COURTS OF TOMORROW:

Evidence from a Nationwide Rollout of Generative AI

Sultan Mehmood ⓡ Christoph Goessmann ⓡ Elliott Ash*

August 9, 2026

Abstract

We present the first large-scale field experiment evaluating the integration of generative AI into a national justice system. In partnership with Pakistan's judiciary, we built *JudgeGPT*, a custom generative AI assistant designed for Pakistan's trial courts. We randomized 1,559 judges serving across 118 courts into one of three arms: (i) AI access with targeted training on its use; (ii) AI access with generic training on technology and law; or (iii) generic training without AI access. Judges who received AI access together with targeted training on the use of the tool were more likely to adopt it, use it more intensively, and continue using it over time. Attitudes toward AI also shift, with treated judges expecting the tool and the targeted training to increase their productivity. Administrative records move in the same direction: districts with greater exposure to treated judges resolve more cases. At median-district exposure, introducing AI with targeted training corresponds to 1,848 additional cases resolved per year, a 6.3 percent increase over the mean. This increase does not appear to come at the expense of reduced quality, as measured by case appeals (which slightly decline) or text-based measures of judicial writing. Chat-log-based usage patterns suggest that judges use AI mainly to clarify legal concepts and support drafting. Targeted training appears to shift use toward tasks where language models are likely to be more useful, such as text improvement, and away from more open-ended legal queries where responses are more costly to verify. 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 Judge Shazia Munawar for her support. We also thank the Pakistan Judicial Academy for their feedback on the AI tool, and Kseniya Palceva, Georgiy Marinichev, Mikhail Ivanov, Andrei Lengler, and Tim Ehrensperger for excellent research assistance. Finally, we thank David Autor, Avi Goldfarb, Erik Brynjolfsson, David Lagakos, W. Bentley MacLeod, Jonah Rockoff, Bocar Ba, and the participants at the NBER Summer Institute 2026, American Economics Association 2026 meetings, Brown Political Economy Seminar, Columbia Applied Micro Seminar, Bocconi AI & Society Conference, and Princeton Seminars 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:** Author 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

Can generative AI support public-sector organizations? Recent theoretical work argues that AI alters task allocation, the organization of production, and the direction of innovation (Aghion et al., 2019; Acemoglu, 2025). Early evidence from the private sector suggests that access to large language models can raise workers' productivity in writing-intensive and information-processing tasks (Noy and Zhang, 2023; Brynjolfsson et al., 2025). Whether similar gains arise in the public sector, and, if so, how large they are, remains less clear. Governments operate under a different set of incentives, produce outputs that are harder to measure, and face constraints that may be more binding than those faced by private organizations (Finan et al., 2017; Acemoglu and Lensman, 2024). These questions are heightened in the courts: judges exercise a form of particularized, expert judgment that recent work suggests may be especially amenable to augmentation by AI (Autor, 2024), while allocating liberty and property rights and shaping the implementation of law and policy. The consequences of AI adoption in this setting may therefore be especially substantial.

Credibly estimating these effects, however, requires overcoming three main challenges. First, judges' adoption of AI is endogenous, making it hard to separate the causal effect of AI from pre-existing differences in judges' ability or motivation. Second, detecting effects on court output requires substantial scale. Administrative measures of case resolution aggregate the work of many judges, so experiments involving only a small share of the judiciary may generate too little variation to identify changes in these aggregate records. Third, understanding why productivity effects arise requires observing how judges use the tool in practice, which is typically unobserved. Knowing that a judge had access to AI reveals little without data on what they asked, how often they used it, and for what purpose.

These challenges motivate our design: random assignment, near-national scale, and direct observation of use. We conduct a large-scale field experiment in partnership with Pakistan's judiciary around *JudgeGPT*, a custom generative AI assistant built for Pakistan's trial courts and grounded in the country's corpus of court rulings and statutes. To address non-random adoption, we randomized 1,559 judges working in trial courts, where most cases are first heard. The sample covers roughly half of the country's trial judges and 70 percent of its district courts. In these courts, civil matters account for roughly 65 percent of the caseload and include disputes over property, contracts, inher-

itance, and family law, with criminal prosecutions making up much of the remainder. Pakistan courts are also highly congested, with roughly 2.26 million cases pending nationwide during our study period, placing considerable pressure on judicial time (Law and Justice Commission of Pakistan, 2024). Further, the experiment began in a judicial setting with limited prior exposure to generative AI: at baseline, only about 25 percent of judges reported familiarity with large language models such as ChatGPT.

Judges were randomly assigned to one of three experimental groups. The first group received access to *JudgeGPT* together with targeted training tailored to its judicial use. The second received access to the same tool but only a generic course on technology and law, without tool-specific guidance. The control group received the generic course without access to *JudgeGPT*. This design separates access to the tool from training on how to use it. The scale of the experiment also allows us to link judge-level assignment to district-level administrative records covering the country and examine effects on case resolution.

We measure outcomes using four complementary data sources. First, through Pakistan's Federal Judicial Academy, we fielded a baseline survey measuring judges' characteristics, prior exposure to AI, and pre-treatment attitudes, followed by a survey roughly three months after rollout measuring post-treatment attitudes and perceived productivity gains from AI access and training. Second, we link experimental assignment to district-level administrative court records to test whether treatment translated into higher case resolution. Third, we use available judicial opinions written before and after the intervention to assess effects on writing quality and whether AI use changed the language judges employ when discussing gender or religion. Finally, we use *JudgeGPT* platform records to measure take-up and engagement, and the associated chat logs to characterize how judges used the tool in practice.

The experiment delivers four main findings. First, access to *JudgeGPT* increases AI use, but sustained engagement depends on targeted training. Tool-specific instruction raises usage at least four-fold relative to generic training on technology and law. Increased usage is accompanied by shifts in attitudes toward generative AI in judicial work. Treated judges report more favorable views of AI assistance in judicial tasks and are more likely to see AI-for-judges training as productivity-enhancing. These changes do not appear to reflect a broader shift in attitudes toward digital technology, as familiarity with Microsoft Office and other non-AI tools does not change across treatment arms.

Second, districts with greater exposure to judges assigned to *JudgeGPT* with targeted training resolve more cases. The estimates are positive across specifications and im-

ply economically meaningful gains in court output. Moving from zero exposure to the 25th percentile of the district exposure distribution corresponds to roughly 616 additional cases resolved per district-year. Moving to median district exposure corresponds to roughly 1,848 additional cases resolved per district-year—a 6.3 percent increase over the mean. These gains do not appear to come at the expense of more legal challenges to judicial decisions: appeal rates, if anything, decline slightly with greater AI exposure.

The implied savings are large, though necessarily approximate. Because greater exposure to *JudgeGPT* increases the number of cases resolved, achieving the same output without the tool would otherwise require additional judicial appointments. We can make a back-of-the-envelope calculation from *JudgeGPT* as the salary-equivalent cost of that capacity. At observed infrastructure costs, the salary-equivalent value of the additional judicial capacity implied by our estimates is approximately $41.90 for each dollar spent on *JudgeGPT*. Even with conservative benchmarks, the potential savings are at least $10 for every dollar invested in *JudgeGPT*.

Third, we look at how AI adoption affects judicial writing and reasoning, in particular to explore whether higher case resolution comes at the cost of lower writing quality or greater bias toward particular groups. Using the available data on opinions written before and after the intervention, we study several measures of judicial writing. As expected, treated judges' post-treatment opinions contain more text classified as AI-generated relative to the control group. But we find little evidence that this translated into a deterioration in writing quality, as measured by readability, judgment length, and an expert-validated LLM-based quality assessment. We likewise find little evidence of systematic changes in gender or pro-Muslim bias in judicial language.

Fourth, chat-log evidence helps interpret these results by showing how judges used *JudgeGPT*. Because the tool was deployed through a custom platform, we observe not only whether judges used it, but also the tasks they asked it to perform. Overall, use was concentrated in legal research and writing assistance, with open-ended legal questions accounting for the largest share of prompts. Targeted training changed the composition of use: judges made more requests for text improvement and summarization and fewer open-ended legal queries. The remaining task categories changed little. These patterns suggest that targeted training shifted use toward more bounded forms of assistance, potentially reducing time spent on routine editing and drafting tasks.

We assess the robustness of these findings in several ways. First, random assignment appears to produce comparable groups, with the treatment arms balanced on observed

judge characteristics and pretreatment attitudes. Second, we rule out a simple effort-based explanation for the results (i.e., treated judges working harder, for example due to experimenter demand, rather than using AI) by showing that treated judges do not report spending more time on legal work or having a higher workload. Third, we test for the extent of spillovers between treated and control judges working in the same courts by testing whether outcomes among no-access judges vary with the number of with-access judges in their district. The estimated relationships are close to zero and statistically insignificant, consistent with such spillovers being limited. Fourth, our estimates remain similar after adjusting for attrition using Lee bounds (Lee, 2009) and more recent methods such as generalized bounding methods (Semenova, 2025), and after adjusting for multiple testing using Romano–Wolf stepdown p-values (Clarke et al., 2020). Fifth, for district-level case-resolution outcomes, we additionally report local court-level balance tests, randomization inference, and bootstrap-based standard errors. Across these exercises, the main conclusions remain similar.

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, while also highlighting heterogeneity across workers, tasks, and educational backgrounds (Autor, 2024; Brynjolfsson et al., 2025; Dell'Acqua et al., 2026; Cruces et al., 2026).[1] A natural next question is whether these gains extend to public-sector organizations, where production is governed by distinct rules and oversight mechanisms (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, to the best of our knowledge, 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. Our results show that AI can reconfigure task

[1] Positive productivity effects are also documented in software development (Peng et al., 2023), writing (Noy and Zhang, 2023), IT security (Bono and Xu, 2024), and creative problem-solving (Boussioux et al., 2024; Lee and Chung, 2024), though effects vary across workers and tasks (Kreitmeir and Raschky, 2024).

content and skill demands even when aggregate or long-run productivity effects are uncertain (Acemoglu et al., 2022; Brynjolfsson et al., 2017; Chen et al., 2025). In the judicial context, AI access with targeted training could affect court output.

Second, 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 LLMs 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). Our intervention pairs AI access with targeted instruction designed to promote tool-specific responsible use. This feature is important because opacity, overreliance, and bias are central concerns in both the broader AI ethics literature and applications of AI to legal decision-making (Buolamwini and Gebru, 2018; Yang and Dobbie, 2020; Vicente and Matute, 2023; Acemoglu et al., 2023). Related evidence highlights the importance of implementation: judicial discretion can substantially mediate the effects of algorithmic risk assessments (Stevenson and Doleac, 2024), while take-up of predictive tools can remain low even among legal professionals who express strong ex ante interest (Abrams et al., 2026). We complement this literature by studying a large institutional deployment across 118 district courts in Pakistan, where AI access was paired with different forms of training and the gains in court output are concentrated among judges receiving targeted, tool-specific training.

Third, we contribute to work on state effectiveness by studying technological change in a core state institution. Limited state capacity in the judiciary is often tied to limited information technology and inadequate support staff (Papaioannou and Karatza, 2018). Prior research shows that institutional design, personnel systems, incentives, monitoring, and training can shape public-sector performance (Mehmood et al., 2024), including in the judiciary (Landes and Posner, 2009; Liu et al., 2022; Mehmood, 2022; de Janvry et al., 2023; Ash and MacLeod, 2024; Chemin et al., 2024; Ash et al., 2026). Yet causal evidence on how new technologies alter the production of justice remains relatively limited, especially for tools that enter judges' day-to-day research and writing workflows. We study these questions through an experiment that introduced a generative AI assistant into frontline trial courts in Pakistan under different training regimes. We measure its effects using administrative records on case resolution and complement these outcomes

with AI-based measures of judicial text and court output. We therefore also speak to recent work in computational social science emphasizing the potential for AI to measure judicial output, while addressing the access and data-structure barriers that have limited these applications (Pah et al., 2022).

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

2 Institutional Background

Colonial Origins. Pakistan's legal institutions trace their roots to the British colonial era, 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 practices (Lau, 2006).

Judicial Hierarchy. Pakistan has a three-tier court hierarchy. That includes 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. Appendix Figure A1 illustrates the organization of the court system and the corresponding cases at each level. The caseload is heavily concentrated at the bottom of this hierarchy. In 2024, Trial Courts accounted for about 82 percent of the 2.26 million pending cases, compared with 15 percent in the High Courts and 2 percent in the Supreme Court.

As courts of first instance, they form 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). For most citizens, these courts are the main point of contact with the formal justice system (Law and Justice Commission of Pakistan, 2024).

This concentration reflects the institutional role of trial courts. Hearings, witness examination, evidence recording, and much of the fact-finding process occur at this level, making it the main site where delay accumulates. Yet the district judiciary operates with limited digital infrastructure, weak research and drafting support, and unusually low staffing: Pakistan has fewer than 2 judges per 100,000 inhabitants, compared with 22 in the European Union (Council of Europe, 2024), 30 in England and Wales (Ministry of Justice and Judicial Office, 2025), and 8 in Brazil (Conselho Nacional de Justiça, 2024).

Judicial Careers. Trial court judges in Pakistan are recruited into a career judiciary. Entry-level Civil Judges and Judicial Magistrates are selected through competitive examinations, while higher district-court positions are filled largely through promotion. Career progression follows a hierarchical ladder from Civil Judge to Senior Civil Judge, Additional District and Sessions Judge, and District and Sessions Judge.

Compensation is set by provincial pay scales and judicial-service rules. In Khyber Pakhtunkhwa, Civil Judges are BPS-18 and District and Sessions Judges BPS-21; 2024–25 budget estimates imply basic pay of roughly PKR 200,000–250,000 per month, or about $720–900, for District and Sessions Judge posts (Government of Khyber Pakhtunkhwa, 2024). In Sindh, a 2025 recruitment notice advertised gross monthly pay of PKR 210,029, or about $750, for newly appointed Civil Judges (High Court of Sindh, 2025). By comparison, a first-instance judge in Brazil received a monthly *subsídio* of R$37,731.80, roughly $7,600 (State of Ceará, 2023). These service rules also restrict outside legal practice while in service and set retirement at age 60 (Government of Punjab, 1974; Pakistan Bar Council, 1976).

Context for Technological Reform. Pakistan's judiciary has recently pursued reforms to modernize court operations and reduce longstanding delays. Severe backlog and staffing constraints have made technology-based solutions especially salient, as courts face pressure to expand capacity without comparable increases in personnel. These efforts include the judiciary's "Technology and Law" program and the reconstitution of the National Judicial Automation Committee (NJAC) in 2023, with Justice Syed Mansoor Ali Shah, then a Supreme Court Justice, playing a prominent leadership role (Press Information Department, Government of Pakistan, 2023). Around that time, the spread of generative AI tools such as ChatGPT piqued increasing interest within the judiciary. Against this backdrop, the judiciary began preparing for a phased introduction of AI-based legal tools to support

judges in their work.

Our study was implemented within this broader reform context. In partnership with the Federal Judicial Academy of Pakistan (FJA), we co-developed *JudgeGPT*, a customized generative AI legal assistant built on retrieval-augmented generation and grounded in Pakistani case law and statutes.[2] The rollout paired access to the tool with structured training on effective and responsible use of this tool. In *Constitutional Petition No. 1010-L/2022*, the Supreme Court of Pakistan described *JudgeGPT* as "tailored to conform with Pakistan's legislative framework, procedural statutes, and domestic jurisprudence." (Supreme Court of Pakistan, 2024).[3]

Baseline Familiarity with LLMs and Digital Tools. *JudgeGPT* was introduced in a judiciary where many judges were familiar with standard digital tools, but exposure to generative AI remained limited. Figure 1 documents this baseline. As of our study start date, only about 25 percent of judges reported familiarity with large language models such as ChatGPT. Familiarity with everyday digital tools was far higher: roughly 70 percent with Microsoft Office or similar office software, about 80 percent with Google or Bing, and over 90 percent with WhatsApp.

This pattern is important for two reasons. First, it shows that the control group was not operating in an environment where generative AI was already widely used. Second, it clarifies why access was paired with training: limited familiarity with LLMs made instruction on effective and responsible use particularly important.

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. The intervention was embedded within ongoing national judicial training and technology initiatives administered by Pakistan's judiciary, and we worked 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. The study received ethical approval from two review bodies: the ETH Zurich Ethics Commission

[2]Retrieval-augmented generation (RAG) links a language model to an external document collection. When a user submits a query, the system first retrieves relevant legal materials from the underlying corpus and then uses those materials to generate a response, reducing reliance on the model's internal training data alone.

[3]The deployment was also noted in a United Nations General Assembly report, which describes it as an ongoing randomized trial of an AI tool for judges and highlights its potential to generate evidence on judicial productivity (United Nations General Assembly, 2025).

(EK-2023-N-343-A) and the Institute of Business Administration (IBA) Karachi Institutional Review Board (IBA-Esub006/rec-24). This dual approval process helped align the research with ethical standards in both international and Pakistani institutional contexts.

These safeguards align with emerging guidance on judicial AI. The Supreme Court of Pakistan cautioned that judges using *JudgeGPT* must verify AI-generated references and preserve human oversight (Supreme Court of Pakistan, 2024). A United Nations General Assembly report similarly stresses that judicial AI should remain under judicial responsibility and be accompanied by AI literacy training (United Nations General Assembly, 2025). Our design reflects these principles: access was paired with training on responsible use, judges retained the decision authority, and the tool was deployed through the Federal Judicial Academy rather than through an external commercial rollout.[4]

In line with recommendations for structured ethics reporting in social science research (Asiedu et al., 2021), Appendix B1 describes the ethical design of the study, including policy equipoise, the researcher's role, potential harms to participants and non-participants, conflicts of interest, and possible misuse of the intervention and its findings. Training and guidance on safe use were included across treatment arms, including a lecture on "legal ethics and technology." For judges assigned to *JudgeGPT*, we additionally explained the nature of large language models, provided examples of helpful and problematic responses, and emphasized the need to verify outputs against authoritative legal sources. The trainings were designed to reduce uncritical reliance on generative AI and to support informed and cautious use in a sensitive institutional setting. While we cannot fully rule out misuse or misunderstanding, the intervention made the tool's limitations explicit and placed responsibility for judicial decisions with the judge.

3 Research Design and Data

3.1 Experimental Setup

Figure 2 summarizes treatment assignment and training timing across arms (see also Table A1). We randomize judges into two treatment arms and one control group. Assign-

[4] A judgment in the Supreme Court of Pakistan noted that "Judge GPT has been conceptualized and extended to the FJA by Professor Elliott Ash and Professor Christoph Goessmann of ETH Zurich, and Dr. Sultan Mehmood of the New Economic School, Russia," and that "the successful implementation of this AI initiative has been overseen by the Director General of the FJA, Mr. Hayat Ali Shah, alongwith Judge Shazia Munawar Makhdoom, former Additional Director (Programs and ICT), FJA" (Supreme Court of Pakistan, 2024).

ment was randomized ex ante in Stata using judges' names and dates of birth.[5]

Judges assigned to Treatment A ($n = 1,197$) receive access to *JudgeGPT* together with structured, practice-oriented training on how to use the tool effectively and responsibly. The training introduces the tool, explains its capabilities and limits, and provides guidance on responsible prompting and judicial applications, including writing support and legal research. Judges assigned to Treatment B ($n = 180$) receive access to *JudgeGPT* and a general course on technology and law. This course covers broader applications of technology in the legal sector, including legal research tools, machine learning, and AI bias, but does not provide tool-specific instruction on *JudgeGPT*. The control group ($n = 182$) receives the same general technology-and-law training but does not receive access to *JudgeGPT*.

At the request of our implementation partner, the Federal Judicial Academy of Pakistan (FJA), and given high demand for *JudgeGPT* access and targeted training, more judges were randomized into the targeted-training treatment arm. The larger size and staggered rollout of Treatment A, which combined *JudgeGPT* access with targeted training, reflected institutional scheduling constraints under Pakistan judiciary's centralized training calendar. Treatment assignment was randomized within these constraints.[6]

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 comprehensive corpus of judicial opinions and laws via retrieval-augmented generation (RAG, henceforth). Appendix Figure A13 shows the *JudgeGPT* interface in Panel A and illustrates how the platform cites judicial opinions and case law in Panel B. The interface would be familiar to anybody who has used an LLM-based chatbot. We beta-tested the system intensively with the Federal Judicial Academy before deployment.

[5]We discuss within-district spillovers later in the paper. The design limited such spillovers through password-protected individual accounts, explicit instructions against credential sharing, and API logs that help detect potential cross-use.

[6]Judges are balanced across waves on baseline characteristics (Figure 4). To evaluate sensitivity of our results to these timing constraints, we run robustness checks where we re-estimate our main regressions specifications using alternative comparisons across groups. We compare each of the three waves of treated groups who received *JudgeGPT* with targeted training, to both of the comparison groups – which received either *JudgeGPT* with generic training or generic training alone. The estimates remain qualitatively similar across these comparisons, though they are less precise because each comparison uses a smaller sample. All specifications include strata fixed effects and follow the main analysis sample restrictions.

Figure 3 summarizes the RAG workflow. The RAG knowledge base contains 129,235 documents: 128,292 judicial opinions and 943 Pakistani statutes from the Pakistan Code. The opinions span 1900–2024 and include decisions 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 easylaw.ai ($n = 109{,}456$), a database commonly used by Pakistani judges when drafting judgments. Because Supreme Court and High Court decisions serve as binding or highly persuasive precedent for trial courts, these materials closely approximate the legal corpus judges rely on in practice.

Each judgment and statute is divided into chunks of roughly 1,000 tokens, with a 128-token overlap across adjacent chunks to preserve local context. Given a user query, *JudgeGPT* searches the indexed corpus using a BM25-based retriever [7] and returns the 10 highest-ranked passages. These passages, totaling roughly 10,000 tokens, are appended to the original prompt before the model generates its response.[8] 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.

The retrieval system is complemented by a context-specific system prompt that further adapts the base model to the judicial setting. The prompt is hidden from users and cannot be edited by them. It steers the model toward the intended tasks, sets response standards, reduces unnecessary repetition, and discourages unsafe or inappropriate uses. 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 prompt defines *JudgeGPT* as a legal assistant for Pakistani judges and directs it toward judicial tasks, including legal research, summarization, drafting, translation, and

[7] BM25 (Best Matching 25) is a standard keyword-based document retrieval algorithm that ranks documents by their relevance to a user's query.

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

administrative support. It also constrains source use by instructing the model to rely on the materials provided, current Pakistani law, and retrieved case decisions, and to cite only authorities available within the platform.

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

3.3 Training Courses

Focus group discussions with the judges suggested that access alone was unlikely to generate effective use, especially given limited prior exposure to LLM-based tools. Judges needed to understand the interface, relevant use cases, and the limitations of generative AI. The intervention therefore paired platform access with training co-developed with the Federal Judicial Academy (FJA), Pakistan's primary institution for training judges and judicial officers.

The training was delivered to judges assigned to the experimental arms. Judges in Treatment A received access to *JudgeGPT* together with targeted training on how to use the tool in routine judicial work. The course combined six 90-minute lectures over three weeks with practical take-home assignments and a final project. Sessions were held outside court hours, from 7:30pm to 9:00pm Pakistan time, and judges participated by laptop or desktop on Zoom.

Appendix Table A1 summarizes the training conditions and provides the instructional materials for each arm. The Treatment A training introduced *JudgeGPT* for judicial work and emphasized responsible use, including risks of bias and hallucinated citations. The curriculum introduced the capabilities and limits of generative AI in legal settings and emphasized learning by doing. Judges practiced using *JudgeGPT* for legal research, summarization, and drafting, while repeatedly applying safeguards against hallucination, including close attention to source material, verification, and double-checking.

Treatment B held fixed access to *JudgeGPT* while varying the content of the training. Judges in this arm received access to the same tool but completed a more general technology-and-law course. The course covered topics such as legal research, machine learning, search tools, and the limits and biases of AI, but did not include targeted training on using *JudgeGPT* for concrete judicial tasks.

The control group received the same generic training course as Treatment B but no access to *JudgeGPT*, providing a training-only placebo. This holds fixed pre-treatment interest in AI, participation in the Judicial Academy program, exposure to broad instruction on technology and law, and interaction with instructors.

To summarize, Treatment B therefore adds *JudgeGPT* access to the generic training benchmark, while Treatment A adds both access and instruction tailored to judicial work. The targeted course focused on how judges could use the tool in adjudication, including legal search, citation checking, record summarization, drafting, and verification against source material. The design therefore compares AI access and targeted AI instruction against a placebo-training condition, rather than against a passive control group.

Across all groups, the training course included incentives to sustain engagement and keep participation comparable across arms. Successful participants received certificates jointly issued by ETH Zurich and the Federal Judicial Academy of Pakistan (FJA), giving judges formal recognition from both an international academic institution and Pakistan's national judicial training body. We announced in advance that the top 1% of judges in the courses (including judges in the no-AI access group) would be invited to an AI and Judging Symposium in Zurich based on engagement.

These incentives contributed to high participation and engagement. 1,559 judges enrolled, corresponding to about 50 percent of Pakistan's trial judiciary and covering judges from roughly 70 percent of the country's district courts. More than 95 percent of enrolled judges submitted at least one assignment.

3.4 Data and Descriptive Statistics

The analysis combines four complementary data sources. Surveys record judges' baseline characteristics, prior exposure to AI, and post-treatment attitudes toward the use of generative AI in judicial work. Administrative records from the Law and Justice Commission of Pakistan (LJCP) provide systematically recorded measures of district-level case resolution and appeals. Judicial opinions submitted by participating judges allow us to examine dimensions of writing and decision-making that are not observed in administrative data. Finally, platform records capture take-up, intensity of use, and the tasks judges undertook with *JudgeGPT*. Appendix Table A2 defines the variables used in the analysis, Appendix Table A3 reports summary statistics, and Appendix Table A4 summarizes treatment timing and district coverage.

Judge Surveys. The baseline survey provides judge characteristics and records pre-treatment attitudes on the use of AI in judicial work, including support for AI assistance and expectations about its usefulness for judges. The post-training survey, administered approximately three months after course completion, measures judges' assessments of *JudgeGPT*, their evaluations of the training, and their beliefs about whether AI can improve judicial productivity.

Administrative Court Records. To measure district-level court outcomes, we digitize administrative records from the LJCP yearbooks for 2021–2024. These records report lower-court case closures overall and separately for civil and criminal cases, providing a measure of court productivity. They also report within-court appeals, which are heard by judges other than the judge who decided the original trial-court case. These appeals allow us to examine whether changes in case disposal are accompanied by changes in subsequent challenges within the court system.

We use the number of cases resolved to measure case disposal, and appeals per 1,000 resolved cases to scale appeals by the volume of cases disposed in each district. Appendix A2 describes the construction of the district-level administrative panel, documents aggregate trends in district-court case closures and within-court appeals, and explains the aggregation of judge-level assignment into district-level exposure. We use these data to examine whether districts with greater exposure to judges assigned to *JudgeGPT* with targeted training experienced larger changes in court outcomes.

Judicial Opinions. As part of the experiment, we assembled a dataset of roughly 4,000 lower-court judicial opinions written by participating judges before and after the intervention. Because lower-court judgments in Pakistan are not systematically digitized, we collected these texts through course submissions. These opinions complement the administrative records by allowing us to examine dimensions of judicial performance reflected in written decisions. We use them to construct measures of AI-generated content, judgment length, readability, bias related to the defendant's gender or religion, and LLM-based judgment quality. AI-generated content is measured using the Pangram Labs detection tool, while readability is measured using the Gunning Fog Index [9].

[9]The Gunning Fog Index is a standard readability metric that combines average sentence length and the share of complex (three or more syllable) words to estimate the education level required to comprehend a text. Higher scores indicate lower readability.

To measure judgment quality, GPT conducts pairwise comparisons of opinions written by the same judge before and after the intervention, assessing their clarity, presentation of relevant facts, legal reasoning, and overall legal quality. To assess whether this GPT-based measure captures meaningful differences in legal quality, we perform a blinded expert evaluation of the comparisons with two experienced Pakistani lawyers. As shown in Appendix Table A5, the pooled lawyer assessment agrees with GPT on the superior opinion 70.6 percent of the time. This is close to the agreement rate between the two lawyers themselves, 73.0 percent.[10] These results suggest that the GPT-based quality measure captures a meaningful component of expert-assessed relative judgment quality, while leaving scope for disagreement in an inherently subjective task.

Summary Statistics. The summary statistics, reported in Appendix Table A3, situate the intervention in a large, working trial-court system where prior exposure to LLMs was limited. Participating judges are mid-career judicial officers: the average judge is 42.9 years old, has 11.7 years of judicial experience, decides 122.4 cases per month, and manages 137 pending cases. Female judges account for approximately 23 percent of the sample. Judges report spending 16.0 hours per week on legal research and judgment writing and 11.3 hours on administrative work. Before the intervention, only about 25 percent reported familiarity with large language models, with greater familiarity with conventional digital tools. At the district level, courts resolved an average of approximately 29,200 cases per year.

