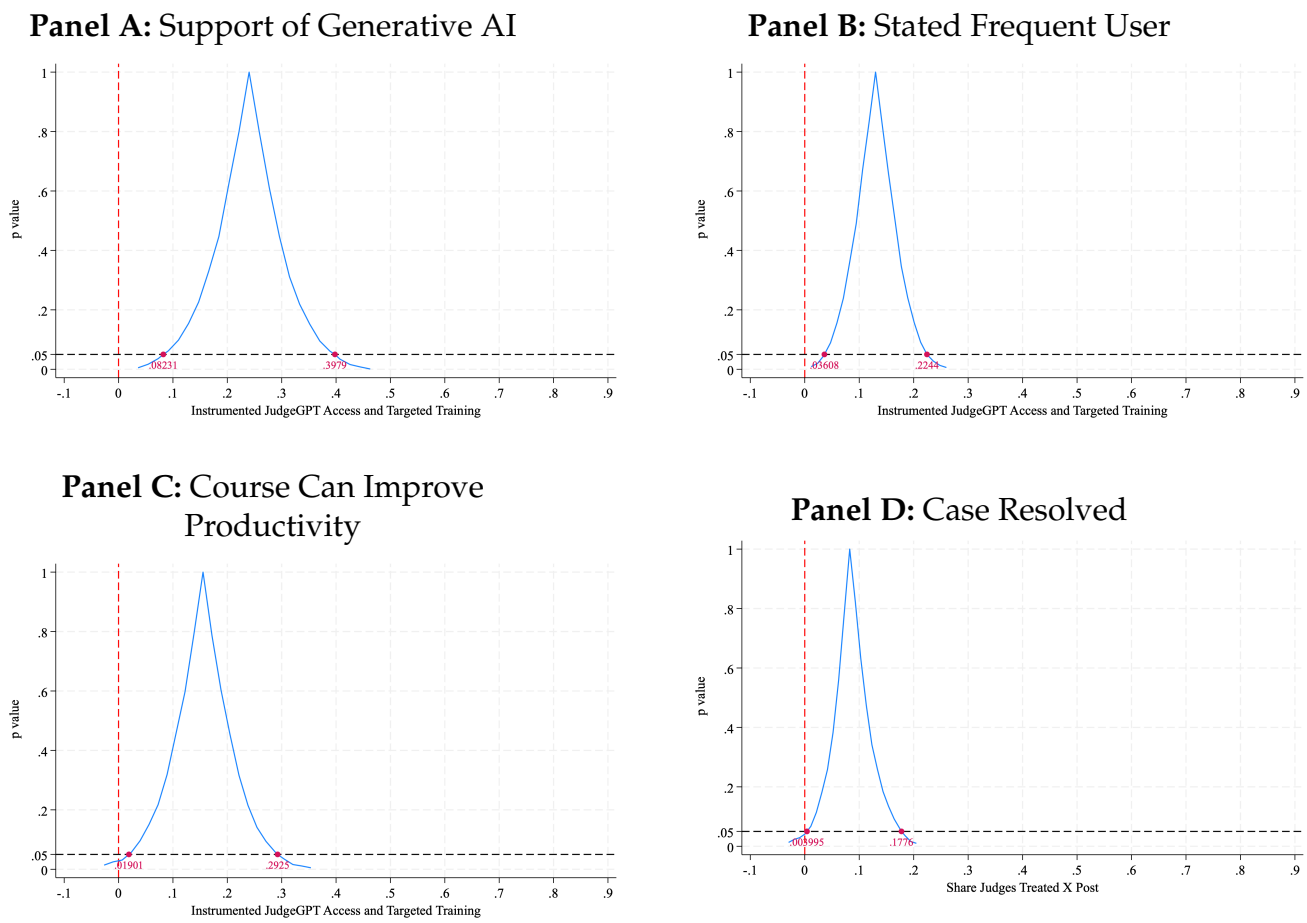


Panel D. Cases Resolved



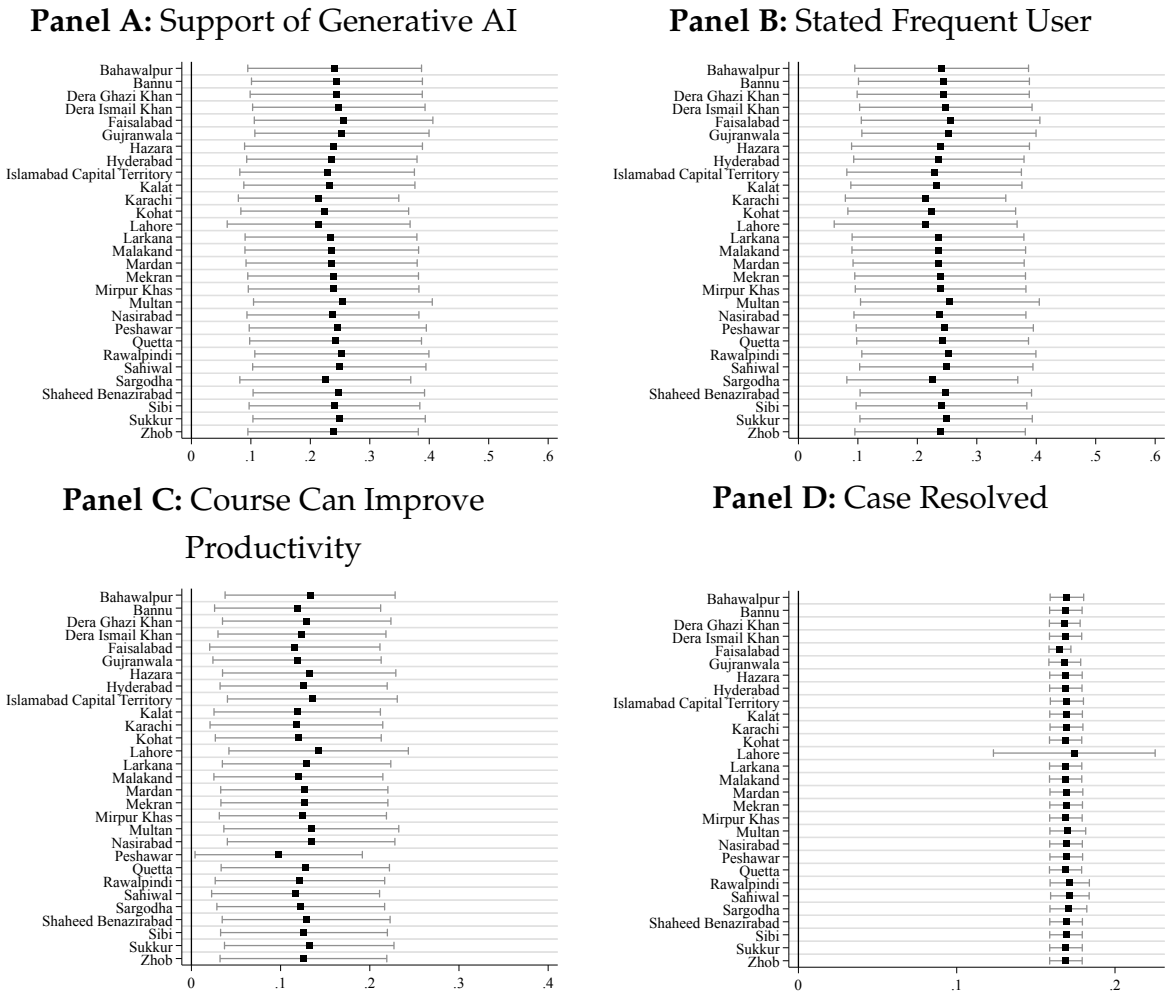
Note: The figure depicts permutation-based inference results. Graphs present the estimated density of permutations of treatment for our main outcomes. For each outcome 1000 permutation of treatment were analyzed. Read line marks the actual OLS estimate of the coefficient of interest.