Platform Records and Chat Logs. We use *JudgeGPT* platform records to measure judges' use of the tool. The records contain login timestamps, IP addresses, and chat histories, from which we construct measures of take-up and engagement, including whether a judge logged in, the number of logins, the number of prompts submitted, and the persistence of use over time.

We use the chat logs to measure how judges use AI. First, we classify *tasks* as what judges asked the tool to do, including text improvement, text generation, legal research, summarization, translation, decision support, and critique of legal reasoning. Second, *intention* records the apparent role assigned to the tool in the judicial workflow, distinguishing requests for information or feedback from prompts seeking assistance with drafting,

[10] Appendix Table A5 Panel C shows that agreement with GPT rises to 75.9 percent among pairs for which both lawyers independently identify the same superior opinion. Panel D reports that GPT and Gemini agree on the direction of quality differences in 70.0 percent of non-tied comparisons.

reasoning, or case assessment. The latter includes *Substantive AI Delegation*, which captures requests that move beyond information retrieval or language support toward more substantive elements of judicial work. The chat logs allow us to examine whether targeted training shifted use toward different tasks, including requests closer to the substantive core of adjudication. The full classification prompts are reported in Appendix Section C2, and Appendix Table A6 provides illustrative prompts for each category.

In the main results, we report take-up, engagement, and treatment-induced shifts in the distribution of tasks assigned to *JudgeGPT*. In the appendix, we report the intention-based classifications as additional evidence on how judges incorporated the tool into their workflow.

3.5 Balance and Attrition

Now we assess whether randomized assignment generated comparable treatment and control groups. Figure 4 reports standardized differences in baseline characteristics between judges assigned to either *JudgeGPT* access arm — targeted training with *JudgeGPT* access or generic training with *JudgeGPT* access — and judges assigned to the control group, which received generic training without *JudgeGPT* access. The figure presents 95% confidence intervals for differences in demographics, judicial experience, baseline workload and productivity, and pre-treatment support for generative AI and expectations about its productivity effects. The estimated differences are small and generally statistically indistinguishable from zero, consistent with the balance expected under judge-level randomization.

Appendix Table A7 reports the corresponding pairwise comparisons separately for each access arm relative to the control group and yields similar conclusions. Because the administrative court-record outcomes analyzed later in the paper are available only at the district-court level, Appendix Figure A4 additionally assesses balance after aggregating randomized exposure to the district level. The district-level characteristics are also balanced at baseline.

Attrition also appears unlikely to account for the results. Appendix Table A8 reports the attrition bounds of Lee (2009) and Semenova (2025). For the targeted-training arm, both the conventional Lee bounds and the more demanding generalized bounds exclude zero across all three outcomes. [11]

[11] 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.

Last, the results do not appear to be driven by any particular wave of Treatment A. Assignment to the three experimental arms was randomized ex ante across 1,559 judges and was blinded to the individual judges. The sequencing of training cohorts was likewise determined at the time of randomization together with the Federal Judicial Academy, with sessions staggered to minimize disruption to routine judicial work across the country. Groups 1 and 2 received training in March and September 2024, respectively, while Groups 3A, 3B, and 3C were trained together in February 2025. Appendix Table A9 re-estimates the main survey outcomes separately for each targeted-training cohort relative to the same generic-training and control groups. Appendix Table A10 reports the corresponding exercise for other outcomes, including tool familiarity and the available district-level measures of case resolution and appeals. Survey-based outcomes can be examined for each cohort, but the district-level case-resolution and appeal outcomes cannot: the administrative data end in 2024, so only Group 1 was exposed early enough to contribute post-treatment variation. Group 3A, thus, provides the cleanest contemporaneous three-arm comparison, whereas estimates for Groups 1 and 2 may also reflect differences in cohort timing and length of exposure. The estimated effects on the main survey outcomes are positive across all three targeted-training cohorts. Group-by-group balance tests likewise show no systematic differences in pre-treatment characteristics across experimental arms, supporting the interpretation that these patterns do not reflect differential sorting across training waves.

3.6 Estimation Approach

We use two empirical designs, dictated by the level at which outcomes are observed. For survey outcomes, we can exploit individual-level random assignment of judges to treatment arms. For administrative outcomes, which are observed at the district-year level, we aggregate judge-level assignment into district-level treatment intensity and compare changes in court outcomes across districts with higher versus lower exposure to treated judges.

Judge-Level Survey Outcomes. We begin with outcomes observed at the judge level. These outcomes come from the follow-up survey and capture judges' attitudes toward AI and familiarity with digital tools. We estimate the effects of *JudgeGPT* access and training using randomized assignment. Because assignment does not mechanically imply use, we

report both intent-to-treat (ITT) effects and treatment-on-the-treated effects using instrumental variables (2SLS).

Our baseline ITT specification is:

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

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 generic 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:

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

where the endogenous regressors are instrumented with their corresponding assignment indicators. 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 baseline controls, $\mathbf{X}_i$.

Court-Level Outcomes. We next turn to administrative outcomes observed at the district-year level. To link judge-level randomization to these court-level outcomes, we construct a district-level measure of treatment intensity: the share of 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 justice department (LJCP) statistics we observe are reported at year-end 2024. We relate treatment inten-

sity to district-level outcomes measured in December 2024, capturing effects roughly nine months after treatment, relative to the previous year's case closures (and appeals). Specifically, we estimate the following difference-in-differences specification:

$$\begin{aligned} Y_{it} = \beta \times \text{Share Treated Judges}_i \times \mathbb{1}[\text{Year} = 2024]_t + \\ \gamma_i + \theta_t + \Gamma' \mathbf{X}_{it} + \varepsilon_{it} \end{aligned} \tag{3}$$

In the case-disposal specifications, Y_{it} is the number of cases resolved in district i and year t. In the appeals specifications, Y_{it} is the number of appeals per 1,000 resolved cases, which scales appeals by the volume of cases disposed in the district.

Share Treated Judges$_i$ is the share of early-treated 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 $\mathbf{X}_{it}$, including measures of judicial capacity, the sitting judges, and pending vacancies. Standard errors are clustered at the district level. Appendix A3 reports alternative specifications that replace treatment shares with treatment-arm count, controlling separately for the total number of judges in each district.

Pre-analysis Plan. The experimental design, treatment arms, primary hypotheses, and planned outcome families were pre-specified in the AEA RCT Registry prior to analysis (RCT ID AEARCTR-0012906). Following the approach in Banerjee et al. (2020), which treats the pre-analysis plan and the implemented paper as related but distinct objects, Appendix B2 reproduces the pre-analysis plan, maps the paper's analyses to the registered design, and documents departures from the original plan. These departures arose primarily because implementation with the judiciary generated administrative outcomes and platform-based measures that were not available when the trial was registered. In addition, some features of the design, such as providing general technology-and-law training to the control group, had to be adapted to the institutional requirements and needs of the judiciary under its technology and law initiative.

4 Main Results

This section presents the main results. Robustness analyses are reported in Section 5.

4.1 JudgeGPT Take-Up and Platform Engagement

We begin by examining whether assignment to *JudgeGPT* changed actual platform use. We estimate these first-stage effects using the judge-level specification in Equation 1, replacing Y_i with platform-recorded measures of use, including logins and prompts. The two assignment coefficients compare judges assigned to targeted training with access and generic training with access to the control group, which received generic training without *JudgeGPT* access. These estimates provide the first stage for the IV analysis in Equation 2. Since control judges did not receive platform access, assignment generates one-sided variation in tool availability, allowing randomized assignment to instrument for realized *JudgeGPT* use in the treatment-on-the-treated specifications.

Table 1 reports the platform-use first stage. Assignment to *JudgeGPT* substantially increases use, especially when access is paired with targeted, tool-specific training. Judges assigned to targeted training log in about 56 more times and submit about 212 more prompts than judges in the control group. Judges assigned to generic training with access also use the platform, but at much lower rates: about 10 additional logins and 25 additional prompts. The difference between the targeted-training and general-training access arms is statistically significant in all columns ($p < 0.01$). The IV estimates are similar in scale, with targeted training generating about 72 additional logins and 270 additional prompts.

Figure 5 shows the corresponding dynamics over time. Judges assigned to targeted training use *JudgeGPT* more from the outset, and the gap relative to generic training widens over the post-assignment period. By week 40, targeted-training judges reach nearly 60 cumulative logins and more than 200 cumulative prompts per judge, compared with about 20 logins and fewer than 50 prompts in the general-training arm. Both series are concave, indicating heavier use soon after assignment followed by slower growth. Usage, however, does not flatten completely in either access arm, suggesting continued engagement after the initial training period, especially when access was paired with targeted training.

We next examine whether access and training also changed broader self-reported use of generative AI. Table 2 reports effects on whether judges state that they are frequent users of generative AI. Unlike platform logins, this outcome is observed for all survey respondents, including control judges who did not receive access to *JudgeGPT*, so it captures a broader adoption margin rather than mechanical use of the study platform alone.

Assignment to *JudgeGPT* access with targeted training increases stated frequent use by about 14 percentage points in the ITT specifications, about a 25 percent increase over the control mean of 54%. The corresponding 2SLS estimates are similar in magnitude, while the estimates for generic training with access are smaller and statistically insignificant.

One concern is that the pattern partly reflects a broader increase in digital familiarity, rather than adoption of generative AI specifically – for example, judges exposed to AI may also become more comfortable using office software or search engines. Table 3 provides little support for this interpretation. The treatment has small and statistically insignificant effects on familiarity with Microsoft Office, Google or Bing search, and WhatsApp, and the targeted-training and general-training access arms are not statistically distinguishable on these outcomes. The adoption response therefore appears specific to generative AI rather than to common digital tools more generally.

4.2 Effect on Attitudes

The next question is whether the treatment shifted judges' attitudes about AI in judicial work. 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 support for AI by about 0.20 points, with similar magnitudes in the 2SLS estimates. Relative to the control mean of 3.350, this corresponds to roughly a 6 percent increase in support for using generative AI in judicial work. This corresponds to an increase of approximately 0.38 standard deviations.

Table 5 turns to judges' *Support for AI Training*, measured by whether they believe such training can improve their productivity. Targeted training raises agreement by 0.14–0.15 points, or about 4 percent relative to the control mean of 3.50, with similar estimates in the 2SLS specifications. The corresponding estimates for generic training are smaller and not statistically significant.[12]

[12]For ease of interpretation, we also report standardized effect sizes for both *Support for Using GPT* and *Support for AI Training*, rescaling each outcome to have mean zero and standard deviation one. These correspond to effects of approximately 0.34 standard deviations for support for using GPT and 0.25 standard deviations for support for AI-for-judges training. Appendix Figure A3 summarizes averages of outcomes across treatment arms. Judges assigned to *JudgeGPT*, report higher agreement that GPT can improve productivity, greater support for using GPT in judicial work, and more frequent use of generative AI. These

4.3 Effect on Case Resolution and Appeals

The survey and platform-usage results speak to use, support, and perceived usefulness, but they cannot show whether the intervention changed the observable work of the courts. Addressing this question requires a different data source and estimation strategy: because case-disposal records are aggregated at the district rather than the judge level, we shift from the judge-level specifications above to the district-level difference-in-differences design in equation (3), using the share of early-2024-class judges assigned to targeted training as a measure of district-level treatment intensity. To move from stated use to realized court activity, we turn to administrative records compiled by the Justice Department (LJCP), focusing on cases closed and appeals filed per 1,000 resolved cases. Because these records cover district courts across Pakistan, they allow us to examine whether the intervention is reflected in aggregate court activity at scale.

These administrative outcomes capture two distinct dimensions of court performance. Cases closed measure output: the volume of cases disposed by district courts. Appeals per 1,000 resolved cases measure downstream contestation: whether resolved cases are more likely to be challenged within the trial-court system.[13] Since the records are available at the district-court level rather than the judge level, we aggregate treatment exposure to the district level. The analysis asks whether districts with greater exposure to targeted *JudgeGPT* training closed more cases after the intervention, and whether this increase in disposal was accompanied by higher appeal rates.[14]

Targeted *JudgeGPT* exposure increases case resolution. Table 6 estimates Equation 3 using the number of cases closed as the dependent variable. Across specifications, districts with greater exposure resolve significantly more cases. Moving from zero to the median exposure level (0.20) is associated with a 6.3 percent increase in resolved cases.

To put the magnitude in more concrete terms, moving from a district with no exposure, such as Sibi, Rajanpur, or Tank, to the 25th percentile of exposure, around 6.7 percent treated judges, implies roughly 616 additional resolved cases. Moving to the median exposure district, around 20 percent treated judges, as in district courts of Buner, Jacob-

averages mirror the regression results.

[13] We cannot yet examine more upstream review in High Court or case reversals because they are observed only after appeals have moved through the court system. Given delays in appellate review, it may take several years before reversal outcomes are meaningfully observed.

[14] The district-level administrative analysis focuses on the targeted-training arm because of the structure of the available administrative data. The LJCP data are published annually with a lag, so the current post-treatment window contains only one post-intervention year.

abad, or Kohistan, implies roughly 1,848 additional resolved cases. The administrative data thus indicate that the intervention affected not only stated use and perceived usefulness, but also realized case resolution in district courts. Appendix Table A11 reports the same outcomes in standard-deviation units. In the saturated specification, the coefficient for cases resolved is 0.171, so moving from zero to the median exposure level, 0.20, corresponds to a 0.034 standard-deviation increase in cases resolved.

We next examine whether the additional cases closed generated greater appellate review. A natural concern is that faster resolution could come from cases resolved on weaker grounds, which litigants would then challenge more often. Table 7 therefore reports effects on appeals per 1,000 resolved cases. This normalization matters because exposure itself increases the number of cases resolved; the outcome measures appeal intensity among closed cases rather than the total number of appeals. The coefficient is negative in each specification. The estimates imply that moving from zero to the median exposure level (0.20) lowers the appeal rate by about 0.15 appeals per 1,000 resolved cases, less than 1 percent of the mean appeal rate. In standardized units, the corresponding coefficient indicates a 0.01 standard-deviation decline for the same increase in exposure. The estimates should therefore be read together: moving from zero to the median exposure level raises the number of cases resolved by about 6 percent, while the appeal rate per resolved case changes by less than 1 percent and, if anything, falls. The expansion in cases resolved is therefore not accompanied by a higher rate of appellate challenge. Targeted *JudgeGPT* training appears to increase case resolution without generating disproportionate appeals among the additional cases resolved.

These estimates also imply a simple cost benchmark. Because *JudgeGPT*-equipped judges resolve more cases per month, producing the same output without the tool would require additional judges which we can value using the annual salary of an average district judge. Valuing the additional case-resolution capacity at the salary cost of equivalent judicial capacity, our estimates imply approximately $41.9 in additional capacity per dollar of recurring *JudgeGPT* expenditure. Under a substantially higher-cost scenario, in which per-user costs are roughly tripled, this ratio remains about $11.8 per dollar. These calculations do not incorporate any potential gains associated with lower appellate contestation and are intended as a cost-effectiveness benchmark rather than a broader welfare measure. Appendix C4 reports the full calculation and sensitivity to alternative cost assumptions.

Appendix Figure A5 provides a distributional analogue to the regression estimates

in Table 6 and Table 7. The figure plots kernel densities of the two administrative outcomes after partialling out district and year fixed effects. In both panels, a value of zero means that a district is exactly at its own historical average, while negative values indicate below-average outcomes. High- and low-treatment districts are defined as above and below the median share of judges assigned to *JudgeGPT* with targeted training.

Panel A shows the distribution of residualized cases resolved. Low-treatment districts cluster tightly around −10,000, with most mass between −20,000 and 0. High-treatment districts shift rightward, with substantial mass between 0 and 20,000 and a longer right tail. This pattern is consistent with the case-resolution estimates: high-treatment districts resolve more cases relative to their own historical baseline, and the shift does not appear to be driven by a small number of outliers.

Panel B plots the corresponding distribution for appeals per 1,000 resolved cases. Here the high-treatment distribution is shifted to the left of the low-treatment distribution and has more mass below zero. This mirrors the regression estimates in Table 7: appeal intensity does not rise in high-treatment districts and, if anything, is modestly lower. The two panels show that greater exposure to targeted *JudgeGPT* training is associated with a rightward shift in case resolution, but not with a corresponding increase in appeal intensity.

Alternative Explanations. One explanation for the increase in cases resolved is that treated judges worked more intensely, rather than becoming more productive in how they handled existing tasks. Appendix Table A12 provides little support for this interpretation across four reported dimensions. First, treated judges do not report spending meaningfully more time on legal research: the estimate for *JudgeGPT* access with associated training is small relative to the mean of about 16 hours per week and statistically indistinguishable from zero. Second, the gains do not appear to come from additional administrative work. The estimate for administrative hours is also small relative to the mean of about 11 hours per week and imprecisely estimated. Third, treated judges do not report higher workload. The point estimate for the targeted-training arm is essentially zero, despite a mean workload rating of about 7 on a 1–10 scale. Finally, reported work-life balance does not deteriorate – the estimate is again close to zero.

These patterns make it difficult to explain the case-resolution gains as judges simply working longer hours, taking on heavier workloads, or shifting substantially more time into research or administration. This interpretation also fits the results on appeals: if

higher disposal reflected rushed or lower-quality resolution, one might expect more appellate challenges, whereas appeal intensity, if anything, decreases. The evidence therefore does not suggest that the increase in case resolution was driven by greater reported work effort or workload. Instead, treated judges appear to close more cases without detectable increases in reported research time, administrative time, workload, or deterioration in work-life balance.

Appeals capture contestation within the legal system, but say little about the quality of judicial writing. We next turn to written opinions, which allow us to assess whether *JudgeGPT* changed drafting behavior and observable features of decisions.

4.4 Effects on Opinion Texts and Judge Bias

Because lower-court cases in Pakistan often take years to reach resolution on appeal (Law and Justice Commission of Pakistan, 2024), text-based outcomes provide an earlier window into potential changes in written decisions. We examine both direct evidence of AI-assisted drafting and downstream features of opinion quality and bias using the available judgment texts. This analysis is necessarily more limited than our main experimental results: the texts come from course submissions, submission rates vary across treatment arms, and the resulting sample is small relative to the full experiment. Due to limited statistical power, these outcomes are also not included in the multiple-testing corrections reported in Table A15. We therefore interpret these estimates cautiously and as exploratory evidence, focusing primarily on whether the intervention generated detectable deterioration in written decisions rather than on establishing improvements in quality. The results, reported in Table 8 and Table A13, cover the following outcomes.

Our first measure is the fraction of each judgment classified as AI-generated using Pangram Labs (Emi and Spero, 2024). This measure provides a useful validation check: if judges assigned to *JudgeGPT* used the tool in drafting, treated opinions should contain a larger AI-generated share. Table 8 Column 1 shows that, as expected, post-treatment opinions by treated judges contain more AI-classified text in both *JudgeGPT* arms. That indicates judges took up the tools and used LLMs in drafting opinions. The targeted-training point estimate is positive, but the Anderson–Rubin test does not reject a zero effect ($p = 0.16$).

We then examine textual features of written opinions, including judgment length, the number of legal references, and a Gunning Fog readability index. Columns 2–4 show lit-

tle evidence of deterioration in the surface features of written opinions. Judgment length is statistically unchanged, the number of references does not fall significantly, and readability is essentially unchanged.

The next opinion measure is overall decision quality, measured using blinded pairwise comparisons of pre- and post-treatment opinions with LLM evaluations. For each judge, we compare pre-treatment and post-treatment opinions and ask the LLM evaluator which opinion is superior on a -2 to $+2$ scale. This relative-comparison approach is well suited to assessing document quality because it avoids requiring the model to assign stand-alone absolute scores to legal texts (Ash and Hansen, 2023). As described in Section 3.4 and Appendix Section C1, the measure is validated against expert annotations by two licensed Pakistani attorneys, who independently labeled the same sample of 90 pairs. Column 5 shows a targeted-training estimate of 0.169 (control mean: 0.42): in the treatment group, the post-treatment opinions are preferred in about 58 percent of pairwise comparisons, versus 42 percent in the control group. The estimate for the placebo-training arm is positive but noisy. Overall, the quality evidence points more toward improvement than deterioration. The estimates are not uniformly precise, however, and should therefore be interpreted as suggestive.

Another relevant dimension of judge writing is potential bias or differential treatment according to identity. Given the discussion of bias potentially induced by large language models (e.g. Vicente and Matute, 2023), we test for an effect of AI usage on bias in opinions. Because LLMs are relatively socially progressive, we examine whether *JudgeGPT* changed the language judges used when discussing gender or religion. Using an LLM-based approach, we score each opinion on a 0–3 scale as either anti-women or pro-Islam based on the language of the opinion itself, adjusting for the direction of the decision (see Appendix Section C3). Figure A6 reports the results, showing null estimates for the effects of training. Overall, we find little evidence that access to *JudgeGPT* increased biased language in judicial opinions. These results do not rule out all forms of decision bias, but they make it less likely that the intervention generated large bias effects that are visible in the written record.

4.5 How Judges Use AI

The evidence so far suggests that *JudgeGPT* increased favorable beliefs about AI and raised case resolution, with no evidence of higher reported work burden or lower quality

in written opinions. We now use platform logs to examine how judges actually engaged with the tool. These data are useful because they capture revealed use of *JudgeGPT* and may help discipline possible interpretations of the increase in case resolution and lower appeal rate.

As detailed in Appendix Section C2, we measure two sets of outcomes from the AI chat logs. First, we analyze the *tasks* undertaken by the judges – e.g. legal research, text improvement, translation. Second, we analyze *intention*, the role of the task in the judicial workflow – e.g. seeking information, asking the AI to execute a judge-determined decision, or asking the AI to take a substantial role in making the decision.

First, we ask, what do judges do with AI? Figure 6 Panel A shows the overall share of tasks. The most common tasks are legal research, text improvement, and text generation. There is relatively low usage on self-critique, translation, and personal use.

Second, we ask how training influenced the AI tasks. Figure 6 Panel B shows that targeted AI training (Treatment A relative to Treatment B) increased the usage of AI for text improvement and summarization. This is somewhat intuitive and encouraging, as the coursework specifically identified those tasks as high-value, low-risk uses of LLMs. On the other hand, targeted training reduced AI usage on open question answering. That is also intuitive, since the coursework specifically warned that open question answering – i.e., generic questions untied to specific laws or precedents – is a risky use of LLMs that is especially prone to hallucination. These results show that judges responded to the training and engaged the LLMs in higher-value work tasks.

Using the intention classification, we can get closer to a more substantive question: whether judges used *JudgeGPT* mainly to obtain information and support their own writing and decisions, or whether instead the judges relied strongly on the AI and turned over high-agency writing and decision tasks. Figure A7 Panel A suggests that most prompts are information-seeking: nearly 60 percent ask the model to explain law, procedure, or legal concepts. The figure summarizes the composition of prompts submitted on the platform and therefore describes realized use rather than treatment effects on particular types of use. Because judges who submit more prompts contribute more observations, the prompt shares also place greater weight on more intensive users. Shares in Figure 6 and Figure A7 are computed over prompts submitted by judges who used the platform, so they describe how the tool was used rather than the effect of assignment on use. Because targeted training raises both the probability of use and the number of prompts, the two arms' prompt pools are composed of different sets of judges. About one-fifth involve

judge-specified tasks, where the judge appears to have already fixed the relevant legal conclusion or output and asks the model to produce, edit, or format the text. Another one-fifth of prompts involve substantive AI delegation, where judges ask the model to do high-agency tasks, including assessing the right decision, producing reasoning, or writing opinions or orders without substantial judge input. Panel B suggests that targeted training increased the share of prompts in which judges had already specified the substantive task and asked the AI to implement it. Thus, the results suggest that targeted training encouraged judges to use *JudgeGPT* for more bounded tasks where the user maintained agency.

5 Robustness Analysis

We complement the main analysis with a set of robustness checks for spillovers, balance, attrition, multiple testing, randomization inference, bootstrap inference, leave-one-out regional exclusions, and alternative specifications. We also discuss concerns related to experimenter demand and external validity. These exercises generally support the interpretation of the main results. Ethical considerations are discussed in Appendix B1. Appendix B2 reproduces the pre-analysis plan, maps the analyses to the registered design, and documents departures.

Spillovers. To limit unauthorized access, JudgeGPT was provided through password-protected individual accounts, and judges were informed that sharing credentials would lead to disqualification. Nevertheless, we assess potential spillovers to control judges using two district-level exposure measures. Table A14 interacts assignment to the no-access group with average platform usage in the district, measured by unique API calls per judge. If control judges benefit from shared outputs or informal diffusion, their outcomes should improve with local platform usage. The estimates are close to zero across outcomes.

Minimum Detectable Effects. We next assess the power of the access-only comparison. Appendix Figure A8 reports minimum detectable effects for judges assigned *JudgeGPT* access with general technology-and-law training relative to judges assigned generic training only. At the realized sample size, the design can detect effects of about 0.17 for AI usage, 0.21 for AI support, and 0.2 for course support. These detectable effects are small enough

to make the access-only null results informative about meaningful changes in adoption and attitudes. The pattern is thus more consistent with targeted training being necessary for sustained adoption than with access alone generating effects that the study is simply unable to detect.

Multiple Hypothesis Testing Adjustments. Because we examine related outcomes, we also report several adjustments for multiple testing. Following standard practice, we group outcomes into conceptually related families and report adjusted p-values alongside the conventional ones. The pattern in Table A15 is clear. The targeted JudgeGPT arm remains robust across most corrections: effects on frequent generative-AI use and support for JudgeGPT remain statistically significant under all reported adjustments, while the AI-course outcome remains significant under most procedures, though it is weaker under the Romano-Wolf correction. The district-level result on total disposals is especially stable, with adjusted p-values close to zero across methods. In contrast, the decline in appeals per 1,000 resolved cases is more suggestive: it is marginally significant conventionally, and survives the sharpened q-value adjustment, but not the most conservative family-wise corrections. The placebo-training arm shows a more limited pattern, with little evidence of increased frequent use or AI-course support, but a positive effect on support for JudgeGPT that survives all reported corrections. These adjustments reinforce the main interpretation: the most stable effects are in the targeted-training arm, with clear gains in attitudes toward JudgeGPT and total cases resolved.

Randomization Inference. With over 1,500 judges, this study is among the largest randomized evaluations in the public sector. However, effect sizes on quality outcomes are modest and precision varies. To strengthen inference, we also assess inference using permutation tests. For each outcome in Appendix Figure A9, we repeatedly reshuffle treatment assignment 1,000 times and compare the observed coefficient to the resulting distribution of placebo estimates. The red vertical line marks the estimate from the actual assignment. Across the main survey outcomes, tool-familiarity outcome, cases resolved, and appeals, the observed estimates lie in the tails of the permutation distributions. This indicates that the estimated effects are unlikely to be generated by chance assignment alone. The exercise is especially reassuring for the attitude and usage outcomes and for cases resolved, where the observed effects are far from the mass of placebo estimates. For appeals, the observed coefficient lies in the lower tail, consistent with the negative

estimate reported in the main tables. These permutation tests support the main interpretation that the targeted JudgeGPT intervention shifted AI-related attitudes and court output, rather than reflecting idiosyncratic features of the realized randomization.

Wild Bootstrap Inference. We complement conventional clustered inference with wild-cluster bootstrap confidence curves, following MacKinnon and Webb (2017). This check is useful because some outcomes are estimated with clustered or district-level variation, where finite-sample inference may be sensitive to asymptotic approximations. Figure A10 plots bootstrap p-value curves for the main outcomes, with the horizontal line marking the 5 percent significance threshold and the vertical line marking the zero-effect benchmark. The bootstrap evidence supports the main survey results: the null of no effect is rejected for support for generative AI, stated frequent use, and beliefs that the AI course can improve productivity. The tools-familiarity outcome remains centered near zero, consistent with the interpretation that the intervention shifted AI-specific attitudes and use rather than general familiarity with digital tools. The court-output result is also robust: cases resolved remains positive under the bootstrap procedure. The appeals estimate is weaker. Overall, the bootstrap exercise supports the main interpretation while preserving the distinction between the strongest results, on AI attitudes, use, and case resolution, and the more suggestive evidence on appeals.

Excluding Regions. To assess whether the results are driven by any particular region, we re-estimate the main specifications after excluding one administrative region at a time. Figure A11 reports the resulting leave-one-region-out estimates. The results are stable across these exclusions. The estimates for support for generative AI, stated frequent use, and beliefs that the AI course can improve productivity remain positive and close to the baseline throughout. In contrast, the tools-familiarity outcome remains centered near zero, as in the main results, suggesting that the intervention affected AI-specific attitudes and use rather than general familiarity with digital tools. The court-output estimates show the same pattern. Total cases resolved remains positive across leave-one-region-out samples, while the appeals estimate is generally negative but less precise. This exercise suggests that the main findings are not driven by any single administrative region or by the inclusion of one unusually influential local court system.