Figure A10: Wild-Bootstrap Confidence Intervals for Treatment Effects



Note: The figure shows wild-bootstrap confidence interval curves for four regression-related outcomes using the clustered wild-bootstrap procedure of MacKinnon and Webb (2017). For each outcome, the curve plots the bootstrap distribution of the estimated coefficient on JudgeGPT Access and Targeted Training (Panels A, B, C) or Share Treated × Post and marks the 5% significance cutoff. Panel A reports results for support of using generative AI; Panel B for indicator if stated frequent user of AI tools; Panel C for Course Expectations; Panel D for normalized cases resolved counts. All specifications include the full set of fixed effects and controls used in the baseline regressions. Standard errors are clustered at the district level.

Figure A11: Robustness to Excluding One Region at a Time



Note: This figure presents the results of estimation of the main specification with one region excluded at a time. "Region" denotes Pakistan's official administrative division, the tier between province and district. Data reflect the structure as of 2024, incorporating the December 2024 Punjab reorganization into 10 divisions, the 2018 FATA merger into Khyber Pakhtunkhwa, and approximately 169 districts in total. Azad Jammu and Kashmir and Gilgit-Baltistan are Pakistan-administered territories with disputed international status. Sources: Wikipedia, Districts of Pakistan; UN OCHA Common Operational Dataset for Administrative Boundaries, Pakistan (reviewed September 2024); Pakistan Bureau of Statistics, Census 2023. 95% confidence intervals along with the point estimates are reported for the district on y axis excluded.

TABLE A1: Treatment Description and Slides of the Trainings

Treatment Group	**Treatment Description**	**Treatment Details**	**Link to the Slides**
Judge GPT and Targeted 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 The course on Law and AI consisting of 6 lectures 5 home assignments requiring the use of AI and quizzes in class Final project on AI application in legal process Technical support throughout the course Certificates and invitation to Law and Technology Workshop in Zurich for successful participants	Link
Judge GPT with Placebo Course	This treatment group received JudgeGPT and 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.	Fine-tuned on the corpus of judgments AI tool The course on Law and Technology consisting of 6 lectures 5 home assignments and quizzes in class Final project on legal research 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 5 home assignments and quizzes in class Final project on legal research Certificates and invitation to Law and Technology Workshop in Zurich for successful participants	Link

Note: Training conditions and course materials. The table describes the content and delivery of the training courses assigned to each experimental arm. The first row corresponds to Treatment A, which combines access to *JudgeGPT* with targeted training on law-and-AI use in judicial tasks, including modules on verification and hallucination risk management. The second row corresponds to Treatment B, which provides access to *JudgeGPT* but pairs it with a placebo course that covers general digital and AI topics without tool-specific instruction. The third row corresponds to the control group, which receives the placebo course only and no access to *JudgeGPT*. The final column reports links to

TABLE A2: Description of variables

Variable	Description
Pretreatment Characteristics	
Age	Judge age.
Female	Indicator, equals 1 if female.
Years of Experience	Years of experience as judge
Support of AI (1-4)	Answer on "Do you support the use of Generative AI assistants such as GPT in your job as a judge? - choose" on from 1 to 4 scale
Number of Cases Pending	Answer on "How many cases do you currently have pending on your desk?"
Number of Cases Decided per Month	Answer on "How many cases did you decide in the last month?"
Number of Cases Comfortable Managing at the Same Time	Answer on "How many cases are you comfortable managing at the same time?"
Hours per Week Spent on Legal Research	Answer on "How many hours a week do you spend on legal research and writing judgments?"
Hours per Week Spent on Administrative Work	Answer on "How many hours a week do you spend on administrative tasks?"
Confidence in Public Speaking (1-10)	Answer on "On a scale from 1 to 10, how confident do you feel in public speaking in court? - choose" on from 1 to 10 scale
Workload (1-10)	Answer on "On a scale from 1 to 10, how would you rate your current workload? - choose" on from 1 to 10 scale
Worklife Balance (1-10)	Answer on "On a scale from 1 to 10, how would you rate your work/life balance, in the sense of balancing family and entertainment, versus time and stress at work? - choose" on from 1 to 10 scale
Self-Assessed Public Speaking (1-10)	Answer on "On a scale from 1 to 10, how confident do you feel in public speaking in court? - choose" on from 1 to 10 scale
Hours Spent on Administrative Needs (1-10)	Answer on "On a scale from 1 to 10, how confident do you feel in dealing with court administrative needs? - choose" on from 1 to 10 scale
Expectations from the Course (1-4)	Answer on "Do you expect that an AI for Judges Course can improve your productivity? - choose" on from 1 to 4 scale

Note: Continued on the next page.

Description of variables

Variable	Description
Familiar with AI	Indicator equal one if AI chosen in question "Which of these technologies are you familiar with and use regularly?"
Familiar with WhatsApp	Indicator equal one if WhatsApp chosen in question "Which of these technologies are you familiar with and use regularly?"
Judge Level Outcomes	
JudgeGPT and Targeted Training	Indicator equals 1 if the judge received access to JudgeGPT and Targeted Training. Slides Here
JudgeGPT without Training	Indicator equals 1 if judge received access to JudgeGPT and Placebo Training. Slides Here
Frequent AI User	Indicator equals 1 if answer on "Which of the following best describes how often you use GPT or other AI systems in your judicial work?" is one from "Almost every day", "Once or twice a week", "A few times a month"
Support of Generative AI Usage	Answer on "Do you support the use of Generative AI assistants such as GPT in your job as a judge? - choose" on a scale from 1 to 4
Support of AI for Judges Course	Answer on "Do you expect that an AI for Judges Course can improve your productivity? - choose" on a scale from 1 to 4
Familiar with AI	Indicator equal one if AI chosen in question "Which of these technologies are you familiar with and use regularly?"
Familiar with WhatsApp	Indicator equal one if WhatsApp chosen in question "Which of these technologies are you familiar with and use regularly?

Note: Continued on the next page.

Description of variables

Variable	Description
Case Analysis Outcomes	
Judgments Readability, post	Gunning fog readability metrics of post treatment judgments submitted by judges
# of Legal Arguments, post	Number of references to legal statutes in post treated judgments
District Level Panel (varies by district-year)	
# of Cases Resolved	Number of cases resolved
# of Civil Cases Resolved	Number of civil cases resolved
# of Criminal Cases Resolved	Number of criminal cases resolved
# of Judges in Districts	Number of judges
# of Vacant Judge Positions	Number of vacant judge positions
# of Cases on Balance	Number of cases on balance
# of Cases in Civil Appeal	Number of civil cases appealed
# of Cases in Criminal Appeal	Number of criminal cases appealed
Share of Criminal Cases Disposed	Number of criminal cases disposed
Share of Criminal Cases Instituted	Number of criminal cases instituted
Share Treated Judges with Targeted Training	Count of judges treated in batch 1 divided on total working strength in the district

Note: The table presents descriptions for the key variables used in the analysis.