Additional Attrition Tests. Building on the attrition analysis above, we relax the unconditional monotonicity assumption underlying the conventional (Lee, 2009) bounds by allowing selection to vary with pre-treatment covariates (Semenova, 2025). Table A8 reports inverse-probability weighted ITT estimates, conventional Lee bounds, and generalized Lee bounds. The targeted-training results remain stable. For all three outcomes, the IPW estimates are close to the baseline estimates, and both the conventional Lee bounds and generalized Lee bounds remain strictly positive. The evidence is weaker for the general-training arm. The IPW estimate remains positive for support for GPT use, but the most demanding generalized Lee bounds include zero for all three outcomes. Thus, the targeted-training results remain robust to a more flexible treatment-selection relationship, while the placebo-training results do not.

Heterogeneity by Province. Table B4 reports treatment effects separately by province as suggested in the pre-analysis plan. The main pattern is that JudgeGPT access with targeted training increases actual platform engagement across all provincial subsamples. The login coefficients are positive in Punjab, Sindh, Khyber Pakhtunkhwa, Balochistan, and the residual other-provinces category, ranging from about 67 additional logins in Punjab to about 167 in Balochistan. Effects on attitudes and stated frequent use are generally positive across provinces. The AI-for-judges course outcome is generally positive but imprecise. The general-training arm produces some positive attitudinal effects, especially in Sindh and Balochistan, but its effects on platform engagement are consistently smaller than those in the targeted-training arm. These estimates suggest that the core usage result is not concentrated in a single province.

District-Courts Event Study. Appendix Figure A12 provides a graphical complement to the table estimates by plotting coefficients from an event-study specification, normalized to 2023 as the omitted year for both cases resolved and appeals. We view this exercise mainly as a robustness check, rather than as the primary basis for identification. Our preferred design rests on the randomized assignment of judges and the associated balance tests. The event-study figure is therefore included to show whether the estimated effects on resolved cases are also visible under an alternative implementation of the same empirical idea. Across both approaches, the pre-period estimates are small. The 2024 coefficient for cases resolved is positive and statistically significant. The coefficient for appeals is also directionally consistent with the main estimates, though less precisely estimated.

Experimenter Demand. We also consider whether the results could reflect experimenter demand. This concern is most relevant for survey outcomes, where judges may infer that researchers or the Judicial Academy expect favorable views of AI. Several features of the design limit this interpretation. The intervention was embedded in an official judicial-training program rather than presented primarily as a researcher-led evaluation of judge performance. Judges were not told the specific outcomes used in the analysis, including the administrative case-resolution outcomes, or text-based measures of judicial writing. More importantly, the effects are concentrated on AI-specific outcomes. Judges report greater support for JudgeGPT and more frequent use of generative AI, but we do not observe comparable treatment-induced shifts in familiarity with unrelated digital tools such as Microsoft Office, search engines, or messaging applications. This pattern is difficult to reconcile with a broad demand-response story in which treated judges simply report more favorable views of technology in general. The administrative and platform outcomes further reduce this concern: login and prompt records are measured passively, and case-resolution outcomes come from administrative court data rather than from judge self-reports. While these checks cannot rule out demand effects entirely, they suggest that demand is unlikely to account for the main pattern of results.[15]

External Validity. The study takes place in Pakistan's trial courts, a setting that shares features with many lower-tier judicial systems in South Asia and other developing-country contexts: common-law foundations, hierarchical judicial careers, high caseloads, and limited courtroom technology. The results may not translate directly to high-income judiciaries or to courts with very different staffing, digital infrastructure, or legal workflows. Still, the setting is useful for understanding AI adoption in constrained public-sector institutions.

Following the SANS framework of List (2020), which emphasizes selection, attrition, naturalness, and scalability, several features of the setting speak to external relevance. The enrolled sample includes nearly half of Pakistan's sitting trial-court judges or about 70% of district courts, reducing concerns about unusually narrow selection. The intervention was delivered through official Judicial Academy channels and incorporated into judges' regular work rather than an artificial experimental task. Finally, the tool and training relied on existing judicial-training infrastructure, making deployment at similar scale

[15]Recent survey of evidence also suggests that, in most experiments, experimenter demand effects are unlikely to be large enough to overturn qualitative conclusions (de Quidt et al., 2026).

plausible in court systems with comparable institutional capacity. Randomization, however, occurred within the enrolled sample, so the experimental estimates pertain directly to participating judges. Because enrollment was voluntary, extrapolation to judges who did not participate requires some caution.

6 Conclusion

This paper studies whether generative AI can raise productivity in a core state organization. In partnership with Pakistan's judiciary, we built *JudgeGPT*, a custom assistant for the country's trial courts. We randomized 1,559 judges into three arms: access to *JudgeGPT* with targeted training for judicial use, access with only generic training on technology and law, and generic training without *JudgeGPT* access. This design separates the effect of access from the effect of targeted instruction in how to use the tool. We combine survey data, administrative court records, judicial opinions, and platform logs to study adoption, case resolution, writing, and how judges use the tool in practice.

The experiment delivers three main conclusions. First, access to *JudgeGPT* increases AI use, but sustained engagement depends strongly on targeted training. Judges who receive tool-specific training use the assistant more intensively and continue using it over time, while use in the generic-training access arm is lower and more front-loaded. Second, administrative records show that districts more exposed to judges assigned to *JudgeGPT* with targeted training resolve more cases. An increase in targeted-exposure intensity from zero to median value raises case resolutions by roughly 6 percent, or about 1,848 additional cases per year at median-district exposure. Appeals per resolved case show no significant increase and, if anything, decline slightly. Third, we find little evidence that these productivity gains come at the cost of weaker written decisions or greater measured bias. Treated judges' opinions contain more text classified as AI-generated, but we do not detect systematic declines in readability, legal referencing, judgment length, or an expert-validated measure of writing quality. Nor do we find evidence of increased gender or pro-Muslim bias in the language of judicial opinions.

The usage data provide a possible interpretation for why the effects are concentrated in the targeted-training arm. Judges appear to use *JudgeGPT* primarily as an aid to legal research and writing rather than as a substitute for judicial decision-making. Use is concentrated in clarifying legal concepts, organizing and improving text, summarizing material, and assisting with drafting. Targeted training not only increases use, but

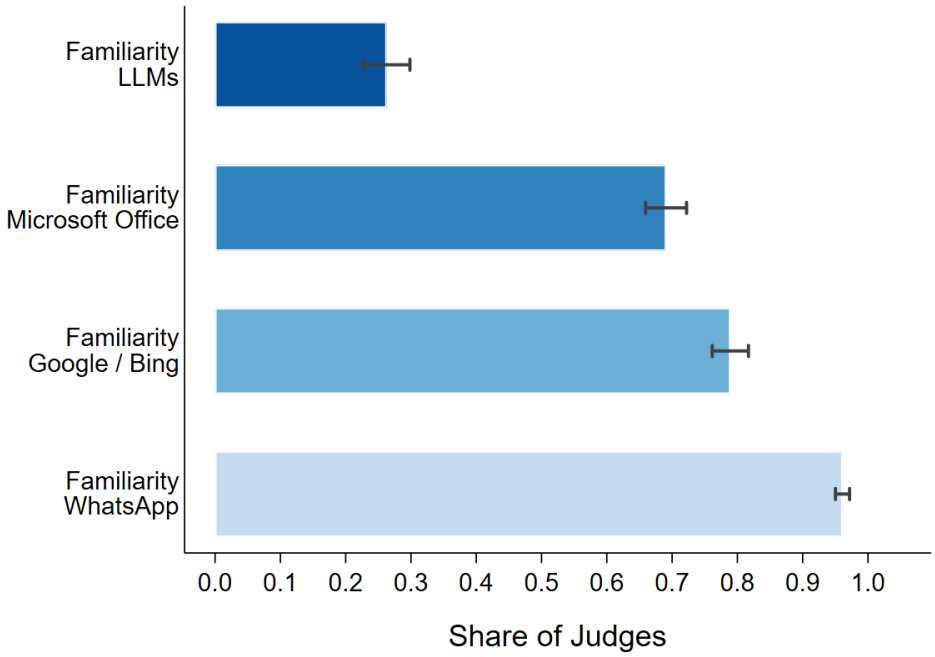
also changes its composition, shifting activity toward text improvement and summarization and away from more open-ended legal queries. These uses overlap with recurring, writing-intensive parts of trial-court work, including organizing facts, structuring arguments, and preparing draft text. In a setting with high caseloads and limited support staff, reductions in the time required for such tasks could contribute to higher case resolution. This interpretation is also consistent with the absence of detectable increases in appeals or deterioration in the writing-quality measures we observe.

The findings speak to the broader question of AI in public-sector organizations. Existing evidence on generative AI has focused largely on private-sector workers and tasks where output is relatively easier to measure. Public institutions differ in important ways. They face weaker market discipline, more rigid procedures, and perhaps a higher cost of error. The judiciary is an especially demanding setting because judges distribute rights, liberty, and property. The fact that a generative AI assistant can raise measured court output in this environment suggests that AI may have practical public-sector applications. But the results also caution against a simple access-based view of technology adoption. The strongest effects appear when access is paired with training that helps judges understand how to use the tool in their regular workflow, which may explain why use is more sustained in the targeted-training arm.

More broadly, the experiment provides a template for evaluating frontier technologies in public institutions. That broader relevance has been recognized beyond the setting we study. Referring directly to our experiment, the UN Secretary General noted in a statement to the General Assembly that "studies of this kind, when administered under judicial supervision, could objectively demonstrate the impact of AI on judges' productivity and the quality of their rulings" (United Nations General Assembly, 2025). Consistent with this objective, we do not study AI as a replacement for judges. We study it as a tool that may change how judges carry out recurring parts of their work, especially when they are trained to use it for tasks such as clarification, summarization, and text organization. For judiciaries facing persistent backlogs, AI is therefore not a panacea. But when a tool is built around relevant legal materials and paired with training that directs use toward appropriate tasks, it can become a practical tool for improving state effectiveness.

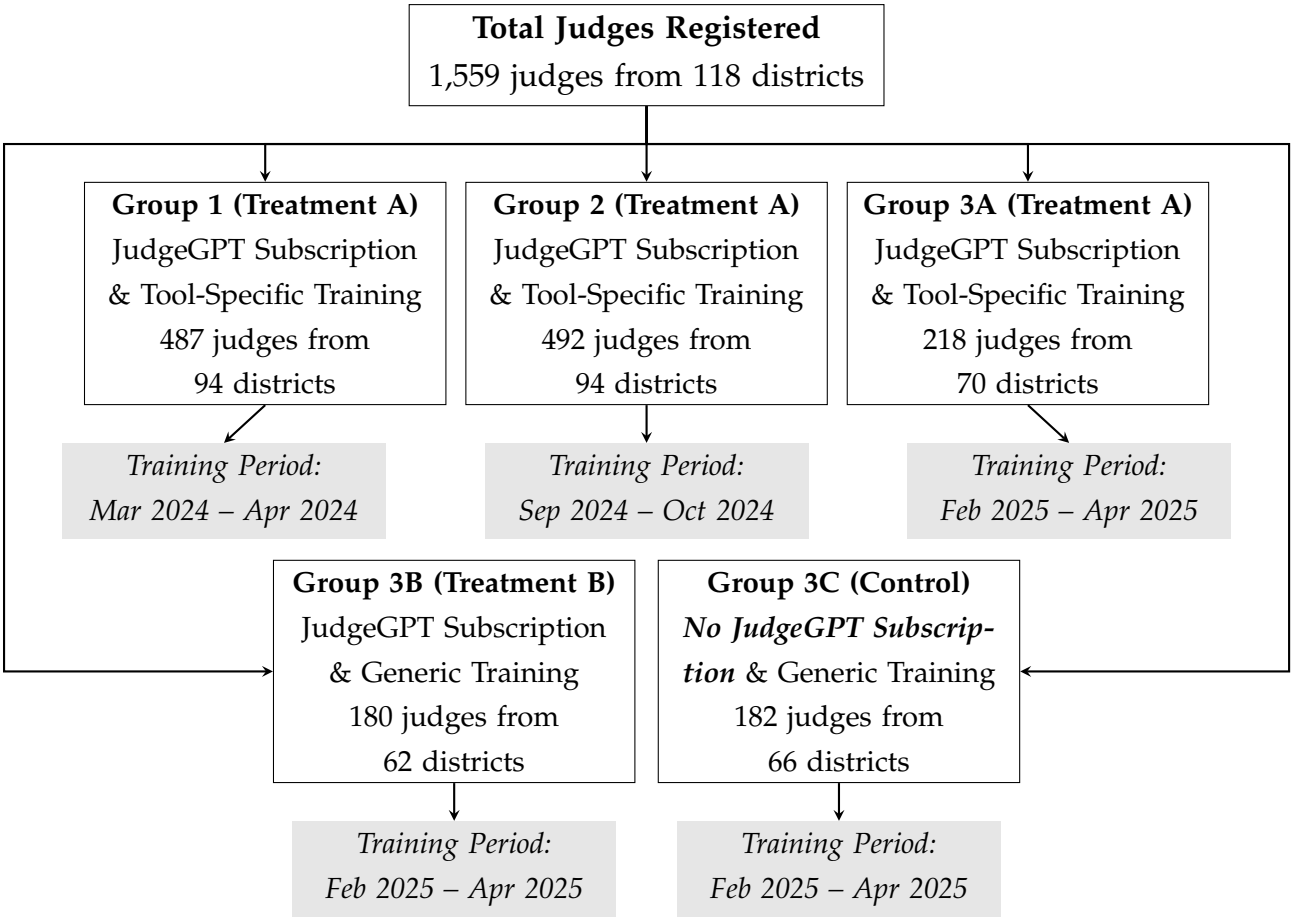
TABLES AND FIGURES

Figure 1: Judges' Baseline Familiarity with Generative AI and Common Digital Tools



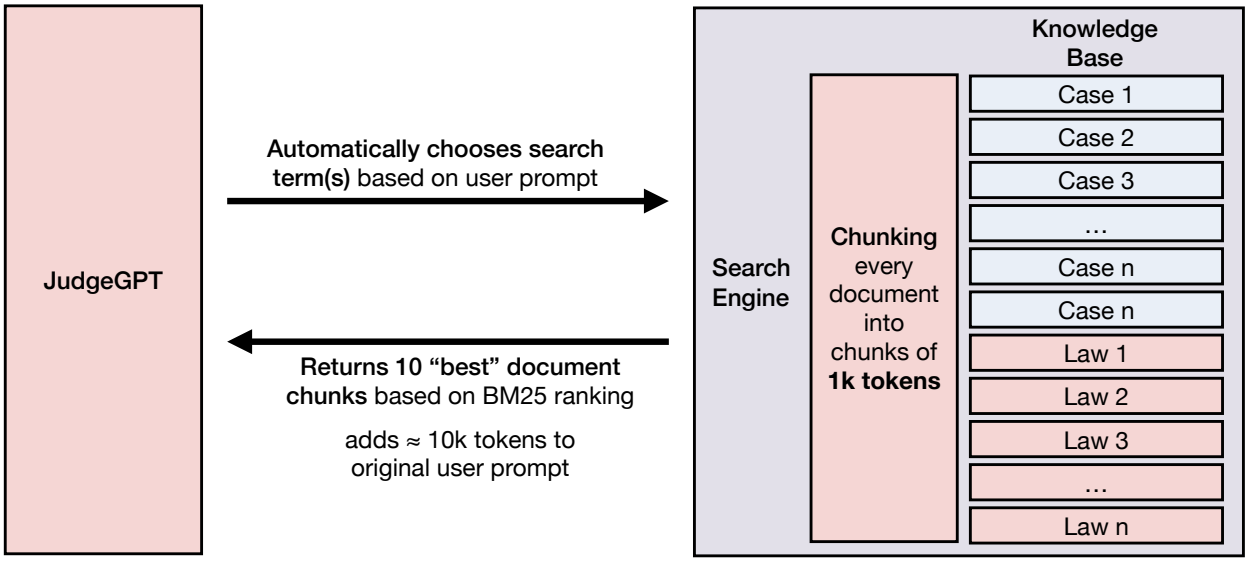
Note: This figure reports the share of judges who reported familiarity with each technology in a pre-treatment baseline survey. "LLMs" equals one if the judge reported familiarity with at least one large language model, including ChatGPT, Gemini, Claude, or another LLM. "Microsoft Office" refers to familiarity with Microsoft Office or similar office software, while "Google/Bing" refers to familiarity with Google Search or Bing. Bars are reported with 95 percent confidence intervals.

Figure 2: Experimental Design and Treatment 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 118 districts were enrolled and assigned across treatment and control groups. The first treatment arm (Groups 1, 2, and 3A; 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 (Group 3B; 180 judges) received access to JudgeGPT alongside a generic technology-and-law course that did not provide instruction on the use of the tool itself. The control group (Group 3C; 182 judges) received the same generic course without access to JudgeGPT.

Figure 3: 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.

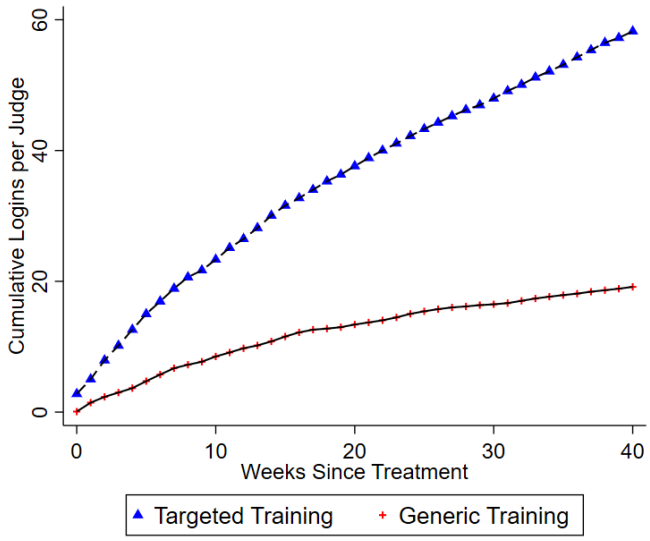
Figure 4: Treatment Balance by Judge Characteristics



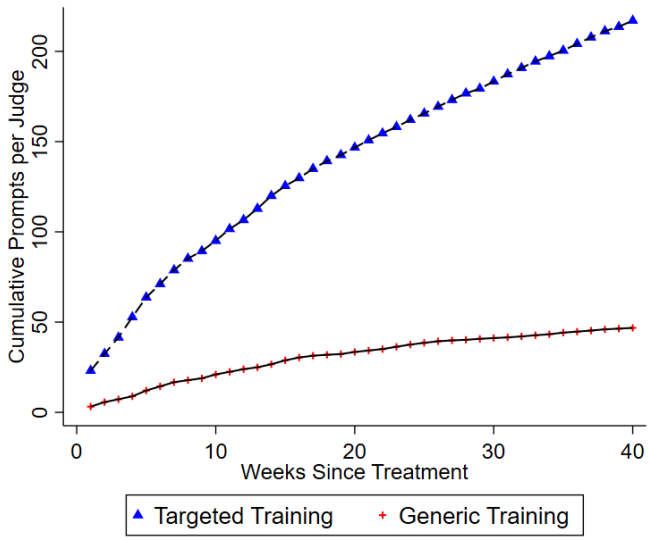
Note: This figure assesses baseline balance between judges assigned to the *JudgeGPT* treatment arms and those assigned to the control group. The treatment group combines Treatment A (*JudgeGPT* with targeted training) and Treatment B (*JudgeGPT* with generic training); the control group received generic 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 5: Dynamics of JudgeGPT Usage, by Training Type

Panel A: Cumulative Logins per Judge



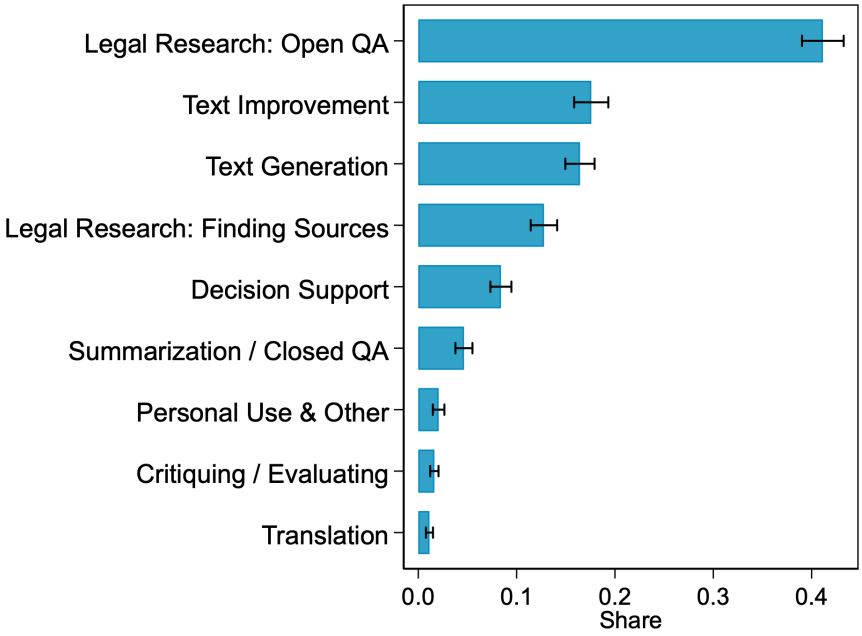
Panel B: Cumulative Prompts per Judge



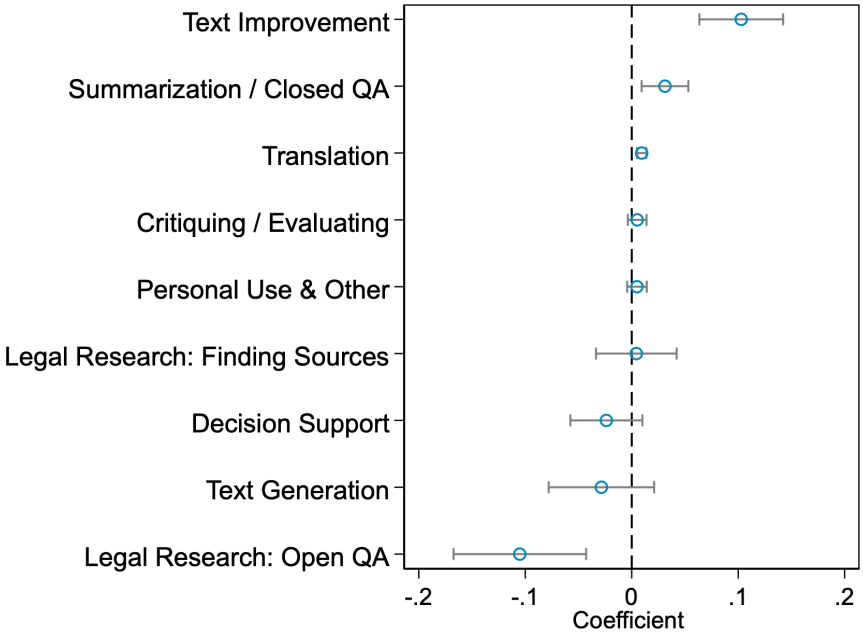
Note: This figure reports cumulative *JudgeGPT* use by weeks since treatment. The x-axis measures weeks elapsed since judges received access to *JudgeGPT*. The y-axis reports cumulative use per judge. Panel A plots cumulative logins per judge, and Panel B plots cumulative prompts per judge. Each panel compares judges assigned to targeted training with judges assigned to generic training; both groups received access to *JudgeGPT*.

Figure 6: JudgeGPT Usage and Training Effects

Panel A: Task Shares



Panel B: Effect of Targeted Training on Tasks



Note: Panel A reports the share of prompts in each category. Panel B reports treatment effects with 95% confidence intervals from judge-level regressions. The omitted category is JudgeGPT Access and Generic Training. All specifications include strata fixed effects and baseline controls.

TABLE 1: JudgeGPT Access and Platform Usage

	# of Logins		# of Prompts	
	ITT	2SLS	ITT	2SLS
	(1)	(2)	(3)	(4)
Assigned JudgeGPT Access and Targeted Training	56.453***		211.590***	
	(2.799)		(13.358)	
Assigned JudgeGPT Access and Generic Training	9.519***		25.331***	
	(2.047)		(9.371)	
JudgeGPT Access and Targeted Training		72.011***		269.879***
		(3.311)		(16.220)
JudgeGPT Access and Generic Training		15.422***		40.801***
		(3.049)		(14.624)
Strata Fixed Effects	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Adj. R^2	.094	.186	.079	.14
Kleibergen-Paap F-statistic		896.935		896.935
p-val. (Targeted Training = Generic Training)	< 0.001	< 0.001	< 0.001	< 0.001
Mean Dep. Var.	46.535	46.535	173.606	173.606
Observations	1559	1559	1559	1559

Note: Robust standard errors are in parentheses. The dependent variables are the number of logins to *JudgeGPT* and the number of prompts (messages) sent by the judge over the post-assignment period. Columns (1) and (3) report intent-to-treat effects of randomized assignment to *JudgeGPT* access and training. *Assigned JudgeGPT Access and Targeted Training* equals one for judges assigned access with tool-specific training, and *Assigned JudgeGPT Access and Generic Training* equals one for judges assigned access with generic technology-and-law training. Columns (2) and (4) report 2SLS estimates in which *JudgeGPT Access and Targeted Training* and *JudgeGPT Access and Generic Training* are instrumented with their corresponding assignment indicators. The control group received placebo training only and no access to *JudgeGPT*. All specifications include randomization-strata fixed effects and controls for 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). Adjusted R^2 and the Kleibergen–Paap first-stage F-statistic are reported. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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	0.135***	0.137***		
	(0.051)	(0.050)		
Assigned JudgeGPT Access and Generic Training	0.065	0.074		
	(0.071)	(0.068)		
JudgeGPT Access and Targeted Training			0.146***	0.148***
			(0.055)	(0.053)
JudgeGPT Access and Generic Training			0.076	0.086
			(0.082)	(0.078)
Strata Fixed Effects	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes
Adj. R^2	.021	.055	.04	.074
Kleibergen-Paap F-statistic			3136.556	2971.268
p-val. (Targeted Training = Generic Training)	0.190	0.219	0.257	0.291
Control Mean	0.544	0.544	0.544	0.544
Observations	1065	1065	1065	1065

Note: Newey-West standard errors are in parentheses. The dependent variable is an indicator equal to one if the judge reports using generative AI at least once or twice a week in the post-intervention survey, and zero otherwise. Assigned JudgeGPT Access and Targeted Training is a binary indicator equal to one for a judge assigned to the targeted course who logged into JudgeGPT at least once, while Assigned JudgeGPT Access and Generic Training is a binary indicator equal to one for a judge logging into GPT with generic course (as in Group 3B). The control group did not receive GPT access while receiving a generic training 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 treatment B (Group 3B) are instrumented with the initial assignment of treatment and treatment B (Group 3B). R^2 and the Kleibergen-Paap F-statistic are reported. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE 3: JudgeGPT Access and Familiarity with Other Tools

	Familiar with Microsoft Office		Familiar with Google / Bing		Familiar with WhatsApp	
	(1) ITT	(2) 2SLS	(3) ITT	(4) 2SLS	(5) ITT	(6) 2SLS
Assigned JudgeGPT Access and Targeted Training	-0.018		0.051		-0.006	
	(0.045)		(0.044)		(0.018)	
Assigned JudgeGPT Access and Generic Training	-0.090		-0.027		0.014	
	(0.065)		(0.064)		(0.022)	
JudgeGPT Access and Targeted Training		-0.019		0.055		-0.006
		(0.048)		(0.047)		(0.019)
JudgeGPT Access and Generic Training		-0.106		-0.032		0.017
		(0.076)		(0.074)		(0.025)
Strata Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Adj. R^2	0.050	0.044	0.025	0.030	0.051	0.050
Kleibergen-Paap F-statistic		2971.268		2971.268		2971.268
p-val. (Targeted Training = Generic Training)	0.167	0.159	0.115	0.128	0.217	0.217
Control Mean	0.718	0.718	0.748	0.748	0.971	0.971
Observations	1065	1065	1065	1065	1065	1065

Note: Robust standard errors are in parentheses. Dependent variables are binary indicators for self-reported familiarity with Microsoft Office (columns 1–2), Google/Bing search (columns 3–4), and WhatsApp (columns 5–6); these are placebo outcomes not targeted by the training. *Assigned JudgeGPT Access and Targeted Training* equals one for judges assigned *JudgeGPT* access with tool-specific training; *Assigned JudgeGPT Access and Generic Training* equals one for judges assigned *JudgeGPT* access with the generic technology-and-law training. Columns (2), (4), and (6) report 2SLS estimates. All treatment indicators are instrumented with their corresponding assignment indicators. All specifications include strata fixed effects and controls for 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). Adjusted R^2 and Kleibergen-Paap first-stage F-statistics are reported. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE 4: JudgeGPT Access and Support of Generative AI Usage

	ITT		2SLS	
	(1)	(2)	(3)	(4)
	Support the Usage of GPT			
Assigned JudgeGPT Access and Targeted Training	0.216***	0.221***		
	(0.072)	(0.071)		
Assigned JudgeGPT Access and Generic Training	0.232**	0.230***		
	(0.091)	(0.089)		
JudgeGPT Access and Targeted Training			0.233***	0.239***
			(0.077)	(0.075)
JudgeGPT Access and Generic Training			0.272**	0.269***
			(0.106)	(0.102)
Strata Fixed Effects	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes
Adj. R^2	.014	.073	.02	.079
Kleibergen-Paap F-statistic			3136.556	2971.268
Control Mean	3.350	3.350	3.350	3.350
p-val. (Targeted Training = Generic Training)	0.797	0.877	0.592	0.667
Observations	1065	1065	1065	1065

Note: Robust standard errors are reported in parentheses. The dependent variable is the response to the survey 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. Columns (1)-(2) report intent-to-treat estimates. *Assigned JudgeGPT Access and Targeted Training* equals one for judges randomly assigned *JudgeGPT* access with tool-specific training, while *Assigned JudgeGPT Access and Generic Training* equals one for judges randomly assigned *JudgeGPT* access with generic technology-and-law training. Columns (3)–(4) report 2SLS estimates. *JudgeGPT Access and Targeted Training* and *JudgeGPT Access and Generic Training* indicate realized treatment in the corresponding arm and are instrumented with their respective random-assignment indicators. All specifications include randomization-strata fixed effects. Columns with controls additionally include age, gender, years of judicial experience, baseline support for AI, measures of baseline workload and productivity (cases decided, cases concurrently managed, and hours spent on administrative work), and baseline perceptions of work–life balance, confidence in public speaking, administrative workload, and the potential usefulness of 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 5: JudgeGPT Access and Attitudes Towards AI and Judicial Productivity

	ITT		2SLS	
	(1)	(2)	(3)	(4)
	GPT Course Can Improve Productivity			
Assigned JudgeGPT Access and Targeted Training	0.144**	0.154**		
	(0.065)	(0.063)		
Assigned JudgeGPT Access and Generic Training	0.071	0.081		
	(0.088)	(0.087)		
JudgeGPT Access and Targeted Training			0.156**	0.166**
			(0.069)	(0.067)
JudgeGPT Access and Generic Training			0.084	0.094
			(0.102)	(0.100)
Strata Fixed Effects	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes
Adj. R^2	.007	.06	.013	.066
Kleibergen-Paap F-statistic			3136.556	2971.268
p-val. (Targeted Training = Generic Training)	0.262	0.264	0.336	0.334
Control Mean	3.495	3.495	3.495	3.495
Observations	1065	1065	1065	1065

Note: Robust standard errors are in parentheses. The dependent variable is the response to the survey question "Do you expect that an AI for Judges Course can improve your productivity?", 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 Generic Training* equals one for judges assigned *JudgeGPT* access with the generic technology-and-law training. Columns (3) and (4) report 2SLS estimates in which *JudgeGPT Access and Targeted Training* and *JudgeGPT Access and Generic 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 6: JudgeGPT Access and Case Resolution

	(1)	(2)	(3)	(4)
	Cases Resolved			
Share JudgeGPT and Targeted Training × Post	8475.311***	8282.470***	9371.122***	9239.049***
	(481.224)	(397.146)	(469.432)	(600.440)
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	0.857	0.874	0.894	0.900
Mean Dep. Var.	29200.72	29200.72	29200.72	29200.72
# 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. *Share JudgeGPT and Targeted Training × Post* equals the number of judges assigned JudgeGPT with targeted training divided by the number of sitting judges in the district court in the baseline denominator year, interacted with a post indicator for 2024. All specifications include district fixed effects; columns (2)–(4) additionally include year fixed effects. Column (3) controls for the number of judges in the district. Column (4) additionally controls for the number of vacant judge positions. The sample includes district-years from 2021 to 2024 with cases resolved as the outcome, district and year fixed effects, and controls. The table reports R^2, the mean of the dependent variable, and the number of clusters and observations. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

TABLE 7: JudgeGPT Access and Appeals

	(1)	(2)	(3)	(4)
	Appeals per 1,000 Resolved Cases			
Share JudgeGPT and Targeted Training × Post	-0.938**	-0.594*	-0.888**	-0.846*
	(0.455)	(0.309)	(0.436)	(0.482)
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	0.336	0.448	0.457	0.461
Mean Dep. Var.	19.352	19.352	19.352	19.352
# Clusters	118	118	118	118
Observations	445	445	445	445

Note: Robust standard errors, clustered by district, are in parentheses. The unit of observation is district-year. The dependent variable is the total appeals per 1,000 cases resolved in a district-year. Total appeals is defined as the sum of civil and criminal appeals. *Share JudgeGPT and Targeted Training × Post* equals the number of judges assigned JudgeGPT with targeted training divided by the number of sitting judges in the district court in the baseline denominator year, interacted with a post indicator for 2024. All specifications include district fixed effects; columns (2)–(4) additionally include year fixed effects. Column (3) controls for the number of judges in the district. Column (4) additionally controls for the number of vacant judge positions. The sample includes district-years from 2021 to 2024 with appeals as the outcome, district and year fixed effects, and controls. The table reports R^2 and the number of clusters and observations. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

TABLE 8: JudgeGPT Access and Text Quality

	2SLS				
	(1)	(2)	(3)	(4)	(5)
	Fraction AI Generated	Length of Judgements	# of References	Gunning Fog Readability	LLM Quality Assessment
JudgeGPT Access and Targeted Training	0.085**	-104.480	-2.924	0.000	0.169*
	(0.042)	(220.903)	(3.052)	(0.043)	(0.088)
JudgeGPT Access and Generic Training	0.283*	427.141	-2.576	0.013	0.399
	(0.163)	(870.547)	(7.040)	(0.095)	(0.339)
Strata Fixed Effects	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes
Adj. R^2	.497	.327	.423	-.011	.014
Kleibergen-Paap F-statistic	8.190	8.026	8.049	7.913	7.913
Anderson-Rubin P-val.	.16	.726	.037	.985	.194
Control Mean	.152	1848.244	7.384	.015	.42
P-value (Access with Training = Access without Training)	.199	.516	.953	.864	.467
Observations	191	259	259	259	259

Note: The table reports 2SLS estimates of the effect of JudgeGPT use on judgment characteristics and text quality. The outcomes are the Pangram-estimated share of AI-generated text, judgment length, the number of statutory and precedential references, standardized Gunning Fog readability, and an LLM-based quality measure. In column (5), GPT-5-mini compares each judge's post-treatment judgments with that judge's pre-treatment judgments and records the share of pairwise comparisons in which the post-treatment judgment is preferred. The two endogenous indicators denote judges who used JudgeGPT after assignment to targeted training or generic training, respectively; both are instrumented by their initial randomized assignments. The omitted group received generic training but no JudgeGPT access. All specifications include the corresponding pre-treatment outcome, baseline judge characteristics, and randomization-strata fixed effects. Newey–West standard errors are reported in parentheses. Adjusted R^2, the Kleibergen–Paap F-statistic, and the Anderson–Rubin p-value for joint nullity of the coefficients are also reported. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

REFERENCES

Abrams, David S., Brian M. Collopy, Aurélie Ouss, Megan Stevenson, and Colin D. Sullivan, "Barriers to Adopting Predictive Algorithms: A Criminal Justice Field Experiment," *American Criminal Law Review*, 2026, *63* (1), 49–87.

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.

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.

_ **,** _ **, and Suresh Naidu**, "Ideas have consequences: The impact of law and economics on american justice," *The Quarterly Journal of Economics*, 2026, *141* (1), 845–887.

_ , **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.

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.

Autor, David, "Applying AI to Rebuild Middle Class Jobs," NBER Working Paper 32140, National Bureau of Economic Research February 2024.

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.

Bono, James and Alec Xu, "Randomized Controlled Trials for Security Copilot for IT Administrators," 2024.

Boussioux, Léonard, Jacqueline N. Lane, Miaomiao Zhang, Vladimir Jacimovic, and Karim R. Lakhani, "The Crowdless Future? Generative AI and Creative Problem-Solving," *Organization Science*, 2024, *35* (5), 1589–1607.

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.

Chemin, Matthieu, Daniel L. Chen, Vincenzo Di Maro, Paul Kimalu, Momanyi Mokaya, and Manuel Ramos-Maqueda, "Data Science for Justice: Evidence from a Nationwide Randomized Experiment in Kenya," 2024. Working paper.

Chen, Yvonne Jie, Jie Gong, Jin Li, and Zibo Zhao, "Better Technology, Worse Motivation: GenAI's Mediocrity Trap," *Available at SSRN 5208163*, 2025.

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

_ , Joseph P Romano, and Michael Wolf, "The Romano–Wolf multiple-hypothesis correction in Stata," *The Stata Journal*, 2020, *20* (4), 812–843.

Conselho Nacional de Justiça, "Justice in Numbers," Technical Report, National Council of Justice (CNJ) 2024. Figure 36: Judicial positions filled per 100,000 inhabitants, by branch of justice.

Council of Europe, "European Judicial Systems: Efficiency and Quality of Justice," Report 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.

de Quidt, Jonathan, Lise Vesterlund, and Alistair J. Wilson, "How Serious Are Experimenter Demand Effects? Evidence and Mitigation Strategies," NBER Working Paper 35350, National Bureau of Economic Research June 2026.

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.

Emi, Bradley and Max Spero, "Technical Report on the Pangram AI-Generated Text Classifier," *arXiv preprint arXiv:2402.14873*, 2024.

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

Government of Khyber Pakhtunkhwa, "Demands for Grants 2024–2025: Administration of Justice," Provincial Assembly of Khyber Pakhtunkhwa 2024. Reports judicial-service grades and annual basic-pay budget allocations.

Government of Punjab, "The Punjab Civil Servants Act, 1974," Government of Punjab, Regulations Wing 1974. Section 12 establishes retirement on completion of the sixtieth year of age unless earlier retirement is directed.

High Court of Sindh, "Advertisement: Appointment of Civil Judges and Judicial Magistrates," March 2025. Advertisement dated 20 March 2025; gross monthly salary stated as PKR 210,029.

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.

Kreitmeir, David and Paul Raschky, "The Heterogeneous Productivity Effects of Generative AI," Working Paper 4745624, SSRN 2024.

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

Lau, Martin, *The Role of Islam in the Legal System of Pakistan*, Vol. 9 of *London-Leiden Series on Law, Administration and Development*, Leiden: Martinus Nijhoff, 2006.

Law and Justice Commission of Pakistan, "Judicial Statistics of Pakistan 2024," 2024.

Lee, Byung Cheol and Jaeyeon Chung, "An Empirical Investigation of the Impact of ChatGPT on Creativity," *Nature Human Behaviour*, 2024, *8* (10), 1906–1914.

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.

List, John A, "Non est disputandum de generalizability? A glimpse into the external validity trial," Technical Report, National Bureau of Economic Research 2020.

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.

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

_ , Shaheen Naseer, and Daniel L Chen, "Altruism in governance: Insights from randomized training for Pakistan's junior ministers," *Journal of Development Economics*, 2024, *170*, 103317.

Ministry of Justice and Judicial Office, "Diversity of the Judiciary: Legal Professions, New Appointments and Current Post-Holders — 2025 Statistics," Technical Report, Ministry of Justice, UK Government 2025.

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.

Pakistan Bar Council, "Pakistan Legal Practitioners and Bar Councils Rules, 1976," 1976. Rules 108-O and 175 concern suspension of practice upon entering government or other service.

Papaioannou, Elias and Stavroula Karatza, "The Greek Justice System: Collapse and Reform," CEPR Discussion Paper DP12731, Centre for Economic Policy Research February 2018.

Peng, Sida, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer, "The Impact of AI on Developer Productivity: Evidence from GitHub Copilot," 2023.

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.

State of Ceará, "Lei Estadual No. 18.324, de 24 de março de 2023," Assembleia Legislativa do Estado do Ceará 2023. Sets the monthly subsidy of a Juiz de Direito de Entrancia Final at BRL 37,731.80 from February 2024.

Stevenson, Megan T. and Jennifer L. Doleac, "Algorithmic Risk Assessment in the Hands of Humans," *American Economic Journal: Economic Policy*, 2024, *16* (4), 382–414.

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.

United Nations General Assembly, "Artificial Intelligence in Judicial Systems: Promises and Pitfalls," Report of the Special Rapporteur on the Independence of Judges and Lawyers A/80/169, United Nations 2025. Margaret Satterthwaite; transmitted by the Secretary-General to the General Assembly at its eightieth session, item 72(b), 16 July 2025.

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

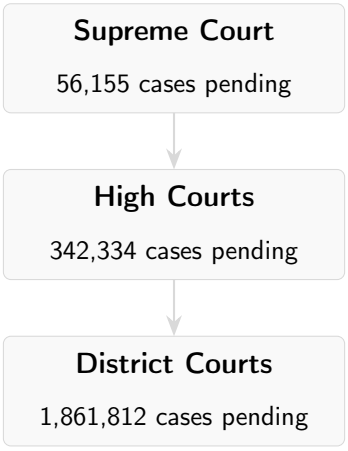
Yang, Crystal S. and Will Dobbie, "Equal Protection Under Algorithms: A New Statistical and Legal Framework," *Michigan Law Review*, 2020, *119* (2), 291–358.

APPENDIX A. FIGURES, TABLES AND DATA DETAILS

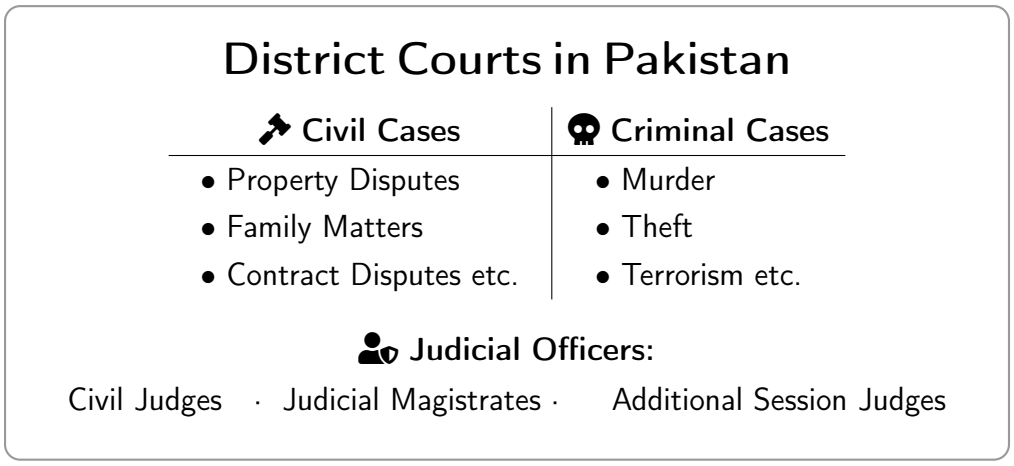
Appendix A1. Additional Figures and Tables

Figure A1: Institutional background

Panel A: Courts Tiers in Pakistan



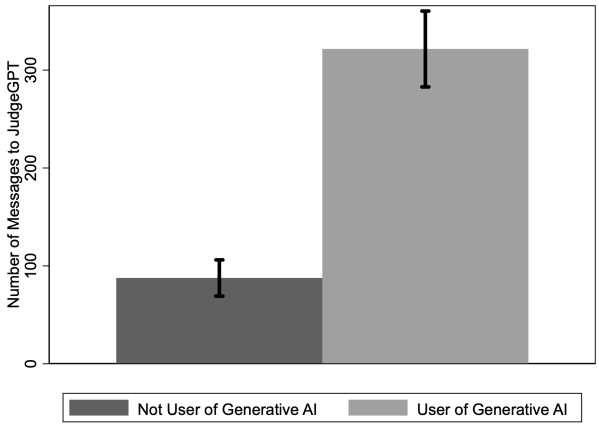
Panel B: District Courts in Pakistan



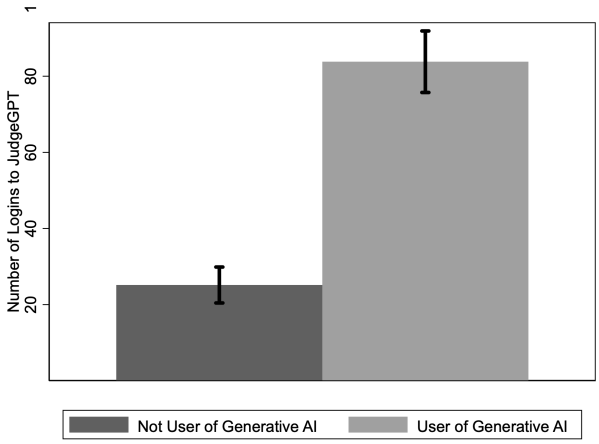
Note: Panel A illustrates Pakistan's three-tier court hierarchy: District Courts (courts of first instance), High Courts (provincial appellate courts), and the Supreme Court (final court of appeal). Case counts report pending matters at each level in 2024 (Law and Justice Commission of Pakistan, 2024). 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. Our experiment is conducted in this first-instance judiciary.

Figure A2: First-Stage: Stated and Actual Generative AI Usage

(a) Panel A. JudgeGPT Message Frequency



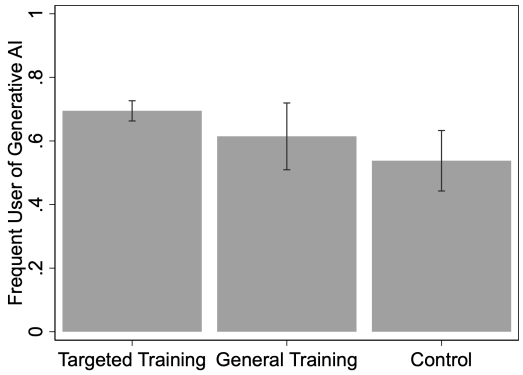
(b) Panel B. JudgeGPT Actual Logins



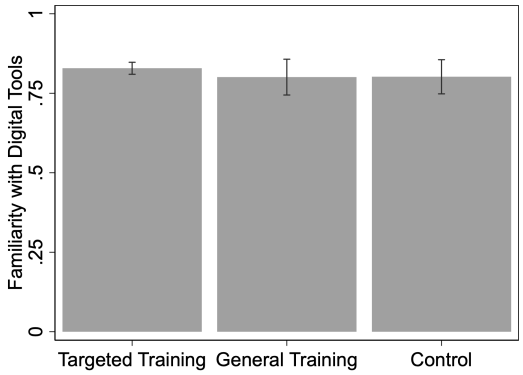
Note: The figure depicts means and standard deviations for usage of JudgeGPT, specifically, number of logins and messages to JudgeGPT among people who described themselves as frequent generative AI users after the treatment and those who did not. Specifically, these individuals answered that they use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

Figure A3: Distribution of Main Outcome across Treatment Arms

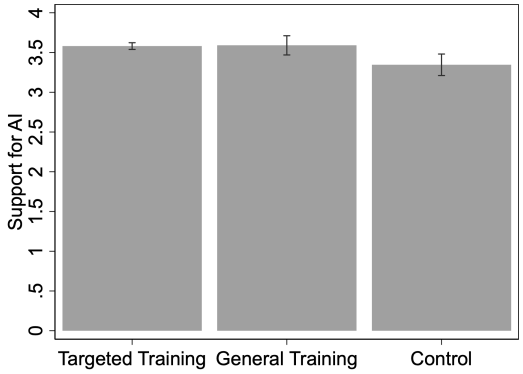
Panel A: Stated Frequent User of Generative AI



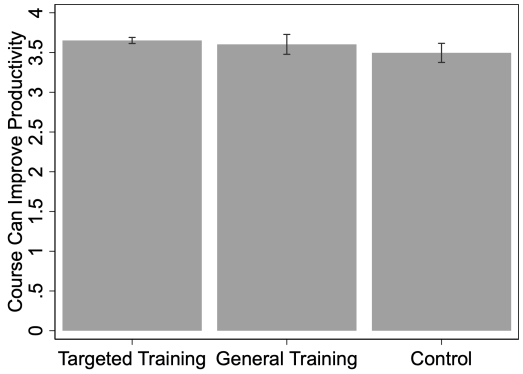
Panel B: Familiarity with Digital Tools



Panel C: Support the Usage of GPT

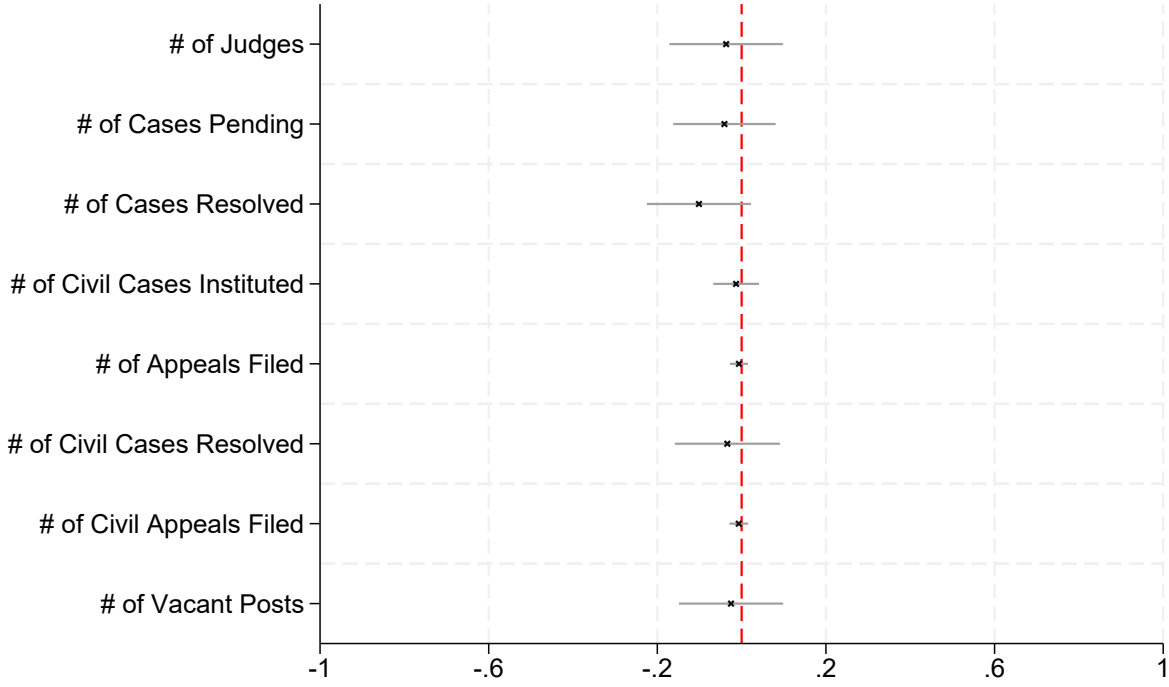


Panel D: GPT Course Can Improve Productivity



Note: Panel A reports the share of frequent users of generative AI, defined as judges who report using GPT or other AI systems in their judicial work at least once or twice a week. Panel B measures the mean of three familiarity indicators: WhatsApp, Microsoft Office, Google/Bing . Panel C measures institutional support for AI integration using 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-4 scale. Panel D reports mean responses to whether the assigned course can improve judicial productivity. The control group received generic training only and no access to *JudgeGPT*. The treatment arms either received targeted *JudgeGPT* training, on effective use of the tool, or received the same generic training as the control group together with access to *JudgeGPT*.

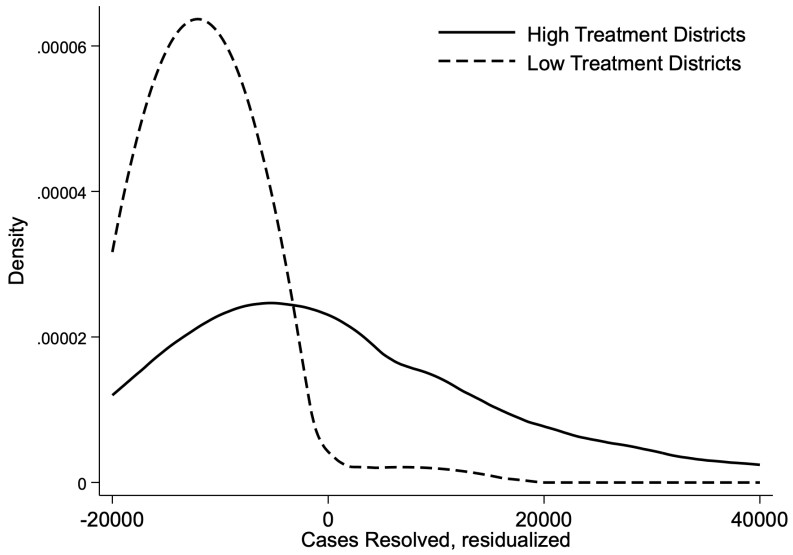
Figure A4: Baseline District Characteristics and the Share of Judges Assigned to JudgeGPT



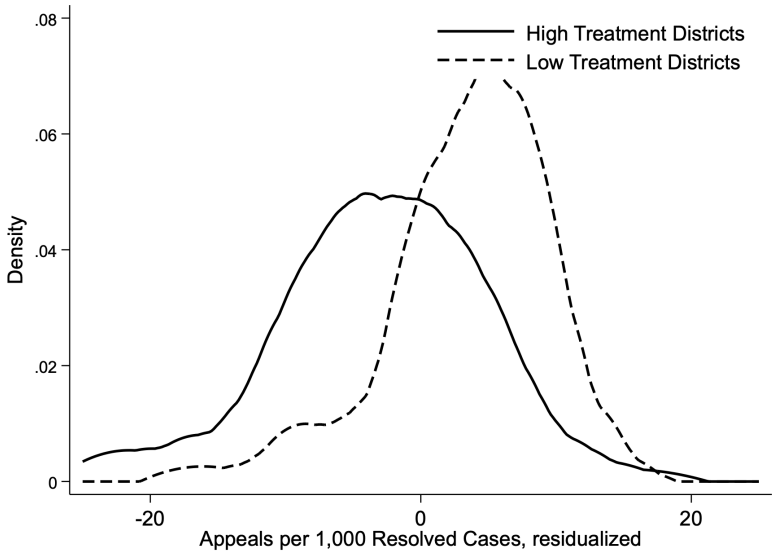
Note: Each point reports the coefficient from a separate regression of a standardized pre-treatment district characteristic on the share of judges assigned to *JudgeGPT* in that district. The outcomes, shown in the order presented in the figure, are the number of judges, pending cases, resolved cases, civil cases filed, appeals filed, civil cases resolved, civil appeals filed, and vacant judicial posts. Outcomes are residualized using district and year fixed effects in the pre-treatment administrative panel, and the reported regressions use the 2023 cross-section. Horizontal lines show 95% confidence intervals. Because randomization occurred at the judge rather than district level, the figure assesses baseline district balance and is descriptive.

Figure A5: Distribution of Cases Resolved and Appeals

Panel A: Cases Resolved

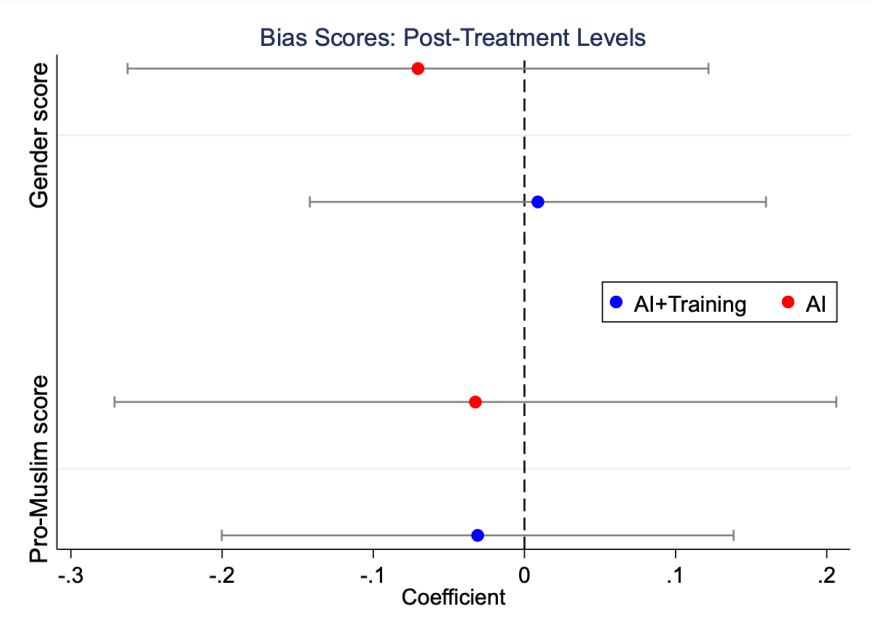


Panel B: Appeals per 1000 Resolved Cases



Note: The figure plots the distribution of residualized cases resolved and appealed, where outcomes are first netted of district and year fixed effects. The residuals capture whether a district resolves or appeals more or fewer cases than predicted by its average level and the national year effect. Positive values indicate outcomes above the predicted level and negative values outcomes below it. We plot these distributions separately for districts above and below the median share of judges assigned to *JudgeGPT* with targeted training in March 2024. Since the residuals are centered at zero, the comparison speaks to differences in the shape, spread, and skewness of the distributions rather than their mean levels.