TABLE A3: 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
Female	0.205	0	0.404	0	1	1559
Years of Experience	11.47	9.200	6.935	0	35	1559
Support of AI (1-4)	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	16.22	14	12.22	0	100	1559
Hours per Week Spent on Administrative Work	11.49	8	11.13	0	100	1559
Confidence in Public Speaking (1-10)	6.506	7	2.115	1	10	1559
Workload (1-10)	6.320	6	2.183	1	10	1559
Worklife Balance (1-10)	5.746	5	2.111	1	10	1559
Self-Assessed Public Speaking(1-10)	7.528	8	2.047	1	10	1559
Hours Spent on Administrative Needs (1-10)	7.502	8	1.891	1	10	1559
Expectations from the Course (1-4)	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	0.00764	-0.0488	0.668	-0.178	9.484	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	28499.3	10578.5	53390.1	0	682933	484
# of Civil Cases Resolved	11799.6	3162.5	20844.5	0	184715	484
# of Criminal Cases Resolved	16316.7	6216	34071.3	0	503817	484
# of Judges in Districts	20.04	12	29.74	0	272	484
# of Vacant Judge Positions	6.975	3	12.93	0	113	484
# of Cases on Balance	14635.0	3746	28812.8	0	351174	484
# of Cases in Civil Appeal	449.1	195.5	673.0	0	5138	484
# of Cases in Criminal Appeal	9.804	4	22.19	0	348	484
Share of Criminal Cases Disposed	0.549	0.623	0.269	0	0.948	484
Share of Criminal Cases Instituted	0.565	0.627	0.254	0	0.951	484
Share Treated Judges with Targeted Training	1.680	0.200	4.888	0	48	472

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 A4: 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 2021 - Jan 2024	0	0	0	0	0	0
March 2024	487	94	0	0	0	0
Sep 2024	979	108	0	0	0	0
Feb 2025	1197	110	180	62	182	66
End of Treatment	Apr 2025					

Note: This Table summarizes the staggered rollout of the experimental groups across judge cohorts from March 2024 to April 2025. In total, 1,559 judges from 110 districts were registered. Batch 1 (487 judges from 94 districts) began in March 2024 and received Treatment A: *JudgeGPT* access with tool-specific training. Batch 2 (492 judges from 94 districts), launched in September 2024, received the same treatment. Batch 3 (218 judges from 70 districts), launched in February 2025, also received Treatment A. In February 2025, Batches 4a and 4b were implemented. Batch 4a (180 judges from 62 districts) received Treatment B: *JudgeGPT* access with generic training. Batch 4b (182 judges from 66 districts) served as the control group and received the same generic training but without *JudgeGPT* access. Training periods were March–April 2024 for Batch 1, September–October 2024 for Batch 2, and February–April 2025 for Batches 3, 4a, and 4b. Randomization was conducted at the judge level, stratified by province and age group, prior to training. Additional details are provided in Figure 1.

TABLE A5: Balance Test Over Judge Characteristics and Pre-Treatment Outcomes

	(1)		(2)		(3)		(1)-(2)		(1)-(3)		(2)-(3)	
	JudgeGPT And Training		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	1195	42.713	181	42.856	183	42.432	1376	0.411	1378	0.939	364	0.448
		(0.205)		(0.511)		(0.505)						
Gender	1195	1.788	181	1.818	183	1.820	1376	0.438	1378	0.363	364	0.752
		(0.012)		(0.029)		(0.028)						
Years of Experience	1195	24.899	181	66.718	183	33.666	1376	0.099*	1378	0.556	364	0.261
		(4.727)		(24.414)		(15.441)						
Support of AI	1195	3.430	181	3.409	183	3.432	1376	0.689	1378	0.983	364	0.711
		(0.024)		(0.063)		(0.061)						
Number of Decided Cases	1195	117.556	181	109.276	183	103.716	1376	0.269	1378	0.102	364	0.600
		(4.362)		(7.692)		(8.275)						
Number of Cases Concurrently Managed	1195	140.787	181	175.680	183	121.530	1376	0.414	1378	0.232	364	0.142
		(9.603)		(38.552)		(17.581)						
Hours spent on Legal Research and Writing Judgments	1195	17.094	181	15.635	183	17.328	1376	0.295	1378	0.663	364	0.218
		(0.678)		(0.863)		(1.025)						
Hours spent on Administrative Work	1195	11.763	181	12.823	183	11.230	1376	0.325	1378	0.553	364	0.242
		(0.412)		(1.078)		(0.794)						
Work-Life Balance	1195	5.762	181	5.663	183	5.721	1376	0.640	1378	0.782	364	0.987
		(0.061)		(0.157)		(0.150)						
Confidence in Public Speaking	1195	7.562	181	7.320	183	7.508	1376	0.196	1378	0.795	364	0.416
		(0.058)		(0.162)		(0.154)						
Confidence in Administrative Work	1195	7.517	181	7.409	183	7.497	1376	0.576	1378	0.995	364	0.705
		(0.055)		(0.142)		(0.142)						
Expectations from AI for Judges	1195	3.722	181	3.740	183	3.716	1376	0.666	1378	0.976	364	0.787
		(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 A6: Balance Test Over Judge Characteristics and Pre-Treatment Outcomes: Responded Subsample

	(1)		(2)		(3)		(1)-(2)		(1)-(3)		(2)-(3)	
	JudgeGPT And Training		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	271	42.871	6	41.833	16	43.125	277	0.834	287	0.949	22	0.027**
		(0.414)		(4.126)		(1.486)						
Gender	271	1.734	6	1.833	16	1.812	277	0.708	287	0.533	22	0.483
		(0.027)		(0.167)		(0.101)						
Years of Experience	271	18.879	6	11.000	16	12.669	277	0.520	287	0.466	22	0.263
		(7.338)		(3.033)		(1.467)						
Support of AI	271	3.465	6	3.500	16	3.875	277	0.753	287	0.001***	22	0.365
		(0.051)		(0.342)		(0.085)						
Number of Decided Cases	271	105.517	6	86.667	16	99.562	277	0.939	287	0.978	22	0.615
		(6.208)		(44.864)		(24.182)						
Number of Cases Concurrently Managed	271	121.454	6	40.000	16	113.750	277	0.013**	287	0.700	22	0.752
		(13.964)		(14.201)		(50.897)						
Hours spent on Legal Research and Writing Judgments	271	16.819	6	18.000	16	22.125	277	0.869	287	0.368	22	0.346
		(0.788)		(6.792)		(4.011)						
Hours spent on Administrative Work	271	12.295	6	29.167	16	11.938	277	0.433	287	0.713	22	0.458
		(0.774)		(18.382)		(3.042)						
Work-Life Balance	271	5.841	6	6.167	16	5.500	277	0.693	287	0.522	22	0.023**
		(0.127)		(0.872)		(0.456)						
Confidence in Public Speaking	271	7.731	6	6.500	16	7.812	277	0.210	287	0.657	22	0.343
		(0.124)		(0.847)		(0.579)						
Confidence in Administrative Work	271	7.583	6	6.333	16	7.562	277	0.128	287	0.890	22	0.428
		(0.116)		(0.843)		(0.532)						
Expectations from AI for Judges	271	3.731	6	3.333	16	3.875	277	0.300	287	0.317	22	0.697
		(0.036)		(0.422)		(0.085)						