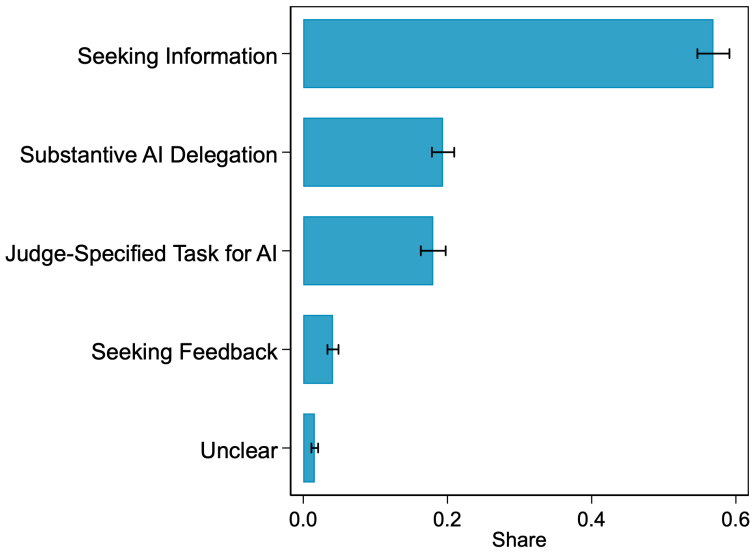
Figure A6: Treatment Effects on Biases



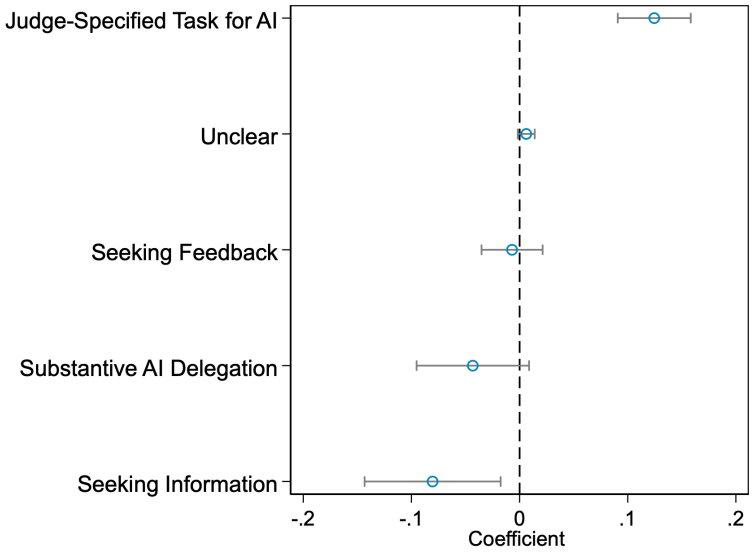
Note: The figure reports 2SLS estimates of the effect of *JudgeGPT* access on post-treatment judicial bias. We measure bias from the text of judicial opinions using a structured LLM evaluation, following the use of language models to recover attitudes and evaluative content from text (Ash et al., 2023, 2024). For each opinion, we provide the full text to GPT-5-mini and separately ask it to assess two dimensions on a 0–3 scale: gender bias, defined as anti-women reasoning or endorsement of conservative gender norms, and pro-Muslim bias, defined as invocation of Islamic legal or moral authority beyond what is required by the case. The prompt instructs the model to rely only on the opinion text, not to infer bias from case type, and to distinguish bias from the routine application of Islamic personal law. All specifications include randomization-strata fixed effects and the baseline survey controls used in the main analysis. The omitted group is judges assigned no *JudgeGPT* access. Bars report 95 percent confidence intervals based on robust standard errors.

Figure A7: JudgeGPT Usage and Training Effects

Panel A: Intention Shares



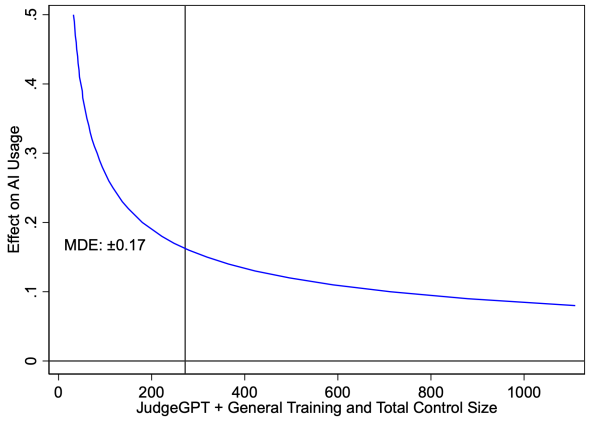
Panel B: Effects on Intentions



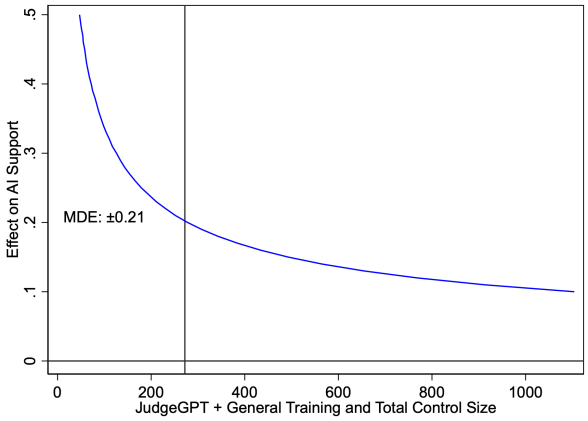
Note: The panel A reports the share of prompts in each category. The panel B reports treatment effects with 95% confidence intervals. In Panel B, the omitted treatment category is JudgeGPT Access and Generic Training, which serves as the reference group in each regression. The pure control arm is excluded because it did not receive platform access and therefore has no JudgeGPT messages. All specifications include strata fixed effects and baseline controls. Unit of observation is an individual JudgeGPT message.

Figure A8: Minimum Detectable Effects for JudgeGPT with Generic Training vs No JudgeGPT with Generic Training

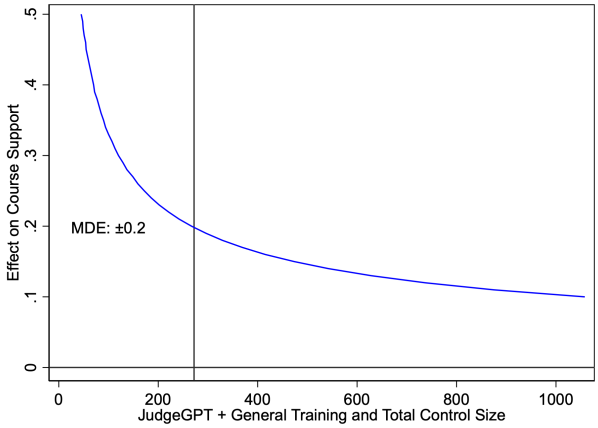
Panel A: AI Usage



Panel B: AI Support



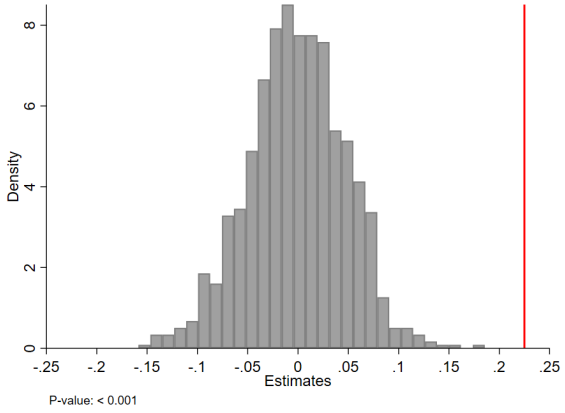
Panel C: Course Support



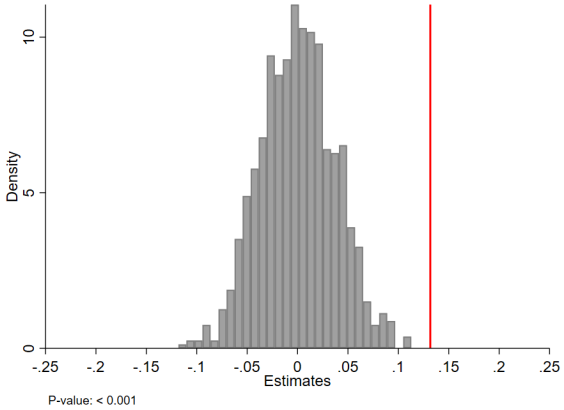
Note: The figure presents the minimum detectable effect (MDE, two-sided) for three outcomes: (a) AI usage, (b) support for generative AI, and (c) perceived improvement in judicial skills (course support). MDE is calculated as the effect size that can be detected with 80% power and 10% significance level (two-sided) for the comparison between Treatment B (JudgeGPT + generic training, n=83) and Control (placebo only, n=182). The vertical line indicates the actual total sample size (N = 265) in the control group (both Treatment B and Control are in the control arm of the full experiment). The curve shows the total sample size required to detect each effect size (keeping the Treatment B / Control ratio). The intersection of the vertical line with the curve gives the MDE. Calculations use a two-sample t-test; standard deviations are taken from the Control and Treatment B groups. The analysis does not adjust for covariates, so these estimates are conservative; models with baseline controls would yield smaller MDEs.

Figure A9: Permutation Inference Test

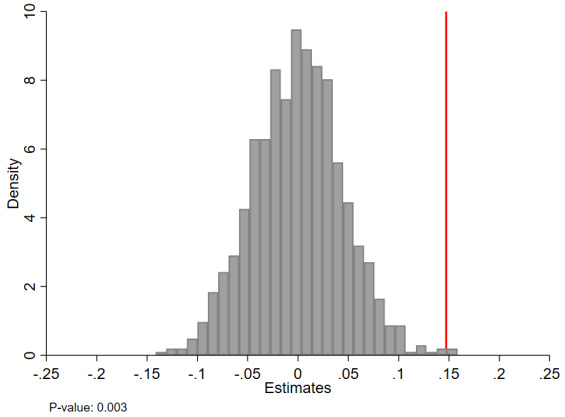
Panel A: Support of Generative AI Usage



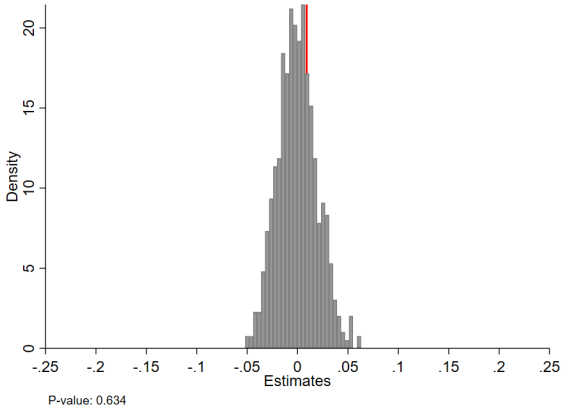
Panel B: Stated Frequent User of Generative AI



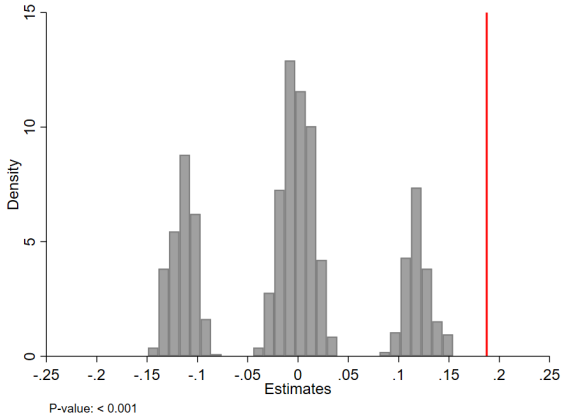
Panel C: Support of AI for Judges Course



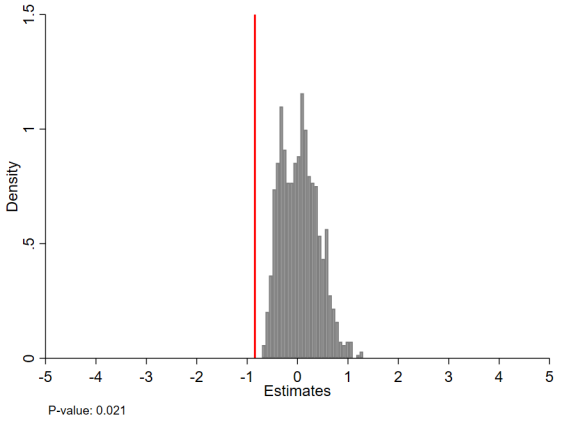
Panel D: Tools Familiarity



Panel E: Cases Resolved



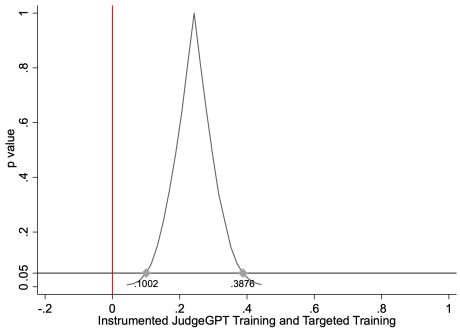
Panel F: Appeals



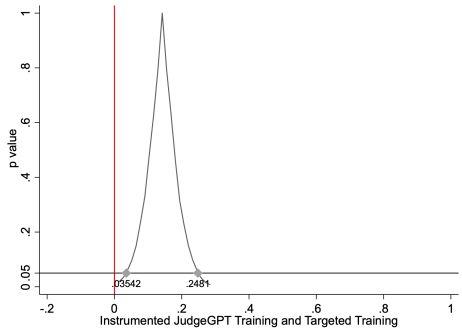
Note: The figure depicts permutation-based inference results. Graphs present the estimated density of permutations of treatment for our main outcomes. For each outcome, 1000 permutations of treatment were analyzed. The red line marks the actual OLS estimate of the coefficient of interest.

Figure A10: Wild-Bootstrap Confidence Intervals for Treatment Effects

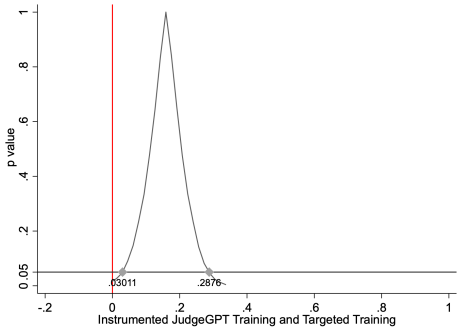
Panel A: Support of Generative AI



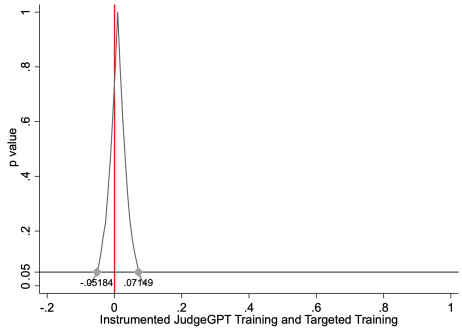
Panel B: Stated Frequent User



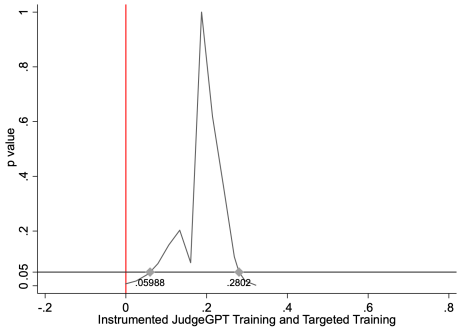
Panel C: Course Can Improve Productivity



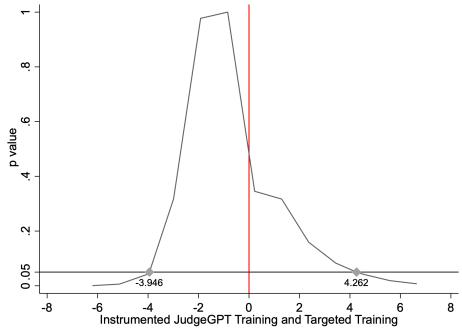
Panel D: Tools Familiarity



Panel E: Case Resolved



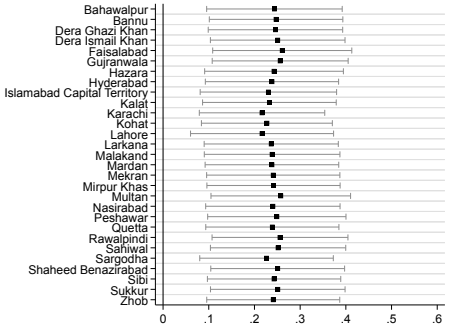
Panel F: Appeals



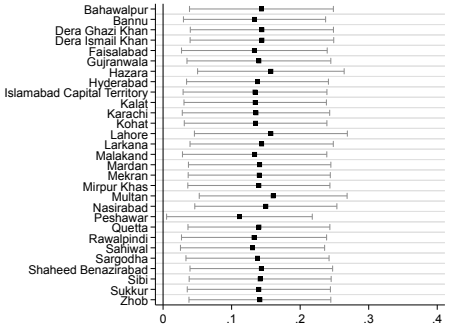
Note: The figure shows wild bootstrap confidence interval curves for six regression-related outcomes using the clustered wild bootstrap procedure of MacKinnon and Webb (2017). For each outcome, the curve plots the bootstrap distribution of the estimated coefficient on JudgeGPT Access and Targeted Training (Panels A, B, C) or Share Treated $\times$ Post and marks the 5% significance cutoff. Panel A reports results for support of using generative AI; Panel B reports results for the indicator of whether the respondent stated they are a frequent user of AI tools; Panel C reports results for course expectations; Panel D reports results for tools familiarity; Panel E reports results for normalized cases resolved counts; Panel F reports results for appeals per 1,000 resolved cases. 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

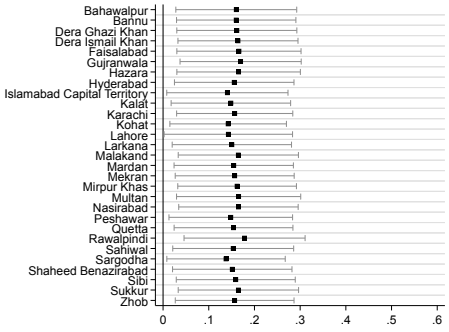
Panel A: Support of Generative AI



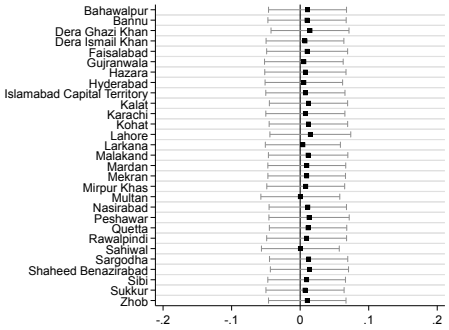
Panel B: Stated Frequent User



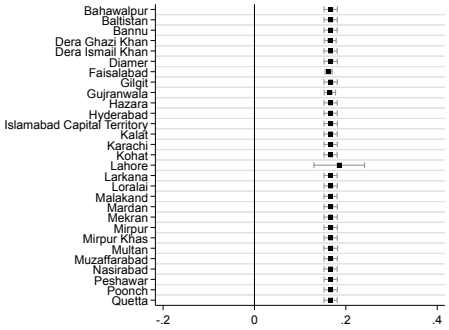
Panel C: Course Can Improve Productivity



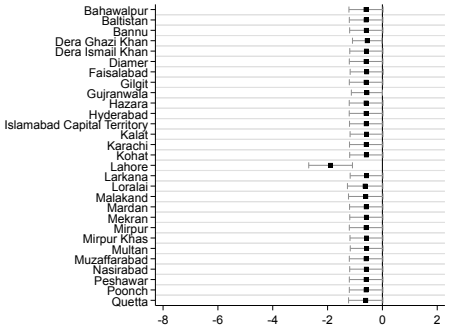
Panel D: Tools Familiarity



Panel E: Case Resolved



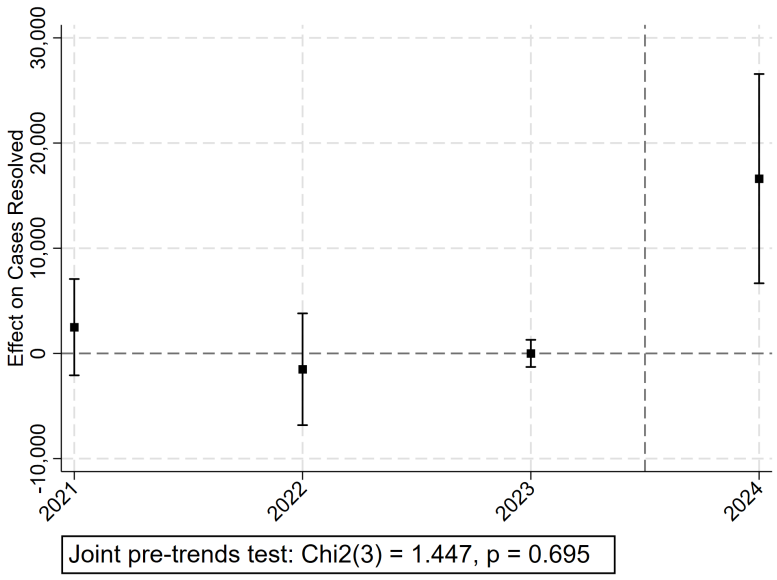
Panel F: Appeals



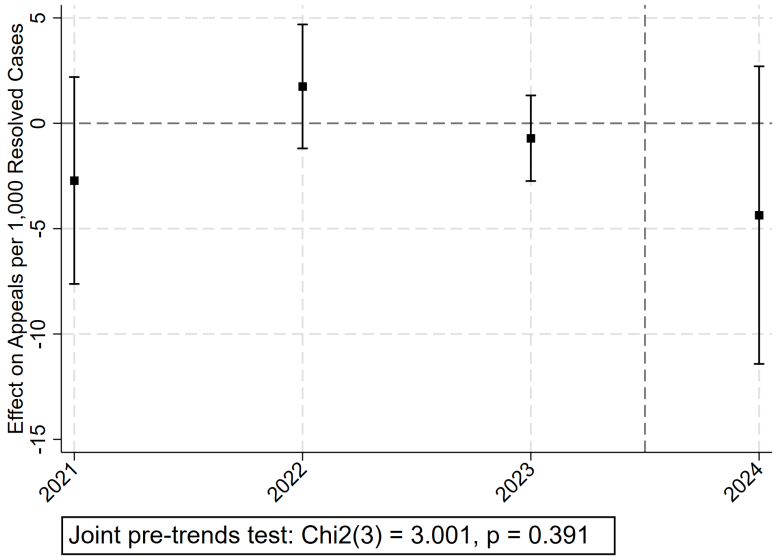
Note: This figure presents the results of estimating the main specification with one region excluded at a time. "Region" denotes Pakistan's official administrative division, the tier between province and district. Data reflect the structure as of 2024, incorporating the December 2024 Punjab reorganization 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 region excluded on the y-axis.

Figure A12: Share Judges Treated and Court Productivity

Panel A: Cases Resolved



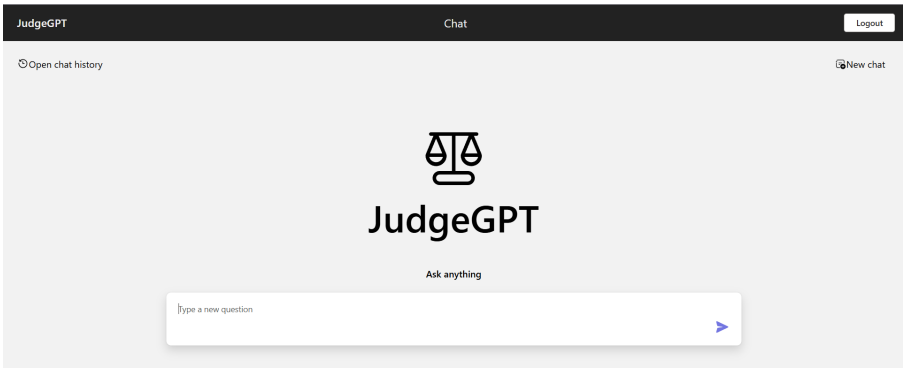
Panel B: Appeals per 1000 Cases Resolved



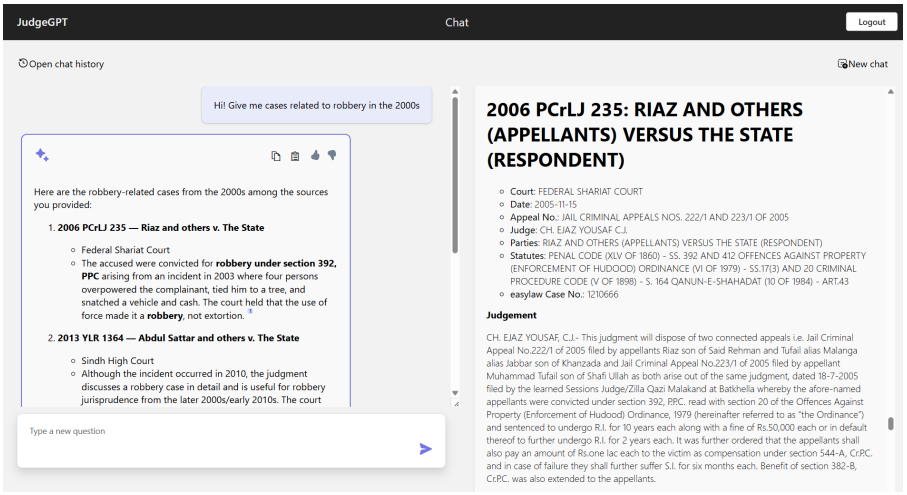
Note: The figure reports coefficients and 95% CI from regressions of share judges treated above median. Following Clarke et al. (2023) procedure, we conduct 1000 bootstrap iterations. In panel A dependent variable is standardized number of cases resolved. In panel B dependent variable is appeals per 1000 cases resolved in the same district-year. Sitting judges, share criminal cases in institution, and share criminal cases disposed are added as covariates. The vertical dashed line marks the JudgeGPT course. Joint Pre-Trends Significance Test reports statistics and corresponding p-value for $\beta_{2021}, \beta_{2022}, \beta_{2023}$ equling zero.

Figure A13: JudgeGPT Interface

Panel A: Start Screen



Panel B: Chat with Case Retrieval



Note: The figure shows the JudgeGPT user interface. Panel A displays the start screen, where the user can type a new question. Panel B shows an example chat session: the user asks for robbery-related cases, and the assistant returns a list of relevant cases (left) along with the full text of the selected judgment (right).

TABLE A1: Treatment Description and Slides of the Trainings

Treatment Group	**Treatment Description**	**Treatment Details**	**Link to the Slides**
JudgeGPT and Targeted Training (AI with tool-specific course)	This group received our full treatment, consisting 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.	• AI tool with custom knowledge base and system prompt • 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
JudgeGPT with Generic Training (AI with technology and law course)	This treatment group received JudgeGPT and a generic 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 citations for legal research, topics in machine learning, such as general overview of machine learning tools and biases of AI.	• AI tool with custom knowledge base and system prompt • 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
Generic Training Only (technology and law course)	This treatment group did not receive an access to JudgeGPT 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 the slide decks used in each condition.

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 a judge.
Support of AI (1–4)	Answer to "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.
Number of Cases Pending	Answer to "How many cases do you currently have pending on your desk?"
Number of Cases Decided per Month	Answer to "How many cases did you decide in the last month?"
Number of Cases Comfortable Managing at the Same Time	Answer to "How many cases are you comfortable managing at the same time?"
Hours per Week Spent on Legal Research	Answer to "How many hours a week do you spend on legal research and writing judgments?"
Hours per Week Spent on Administrative Work	Answer to "How many hours a week do you spend on administrative tasks?"
Confidence in Public Speaking (1–10)	Answer to "On a scale from 1 to 10, how confident do you feel in public speaking in court? – choose" on a scale from 1 to 10.
Workload (1–10)	Answer to "On a scale from 1 to 10, how would you rate your current workload? – choose" on a scale from 1 to 10.
Work-Life Balance (1–10)	Answer to "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 a scale from 1 to 10.
Post-treatment Confidence in Public Speaking (1-10)	Post-treatment Answer to "On a scale from 1 to 10, how confident do you feel in public speaking in court? – choose" on a scale from 1 to 10.
Hours Spent on Administrative Needs (1–10)	Answer to "On a scale from 1 to 10, how confident do you feel in dealing with court administrative needs? – choose" on a scale from 1 to 10.
Expectations from the Course (1–4)	Answer to "Do you expect that an AI for Judges Course can improve your productivity? – choose" on a scale from 1 to 4.

Note: Continued on the next page.

Description of Variables

Variable	Description
Familiar with AI	Indicator equals 1 if AI is chosen in the question "Which of these technologies are you familiar with and use regularly?"
Familiar with WhatsApp	Indicator equals 1 if WhatsApp is chosen in the question "Which of these technologies are you familiar with and use regularly?"
Familiar with MS Office	Indicator equals 1 if MS Office is chosen in the question "Which of these technologies are you familiar with and use regularly?"
Familiar with Google/Bing	Indicator equals 1 if Google or Bing is chosen in the 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.
JudgeGPT and Generic Training	Indicator equals 1 if the judge received access to JudgeGPT and Generic Training.
Frequent AI User	Indicator equals 1 if the answer to "Which of the following best describes how often you use GPT or other AI systems in your judicial work?" is one of "Almost every day", "Once or twice a week", and zero otherwise.
Support of Generative AI Usage	Answer to "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 to "Do you expect that an AI for Judges Course can improve your productivity? – choose" on a scale from 1 to 4.

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.
LLM Quality Assessment	Assessed with LLM quality score, pairwise
District Level Panel (varies by district-year)	
# of Cases Resolved	Number of cases resolved.
Appeals per 1000 Resolved Cases	Number of Appeals Divided by Thousands 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 in Civil Appeal	Number of civil cases appealed.
# of Cases in Criminal Appeal	Number of criminal cases appealed.
Share Treated Judges with Targeted Training	Count of judges treated in March 2024 divided by the total working strength in the district.
Chatlogs Outcomes (varies by judge)	
Text Improvement	Share of conversations in which the judge asks AI to edit or improve text already written (grammar, style, formality) rather than generate new content.
Text Generation	Share of conversations in which the judge asks AI to produce new written content from scratch (e.g., draft judgments, orders, letters).
Decision Support	Share of conversations in which the judge asks AI to determine the appropriate ruling in a specific case (e.g., who should win, whether to grant bail), based on case-specific facts.
Legal Research: Finding Sources	Share of conversations in which the judge asks AI to identify specific legal authorities: statutes, case law, citations, or precedents.

Note: Continued on the next page.

Description of Variables

Variable	Description
Legal Research: Open QA	Share of conversations in which the judge asks a general legal question to understand a concept, definition, or procedure, without seeking citations or applying the law to case facts.
Summarization / Closed QA	Share of conversations in which the judge asks AI to summarize, extract facts from, or answer questions about a document pasted into the prompt.
Translation	Share of conversations in which the judge asks AI to translate text between languages, typically English and Urdu.
Critiquing / Evaluating	Share of conversations in which the judge asks AI to assess the soundness or weaknesses of legal reasoning or arguments in a draft judgment or a party's position.
Personal Use & Other	Share of conversations unrelated to judicial or legal work (personal queries, casual conversation) or that do not fit any other task category.
Judge-Specified Task for AI	Share of conversations in which the judge has already made all judgment calls and directs AI only to produce the specified output.
Seeking Feedback	Share of conversations in which the judge shares their own draft, decision, or reasoning and asks AI to evaluate, critique, or improve it.
Seeking Information	Share of conversations in which the judge asks AI to explain the law or a legal concept in the abstract, without case-specific facts or a request for a judgment call.
Substantive AI Delegation	Share of conversations in which the judge hands off a high-agency task (deciding the outcome, generating the reasoning, or drafting the opinion) with minimal guidance beyond the case facts.
Unclear	Share of conversations too brief, ambiguous, or fragmentary for the classifier to determine a predominant task or intention.

Note: The table provides short definitions of the variables used in the analysis. Descriptive statistics are provided in the following table.

TABLE A3: Summary Statistics of Main Variables

Panel A: Pretreatment Characteristics						
	Mean	**Median**	**SD**	**Min**	**Max**	**N obs**
Age	42.92	43	6.998	27	60	1076
Female	0.226	0	0.418	0	1	1076
Years of Experience	11.67	10	6.889	0	35	1076
Support of AI (1-4)	3.452	4	0.804	1	4	1076
Number of Cases Pending	566.4	366	900.0	0	24000	1076
Number of Cases Decided per Month	122.4	80	154.6	0	2500	1076
Number of Cases Comfortable Managing	137.4	40	308.3	0	6007	1076
Hours per Week Spent on Legal Research	16.04	14	11.75	0	100	1076
Hours per Week Spent on Administrative Work	11.33	8	10.72	0	100	1076
Confidence in Public Speaking (1-10)	6.583	7	2.078	1	10	1076
Workload (1-10)	6.414	7	2.124	1	10	1076
Worklife Balance (1-10)	5.787	5	2.087	1	10	1076
Self-Assessed Public Speaking(1-10)	7.533	8	2.028	1	10	1076
Confidence in Administrative Work (1-10)	7.468	8	1.872	1	10	1076
Expectations from the Course (1-4)	3.740	4	0.588	1	4	1076
Familiar with AI	0.263	0	0.440	0	1	1076
Familiar with WhatsApp	0.962	1	0.192	0	1	1076
Familiar with MS Office	0.690	1	0.463	0	1	1076
Familiar with Google/Bing	0.789	1	0.408	0	1	1076
Panel B: Judge Level Outcomes						
	Mean	**Median**	**SD**	**Min**	**Max**	**N obs**
JudgeGPT and Related Training	0.750	1	0.433	0	1	1076
JudgeGPT without Training	0.0762	0	0.265	0	1	1076
Frequent AI User	0.650	1	0.477	0	1	1076
Support of AI for Judges Course	3.622	4	0.575	1	4	1076
Familiar with AI	0.626	1	0.484	0	1	1076
Panel C: Case Analysis Outcomes						
	Mean	**Median**	**SD**	**Min**	**Max**	**N obs**
Judgments Readability, post	0.00766	-0.0488	0.668	-0.178	9.485	293
# of Legal References, post	7.311	2.500	10.46	0	61	293
Panel D: District Level Panel						
	Mean	**Median**	**SD**	**Min**	**Max**	**N obs**
# of Cases Resolved	29200.7	11731	53881.2	0	682933	472
Appeals per 1000 Resolved Cases	19.35	13.72	19.66	0	177.5	445
# of Civil Cases Resolved	12096.7	3322.5	21023.6	0	184715	472
# of Criminal Cases Resolved	16717.2	6465.5	34408.0	0	503817	472

Note: Continued on next page

Table A3 – continued from previous page

	Mean	Median	SD	Min	Max	N obs
# of Sitting Judges	20.42	13	30.02	0	272	472
# of Vacant Judge Positions	7.097	3	13.07	0	113	472
# of Cases on Balance	15002.2	4029	29083.9	0	351174	472
# of Cases in Civil Appeal	460.3	202	677.8	0	5138	472
# of Cases in Criminal Appeal	10.02	4	22.42	0	348	472
Share of Criminal Cases Disposed	0.553	0.623	0.263	0	0.948	472
Share of Criminal Cases Instituted	0.570	0.629	0.248	0	0.951	472
Panel E: Chatlogs Outcomes						
	Mean	**Median**	**SD**	**Min**	**Max**	**N obs**
Text Improvement	0.167	0.0507	0.243	0	1	805
Text Generation	0.152	0.0741	0.199	0	1	805
Decision Support	0.0777	0.00922	0.142	0	1	805
Legal Research: Finding Sources	0.118	0.0476	0.181	0	1	805
Legal Research: Open QA	0.397	0.347	0.306	0	1	805
Summarization / Closed QA	0.0424	0	0.115	0	1	805
Translation	0.0109	0	0.0506	0	0.692	805
Critiquing / Evaluating	0.0149	0	0.0553	0	1	805
Personal Use & Other	0.0197	0	0.0824	0	1	805
Judge-Specified Task for AI	0.172	0.0625	0.242	0	1	805
Seeking Feedback	0.0384	0	0.105	0	1	805
Seeking Information	0.548	0.545	0.326	0	1	805
Substantive AI Delegation	0.181	0.125	0.211	0	1	805
Unclear	0.0151	0	0.0640	0	1	805

Note: This table reports mean, median, standard deviation, minimum, maximum, and number of observations 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. Panel E summarizes chatlogs variables (judge level).

TABLE A4: Treatment Roll-out Summary

	JudgeGPT subscription & GPT instruction course		Judge GPT Subscription & Generic Technology and Law Course		Generic Technology & Law Course		% of Districts Covered
	Judges	Districts	Judges	Districts	Judges	Districts	
Jan 2021 – Jan 2024	0	0	0	0	0	0	0.00
March 2024	487	94	0	0	0	0	55.62
Sep 2024	979	108	0	0	0	0	63.91
Feb 2025	1197	110	180	62	182	66	69.82
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 118 districts were registered. Group 1 (487 judges from 94 districts) began in March 2024 and received Treatment A: *JudgeGPT* access with tool-specific training. Group 2 (492 judges from 94 districts), launched in September 2024, received the same treatment. Group 3A (218 judges from 70 districts), launched in February 2025, also received Treatment A. In February 2025, Groups 3B and 3C were implemented. Group 3B (180 judges from 62 districts) received Treatment B: *JudgeGPT* access with generic training. Group 3C (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 Group 1, September–October 2024 for Group 2, and February–April 2025 for Groups 3A, 3B, and 3C. Randomization was conducted at the judge level, stratified by province and age group, prior to training. Additional details are provided in Figure 2.

TABLE A5: Opinion Quality: Expert Validation and Cross-Model Reliability

Comparison	Directional Agreement
Panel A. Human Benchmark	
Lawyer A vs. Lawyer B	73.0%
Panel B. Validation of the Main GPT Measure	
Pooled Lawyer Ratings vs. GPT	70.6%
Panel C. GPT Agreement When Lawyers Concur	
Consensus Lawyer Rating vs. GPT	75.9%
Panel D. Cross-Model Reliability	
GPT vs. Gemini	70.0%

Note: Two independent lawyers each evaluated the same 90 anonymized pre/post judgment pairs using the same -2 to $+2$ quality scale applied by GPT. GPT is the model used to construct the main full-sample quality measure. Directional agreement is the share of non-tied comparisons for which the two relevant ratings identify the same judgment as superior; comparisons in which either relevant rating assigns a tie are excluded. Pooled lawyer ratings combine the separate evaluations of the two lawyers. Consensus Lawyer Rating conditions on comparisons for which both lawyers identify the same superior judgment. The GPT–Gemini comparison is reported as a cross-model reliability diagnostic and is not used to construct the main quality measure.

TABLE A6: Sample User Queries for Each Classification Category

Category	**User Query (Anonymized)**
Judge-Specified Task for AI	"Simplify the following judgment for litigants: ..."
Seeking Feedback	"What do you think about these critical comments of your work? Do you agree with any of the points? Would you rewrite anything in response to these points?"
Seeking Information	"Precedents for dismissal of Injunction Application under order XXXIX Rule 39 Civil Procedure Code"
Substantive AI Delegation	"Whether the accused is entitled for bail in smuggling case wherein his name was inserted by IO [Investigator Officer]..."
Translation	"Translate the following into legal English: [URDU QUERY]
Unclear	"To determine whether your assignments have been accepted, you may consider the following steps based on legal principles and practices: 1. Official Communication ... 6. Seek Legal Advice ..."
Legal Research: Open QA	"Whether appearance of plaintiff in person is mandatory before the family court at the time of filling a fresh family suit? guide with help of concern laws"
Text Improvement	"fix grammar mistakes: IN THE COURT OF..."
Legal Research: Finding Sources	"Find judgment if location of property are disputed"
Summarization / Closed QA	"summarize this decision: ORDER SHEETIN [COURT NAME] KARACHISMA No.55 of..."
Text Generation	"Rewrite the order in light of the judgment PLD 2023 Peshawar 12, emphasising the cut off date as established by the law and case laws, and also highlight that the suit of such nature cannot be entertained now"
Decision Support	"[NAME] filed suit for rendition of accounts against his brother [NAME], claiming that the business owned by their father has been taken over by [NAME] since the death of their father. [NAME] therefore prayed for the rendition of accounts of business. [NAME] denied this fact alleging that the land was owned by their deceased father but the business was established by him. **what should be the fate of case?**"
Personal Use & Other	"how can i become the Chief Justice in Pakistan?"

Continued from previous page	
Category	**User Query (Anonymized)**
Critiquing / Evaluating	"critically analyze following judgment in view of case laws of superior courts of [COUNTRY] and make necessary grammatical corrections: IN THE COURT OF SENIOR CIVIL JUDGE, [CITY] FC Suit No.[CASE NO]/[YEAR] (Before: [JUDGE TITLE] [SURNAME]) [NAME] son of [NAME] ... Plaintiff. Versus [NAME] son of [NAME] ... Defendant. ... This judgment will dispose of an FC suit No.[CASE NO]/[YEAR] filed by Plaintiff against defendant for Declaration, Specific Performance of Contract, Possession, Recovery and Permanent Injunction. ..."

Note: All personal data, case numbers, dates, locations, court names, and other identifying information have been replaced with [REDACTED] or generic placeholders. The queries are reproduced exactly as they appear in the original logs, with only anonymization changes.

TABLE A7: Balance Test Over Judge Characteristics and Pre-Treatment Outcomes

	(1)		(2)		(3)		(1)-(2)		(1)-(3)		(2)-(3)	
	JudgeGPT And Training		Only JudgeGPT		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	868	42.942	93	43.333	104	42.231	961	0.640	972	0.939	197	0.494
		(0.239)		(0.742)		(0.617)						
Female	868	0.224	93	0.237	104	0.231	961	0.619	972	0.942	197	0.344
		(0.014)		(0.044)		(0.042)						
Years of Experience	868	11.648	93	11.991	104	11.560	961	0.850	972	0.285	197	0.913
		(0.237)		(0.744)		(0.613)						
Support of AI	868	3.439	93	3.495	104	3.481	961	0.618	972	0.660	197	0.955
		(0.027)		(0.083)		(0.081)						
Number of Decided Cases	868	124.588	93	117.581	104	111.692	961	0.333	972	0.094*	197	0.627
		(5.494)		(11.214)		(12.379)						
Number of Cases Concurrently Managed	868	138.649	93	130.000	104	140.567	961	0.519	972	0.590	197	0.994
		(10.888)		(24.491)		(26.727)						
Hours spent on Legal Research and Writing Judgments	868	16.768	93	15.753	104	16.500	961	0.514	972	0.997	197	0.847
		(0.702)		(1.054)		(1.164)						
Hours spent on Administrative Work	868	11.386	93	14.667	104	10.356	961	0.067*	972	0.286	197	0.035**
		(0.390)		(1.710)		(0.870)						
Work-Life Balance	868	5.795	93	5.688	104	5.798	961	0.812	972	0.884	197	0.740
		(0.071)		(0.226)		(0.190)						
Confidence in Public Speaking	868	7.552	93	7.140	104	7.702	961	0.051*	972	0.532	197	0.108
		(0.068)		(0.227)		(0.201)						
Confidence in Administrative Work	868	7.485	93	7.151	104	7.587	961	0.084*	972	0.515	197	0.137
		(0.064)		(0.198)		(0.178)						
Expectations from AI for Judges	868	3.727	93	3.849	104	3.721	961	0.014**	972	0.942	197	0.176
		(0.020)		(0.043)		(0.060)						

Note: The table presents the balance test. The treatment group was assigned a JudgeGPT subscription and training explaining the use of JudgeGPT (Groups 1-3A); Group 3B received JudgeGPT access and generic technology-and-law training (no targeted training); Group 3C received the generic technology-and-law training only, with no JudgeGPT access. Table reports mean for pretreatment covariates as well as 3 pairwise t-tests p-value. Heteroskedasticity-robust standard errors are in parenthesis. 18 strata fixed effects are included. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

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	0.134***	0.218***	0.139**
	(0.048)	(0.064)	(0.057)
JudgeGPT Access and Generic Training	0.045	0.203**	0.042
	(0.069)	(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	0.137***	0.225***	0.142**
	(0.048)	(0.064)	(0.057)
Lee Bounds	[0.019 , 0.324]	[0.077 , 0.454]	[0.019 , 0.363]
Generalized Lee Bounds	[0.047 , 0.331]	[0.134 , 0.851]	[0.078 , 0.542]
Panel C: JudgeGPT + Generic 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 Generic Training IPW ITT	0.046	0.205**	0.043
	(0.069)	(0.089)	(0.085)
Lee Bounds	[0.005 , 0.082]	[0.077 , 0.261]	[-0.065 , 0.090]
Generalized Lee Bounds	[-0.052 , 0.098]	[-0.194 , 0.472]	[-0.247 , 0.256]
Mean dep. var.	0.649	3.544	3.619
Observations	1082	1084	1084

Note: Panel A reports baseline ITT treatment estimates. Panels B and C report treatment effects for Arm A (JudgeGPT + targeted training) and Arm B (JudgeGPT + generic 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 (2025) 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 Main Outcomes by Groups

	AI Usage	AI Support	Course Support
	(1)	(2)	(3)
Panel A: Treatment A uses only Group 1, Treatment B and Control are Same			
JudgeGPT Access and Targeted Training	0.208***	0.314***	0.207***
	(0.052)	(0.072)	(0.065)
JudgeGPT Access and Generic Training	0.076	0.296***	0.108
	(0.079)	(0.102)	(0.098)
Panel B: Treatment A uses only Group 2, Treatment B and Control are Same			
JudgeGPT Access and Targeted Training	0.068	0.216***	0.144**
	(0.060)	(0.080)	(0.072)
JudgeGPT Access and Generic Training	0.046	0.248**	0.068
	(0.079)	(0.099)	(0.095)
Panel C: Treatment A uses only Group 3A, Treatment B and Control are Same			
JudgeGPT Access and Targeted Training	0.077	0.161*	0.072
	(0.067)	(0.092)	(0.081)
JudgeGPT Access and Generic Training	0.068	0.223**	0.042
	(0.078)	(0.102)	(0.096)
Strata Fixed Effects	Yes	Yes	Yes
Controls	Yes	Yes	Yes
Adj. R^2	.043	.086	.069
Mean Dep. Var.	.588	3.474	3.552
Control Mean	.471	3.372	3.492
Observations	345	346	346

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

TABLE A10: JudgeGPT Access and Secondary Outcomes by Groups

Panel A: Treatment A uses only Group 1, Treatment B and Control is Same			
	Tools Familiarity	Cases Resolved	Appeals
	(1)	(2)	(3)
JudgeGPT Access and	0.032		
Targeted Training	(0.030)		
JudgeGPT Access and	-0.034		
Generic Training	(0.045)		
Share JudgeGPT and Targeted		9239.049***	-0.846*
Training × Post		(600.440)	(0.482)
Panel B: Treatment A uses only Group 2, Treatment B and Control are Same			
JudgeGPT Access and	0.003		
Targeted Training	(0.034)		
JudgeGPT Access and	-0.029		
Generic Training	(0.046)		
Panel C: Treatment A uses only Group 3A, Treatment B and Control are Same			
JudgeGPT Access and	-0.029		
Targeted Training	(0.040)		
JudgeGPT Access and	-0.038		
Generic Training	(0.046)		
Strata Fixed Effects	Yes		
Controls	Yes	Yes	Yes
Adj. R^2	-.008	.9	.461
Mean Dep. Var.	.782	29200.72	19.352
Control Mean	.751		
Observations	343	472	445

Note: The table reports treatment-effect estimates separately by training group. In each panel, the targeted-training arm is restricted to judges in the indicated group, while the generic-training and control groups are unchanged. Panel A uses Group 1 for targeted training, Panel B uses Group 2, and Panel C uses Group 3A. Cases resolved and appeals are reported only in Panel A because these district-level outcomes come from the 2024 judicial statistics. Only the March 2024 training group was exposed early enough for its district treatment share to be linked to outcomes measured in December 2024; the later groups were trained afterward. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE A11: District-Level Outcomes in Standard-Deviation Units

Panel A: Cases Resolved				
	(1)	(2)	(3)	(4)
Share JudgeGPT and Targeted Training × Post	0.157***	0.154***	0.174***	0.171***
	(0.009)	(0.007)	(0.009)	(0.011)
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	0.857	0.874	0.894	0.900
Mean Dep. Var.	0	0	0	0
# Clusters	118	118	118	118
Observations	472	472	472	472
Panel B: Appeals per 1,000 Resolved Cases				
Share JudgeGPT and Targeted Training × Post	-0.048**	-0.030*	-0.045**	-0.043*
	(0.023)	(0.016)	(0.022)	(0.025)
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	0.336	0.448	0.457	0.461
Mean Dep. Var.	0	0	0	0
# Clusters	118	118	118	118
Observations	445	445	445	445

Note: Robust standard errors, clustered by district, are in parentheses. The unit of observation is district-year. Panel A uses the number of cases resolved in a district-year as the dependent variable. Panel B uses the total appeals per 1,000 cases resolved in a district-year as the dependent variable. In both panels, the dependent variable is transformed to have mean zero and standard deviation one within the estimation sample. Total appeals is defined as the sum of civil and criminal appeals. *Share JudgeGPT and Targeted Training × Post* equals the number of judges assigned JudgeGPT with targeted training divided by the number of sitting judges in the district court in the baseline denominator year, interacted with a post indicator for 2024. All specifications include district fixed effects; columns (2)–(4) additionally include year fixed effects. Column (3) controls for the number of judges in the district. Column (4) adds district-year controls for the number of sitting judges and the number of vacant judge positions. The table reports adjusted R^2, the mean of the dependent variable, and the number of clusters and observations. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

TABLE A12: Effect of JudgeGPT Access and Associated Training on Time Allocation

	2SLS			
	(1)	(2)	(3)	(4)
	Hours Spent on Legal Research	Hours Spent on Administrative Work	Reported Workload	Reported Worklife Balance
JudgeGPT Access and Associated Training	0.425	0.296	-0.035	-0.058
	(1.109)	(0.789)	(0.180)	(0.188)
JudgeGPT Access without Training	2.550	-0.262	0.093	-0.121
	(2.042)	(1.466)	(0.263)	(0.299)
Strata Fixed Effects	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Adj. R^2	.027	.095	.109	.053
Kleibergen-Paap F-statistic	3013.392	3013.392	3013.392	3013.392
P-value (Access with Training = Access without Training)	.248	.67	.511	.795
Mean Dep. Var.	16.249	10.728	6.968	6.481
Observations	1065	1065	1065	1065

Note: Newey-West standard errors in parentheses. The dependent variables are defined as answers to the following questions: How many hours a week do you spend on legal research? , How many hours a week do you spend on dealing with administrative needs? , Rate your workload on a scale 1-10 , Rate your worklife balance on a scale 1-10. 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 (Group 3b). Control group did not receive GPT access while getting a general training course. Controls include pre-treatment characteristics reported in the balance test and strata fixed effects. In columns (3)-(4) treatment and control are instrumented with initial assignment of treatment and control. R^2 and Kleibergen-Paap F-statistic are reported. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE A13: JudgeGPT Access and Biases

	ITT	2SLS	ITT	2SLS
	(1)	(2)	(3)	(4)
	Gender Bias Score		Muslim Bias Score	
Assigned JudgeGPT Access and Targeted Training	0.009		-0.030	
	(0.076)		(0.085)	
Assigned JudgeGPT Access and General Training	-0.065		-0.030	
	(0.092)		(0.115)	
JudgeGPT Access and Targeted Training		0.009		-0.031
		(0.077)		(0.086)
JudgeGPT Access and General Training		-0.070		-0.032
		(0.098)		(0.122)
Strata Fixed Effects	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Adj. R^2	.025	.025	.047	.046
Kleibergen-Paap F-statistic		1071.45		1071.45
p-val. (Targeted Training = General Training)	0.229	0.214	1.000	0.987
Control Mean	0.261	0.261	0.415	0.415
Observations	487	487	487	487

Note: Newey-West standard errors are in parentheses. The dependent variable is LLM assessed bias scores: gender bias in columns (1) and (2); Muslim bias in columns (3) and (4). JudgeGPT Access and Targeted Training is a binary indicator equal to one for a judge assigned to the course who logged into JudgeGPT at least once, while Used JudgeGPT Access and General Training is a binary indicator equal to one for a judge logging into GPT without having the course (as in Group 3B). The control group did not receive GPT access while receiving a general training 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 (2) and (4), treatment and control A are instrumented with the initial assignment of treatment and control A. R^2 and the Kleibergen-Paap F-statistic are reported. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE A14: Spillover effect of JudgeGPT Access and Associated Course

	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	0.003	0.002	0.002
	(0.002)	(0.002)	(0.002)
JudgeGPT Access and Associated Training	0.278***	0.241***	0.201***
	(0.045)	(0.080)	(0.068)
JudgeGPT Access with Generic Training	0.189**	0.245**	0.125
	(0.074)	(0.104)	(0.099)
Strata Fixed Effects	Yes	Yes	Yes
Controls	Yes	Yes	Yes
Adj. R^2	.086	.071	.065
Mean Dep. Var.	.658	3.544	3.618
Observations	927	927	927

Note: Standard errors clustered on a district level in parenthesis. The dependent variable is an indicator of a judge reporting the use of any AI tool, support of GPT usage rated on the scale 1-4, 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 1-4. 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 training as first Control Group (Group 3C). Control group did not received GPT access while getting a general training course. Interaction of a judge being in the control group with the number of unique APIs per judge 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 A15: Treatment Effects with Multiple Hypothesis Adjustments

	(1) Frequent Gen AI	(2) Support JudgeGPT	(3) AI Course	(4) Total Disposal	(5) Appeals per 1000
JudgeGPT Access	0.142	0.243	0.159	–	–
and Targeted Training	(0.053)	(0.074)	(0.066)	–	–
p-value	0.007	0.001	0.016	–	–
Romano-Wolf p	0.043	0.008	0.050	–	–
Sharpened q	0.008	0.004	0.011	–	–
Westfall-Young p	0.016	0.004	0.016	–	–
Bonferroni p	0.014	0.003	0.016	–	–
Šidák p	0.014	0.003	0.016	–	–
JudgeGPT Access	0.069	0.266	0.080	–	–
and Generic Training	(0.078)	(0.102)	(0.099)	–	–
p-value	0.379	0.009	0.421	–	–
Romano-Wolf p	0.611	0.043	0.611	–	–
Sharpened q	0.391	0.028	0.391	–	–
Westfall-Young p	0.611	0.032	0.611	–	–
Bonferroni p	0.758	0.027	0.758	–	–
Šidák p	0.614	0.027	0.614	–	–
Share	–	–	–	9239.049	-0.846
Judges Targeted	–	–	–	(600.440)	(0.482)
p-value	–	–	–	0.000	0.082
Sharpened q	–	–	–	0.001	0.043
Westfall-Young p	–	–	–	0.000	0.464
Bonferroni p	–	–	–	0.000	0.082
Šidák p	–	–	–	0.000	0.082
Strata FE / District & Year FE	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes
Observations	1070	1070	1070	472	445
Control Mean / Mean Dep. Var.	0.548	3.346	3.500	29200.720	19.352

Note: The first three outcomes are individual-level IV estimates (Arm A = JudgeGPT + targeted training, Arm B = JudgeGPT + generic training), with strata fixed effects and controls. The last two outcomes are district-level reduced-form estimates of the share of judges with JudgeGPT access on court outcomes, with district and year fixed effects and court-level controls. P-values are reported with stars based on conventional thresholds. Multiple-hypothesis adjustments: Romano-Wolf stepdown p-values (only for the three individual outcomes), Westfall-Young stepdown p-values, Bonferroni and Šidák corrections, and sharpened q-values (Benjamini-Krieger-Yekutieli). All resampling procedures use 999 replications. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Appendix A2. District Court Statistics

Overview

The administrative analysis links individual-level treatment assignments to district-level court outcome data from the Law and Justice Commission of Pakistan (LJCP). This appendix describes, step by step, how the district-level treatment-exposure variable is constructed, what data sources underlie the denominator (total judges per district), and why the non-targeted-training arm cannot be separately identified in the administrative analysis.

Step 1: Assigning Experimental Judges to Districts

Each judge participating in the experiment completed a baseline survey that included, among other items, the name of the court in which they currently serve. Because the overwhelming majority of trial courts in Pakistan are named after the district in which they are located (e.g., "Karachi," "Lahore," "Islamabad"), the court field could in most cases be used directly as the district identifier. A minority of respondents entered a sub-district court name, a city name, or an abbreviation that did not map unambiguously to a district. These entries were resolved through manual review: each non-standard court name was cross-referenced with the official list of districts published by each provincial government, and the corresponding district was assigned by hand. Through this procedure, every judge in the experimental sample — whether assigned to targeted training with *JudgeGPT* (Treatment A), generic training with *JudgeGPT* (Treatment B), or the active control group (Group 3C) was linked to exactly one district.