Note: The table presents the balance test on subsample which includes only people answered on questions after treatment. The treatment group was assigned a JudgeGPT subscription and training 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 A7: Spillover effect of JudgeGPT Access and Associated Course

Panel A: Spillovers measured by observed usage (API intensity)			
	Stated Frequent User of Generative AI	Support the Usage of GPT	GPT Course Can Improve Productivity
No Access to GPT ×			
# of Unique APIs per Judge in District	0.001	0.001	0.001
	(0.001)	(0.001)	(0.001)
JudgeGPT Access and Targeted Training	0.223***	0.268***	0.206***
	(0.050)	(0.079)	(0.068)
JudgeGPT Access with Placebo Training	0.230***	0.290***	0.159
	(0.074)	(0.096)	(0.098)
Panel B: Spillovers measured by treatment intensity (share treated)			
	Stated Frequent User of Generative AI	Support the Usage of GPT	GPT Course Can Improve Productivity
No Access to GPT ×			
Share of Judges with GPT Access in District	0.128	0.079	0.121
	(0.078)	(0.109)	(0.085)
JudgeGPT Access and Associated Training	0.263***	0.272***	0.225***
	(0.054)	(0.084)	(0.071)
JudgeGPT Access with Placebo Training	0.269***	0.293***	0.179*
	(0.076)	(0.099)	(0.099)
Strata Fixed Effects	Yes	Yes	Yes
Controls	Yes	Yes	Yes
Adj. R^2	.067	.046	.041
Mean Dep. Var.	.823	3.542	3.616
Observations	937	939	939

Note: Standard errors clustered on a district level in parenthesis. The dependent variable is a an indicator of a judge reporting the use of any AI tool, 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? and the support for the AI for Judges course rated on a 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 training as first Control Group (batch4a). Control group did not received GPT access while getting a placebo course. Interaction of a judge being in the control group with the number of unique APIs per judge and share of judges with JudgeGPT access in the district are added to test for spillovers. Controls include pre-treatment characteristics reported in the balance test and strata fixed effects. R^2 and Kleibergen-Paap F-statistic are reported. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE A8: Estimates of ITT Effects on AI-Related Outcomes

Panel A: Baseline Estimates	Stated Frequent User of Generative AI	Support the Usage of GPT	GPT Course Can Improve Productivity
	(1)	(2)	(3)
JudgeGPT Access and Targeted Training	0.113***	0.218***	0.139**
	(0.039)	(0.064)	(0.057)
JudgeGPT Access and Placebo Training	0.129**	0.203**	0.042
	(0.058)	(0.089)	(0.085)
Panel B: JudgeGPT + Targeted Training vs Control	**Stated Frequent User of Generative AI**	**Support the Usage of GPT**	**GPT Course Can Improve Productivity**
	(1)	(2)	(3)
JudgeGPT Access and Targeted Training IPW ITT	0.114***	0.225***	0.142**
	(0.039)	(0.064)	(0.057)
Lee Bounds	[0.049 , 0.283]	[0.077 , 0.454]	[0.019 , 0.363]
Generalized Lee Bounds	[0.078 , 0.542]	[0.134 , 0.904]	[0.078 , 0.705]
Panel C: JudgeGPT + Placebo Training vs Control	**Stated Frequent User of Generative AI**	**Support the Usage of GPT**	**GPT Course Can Improve Productivity**
	(1)	(2)	(3)
JudgeGPT Access and Placebo Training IPW ITT	0.130**	0.205**	0.043
	(0.058)	(0.089)	(0.085)
Lee Bounds	[0.081 , 0.155]	[0.077 , 0.261]	[-0.065 , 0.090]
Generalized Lee Bounds	[-0.019 , 0.212]	[-0.065 , 0.472]	[-0.119 , 0.224]
Mean dep. var.	0.820	3.544	3.619
Observations	1082	1084	1084

Note: Panel A reports outcome-specific response rates by assignment group. Panels B and C report treatment effects for Arm A (JudgeGPT + targeted training) and Arm B (JudgeGPT + placebo training), each relative to control. Panel A uses only observed outcomes and includes the main pre-treatment controls and strata fixed effects. "Lee bounds" are Lee (2009) attrition bounds under monotone selection, constructed by trimming the outcome distribution of the higher-response group to match response rates across groups on residualized by controls and strata fixed effects. "Generalized Lee bounds" follow Semenova (2023) and relax the monotone selection assumption by allowing the direction of selection to vary with covariates, combining a flexible first-stage selection model (logistic regression) with quantile regression of the outcome, and use a data-driven trimming threshold based on the sample size. The bounds are estimated via orthogonal moments to reduce bias from first-stage estimation. Standard errors are in parentheses for point estimates; bounds are reported as intervals. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

TABLE A9: JudgeGPT Access and Bias

	2SLS			
	(1)	(2)	(3)	(4)
	Defendant Won			
Female Defendant × JudgeGPT And Targeted Training	0.054	0.054		
	(0.047)	(0.047)		
Female Defendant × JudgeGPT And Placebo Training	-0.131	-0.131		
	(0.245)	(0.245)		
Female Defendant	-0.063**	-0.063**		
	(0.030)	(0.030)		
Muslim Defendant × JudgeGPT And Targeted Training			-0.014	-0.033
			(0.054)	(0.056)
Muslim Defendant × JudgeGPT And Placebo Training			-0.006	-0.084
			(0.004)	(0.113)
Muslim Defendant			0.061**	0.062**
			(0.029)	(0.030)
JudgeGPT Access and Targeted Training	Yes	Yes	Yes	Yes
JudgeGPT Access and Placebo Training	Yes	Yes	Yes	Yes
Judge Fixed Effects	Yes	Yes	Yes	Yes
Time-by-Strata Fixed Effects	No	Yes	No	Yes
Adj. R^2	.596	.596	.596	.594
Mean Dep. Var.	.436	.436	.436	.436
Observations	1690	1690	1690	1690

Note: Robust standard errors in parentheses are clustered by Judge. The unit of observation is case. The dependent variable is an indicator of a Defendant winning the case. JudgeGPT Access and Targeted Training is a binary indicator of the judge with the training assigned logged into JudgeGPT at least once interacted with post dummy, while JudgeGPT Access and Placebo Training is a binary indicator of a judge logging into GPT while not having the training as first Control Group (batch4a) interacted with post dummy. Control group did not received GPT access while getting a placebo course. Muslim (Female) Defendant is an indicator of a Defendant being muslim (female) based on LSTM classification of the names. Judge Fixed Effects and pre-post fixed effects are included. In columns (1)-(4) treatment A and treatment B are instrumented with intial assignment of treatment A and treatment B interacted with post dummy and with corresponding interations with indicators of plaintiffs being muslim (female). Adjusted R^2 is reported. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE A10: Robustness to Adjustments to Multiple Hypothesis Testing

Panel A: Individual Level Outcomes			
	Support JudgeGPT	Frequent User of Gen AI	AI-for-Judges Course
	(1)	(2)	(3)
JudgeGPT Access and	0.239	0.126	0.152
Targeted Training	(0.073)	(0.048)	(0.065)
p-value	0.001***	0.008***	0.020**
Romano-Wolf p-value	0.020**	0.040**	0.059*
Sharpened q-value	0.004***	0.009***	0.013**
Westfall-Young p-value	0.010***	0.030**	0.030**
Bonferroni-corrected p-value	0.003***	0.016**	0.020**
JudgeGPT Access and	0.254	0.164	0.074
Placebo Training	(0.101)	(0.065)	(0.098)
p-value	0.012**	0.011**	0.451
Romano-Wolf p-value	0.059**	0.059**	0.455*
Sharpened q-value	0.019**	0.019**	0.177
Westfall-Young p-value	0.030**	0.030**	0.450
Bonferroni-corrected p-value	0.033**	0.033**	0.451
Panel B: District Level Outcomes			
	Total Disposal	Civil Disposal	Criminal Disposal
	(1)	(2)	(3)
Share Judges JudgeGPT Access	0.083	0.044	0.213
and Targeted Training	(0.031)	(0.009)	(0.007)
p-value	0.008***	0.000***	0.000***
Romano-Wolf p-value	0.089*	0.089*	0.010*
Sharpened q-value	0.003***	0.001***	0.001***
Westfall-Young p-value	0.170	0.060*	0.000***
Bonferroni-corrected p-value	0.008***	0.000***	0.000***

Note: Panel A reports IV estimates of treatment effects for Arm A (*JudgeGPT* + targeted training) and Arm B (*JudgeGPT* + placebo training), each relative to control. Panel B reports reduced-form estimates of the effect of the share of judges with JudgeGPT access on district-level case disposition outcomes. All individual-level models include pre-treatment controls (strata fixed effects, years of experience, pre-treatment AI support, caseload comfort, work-life balance, and other baseline characteristics) and are estimated using two-stage least squares with robust standard errors. District-level models include year and district fixed effects, court workload controls (working strength, vacant posts, case balances, and share of criminal cases), and cluster-robust standard errors at the district level. P-values are reported with significance stars based on conventional thresholds. To address multiple hypothesis testing across the three outcomes within each panel, we report several family-wise error rate (FWER) and false discovery rate (FDR) adjustments: Romano-Wolf stepdown p-values (**?**), Westfall-Young stepdown p-values (**?**), Bonferroni-corrected p-values, Šidák-corrected p-values, and sharpened q-values following **?**. All multiple-testing procedures are implemented with 100 bootstrap replications. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

TABLE A11: JudgeGPT Access and Main Outcomes by Batches

Panel A: Treatment A uses only Batch 1, Treatment B and Control is Same				
	AI Usage	AI Support	Course Support	Cases Resolved
	(1)	(2)	(3)	(4)
JudgeGPT Access and Targeted Training	0.181***	0.304***	0.196***	
	(0.046)	(0.072)	(0.064)	
JudgeGPT Access and Placebo Training	0.163**	0.282***	0.107	
	(0.064)	(0.102)	(0.097)	
Share JudgeGPT and Targeted Training × Post				0.393*
				(0.201)
Panel B: Treatment A uses only Batch 2, Treatment B and Control is Same				
JudgeGPT Access and Targeted Training	0.058	0.206***	0.142**	
	(0.053)	(0.079)	(0.072)	
JudgeGPT Access and Placebo Training	0.153**	0.231**	0.059	
	(0.064)	(0.098)	(0.094)	
Panel C: Treatment A uses only Batch 3, Treatment B and Control is Same				
JudgeGPT Access and Targeted Training	0.108*	0.150*	0.069	
	(0.057)	(0.091)	(0.080)	
JudgeGPT Access and Placebo Training	0.194***	0.212**	0.042	
	(0.064)	(0.101)	(0.095)	
Strata Fixed Effects	Yes	Yes	Yes	
Controls	Yes	Yes	Yes	Yes
Adj. R^2	.1	.082	.071	.803
Mean Dep. Var.	.787	3.479	3.555	0
Observations	352	353	353	483

Note: The table reports treatment-effect estimates separately by batches. In each panel, the targeted-training arm is restricted to judges treated in the indicated batch, while the placebo-training and control groups are kept unchanged. Panel A uses Batch 1 for the targeted-training arm, Panel B uses Batch 2, and Panel C uses Batch 3. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE A12: Human–AI Agreement in Pairwise Opinion Comparison

Comparison	Directional Agreement	N
GPT vs. Lawyer	70.1%	87
GPT vs. Gemini	65.2%	89
GPT+Gemini (agree) vs. Lawyer	83.9%	56

A licensed Pakistani lawyer rated 90 pre/post opinion pairs on the same -2 to $+2$ quality scale used by the LLMs. Directional agreement is the fraction of pairs where both raters agree on which opinion is better; pairs where either rater scored a tie are excluded. "GPT+Gemini (agree)" restricts to pairs where GPT and Gemini agree on direction before comparing to the lawyer.

APPENDIX B. ETHICS, FURTHER DATA, PRE-ANALYSIS PLAN

Appendix B1. Structured Approach to Ethics

Ethical considerations and transparency Consistent with the call for structured ethics appendices in social science research (Asiedu et al., 2021), we document the ethical design choices underlying the *JudgeGPT* experiment in an ethics appendix. We discuss (i) *policy equipoise*, given uncertainty about whether generative AI improves productivity without degrading quality or fairness; (ii) the *role of the researcher*, which in our setting is limited to providing a tool and training rather than intervening in case outcomes; (iii) *potential harms* to participants and nonparticipants, including risks of overreliance, opacity, and identity-linked disparities; (iv) *potential harms from data collection*; (v) *conflicts of interest* and institutional relationships governing access to the courts; (vi) safeguards for *intellectual freedom* and judicial independence, including that the tool is advisory and judges retain full discretion; (vii) *feedback and debriefing*, including communication to participating judges and relevant institutions about aggregate findings; and (viii) foreseeable *misuse of research outputs*, such as deploying AI without training or using performance metrics for punitive monitoring. We also describe the mitigation steps embedded in the intervention, including targeted training, usage guidance, and monitoring for account sharing, as well as the outcome measurement strategy that directly tests for changes in opinion quality and identity-linked decision patterns.

i) Policy Equipoise and Scarcity

There was genuine policy equipoise regarding whether providing generative AI access alone, or providing it with targeted training, would improve or hinder judicial productivity and opinion quality. Nevertheless, to manage scarcity and ensure fairness, control group received a placebo course on technology to ensure all participants received some form of professional development.

ii) Role of the Researchers with Respect to Implementation

The researchers co-developed the custom AI tool, JudgeGPT, in partnership with Pakistan's judiciary and the Federal Judicial Academy. While researchers designed the randomized controlled trial (RCT) and monitored platform logs for integrity, the intervention

was delivered through official judicial training channels to integrate it into day-to-day work processes.

iii) Potential Harms to Participants or Non-Participants from the Intervention

A primary risk to non-participants (litigants) was that AI-assisted decisions might be lower quality or reflect increased bias. The study addressed this by using LLM-based, human-validated measures of opinion quality and administrative data to ensure that productivity gains did not come at the expense of legal reasoning or impartiality. Analysis confirmed no increase in gender or ethnic bias.

iv) Potential Harms from Data Collection or Research Protocols

To mitigate risks related to data privacy, JudgeGPT used password-protected logins, ensuring anonymity. Productivity outcomes were measured using routine administrative statistics from the Justice Department of Pakistan (LJCP) rather than outcomes generated solely within the experiment, reducing experimenter-induced reporting bias.

v) Financial and Reputational Conflicts of Interest

The authors declare no competing financial interests. The study was conducted in collaboration with state institutions, and the Supreme Court of Pakistan formally recognized the use of such tools in trial courts, providing a framework for institutional oversight.

vi) Intellectual Freedom

The researchers maintained full intellectual freedom to report both the benefits (productivity gains) and the risks (deterioration in quality when AI is used without training) of the intervention.

vii) Feedback to Participants or Communities

The study included structured training sessions for all participating judges, including a lecture on "legal ethics and technology". Successful participants received certificates jointly issued by ETH Zurich and the Federal Judicial Academy.

viii) Foreseeable Misuse of Research Results

A foreseeable misuse is the adoption of AI tools by judiciaries without the "targeted training" found to be essential for maintaining decision quality. The paper explicitly warns that AI access alone may reduce opinion quality relative to a no-AI control group.