Step 2: Counting Treated Judges per District

After district assignment, we aggregate the individual-level treatment indicator to the district level. For each district d, we count the number of participating judges assigned to targeted training with *JudgeGPT* (Treatment A, Group 1):

$$N_d^{\text{treated}} = \sum_{j \in \text{Treat. A}} \mathbf{1}[\text{district}(j) = d].$$

Step 3: Obtaining the District-Level Denominator from Judicial Statistics

To express district-level treatment intensity as a *share* rather than a raw count and make the measure comparable across districts of very different sizes, we divide N_d^{treated} by the total number of active judges serving in district d at the relevant point in time.

The denominator is drawn from the *Judicial Statistics of Pakistan*, an annual publication of the Law and Justice Commission of Pakistan (LJCP). Each edition reports, separately for every district, the sanctioned strength of the subordinate judiciary and the number of judges actually working in the field, disaggregated by gender and cadre category. The categories recorded in these reports include: judges working in-field (male and female), judges working ex-cadre (male and female), and the number of vacant sanctioned posts. Figure A14 and Figure A15 show representative pages from the LJCP Judicial Statistics report, illustrating the structure of the raw source data.

Figure A14: Number of Judges in each district from the *Judicial Statistics of Pakistan*

Districts	Civil Judges/Judicial Magistrates/Family Judges etc.				
	Sanctioned Strength	Working in field		Working in ex-Cadre	
		Male	Female	Male	Female
Attack	18	14	2		
Bahawalnagar	45	16	2		
Bahawalpur	52	26	5	2	
Bhakkar	20	11	1		
Chakwal	18	10	4		
Chiniot	20	8	2		
Dera Ghazi Khan	28	14			
Faisalabad	112	45	25		
Gujranwala	71	29	13		

Note: Representative page from the *Judicial Statistics of Pakistan* published by the Law and Justice Commission of Pakistan (LJCP), showing the structure of the raw district-level data on sanctioned strength, judges working in field (male and female), judges working ex-cadre (male and female). These figures are aggregated across cadre categories to obtain the total number of active judges per district.

Figure A15: Case statistics from the *Judicial Statistics of Pakistan*

Panel A: Category-wise Institution, Disposal, and Balance (2024)

1.10 Consolidated statement of category-wise Institution, Disposal and Balance of Cases at the Supreme Court of Pakistan during the Year 2024

Sr. #	Categories	Institution during the year
1	Civil Appeal	1994

Panel B: Annual Caseload, Multan Bench (2015–2024)

Multan Bench - 1st Bench Registry

Year	Disposal
2015	39 503
2016	24 116
2017	36 259
2018	30 875
2019	32 275
2020	27 168
2021	38 782
2022	39 254
2023	33 581
2024	**32 359**

Note: This figure displays representative excerpts from the *Judicial Statistics of Pakistan* published annually by the Law and Justice Commission of Pakistan (LJCP), illustrating the two main types of tables used to construct outcome and control variables in the administrative analysis. Panel A disaggregates caseload by legal category (civil appeals, criminal appeals, constitutional petitions, etc.), providing the raw counts used to construct the appeals controls in the district-level regressions. Panel B presents annual pending cases, institution, and disposal for the Multan Bench (1st Bench Registry), 2015–2024.

To obtain the total number of *active* judges in district d in year t, we sum across all working categories: in-field male, in-field female, ex-cadre male, and ex-cadre female.

Figure A16 displays a representative excerpt from the cleaned panel dataset constructed from these reports. The panel covers all districts for which LJCP statistics were available over the 2021–2024 period.

Figure A16: Excerpt from the cleaned district-year panel constructed from the LJCP

Province	District	Year	Judges_Sanctioned	Judges_Working_Male	Judges_Working_Female	Judges_Ex_cadre_Male	Judges_Ex_cadre_Female
Islamabad Capital Territory	Islamabad East	2023	34	29	8	0	0
Islamabad Capital Territory	Islamabad West	2023	28	23	7	0	0
Khyber Pakhtunkhwa	Abbottabad	2023	35	16	12	3	0
Khyber Pakhtunkhwa	Bajaur	2023	7	7	0	0	0
Khyber Pakhtunkhwa	Bannu	2023	28	18	6	3	0
Khyber Pakhtunkhwa	Battagram	2023	9	5	1	0	0
Khyber Pakhtunkhwa	Buner	2023	14	8	4	1	0

Note: Excerpt from the cleaned district-year panel constructed from the LJCP *Judicial Statistics of Pakistan* reports. Each row corresponds to one district–year observation and contains the province, district name, year, sanctioned strength, the number of judges working in field and ex-cadre (by gender), and the number of vacant posts. The total active judge count used as the denominator in treatment-share construction is the sum of the four working columns.

Step 4: Constructing the District Share of Targeted-Training Judges

The district-level treatment-exposure variable is defined as the fraction of active district judges who were assigned to targeted training with *JudgeGPT*:

$$\text{ShareTreated}_d = \frac{N_d^{\text{treated}}}{N_{d,t^*}^{\text{total}}},$$

where t^* is the year immediately preceding the rollout of the experiment (2023). Using the pre-rollout judge count as the denominator ensures that the exposure measure is not contaminated by any mechanical changes in judicial staffing that might themselves be a consequence of the intervention. This variable is the key regressor in all district-level regressions reported in the main text.

Balance Tests at the District Level

Because treatment was randomized at the judge level rather than the district level, districts in our administrative sample are not guaranteed to be balanced on pre-determined characteristics. We conduct balance tests on a set of district-level covariates observed prior to the intervention. The results, reported

in Appendix Figure A4, show no statistically significant relationship between the district-level treatment share and any of these pre-determined characteristics.

Why the General-Training Arm Is Not Separately Identified in District-Level Analysis

The administrative analysis exploits variation in $\texttt{ShareTreated}_d$ but does not separately estimate the effect of the non-targeted-training arm (Treatment B, i.e., judges receiving *JudgeGPT* access with general rather than targeted training). This is not a modeling choice but a data limitation that arises from two compounding constraints.

First, the LJCP *Judicial Statistics of Pakistan* are published annually with a significant lag. At the time of the administrative analysis, the most recent available edition covers the calendar year 2024. The relevant post-treatment window for the experiment therefore spans 2024 at most, and data for 2025 have not yet been published. This restricts the post-treatment panel to a single post-intervention year.

Second, and more importantly, the non-targeted-training rollout (Treatment B) and the targeted-training rollout (Treatment A) overlapped in calendar time. Both treatment arms were deployed over the same experimental period, meaning that—at the district level—the share of judges receiving targeted training and the share receiving non-targeted training are mechanically correlated. A district that by chance received more experimental judges tends to have higher shares of *both* arms simultaneously. With only a single post-treatment year of administrative outcomes available, it is not possible to separately identify the two treatment-share variables: the design does not provide enough independent variation between the two arms at the district level within the available outcome window to estimate their effects separately. We therefore restrict the administrative analysis to the share of targeted-training judges, which is the arm for which take-up and platform engagement are substantially higher, and for which the individual-level analysis provides the strongest evidence of behavioral change.

Appendix A3. Alternative Measures of Cases Resolved

Relation to Appendix A2

Appendix A2 described in detail how the district-level treatment-exposure variable is constructed. In brief, we aggregated individual-level treatment assignments determined by judge-level randomization to the district level by counting the number of participating judges assigned to targeted training with *JudgeGPT* in each district, and then dividing that count by the total number of active judges in the district as reported in the *Judicial Statistics of Pakistan* published by the Law and Justice Commission of Pakistan (LJCP). The resulting variable, ShareTreated_d, is the fraction of district judges who received targeted training with the AI tool.

Although treatment was randomized at the judge rather than district level, approximate district-level balance is a natural implication of the design: judges were assigned independently within training group, with no district-level targeting. Appendix Figure A4 confirms this intuition. The district share of judges assigned to *JudgeGPT* is not systematically related to the number of judges, pre-treatment filings, pending or resolved cases, appeals, or judicial vacancies. We therefore use district treatment intensity as a source of variation in the administrative analysis while retaining the judge-level assignment as the underlying randomized design when data allows.

Baseline Specification

The main administrative analysis estimates the following difference-in-differences specification:

$$Y_{dt} = \alpha_d + \alpha_t + \beta \left(\text{ShareTreated}_d \times \mathbf{1}[t = 2024]\right) + \mathbf{\Gamma}'\mathbf{X}_{dt} + \varepsilon_{dt}, \tag{4}$$

where Y_{dt} is the number of cases resolved in district d in year t, α_d and α_t are district and year fixed effects, and $\mathbf{X}_{dt}$ is a vector of time-varying district-level controls, including the number of sitting judges and vacant judge positions. Standard errors are clustered at the district level.

The treatment-intensity measure, ShareTreated_d, is the number of judges in district d assigned to JudgeGPT with targeted training in the March 2024 cohort, divided by the district's pre-treatment number of active judges. The coefficient β therefore measures how the change in case resolution in 2024 varies with the district's exposure to the randomized treatment. For example, moving from zero exposure to a treatment share of s corresponds to a change of $s\beta$ cases resolved.

Alternative Count-Based Specification

The baseline treatment measure is a share,

$$\text{ShareTreated}_d = \frac{\text{TreatedJudges}_d}{\text{TotalJudges}_d},$$

and therefore incorporates both the number of treated judges and the size of the district bench. As a robustness check, we examine whether the results are similar when treatment intensity is instead measured using the number of treated judges directly.

Because some districts received no treated judges in the March 2024 cohort, we use $\log(1 + \text{TreatedJudges}_d)$, which retains zero-exposure districts in the estimation sample. A representative specification is

$$\widetilde{Y}_{dt} = \alpha_d + \alpha_t + \beta_{\text{count}} \left(\log(1 + \text{TreatedJudges}_d) \times \mathbf{1}[t = 2024]\right) + \mathbf{\Gamma}'\mathbf{X}_{dt} + \varepsilon_{dt}, \tag{5}$$

where $\widetilde{Y}_{dt}$ denotes the outcome standardized to have mean zero and standard deviation one within the estimation sample. As in the baseline analysis, the specifications control for judicial capacity, including the number of sitting judges and vacant judge positions. We report progressively more saturated specifications with district and year fixed effects and additional district-level controls.

This count-based specification provides a complementary robustness check that does not mechanically scale treatment by the size of the district bench. If the estimated effect reflects greater exposure to JudgeGPT rather than an artifact of the share-based treatment measure, the estimated relationship should remain positive under this alternative parameterization.

TABLE A16: JudgeGPT Access and Court Efficiency

	(1)	(2)	(3)	(4)
	Cases Resolved			
Log JudgeGPT and Targeted Training × Post	0.349***	0.468**	0.480**	0.402*
	(0.110)	(0.205)	(0.212)	(0.207)
Year Fixed Effects	Yes	Yes	Yes	Yes
District Fixed Effects	No	Yes	Yes	Yes
# of Judges in the District	No	No	Yes	Yes
Controls	No	No	No	Yes
Adj. R^2	.751	.769	.77	.804
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. *Log JudgeGPT and Targeted Training × Post* equals log number of judges assigned *JudgeGPT* with targeted training, 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) additionally controls for the share of criminal cases among cases filed and the share of criminal cases among cases resolved. The table also reports the number of clusters and the number of observations. The sample includes district-years from 2021 to 2024 with cases resolved as the outcome, district and year fixed effects, and controls. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE A17: JudgeGPT Access and Court Efficiency by Case Type.

	(1)	(2)	(3)	(4)
	Civil Cases Resolved		Criminal Cases Resolved	
Log JudgeGPT and Targeted Training × Post	0.210***	0.154**	0.631**	0.546*
	(0.075)	(0.068)	(0.281)	(0.279)
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	.894	.912	.659	.702
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. *Log JudgeGPT and Targeted Training × Post* equals log number of judges assigned *JudgeGPT* with targeted training, interacted with a post indicator for 2024. Columns (1) and (3) control for the number of judges in the district, while columns (2) and (4) additionally control for the share of criminal cases among cases filed and the share of criminal cases among cases resolved. District Fixed Effects and year fixed effects are included. The sample includes district-years from 2021 to 2024 with civil cases resolved and criminal cases resolved as the outcome, district and year fixed effects, and controls. Adjusted R^2 is reported. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

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, the 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 religion bias.

iv) Potential Harms from Data Collection or Research Protocols

To mitigate risks related to data privacy, JudgeGPT data were saved on a password-protected server available only to the research team. 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. Pre-Analysis 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 rollout 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 generic training (a general "technology and law" course with information on risk mitigation), 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 B1 under alternative fixed-effects specifications. Second, we study AI tool utilization. These results are reported in Table 2 and, for completeness, in Column (2) of Table B1, 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 and legal arguments – measured by references to cases/statutes. These results are presented in Table B2. As shown in Appendix Table B2, 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. Fourth, we examine perceptions of AI utility, reported in Table 4 and, for completeness, in Column (1) of Table B1 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 (3) of Table B1 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 B1, 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 B1: 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	0.231***	0.137**	0.157**	83.223***
	(0.080)	(0.056)	(0.071)	(3.685)
JudgeGPT Access and Generic Training	0.285***	0.069	0.101	16.987***
	(0.109)	(0.083)	(0.105)	(2.072)
Adj. R^2	.012	.024	.009	.121
Control Mean	3.350	0.544	3.495	0.000
Observations	1065	1065	1065	1065
Panel B: Court FE				
JudgeGPT Access and Targeted Training	0.240	0.100	0.218*	83.800***
	(0.151)	(0.104)	(0.119)	(9.689)
JudgeGPT Access and Generic Training	0.208	0.026	0.144	15.591
	(0.215)	(0.140)	(0.201)	(13.127)
Adj. R^2	-.065	.019	.002	-.059
Control Mean	3.350	0.544	3.495	0.000
Observations	1064	1064	1064	1064
Panel C: Strata + Court FE				
JudgeGPT Access and Targeted Training	0.244	0.110	0.208*	84.146***
	(0.151)	(0.102)	(0.116)	(10.747)
JudgeGPT Access and Generic Training	0.245	0.057	0.165	14.922
	(0.214)	(0.140)	(0.197)	(15.413)
Adj. R^2	-.056	.025	.008	-.046
Control Mean	3.350	0.544	3.495	0.000
Observations	1064	1064	1064	1064

Note: The table reports treatment-effect estimates separately with different fixed effects. Panel A uses no fixed effects, Panel B uses court fixed effects, Panel C uses both strata and court fixed effects. Standard 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 B3), by province (Table B4), and by case type (civil vs. criminal; Table B5).*[16]

[16]For all heterogeneity analyses, we exclude Gunning Fog and legal-referencing outcomes, which display very limited variation. Table B2 reports the corresponding estimates for the full sample and shows null effects of both treatment arms.

TABLE B2: 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 Generic 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 Generic Training			-0.016	0.054
			(0.026)	(0.094)
Anderson-Rubin P-val.			.715	.836
Mean Dep. Var.	.008	.011	.008	.011
	# of Legal References			
Assigned JudgeGPT Access and Targeted Training	-4.522	-10.062		
	(11.130)	(10.404)		
Assigned JudgeGPT Access and Generic 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 Generic 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 Robust Standard Errors in parentheses. The dependent variable is a Gunning Fog readability metric standardized to mean 0 and standard deviation 1, calculated using the mean number of words in sentences and the number of complex words in Panel A and the number of legal references, calculated as the number of references to statutes and precedents in the case. The dependent variable is taken as a mean for pre- and post-treatment cases for each judge. *JudgeGPT Access and Associated Training* is a binary indicator of a judge with the training assigned who 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 (first Control Group). The control group did not receive 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 treatment B (Group 3B) are instrumented with initial assignment of treatment and generic training. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE B3: 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	0.298***	0.182***	0.232***	89.702***
	(0.100)	(0.067)	(0.089)	(5.383)
JudgeGPT Access and Generic Training	0.370**	0.042	0.249*	12.947**
	(0.152)	(0.112)	(0.147)	(6.146)
Adj. R^2	.086	.065	.086	.148
Control Mean	3.299	0.552	3.448	0.000
Observations	590	590	590	590
Panel B: Above Median				
JudgeGPT Access and Targeted Training	0.121	0.095	0.058	71.105***
	(0.114)	(0.090)	(0.103)	(5.551)
JudgeGPT Access and Generic Training	0.149	0.123	-0.055	16.304***
	(0.143)	(0.112)	(0.143)	(4.805)
Fixed Effects	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Adj. R^2	.049	.06	.027	.095
Control Mean	3.444	0.528	3.583	0.000
Observations	475	475	475	475

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

TABLE B4: 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*	0.029	0.110	67.321***
Targeted Training	(0.095)	(0.072)	(0.084)	(4.779)
JudgeGPT Access and	0.157	-0.028	0.036	14.444***
Generic Training	(0.144)	(0.108)	(0.137)	(5.336)
Adj. R^2	.095	.08	.071	.145
Control Mean	3.449	0.539	3.562	0.000
Observations	490	490	490	490
Panel B: Sindh Subsample				
JudgeGPT Access and	0.276	0.251**	0.230	104.468***
Targeted Training	(0.214)	(0.122)	(0.178)	(12.884)
JudgeGPT Access and	0.475*	0.320*	0.360*	20.237
Generic Training	(0.250)	(0.164)	(0.218)	(14.407)
Adj. R^2	.082	.066	.053	.145
Control Mean	3.343	0.514	3.457	0.000
Observations	214	214	214	214
Panel C: Khyber Pakhtunkhwa Subsample				
JudgeGPT Access and	0.252*	0.272**	0.147	91.028***
Targeted Training	(0.133)	(0.111)	(0.141)	(11.319)
JudgeGPT Access and	0.244	0.046	-0.082	31.452***
Generic Training	(0.184)	(0.151)	(0.201)	(10.572)
Adj. R^2	.049	.096	.078	.202
Control Mean.	3.229	0.343	3.371	0.000
Observations	225	225	225	225
Panel D: Balochistan Subsample				
JudgeGPT Access and	0.494*	0.174	0.475*	167.435**
Targeted Training	(0.295)	(0.351)	(0.253)	(84.817)
JudgeGPT Access and	0.466	0.591	0.647***	68.791
Generic Training	(0.290)	(0.375)	(0.244)	(74.846)
Adj. R^2	-.113	-.092	-.09	-.007
Control Mean	3.636	0.364	3.636	0.000
Observations	66	66	66	66
Panel E: Other Provinces Subsample				
JudgeGPT Access and	0.715***	0.431*	0.599***	125.503***
Targeted Training	(0.116)	(0.233)	(0.173)	(23.233)
JudgeGPT Access and	-0.011	0.103	-1.578***	-89.284
Generic Training	(0.376)	(0.583)	(0.367)	(110.492)
Fixed Effects	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Adj. R^2	.373	-.051	.233	.002
Control Mean	3.000	0.333	3.167	0.000
Observations	39	39	39	39

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, and Panel E uses Islamabad (federal territory). Standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE B5: 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	0.246	-0.033	0.143	82.941***
Targeted Training	(0.239)	(0.142)	(0.145)	(19.650)
JudgeGPT Access and	0.368	0.074	0.014	0.921
Generic Training	(0.350)	(0.258)	(0.304)	(37.705)
Adj. R^2	.012	-.051	.066	-.007
Control Mean	3.373	0.593	3.525	0.000
Observations	526	526	526	526
Panel B: Civil				
JudgeGPT Access and	0.218	0.130	0.428**	83.295***
Targeted Training	(0.261)	(0.150)	(0.204)	(14.737)
JudgeGPT Access and	0.088	-0.105	0.287	-1.432
Generic Training	(0.447)	(0.208)	(0.378)	(19.570)
Fixed Effects	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Adj. R^2	-.248	.02	-.099	.176
Control Mean	3.318	0.477	3.455	0.000
Observations	322	322	322	322

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

TABLE B6: Tool Usage and AI Support

	AI Support			
# Logins	0.108***	0.118***	0.104***	0.095***
	(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.	1065	1065	1064	1064
Control Mean	3.350	3.350	3.350	3.350

Note: The table reports treatment-effect estimates of the number of logins on support for generative AI. Standard 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 5, which 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 generic training. The targeted-training 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 0.14 in the ITT specifications, while the corresponding 2SLS estimates for JudgeGPT access with targeted training are about 0.15. Relative to the control mean of 0.54, this corresponds to roughly a 25 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 two licensed Pakistani attorneys. Table 8 reports the results.*

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 B3. Judicial Statistics Yearbook

Figure B1: Judicial Statistics Yearbook and Cases Closed

Panel A: Title Page of Judicial Statistics Year Books



Judicial Statistics of Pakistan

2024

National Judicial (Policy Making) Committee

Law and Justice Commission *of* Pakistan
Supreme Court of Pakistan Building
Constitution Avenue, Islamabad

Panel B: Cases Closed Per Year

	2020	2021	2022	2023	2024
Civil Cases	1016107	1606938	931579	1473642	1698852
Criminal Cases	1071458	2071548	866985	1761082	3197649
Total Cases	2088722	3736477	1800320	3301618	4955233

Note: Panel A shows the title page of Judicial Statistics of Pakistan 2024, published by the National Judicial Policy Making Committee at the Law and Justice Department of Pakistan (LJCP). These reports constitute the core source for our administrative outcome data, recording annual district-level case-resolution statistics. Panel B reports the annual number of cases closed from 2020 to 2024. Civil and criminal cases are those recorded in the administrative data as ordinary civil suits and ordinary criminal proceedings, respectively. These categories do not sum to total cases because the data separately record matters under special jurisdictions, including family cases and proceedings under special statutes.

APPENDIX C. SUPPLEMENTAL ANALYSIS

C1. Lawyer Validation of GPT-Based Judgment Quality Comparisons

To validate the GPT-based judgment quality measure used in our main analysis, we commissioned two independent licensed attorneys in Pakistan to evaluate a random sample of 90 judgment pairs drawn from the submitted-opinion sample in treatment Groups 3A/B/C. Each pair consists of one judgment written before the intervention and one judgment written afterward by the same judge. For each pair, the lawyers received two anonymized texts, labeled A and B, without information identifying the judge or indicating which judgment was written before or after the intervention. They evaluated the relative quality of the two judgments using the same -2 to $+2$ scale applied by GPT: $+2$ if A was much better, $+1$ if A was slightly better, 0 for a tie, -1 if B was slightly better, and -2 if B was much better. The evaluation criteria followed the dimensions used in the GPT prompt, including clarity, conciseness, factual presentation, and legal reasoning.

Agreement Metrics. We link each lawyer evaluation to the corresponding GPT and Gemini ratings using a crosswalk from the annotation files to the underlying judgment-pair identifiers. Directional agreement equals one when two relevant ratings identify the same judgment in a pair as superior and zero when they identify different judgments as superior. Comparisons in which either relevant rating assigns a tie are excluded from the corresponding agreement calculation. Pooled lawyer ratings stack the separate evaluations of the two lawyers when comparing lawyer assessments with GPT. Consensus Lawyer Rating refers to the subset of comparisons for which both lawyers independently identify the same superior judgment.

Results. Appendix Table A5 presents four comparisons. Panel A reports the lawyer benchmark: Lawyer A and Lawyer B identify the same superior judgment in 73.0 percent of non-tied comparisons. Panel B reports the direct validation of the GPT-based measure used in the main analysis: pooled lawyer ratings agree with GPT in 70.6 percent of non-tied comparisons, a rate close to the lawyer benchmark. Panel C restricts attention to comparisons for which both lawyers independently identify the same superior judgment; within this subset, agreement with GPT rises to 75.9 percent. Panel D reports an additional cross-model reliability diagnostic: GPT and Gemini agree on the direction of quality differences in 70.0 percent of non-tied comparisons. The GPT–Gemini comparison is not used to construct the main judgment quality measure.

Interpretation. The lawyer benchmark indicates that relative judgment quality is not evaluated mechanically, even by experienced legal readers. Against this benchmark, the agreement between pooled lawyer ratings and GPT is similar in magnitude, and GPT agreement increases when both lawyers concur on which judgment is superior. The cross-model comparison further shows that GPT–Gemini agreement is of a comparable magnitude. We therefore interpret the GPT-based measure as an informative, but necessarily noisy, proxy for relative judgment quality in the submitted-opinion sample.

C2. Chat Classification Prompt

We use the following complete prompt for the updated task-and-intention chat classification:

```
You will analyze the interactions of a judge with an AI LLM Assistant ("chatbot").

You will be given a sequence of prompts that a judge entered into the chatbot, without
the chatbot's responses.

Your task is to identify (1) which TASKS the judge is performing, and (2) the judge's
overall INTENTION in using the AI.

Base your analysis only on the content of the judge's prompts. Do not infer intent from
imagined chatbot responses.

========================================
PART 1: TASKS
========================================

More than one task category may apply. Select all that are meaningfully represented.
For each selected category, annotate it as primary (central to the conversation) or
secondary. A category is primary if removing it would fundamentally change the
description of what the judge is using the chatbot for.

For each selected category, briefly explain why it applies and provide short verbatim
snippets from the prompts.

### Task categories

1. Text Improvement
Editing or improving text the judge has already written: correcting grammar and
spelling, improving style, register, or flow, making language more formal or legalistic,
simplifying for a non-legal audience, anonymizing or redacting PII.
- The judge provides existing text and asks for it to be improved or modified.
- Examples: "correct the grammar", "make this more formal", "rephrase in better words",
  "simplify for the public", "anonymize this judgment"
- Edge cases: If the judge provides a draft and asks to expand or complete it
  substantially, that may be Text Generation rather than Text Improvement. If asked to
  both correct grammar and rewrite in a better style, classify as Text Improvement.

2. Text Generation
Producing new written content from scratch: drafting judgments, court orders, legal
```

documents, correspondence, or any other written output where the judge does not provide a substantial existing draft.
- Examples: "write a judgment", "draft bail order", "write an order dismissing the suit", "draft a letter to the government official", "write an application"
- Edge cases: Distinct from Text Improvement - here the judge is asking AI to create content, not refine existing content. If the judge provides only a decision or bare facts and asks AI to write the opinion, classify as Text Generation regardless of whether they also ask for Decision Support.

3. Decision Support
Helping the judge determine the appropriate ruling in a specific case: who should win, whether to convict or acquit, whether to grant bail or an injunction, what remedy or sentence to impose. This is about the outcome of the case - not the reasoning that supports it and not the written opinion that expresses it.
- Examples: "what should my decision be given these facts", "who should win this case", "should bail be granted based on these facts", "whether the accused should be acquitted or convicted"
- Edge cases: Requires specific case facts. A purely abstract legal question ("what is the law on bail") is Legal Research - Open QA (5), not Decision Support. Analysis of why a decision is correct belongs in Critiquing or Evaluating Arguments (8). Writing the opinion belongs in Text Generation (2). If the judge asks both what to decide AND to write the opinion, mark both Decision Support and Text Generation.

4. Legal Research - Finding Sources
Searching for specific legal authorities: statutes, case law, citations, precedents. The judge wants to be pointed to sources that exist.
- Examples: "find case law on adverse possession", "provide precedents for bail", "what is the citation for Zainab murder case", "find latest similar cases", "compare these two judgments"
- Edge cases: If the judge asks what the law says on a topic without seeking citations, that is Legal Research - Open QA (5). If the judge asks to synthesize or compare already-identified cases, that still belongs here. If the judge asks for both an explanation and supporting case law, mark both (4) and (5).

5. Legal Research - Open Question Answering
Asking general legal questions to understand the law, concepts, definitions, procedures, or principles - without seeking specific citations and without applying the law to specific facts.
- Examples: "what is the difference between 392 PPC and 395 PPC", "explain mesne profit", "what is the limitation period for review applications", "what is the effect of non-recording a statement under Section 161"

- Edge cases: Distinct from Decision Support (3) - no specific case facts are being analyzed. Distinct from Finding Sources (4) - the judge wants an explanation, not a citation or case reference. If the judge asks for both a legal explanation and supporting case law, mark both (4) and (5).

6. Summarization or Closed Question Answering on a Provided Document
Seeking help in understanding a document the judge has pasted into the prompt: summarizing it, extracting facts, building timelines, answering specific questions about it, finding contradictions in testimony, or evaluating evidence within it.
- Examples: "summarize this judgment", "extract the key facts", "build a chronology", "find contradictions in this deposition", "what are the disputed facts in this record"
- Edge cases: The judge must provide the actual document or a substantial portion of it. Asking about a case by name without pasting text is Legal Research (4 or 5), not summarization. Evaluating witness testimony from a pasted deposition belongs here.

7. Translation
Converting text between languages, typically between English and Urdu.
- Examples: "translate this into Urdu", "translate into English"
- Edge cases: If the judge submits a prompt in Urdu but the goal is legal analysis, classify by the substantive task. Mark Translation only when language conversion is the explicit or obvious goal.

8. Critiquing or Evaluating Arguments
Critically assessing the quality, soundness, or weaknesses of legal reasoning, arguments, or conclusions - whether in a draft judgment, a party's argument, or a legal position.
- Examples: "critically analyze this judgment", "identify weaknesses in this reasoning", "argue against this position", "generate counterarguments", "what are the grounds of appeal", "is this legally sound"
- Edge cases: Distinct from Decision Support (3), which is about reaching a decision. This category is about evaluating the quality of reasoning or arguments. If the judge asks both what to decide and to critique the reasoning, mark both.