Appendix B2. Preanalysis Plan, RCT ID AEARCTR-0012906

Below we present our pre-analysis plan, following Banerjee et al. (2020), who recommend treating the final paper as distinct from the "results of the PAP" and, where helpful, providing a short, publicly available "populated PAP." The original PAP as registered on the AEA RCT Registry is reproduced below in normal font; our additions and responses, including this note, appear in italics.

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.

Response: *Following the first registration campaign launched by Federal Judicial Academy on their online platform, demand for generative AI tools increased sharply and the judiciary asked us to run a second recruitment round. The resulting sample comprises 1559 trial-court judges, compared with 800 planned ex ante. We subsequently randomized the recruited sample into the experimental treatment arms.*

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.

Response: *We ultimately secured agreement from Pakistan's judiciary to evaluate JudgeGPT in a three-arm randomized controlled trial with 1,559 trial-court judges. Assignment was to: (i) generative AI access plus targeted training (tool-specific instruction and risk mitigation), (ii) access plus placebo training (a general "technology and law" course with information on risk mit-*

igation), or (iii) placebo-training-only control (no generative AI access but some information in risk mitigation). Based on input from local counterparts and the ethics review process, we delivered the placebo course to all arms. This choice allows us to provide baseline information on some risks and responsible use of generative AI to all judges, including those in the control group who could still use off-the-shelf tools such as ChatGPT outside the study.

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

Response. *We report results for the five pre-specified PAP outcome variables. Most of these outcomes are already presented in the main paper, and we indicate where the corresponding results appear. In this appendix, we additionally report alternative specifications suggested in the PAP, including versions with and without strata or court fixed effects, to assess the robustness of the main findings. First, we measure attendance and participation using the number of logins. These results are reported in the main text in Table 1; for completeness, we also report them in Column (4) of Table PAP 1 under alternative fixed-effects specifications. Second, we study AI tool utilization. These results are reported in Table 2 using tool-login measures and, for completeness, in Column (3) of Table PAP 1, where we repeat the analysis using the indicator for whether a judge self-reports being a frequent user of Generative AI, again under alternative fixed-effects specifications. Third, following the PAP, we analyze judgment-text outcomes, including readability, argument diversity, and legal argument content. These results are presented in Table PAP 2. For completeness, we also summarize the readability results in Column (5) of Table PAP 1 using the Gunning Fog index. As shown in Appendix Table PAP 2, across both ITT and 2SLS specifications, we find no statistically significant effects of JudgeGPT access, with or without targeted training, on judgment readability. The point estimates are small throughout, and we do not detect systematic differences across treatment arms. In other words, judges with access to the tool, whether trained or not, did not produce opinions that were systematically more readable. Because these judgment-text measures*

were explicitly pre-specified in the PAP, we relabel the current Table A9 as ***Table PAP 2*** *and cite it accordingly in the appendix. Fourth, we examine perceptions of AI utility, reported in Table 4 and, for completeness, in Column (1) of Table PAP 1 using responses to the question, "Do you support the use of Generative AI assistants such as GPT in your job as a judge?" measured on a 1–4 scale. Fifth, we study productivity expectations related to AI, reported in Table 5 and, for completeness, in Column (5) of Table PAP 1 using responses to the question, "Do you expect that an AI for Judges Course can improve your productivity?" also measured on a 1–4 scale.*

Regression Specification

We will estimate OLS regressions with the following specifications:

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

Response*: Below, as Table PAP 1, we report our main outcomes using the specifications checks suggested in the pre-analysis plan: no covariates, court fixed effects (when court identifiers are available), and fixed effects for the stratification unit. Because court identifiers are not consistently observed, we do not include court fixed effects in all main-paper specifications. For the subset of observations where we can assign a court, we report results with court fixed effects here. In the main results, we also present estimates without covariates to adhere to the pre-analysis plan; adding our rich set of baseline controls leaves the estimates largely unchanged, so we also report controlled specifications in the main paper for precision. The table below follows the pre-analysis plan specification.*

TABLE PAP 1: Robustness to Different FEs

Panel A: No FE	Support JudgeGPT	Frequent User of Gen AI	AI-for-Judges Course	# Logins
	(1)	(2)	(3)	(4)
JudgeGPT Access and Targeted Training	0.232***	0.124**	0.147**	82.783***
	(0.079)	(0.050)	(0.070)	(3.657)
JudgeGPT Access and Placebo Training	0.274**	0.159**	0.080	16.671***
	(0.109)	(0.068)	(0.104)	(2.014)
Panel B: Court FE				
JudgeGPT Access and Targeted Training	0.241	0.142*	0.219*	83.944***
	(0.151)	(0.085)	(0.119)	(9.687)
JudgeGPT Access and Placebo Training	0.191	0.185	0.140	14.163
	(0.213)	(0.113)	(0.199)	(13.035)
Panel C: Strata + Court FE				
JudgeGPT Access and Targeted Training	0.246	0.141*	0.208*	84.170***
	(0.151)	(0.081)	(0.116)	(10.722)
JudgeGPT Access and Placebo Training	0.227	0.190*	0.159	13.803
	(0.212)	(0.111)	(0.195)	(15.239)

Note: The table reports treatment-effect estimates separately with different fixed effects. Panel A uses no FE, Panel B uses court fe, Panel C uses both strata and court fe. Standart errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

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.

***Response**: Below, we report heterogeneity analyses for our main outcomes, as pre-specified in the pre-analysis plan, using the full set of baseline covariates and stratification-unit fixed effects. We present heterogeneity by judge age (Table PAP 3) and by province (Table PAP 4). We also examine heterogeneity by case type (civil vs. criminal; Table PAP 5) and by pre-treatment familiarity with AI tools (Table PAP 6).*[8]

[8]For all heterogeneity analyses, we exclude Gunning Fog and legal-arguments outcomes, which display very limited variation. Table PAP 3 reports the corresponding estimates for the full sample and shows null effects of both treatment arms.

TABLE PAP 2: JudgeGPT Access and Text Quality

	ITT		2SLS	
	(1)	(2)	(3)	(4)
	Gunning Fog Readability			
Assigned JudgeGPT Access and Targeted Training	-0.011	0.006		
	(0.013)	(0.042)		
Assigned JudgeGPT Access and Placebo Training	-0.011	0.036		
	(0.017)	(0.067)		
JudgeGPT Access and Targeted Training			-0.011	0.006
			(0.013)	(0.041)
JudgeGPT Access and Placebo Training			-0.016	0.054
			(0.026)	(0.094)
Anderson-Rubin P-val.			.715	.836
Mean Dep. Var.	.008	.011	.008	.011
	# of Legal Arguments			
Assigned JudgeGPT Access and Targeted Training	-4.522	-10.062		
	(11.130)	(10.404)		
Assigned JudgeGPT Access and Placebo Training	17.315	20.366		
	(29.951)	(31.991)		
JudgeGPT Access and Targeted Training			-4.468	-9.993
			(11.061)	(10.140)
JudgeGPT Access and Placebo Training			25.686	30.795
			(39.061)	(42.679)
Strata Fixed Effects	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes
Anderson-Rubin P-val.			.708	.395
Mean Dep. Var.	124.494	125.155	124.494	125.155
Observations	293	275	293	275

Note: Newey-West standard errors in parentheses. The dependent variable is a gunning fod readability metrics standartized to mean 0 and standard deviation 1, 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE PAP 3: Heterogeneity by Judge Age

Panel A: Below Median	Support JudgeGPT	Frequent User of Gen AI	AI-for-Judges Course	# Logins
	(1)	(2)	(3)	(4)
JudgeGPT Access and Targeted Training	0.292***	0.159***	0.207**	89.831***
	(0.097)	(0.060)	(0.088)	(5.112)
JudgeGPT Access and Placebo Training	0.337**	0.185**	0.204	14.046**
	(0.151)	(0.092)	(0.144)	(5.884)
Panel B: Above Median				
JudgeGPT Access and Targeted Training	0.128	0.100	0.056	71.586***
	(0.114)	(0.082)	(0.102)	(5.583)
JudgeGPT Access and Placebo Training	0.160	0.160*	-0.046	17.006***
	(0.143)	(0.094)	(0.142)	(4.561)

Note: The table reports treatment-effect estimates separately by subsamples. Panel A uses judges whose age below median (43), Panel B uses judges whose age above median. Standart errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE PAP 4: Heterogeneity by Province

Panel A: Punjab Subsample	Support JudgeGPT	Frequent User of Gen AI	AI-for-Judges Course	# Logins
	(1)	(2)	(3)	(4)
JudgeGPT Access and	0.167*	0.069	0.109	67.300***
Targeted Training	(0.095)	(0.064)	(0.084)	(4.787)
JudgeGPT Access and	0.156	0.043	0.050	15.044***
Placebo Training	(0.141)	(0.092)	(0.134)	(5.539)
Panel B: Sindh Subsample				
JudgeGPT Access and	0.256	0.169	0.197	106.470***
Targeted Training	(0.217)	(0.107)	(0.183)	(12.368)
JudgeGPT Access and	0.443*	0.283**	0.288	26.377*
Placebo Training	(0.255)	(0.135)	(0.223)	(13.817)
Panel C: Khyber Pakhtunkhwa Subsample				
JudgeGPT Access and	0.252*	0.275**	0.147	91.028***
Targeted Training	(0.133)	(0.110)	(0.141)	(11.319)
JudgeGPT Access and	0.244	0.283**	-0.082	31.452***
Placebo Training	(0.184)	(0.133)	(0.201)	(10.572)
Panel D: Balochistan Subsample				
JudgeGPT Access and	0.586***	0.001	0.117	132.355**
Targeted Training	(0.177)	(0.263)	(0.403)	(53.171)
JudgeGPT Access and	0.564**	0.443	0.229	61.538
Placebo Training	(0.244)	(0.308)	(0.455)	(58.281)
Panel E: Other Provinces Subsample				
JudgeGPT Access and	0.715***	-0.181	0.599***	125.503***
Targeted Training	(0.116)	(0.161)	(0.173)	(23.233)
JudgeGPT Access and	-0.011	-0.549	-1.578***	-89.284
Placebo Training	(0.376)	(0.540)	(0.367)	(110.492)

Note: The table reports treatment-effect estimates separately by subsamples. Panel A uses Punjab province, Panel B uses Sindh province, Panel C uses Khyber Pakhtunkhwa province, Panel D uses Balochistan province, Panel E uses Islamabad (federal territory). Standart errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE PAP 5: Heterogeneity by Judge Type

Panel A: Criminal	Support JudgeGPT	Frequent User of Gen AI	AI-for-Judges Course	# Logins
	(1)	(2)	(3)	(4)
JudgeGPT Access and Targeted Training	0.254	0.064	0.141	83.218***
	(0.239)	(0.099)	(0.145)	(19.587)
JudgeGPT Access and Placebo Training	0.289	0.237	0.016	-3.021
	(0.339)	(0.164)	(0.284)	(34.706)
Panel C: Civil				
JudgeGPT Access and Targeted Training	0.218	0.135	0.428**	83.295***
	(0.261)	(0.131)	(0.204)	(14.737)
JudgeGPT Access and Placebo Training	0.088	0.104	0.287	-1.432
	(0.447)	(0.177)	(0.378)	(19.570)

Note: The table reports treatment-effect estimates separately by subsamples. Panel A uses Criminal Judges Subsample, Panel B uses Civil Judges Subsample. Standart errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE PAP 6: Heterogeneity by Pre Treatment Familiarity

Panel A: Pre-Familiar with AI	Support JudgeGPT	Frequent User of Gen AI	AI-for-Judges Course	# Logins
	(1)	(2)	(3)	(4)
Assigned JudgeGPT Access	-0.267***	-0.004	-0.247***	122.382***
and Targeted Training	(0.099)	(0.049)	(0.079)	(28.992)
Assigned JudgeGPT Access	0.000	0.000	0.000	0.000
and Placebo Training	(.)	(.)	(.)	(.)
Panel B: Pre-UnFamiliar with AI				
Assigned JudgeGPT Access	0.181**	0.077	0.099	55.636***
and Targeted Training	(0.072)	(0.047)	(0.064)	(3.730)
Assigned JudgeGPT Access	0.204**	0.136**	0.046	15.754***
and Placebo Training	(0.088)	(0.057)	(0.085)	(3.161)

Note: The table reports treatment-effect estimates separately by subsamples. Panel A uses judges familiar with AI tools before treatment, Panel B uses judges unfamiliar with AI tools before treatment. Standart errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE PAP 7: Tool Usage and Productivity

	AI Support			
# Logins	0.107***	0.117***	0.105***	0.097***
	(0.019)	(0.019)	(0.032)	(0.032)
Strata FE	No	Yes	Yes	Yes
Court FE	No	No	Yes	Yes
Controls	No	No	No	Yes
N Obs.	1081	1081	1080	1080
Mean Dep. Var.	3.547	3.547	3.547	3.547

Note: The table reports treatment-effect estimates of count logins on support of generative AI using. Standart errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Hypotheses

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

***Response**: We report evidence on H1 in Figure 4 shows that judges assigned JudgeGPT access with targeted training course use the tool more intensively and more persistently over time than judges assigned JudgeGPT access with placebo training. The targeted-training course group remains active at substantially higher rates throughout the post-assignment period, with a noticeably slower decline in usage. Consistent with this pattern, Table 2 shows that treatment increases the probability that judges report being frequent users of generative AI by about 0.10 in the ITT specifications, while the corresponding 2SLS estimates for JudgeGPT access with targeted training are about 0.15. Relative to the sample mean, this corresponds to roughly a 15 percent increase in the likelihood of reporting frequent generative AI use.*

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

***Response**: We assess H2 using blinded pairwise comparisons of pre- and post-treatment judgment quality, evaluated by LLMs and validated against the assessments of a licensed Pakistani Supreme Court lawyer. Figure 6 reports the results. In the JudgeGPT plus targeted-training group, post-treatment judgments are preferred in about 70 percent of comparisons, well above the 50 percent benchmark of no systematic change. By contrast, in the JudgeGPT plus placebo-training group, post-treatment judgments are preferred in only about 35 percent of comparisons, while the active control group (no JudgeGPT + placebo course) remains close to 50 percent. These patterns are consistent with targeted AI assistance improving judgment quality and legal reasoning.*

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

***Response**: The evidence supports H3. As shown in Table 4, judges assigned JudgeGPT report greater support for using generative AI in their judicial work, and this pattern is stable across specifications. These results are consistent with the view that exposure to the tool increased judges' perceived usefulness of AI for judicial tasks*

APPENDIX C. TEXT ANALYSIS APPENDIX

C1. Human Validation of LLM Opinion Quality Comparisons

To validate the LLM-based opinion quality measure, we commissioned human annotations from a licensed Pakistani lawyer from the Supreme Court of Pakistan (With over 10 years of experience) to rate the judgments. The lawyer rated 90 pre/post opinion pairs drawn from the treatment batches (3, 4a, 4b). For each pair, the lawyer received the two opinion texts (labeled A and B) with no indication of which was pre- or post-treatment, and rated which opinion was better on the same -2 to $+2$ scale used by the LLMs: $+2$ (A much better), $+1$ (A slightly better), 0 (tie), -1 (B slightly better), -2 (B much better). Quality was assessed along the same dimensions as the LLM prompt: clarity, conciseness, factual presentation, and legal reasoning.

Agreement metrics. We link each human rating to the corresponding GPT and Gemini scores via a crosswalk that maps annotation folders to case-ID pairs. We then compute directional agreement: the fraction of pairs where both the human and the AI agree on which opinion is better, excluding pairs where either rater scored a tie (since direction is undefined for ties).

Table A12 reports the results for the annotator with stronger AI agreement. GPT and the lawyer agree on direction in 70.1% of the 87 non-tie pairs. GPT and Gemini agree in 65.2% of their 89 non-tie pairs. When GPT and Gemini agree with each other on direction, the AI ensemble agrees with the lawyer in 83.9% of cases. These rates are consistent with the level of agreement typically observed between independent human raters in similar text-evaluation tasks.

Additional metrics. For the same annotator, the Spearman rank correlation between human and GPT scores on the 5-point scale is $\rho = 0.319$ ($p = 0.002$, $N = 90$), and between human and Gemini scores is $\rho = 0.472$ ($p < 0.001$, $N = 89$). Cohen's κ on the 3-category sign (A better / equal / B better) is 0.282 for GPT and 0.461 for Gemini. The linearly weighted κ on the full 5-point scale is 0.223 for GPT. These values indicate fair-to-moderate agreement, which is typical when comparing human raters in subjective text-quality tasks, particularly for legal documents in a multilingual setting where opinions contain mixed Urdu and English text from OCR scans.

Score distributions. Both the human and AI raters tend to favor B (post-treatment) opinions on average: the human mean score is -0.63 and the GPT mean is -0.62 (on the -2 to $+2$ scale where negative means B is preferred). GPT never scores a tie (0), producing a forced-choice distribution concentrated at -1 and -2, whereas the human annotator scores ties in 3 of 90 pairs. Gemini exhibits less negative skew (mean $= 0.04$), reflecting a tendency toward more balanced scoring.

Position-bias control. The main LLM comparison pipeline generates both orderings for each pair (A=Pre/B=Post and A=Post/B=Pre) to control for position bias. The human validation uses only one ordering per pair (A=Pre, B=Post). The agreement metrics reported here therefore reflect agreement on a fixed ordering, and the position-bias controls apply only to the LLM scores used in the main analysis.

Discussion. The second annotator exhibited substantially lower agreement with the LLMs (directional agreement of 43% with GPT), suggesting meaningful inter-annotator variability. This is not uncommon in legal text evaluation, where assessors may weight different quality dimensions differently. The stronger-agreement annotator's results support the use of the LLM-based measure as a reasonable proxy for human judgment of opinion quality, particularly when GPT and Gemini agree. The 84% agreement rate in the AI-consensus subset provides additional confidence that the ensemble measure captures genuine quality differences rather than idiosyncratic model behavior.

C2. LLM-Based Bias Detection in Judicial Opinions

In addition to the vignette-based bias measures reported in the main text, we conduct a direct analysis of bias *within the text* of judicial opinions. We use an LLM to evaluate each opinion along two dimensions: (i) gender bias, defined as anti-women reasoning or endorsement of conservative gender norms; and (ii) pro-Muslim bias, defined as invocation of Islamic legal or moral authority beyond what the case requires. Each opinion receives a score from 0 (no bias detected) to 3 (strong bias), along with a flag indicating whether the dimension is relevant to the case at all.

Methodology. The methodology builds on earlier work analyzing attitudes from text (Ash et al., 2023, 2024), drawing on more recently developed language models for this

task (Ash et al., 2025). We send the full text of each opinion to GPT-5-mini with a structured prompt that asks the model to assess gender and pro-Muslim bias on a 0–3 scale, citing specific language from the opinion. The prompt instructs the model to score based only on the text, not to infer bias from case type alone, and to distinguish routine application of Islamic personal law from biased reasoning. Opinions where OCR artifacts or non-case documents (memos, research notes) were detected by our earlier document classification pipeline are excluded. This yields scores for 4,252 opinions across all experimental batches.

Prevalence. Gender is relevant to approximately 39% of opinions, primarily in family law, maintenance, and cases involving female complainants. Islamic or Muslim identity is relevant in about 32% of opinions.

Regressions. We aggregate bias scores to the judge level by taking the mean across each judge's post-treatment opinions, then estimate OLS regressions following the same specification as the chatlog analysis:

$$\text{Bias}_j = \beta_1 \cdot \mathbf{1}[\text{AI + Training}]_j + \beta_2 \cdot \mathbf{1}[\text{AI, No Training}]_j + \delta_s + X_j'\gamma + \varepsilon_j,$$

where the omitted category is the No AI group (Batch 4b), δ_s are strata fixed effects, X_j are the same survey controls used throughout the paper, and standard errors are HC1 robust. We report results for two sets of outcomes: post-treatment bias scores and post-treatment dimension prevalence (Figure A5).

Results. Neither treatment arm produces a statistically significant effect on either bias measure. For gender bias scores, the AI + Training coefficient is 0.24 ($p = 0.61$) and the AI, No Training coefficient is -0.03 ($p = 0.97$). For pro-Muslim bias scores, the coefficients are -0.25 ($p = 0.39$) and -0.09 ($p = 0.80$), respectively. The prevalence regressions tell a similar story: no treatment arm significantly shifts whether gender or religious dimensions appear in post-treatment opinions. All confidence intervals are wide, reflecting the small sample of 41 judges (20 in Batch 3, 6 in Batch 4a, 15 in Batch 4b) with post-treatment opinions in these batches.

Interpretation. We find no evidence that access to *JudgeGPT* exacerbates or reduces bias in judicial opinions along either dimension. The null results are consistent across bias

scores and dimension prevalence. These text-based measures complement the vignette-based bias analysis in the main text by providing a direct assessment of bias in actual judicial writing, though the small post-treatment sample sizes limit statistical power.