9. Personal Use and Other
Tasks unrelated to judicial or legal work (personal queries, financial advice, casual conversation, creative writing), or tasks that do not fit any other category (junk input, gibberish, unclassifiable content).
- Examples: stock market questions, travel planning, poetry, accidental code pastes

==========
PART 2: INTENTION

```
========================================

Identify the judge's single predominant intention for this conversation.

### Intention categories

Judge-Specified Task for AI
The judge has completed all intellectual work and is directing AI to produce a
specified output. No judgment calls are left to the AI - it is acting as a word
processor or formatter.
- Signals: judge specifies exactly what to produce; decision, reasoning, and content
  are already determined by the judge
- Example: "The defendant is acquitted. Draft the order." or "Correct the grammar."
- Edge case: Distinct from Substantive AI Delegation - in Judge-Specified Task for AI the
  judge has already decided everything; AI only produces the artifact.

Seeking Feedback / Evaluation from AI
The judge has a draft, a decision, or a line of reasoning and asks the AI to evaluate,
critique, validate, or improve it.
- Signals: judge shares their own work or reasoning and asks if it is sound, what is
  missing, or how to improve it
- Example: "Here is my draft judgment - critically analyze it." or "I think bail should
  be granted because of X - is my reasoning sound?"
Seeking Information from AI
The judge wants to understand the law, a legal concept, a procedure, or a principle.
No case-specific facts are provided and no judgment call is requested - the judge is
learning or looking something up.
- Signals: abstract question with no specific case facts; no draft or prior reasoning
  shared; the judge is building their own understanding
- Example: "What is the limitation period for X?" or "Explain mesne profit." or
  "What is the procedure under Section 145 CrPC?"
- Edge case: Distinct from Substantive AI Delegation - here the judge is seeking to
  understand the law themselves, not asking AI to apply it to a specific case.

Substantial Delegation to AI
The judge hands off a high-agency task with minimal guidance: asking AI to choose the
decision (who wins or what the ruling should be), generate the reasoning (why), or
produce the substantive writing (the opinion or order) - without providing their own
analysis, conclusion, or meaningful direction beyond the facts.
- Signals: specific case facts or a document provided + AI asked to decide, reason, or
  write; no judge input on the substance beyond supplying the raw material
```

```
- Example: "Here are the facts. Who should win?" or "Write a judgment for this case."
  with only facts and no stated reasoning or conclusion from the judge.
- Edge case: Distinct from Judge-Specified Task for AI - in Substantive AI Delegation
  the AI is making the substantive choices; in Judge-Specified Task for AI the judge
  already made them.

Unclear / Other
The prompts are too brief, ambiguous, or fragmentary to determine the predominant
intention.

========================================
OUTPUT FORMAT
========================================

Answer in the following JSON format:

{
  "tasks": [
    {
      "category": "<category name>",
      "role": "primary | secondary",
      "explanation": "<brief explanation>",
      "evidence": ["<verbatim snippet>", "<verbatim snippet>"]
    }
  ],
  "intention": "<Judge-Specified Task for AI | Seeking Feedback / Evaluation from AI | Seeking
  Information from AI | Substantial Delegation to AI | Unclear / Other>",
  "intention_explanation": "<one sentence explaining the intention classification>",
  "confidence": <1 | 2 | 3>
}

Here is the numbered sequence of prompts:
```

C3. LLM-Based Bias Detection in Judicial Opinions

In addition to the bias measures reported in the main text, we conduct additional 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 participating judges.

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 the bias measures to the judge level by averaging across each judge's post-treatment opinions. We then estimate the baseline both OLS and 2SLS specifications in Equation 1 and Equation 2:

$$\begin{aligned}\text{Bias}_j = {} & \pi_1 \,\text{JudgeGPT Access and Targeted Training}_j \\ & + \pi_2 \,\text{JudgeGPT Access and Generic Training}_j \\ & + \delta_s + X_j'\gamma + \varepsilon_j,\end{aligned}$$

where realized access and use in the two treatment conditions are instrumented with the corresponding randomized assignment indicators. The omitted category is the group assigned to generic training without access to *JudgeGPT*, δ_s denotes randomization-strata fixed effects, and X_j contains the baseline controls used in the main analysis. Standard errors are HC1 robust.

Results. Figure A6 reports no statistically significant effect of either treatment condition on either measure of bias. For the gender-bias score, the point estimates are negative for both AI with targeted training and AI without targeted training. For the pro-Muslim bias score, the estimate for AI with targeted training is close to zero, while the estimate for AI without targeted training is modestly negative. In each case, however, the 95 percent confidence interval includes zero and is sufficiently wide to encompass effects in both directions. We find no clear evidence that access to *JudgeGPT*, with or without targeted training, changes gender- or religion-related bias in judicial writing.

Appendix C4. Cost-Effectiveness of JudgeGPT

Costs of Running JudgeGPT

As we show in Table C1, the cost of running JudgeGPT fluctuated between 1.3k USD and 5.7k USD per month for infrastructure, averaging 2.4k USD per month for the time frame March 2024 to May 2026 – a period covering 27 months.

These costs include the charges for the API calls that JudgeGPT makes to the *LLM*, the search service that feeds into the *RAG*, the *web app* the judges use including a *database* hosting their conversation history, general *storage*, and the identity provider *Auth0* underlying the login mechanism. Not included are any infrastructure costs accrued during the development of the tool before March 2024 and any labor costs such as for the development of the assistant, obtaining the RAG knowledge base, maintenance, or user support over the entire development and deployment life cycle.

While the LLM charges scale with the number of inference tokens, the RAG, database, web app, storage, and identity service charges scale with a mixture of usage and number but behave more like fixed infrastructure costs when compared to LLM charges. In earlier phases of the RCT, the cost of inference – input tokens sent to the Azure-hosted OpenAI API as well as output tokens received from it – dominated costs. Over time, especially with the release of `GPT-4 family` that JudgeGPT used from August 2024 on, the dominance of inference costs decreased. Other variations in costs were caused by us scaling infrastructure components, such as the web app or the search service, according to the number of registered and monthly active users.

Merging in usage statistics from Table C2, we can analyze total cost in Table C1 on a more granular level. In Table C2, the number of active users is the number of users with at least one login in a given month.

We find that the average cost of running JudgeGPT is 2.73 USD (5.99 USD) per registered (active) user per month – 86% (70%) cheaper than providing a ChatGPT subscription to registered users at a 20 USD per user per month baseline. The average cost per conversation is 0.58 USD, whereby an average conversation consumes 44k tokens and comprises 2.8 user prompts. The token usage includes the user prompt, communication to and from the knowledge base (see Figure 3), as well as output tokens.

The average number of monthly active JudgeGPT users is 428, with 9.2 login sessions and 11.3 conversations per active user per month on average. A login session can last up to 7 days if the user is not inactive for more than 3 days at a time. Multiple devices for the same user require separate login sessions.

The number of registered users in Table C2 differs slightly from the expected number according to Figure 2. This is for three reasons:

First, the number of registered users includes 21 test accounts for RCT organizers. Second, for Group 3A we only distributed JudgeGPT accounts to those participants who responded to a mandatory on-boarding survey before the first Group 3A training session. A total of 71 participants from Group 3A did not satisfy that condition. Third, in November 2025 we provided 22 Private Secretaries from Peshawar High Court with JudgeGPT accounts.

As infrastructure costs are not available separate by user type (RCT participants, test users, and later onboarded Private Secretaries) and as assistants deployed in a similar setting will likely carry a similar

TABLE C1: Cost of Running JudgeGPT

	Cost [USD]							Total Cost per ... [USD]			
Month	LLM	RAG	Database	Web Server	Storage	Auth0	Total	Reg. User	Active User	Conversation	User Prompt
2024-03	2,646.84	1,724.05	1.20	1,118.82	0.43	175.00	5,666.33	11.36	15.07	0.63	0.20
2024-04	1,274.26	1,039.67	0.67	429.91	0.44	175.00	2,919.94	5.85	9.54	0.53	0.22
2024-05	1,936.22	1,088.19	0.85	201.98	0.44	175.00	3,402.68	6.75	10.22	0.51	0.19
2024-06	1,281.46	1,051.53	0.63	194.83	0.44	175.00	2,703.89	5.36	10.28	0.76	0.30
2024-07	1,319.57	1,067.54	0.47	197.79	0.44	175.00	2,760.81	5.48	10.62	0.81	0.32
2024-08	669.50	1,054.75	0.29	195.41	2.77	175.00	2,097.72	4.14	10.18	1.00	0.40
2024-09	516.65	969.60	1.49	327.68	9.33	175.00	1,999.74	2.01	3.22	0.19	0.06
2024-10	745.65	1,008.52	1.12	360.19	0.46	175.00	2,290.94	2.30	4.52	0.29	0.10
2024-11	493.39	1,000.05	0.95	357.16	0.47	175.00	2,027.02	2.03	4.26	0.32	0.11
2024-12	1,894.98	1,056.12	0.90	377.18	0.48	175.00	3,504.67	3.52	7.95	0.67	0.23
2025-01	2,480.85	1,072.66	3.44	383.09	0.49	175.00	4,115.53	3.60	6.70	0.42	0.15
2025-02	2,644.86	974.33	3.52	347.98	0.50	175.00	4,146.18	3.13	6.37	0.54	0.20
2025-03	2,278.91	1,063.43	6.97	379.80	0.49	175.00	3,904.60	2.94	5.64	0.43	0.15
2025-04	1,583.07	1,015.48	5.00	362.67	0.48	175.00	3,141.71	2.37	4.84	0.45	0.16
2025-05	227.84	982.22	4.18	350.79	0.45	175.00	1,740.48	1.31	2.95	0.25	0.09
2025-06	23.07	951.40	3.79	339.79	0.45	175.00	1,493.49	0.99	2.80	0.33	0.13
2025-07	34.85	951.76	4.07	339.92	0.43	175.00	1,506.04	1.00	2.75	0.27	0.10
2025-08	16.54	960.80	3.78	364.66	0.44	175.00	1,521.22	1.01	3.84	0.56	0.21
2025-09	31.47	925.78	3.90	365.72	0.44	175.00	1,502.31	1.00	3.04	0.35	0.13
2025-10	26.67	950.21	4.07	353.36	0.43	175.00	1,509.75	1.00	3.29	0.37	0.13
2025-11	13.79	916.45	3.89	340.81	0.43	175.00	1,450.37	0.95	3.47	0.56	0.23
2025-12	12.07	961.28	3.65	357.48	0.44	175.00	1,509.91	0.99	4.48	0.64	0.25
2026-01	10.32	939.27	4.01	349.29	0.53	175.00	1,478.42	0.97	4.72	0.74	0.27
2026-02	7.67	826.77	3.41	307.45	0.42	175.00	1,320.72	0.86	4.11	0.77	0.29
2026-03	6.30	920.23	3.88	342.21	0.42	175.00	1,448.04	0.95	5.49	1.07	0.41
2026-04	7.57	917.78	4.36	341.30	0.43	175.00	1,446.44	0.94	5.48	0.95	0.34
2026-05	12.83	862.39	3.56	349.60	0.43	**175.00**	1,403.81	0.92	6.00	1.16	0.43
Average	**822.12**	**1,009.34**	**2.89**	**360.62**	**0.87**	**175.00**	**2,370.84**	**2.73**	**5.99**	**0.58**	**0.21**
Total	**22,197.20**	**27,252.26**	**78.05**	**9,736.87**	**23.40**	**4,725.00**	**64,012.76**				

TABLE C2: Cost-Related JudgeGPT Usage Statistics

	Users						Per Active User			
Month	Registered	Active	Logins	Conversations	User Prompts	Tokens [1M]	Logins	Conversations	User Prompts	Tokens [1M]
2024-03	499	376	6,250	9,042	28,703	234.43	16.62	24.05	76.34	0.62
2024-04	499	306	3,845	5,517	13,506	109.48	12.57	18.03	44.14	0.36
2024-05	504	333	4,772	6,732	17,562	163.57	14.33	20.22	52.74	0.49
2024-06	504	263	2,790	3,546	9,008	109.36	10.61	13.48	34.25	0.42
2024-07	504	260	2,903	3,391	8,688	115.38	11.17	13.04	33.42	0.44
2024-08	507	206	1,694	2,101	5,253	65.89	8.22	10.20	25.50	0.32
2024-09	997	621	8,442	10,749	35,473	486.18	13.59	17.31	57.12	0.78
2024-10	997	507	6,232	7,807	22,637	439.30	12.29	15.40	44.65	0.87
2024-11	997	476	5,173	6,283	17,743	278.50	10.87	13.20	37.28	0.59
2024-12	997	441	4,397	5,221	15,167	376.77	9.97	11.84	34.39	0.85
2025-01	1,144	614	7,375	9,753	28,124	555.78	12.01	15.88	45.80	0.91
2025-02	1,325	651	6,491	7,734	20,956	488.55	9.97	11.88	32.19	0.75
2025-03	1,326	692	7,197	9,078	26,495	514.34	10.40	13.12	38.29	0.74
2025-04	1,326	649	5,809	7,037	19,499	421.89	8.95	10.84	30.04	0.65
2025-05	1,326	590	5,870	7,082	18,567	322.92	9.95	12.00	31.47	0.55
2025-06	1,509	533	3,889	4,492	11,658	149.86	7.30	8.43	21.87	0.28
2025-07	1,509	548	4,852	5,622	15,221	232.98	8.85	10.26	27.78	0.43
2025-08	1,509	396	2,079	2,704	7,195	110.37	5.25	6.83	18.17	0.28
2025-09	1,509	494	3,995	4,301	11,683	182.86	8.09	8.71	23.65	0.37
2025-10	1,509	459	3,793	4,093	11,231	175.84	8.26	8.92	24.47	0.38
2025-11	1,531	418	2,955	2,590	6,434	93.01	7.07	6.20	15.39	0.22
2025-12	1,531	337	2,052	2,344	5,955	80.31	6.09	6.96	17.67	0.24
2026-01	1,531	313	1,796	2,006	5,477	72.15	5.74	6.41	17.50	0.23
2026-02	1,531	321	1,773	1,715	4,482	54.55	5.52	5.34	13.96	0.17
2026-03	1,531	264	1,352	1,350	3,497	46.09	5.12	5.11	13.25	0.17
2026-04	1,531	264	1,509	1,520	4,275	53.62	5.72	5.76	16.19	0.20
2026-05	1,531	234	1,179	1,211	3,232	30.64	5.04	5.18	13.81	0.13
Average	**1,175**	**428**	**4,091**	**5,001**	**13,990**	**220.91**	**9.24**	**11.28**	**31.16**	**0.46**
Total			**110,464**	**135,021**	**377,721**	**5,965**				

overhead of additional users, we decided to consistently include the overhead users in our calculations.

Estimated Gains in Case Resolutions

The experimental and administrative results establish that access to *JudgeGPT* with targeted training raises the number of cases resolved without increasing the appeal rate. A natural policy question follows: how does the cost of providing AI-assisted tools to existing judges compare to the cost of hiring additional judges to achieve the same increase in judicial output? This appendix provides a back-of-the-envelope cost-benefit calculation to address this question.[17]

The cost saving from *JudgeGPT* arises through a single channel operating in two steps. First, each judge equipped with the tool resolves more cases per month, raising per-judge productivity above the no-AI baseline. Second, because the same caseload can then be handled by fewer, more productive judges, the bench required to clear a given workload shrinks - reducing the number of judicial positions the system must fund. The net wage-bill saving quantified below is the salary value of the judicial positions that this higher per-judge productivity renders unnecessary.

We calibrate the calculation using the following parameters, drawn from the experimental data, the administrative results, publicly available salary benchmarks, and the observed infrastructure costs of running *JudgeGPT*.

Judicial productivity without *JudgeGPT*. The summary statistics (Panel A of Table A3) report that the average judge in the pre-treatment sample decides 122.4 cases per month and the average number of judges in the district is 20. We take this as the baseline monthly productivity of a judge without access to the AI tool.

Treatment effect per judge. The most demanding specification in the administrative analysis (column 4 of Table 6) estimates that a one-unit increase in the district share of targeted-training judges raises total cases resolved by 9,239.049 per year. With an average district bench of 20 judges, the implied per-judge annual treatment effect is:

$$\Delta_{\text{annual}} = \frac{9{,}239.049}{20} \approx 461.95 \text{ cases per year} \approx 38.5 \text{ cases per month.}$$

A judge equipped with *JudgeGPT* therefore resolves approximately $122.4 + 38.5 = 160.9$ cases per month, a productivity gain of roughly **31.5 percent** relative to baseline.

Annual judge salary. We benchmark judicial compensation using publicly available salary data for Pakistani judges and magistrates.[18] The average gross annual salary of a judge or magistrate in Pakistan is

[17] These calculations are intended as illustrative rather than a complete estimate of the intervention's social return and exclude the fixed costs of developing and delivering the training, including instructor time.

[18] Source: SalaryExpert (https://www.salaryexpert.com/salary/job/judge-and-magistrate/pakistan). All PKR figures are converted to USD at the exchange rate of 1 USD = 278.30 PKR prevailing in June 2026.

reported as PKR 3,429,678, equivalent to approximately **$12,323 per year**, or **$1,027 per month**.

Cost of running *JudgeGPT*. Unlike a standard per-seat software subscription, *JudgeGPT* is a purpose-built system whose infrastructure costs scale with usage rather than with the number of licensed users. Total monthly infrastructure costs - covering the web application, database, LLM API calls via Azure-hosted OpenAI, the retrieval-augmented generation (RAG) search service, general storage, and the identity provider - fluctuated between approximately $1,300 and $5,700 per month over the period March 2024 to May 2026, averaging **$2,371 per month** across 27 months, as reported in Table C1. These figures exclude labor costs such as development, knowledge-base curation, maintenance, and user support.

Averaging the monthly ratio of infrastructure cost to monthly active users (defined as users with at least one login in a given month) across the 27 months yields an average cost of **$5.99 per active user per month**, the figure reported in the final row of Table C1. Pooling instead over the whole observation window - total infrastructure cost divided by total active-user-months - gives a very similar $5.54; we use the higher of the two as the conservative choice throughout. With an average of 428 monthly active users over the observation period, this per-user figure is substantially below the cost of an individual commercial ChatGPT subscription and provides the relevant marginal cost for the present calculation.

Interpretation of the Saving

Before presenting the calculation, we state precisely what it does and does not establish. The experiment identifies a causal increase in cases resolved per judge; it does not directly observe fewer judges, avoided appointments, or reduced budgets. The bench reduction below is therefore a modeled implication of the measured productivity gain, not an observed fiscal outcome, and the gain may in part reflect faster clearance of an existing backlog rather than a permanent contraction of the required bench. Accordingly, the figures below are best read as a *salary-equivalent capacity benchmark* - the wage-bill value of the judicial capacity that the productivity gain frees up - which the system could realise as savings only through corresponding staffing or budgetary decisions. We retain the compact language of "saving" throughout for readability, subject to this interpretation.

Comparing Costs for a Bench of 10 Judges. We illustrate the cost comparison using a representative district bench of 10 judges - a common size in the sample.

Step 1: Total caseload without *JudgeGPT*. A bench of 10 judges resolves:

$$Q_0 = 10 \times 122.4 = 1{,}224 \text{ cases per month.}$$

Step 2: Number of *JudgeGPT*-equipped judges required to match Q_0. Each judge with access to the AI tool resolves 160.9 cases per month. To handle the same caseload of 1,224 cases, the required number of judges is:

$$N^* = \frac{1{,}224}{160.9} \approx 7.6 \approx 8 \text{ judges.}$$

That is, **8 judges equipped with *JudgeGPT*** can resolve the same monthly caseload as **10 judges without it** - a reduction of 2 judicial positions, or 20 percent of the bench.

Step 3: Annual cost comparison. We compute the annual cost of the AI tool for 8 judges assuming all are active users each month, at \$5.99 per active user per month:

$$C_{JudgeGPT} = 8 \times \$5.99 \times 12 \approx \$575 \text{ per year.}$$

Table C5 summarises the full annual cost of each scenario.

TABLE C3: Parameters Used in the Cost-Benefit Calculation

Symbol	Parameter	Value
q_0	Cases resolved per judge per month (no AI)	122.4
Δ	Treatment effect per judge (cases/month)	+38.5
q_1	Cases resolved per judge per month (with AI)	160.9
	Productivity gain	+31.5%
s	Annual judge salary (USD)	\$12,323
c	*JudgeGPT* cost per active user per month	\$5.99
N	Bench size (illustrative district)	10 judges

Note: Baseline productivity q_0 from Panel A of Table A3. Treatment effect Δ derived from column 4 of Table 6 (9,239.049 cases/district-year ÷ 20 judges ÷ 12 months). Salary from SalaryExpert (PKR 3,429,678 at 1 USD = 278.30 PKR, June 2026). Per-user cost c = average of the monthly cost per active user over March 2024–May 2026 (27 months), taken from the final row of Table C1; over the same window average monthly infrastructure cost is \$2,371 and average monthly active users 428.

TABLE C4: Step-by-Step Cost-Benefit Calculation (10-Judge Bench)

Step	Quantity	Formula	Result
1	Baseline caseload	$N \times q_0 = 10 \times 122.4$	1,224 cases/mo
2	Judges needed with AI	$\frac{N \times q_0}{q_1} = \frac{1{,}224}{160.9}$	8 judges
3	Judges saved	$10 - 8$	2 judges
4	Salary cost, no AI	$10 \times \$12{,}323$	\$123,230/yr
5	Salary cost, with AI	$8 \times \$12{,}323$	\$98,584/yr
6	*JudgeGPT* cost	$8 \times \$5.99 \times 12$	\$575/yr
7	Total cost, with AI	$\$98{,}584 + \$575$	\$99,159/yr
8	**Annual saving**	$\$123{,}230 - \$99{,}159$	**\$24,071/yr**

Note: N = bench size, q_0 = baseline cases/judge/month (122.4), q_1 = cases/judge/month with *JudgeGPT* (160.9). Step 2 rounds up from 7.6 to 8. Step 6 is rounded from \$575.04. All monetary figures in USD; see Table C3 for parameter sources. This table decomposes the summary figures reported in Table C5.

TABLE C5: Annual Cost Comparison: *JudgeGPT* vs. Additional Judicial Appointments

	Scenario A **10 judges, no AI**	**Scenario B** **8 judges + *JudgeGPT***
Number of judges	10	8
Annual salary cost (USD)	\$123,230	\$98,584
Annual *JudgeGPT* cost (USD)	—	\$575
Total annual cost (USD)	**\$123,230**	**\$99,159**
Monthly cases resolved	1,224	1,287
Annual saving (USD)	—	**\$24,071**

Note: Salary cost computed using the average gross annual salary of PKR 3,429,678 (\$12,323 at 1 USD = 278.30 PKR, June 2026) per judge (SalaryExpert). *JudgeGPT* infrastructure cost computed at \$5.99 per active user per month, averaged over 27 months (March 2024–May 2026), multiplied by 8 judges and 12 months. This per-user cost excludes labor costs for development, knowledge-base construction, maintenance, and user support. Monthly cases resolved in Scenario B reflect the per-judge monthly productivity of 160.9 cases.

Deploying *JudgeGPT* across an 8-judge bench costs approximately **\$99,159 per year** - a saving of **\$24,071**

(roughly 19.5 percent) relative to maintaining a 10-judge bench without AI assistance, while simultaneously resolving a slightly *higher* number of cases (1,287 vs. 1,224 per month). Crucially, the entire annual cost of the AI infrastructure for 8 judges ($575) amounts to under **five percent** of the annual salary of a single additional judge ($12,323), making the tool extraordinarily cost-effective as a substitute for marginal judicial hiring.

Return per Dollar Invested in *JudgeGPT*

The preceding comparison expresses the advantage of *JudgeGPT* as an annual saving for a representative bench. A complementary and more intuitive way to summarise the same result is to ask: *for every dollar spent on JudgeGPT infrastructure, how many dollars does the judicial system save?* This ratio isolates the efficiency of the AI investment itself, independent of the size of the bench.

We use the 10-judge bench from Table C5. Deploying *JudgeGPT* across the 8 judges required to match the baseline caseload costs $575 per year in infrastructure, while lowering the total wage bill enough to generate an annual saving of $24,071 relative to a 10-judge bench without the tool. The return on each dollar invested is therefore:

$$R = \frac{\text{Annual saving}}{\text{Annual } \textit{JudgeGPT} \text{ cost}} = \frac{\$24{,}071}{\$575} \approx 41.9.$$

That is, **every $1 invested in *JudgeGPT* yields approximately $41.9 in net savings** to the judicial system through reduced judicial-salary requirements. Equivalently, the infrastructure investment recovers its own cost roughly **42 times over** within a single year. Table C6 sets out the calculation step by step.

TABLE C6: Net Savings per Dollar Invested in *JudgeGPT* (10-Judge Bench)

Step	Quantity	Formula	Result
1	Salary cost, no AI (10 judges)	$10 \times \$12{,}323$	\$123,230/yr
2	Salary cost, with AI (8 judges)	$8 \times \$12{,}323$	\$98,584/yr
3	*JudgeGPT* cost	$8 \times \$5.99 \times 12$	\$575/yr
4	Total cost, with AI	$\$98{,}584 + \$575$	\$99,159/yr
5	Annual net saving	$\$123{,}230 - \$99{,}159$	\$24,071/yr
6	**Return per dollar invested**	$\frac{\$24{,}071}{\$575}$	**\$41.9**

Note: The calculation reuses the parameters and worked example of Table C5: baseline productivity of 122.4 cases per judge per month, 160.9 cases per judge per month with *JudgeGPT*, an annual judge salary of \$12,323 (PKR 3,429,678 at 1 USD = 278.30 PKR, June 2026; SalaryExpert), and an infrastructure cost of \$5.99 per active user per month averaged over March 2024–May 2026 (27 months). Step 6 divides the annual net saving (Step 5) by the annual *JudgeGPT* cost (Step 3). The return reflects *net* recurring-cost savings—i.e. reductions in the judicial wage bill net of infrastructure cost—and excludes one-time and ongoing labor costs of development, knowledge-base construction, maintenance, and user support.

This \$41.9-per-dollar figure understates the full benefit of the tool in one respect and rests on a recurring-cost-only accounting in another. It *understates* the benefit because it counts only the wage-bill saving from substituting AI-assisted productivity for marginal judicial hiring, and excludes the value of the accompanying appeal-rate reduction (see above), which would raise the return further. At the same time, it nets only the tool's *recurring infrastructure cost* against the saving: it excludes one-time and ongoing labor costs of development, knowledge-base construction, maintenance, security, and support, so it should not be read as a fully loaded benefit-to-cost ratio. If instead one values the *gross* productivity gain per judge - the salary value of the additional output - against the per-judge subscription cost, the implied ratio rises to the order of **171:1**. The truth for any given jurisdiction lies within this range, but even the recurring-cost net-savings measure leaves *JudgeGPT* an order of magnitude more cost-effective than expanding the bench.

Alternative Calculations

As a sensitivity check, we repeat the calculation under a substantially higher per-user cost of \$20 per active user per month - more than triple the observed \$5.99 - to approximate a scenario in which infrastructure, labor, and administrative overheads are bundled into the recurring per-user charge. All other parameters, including the 27-month observation window (March 2024–May 2026), are unchanged.

TABLE C7: Step-by-Step Cost-Benefit Calculation (10-Judge Bench, $20 per User)

Step	Quantity	Formula	Result
1	Baseline caseload	$N \times q_0 = 10 \times 122.4$	1,224 cases/mo
2	Judges needed with AI	$\frac{N \times q_0}{q_1} = \frac{1{,}224}{160.9}$	8 judges
3	Judges saved	$10 - 8$	2 judges
4	Salary cost, no AI	$10 \times \$12{,}323$	\$123,230/yr
5	Salary cost, with AI	$8 \times \$12{,}323$	\$98,584/yr
6	*JudgeGPT* cost	$8 \times \$20 \times 12$	\$1,920/yr
7	Total cost, with AI	$\$98{,}584 + \$1{,}920$	\$100,504/yr
8	**Annual saving**	$\$123{,}230 - \$100{,}504$	**\$22,726/yr**

Note: N = bench size, q_0 = baseline cases/judge/month (122.4), q_1 = cases/judge/month with *JudgeGPT* (160.9). Step 2 rounds up from 7.6 to 8. All monetary figures in USD; see Table C3 for parameter sources. This table decomposes the summary figures reported in Table C8.

TABLE C8: Annual Cost Comparison: *JudgeGPT* vs. Additional Judicial Appointments ($20 per User)

	Scenario A	**Scenario B**
	10 judges, no AI	**8 judges + *JudgeGPT***
Number of judges	10	8
Annual salary cost (USD)	\$123,230	\$98,584
Annual *JudgeGPT* cost (USD)	—	\$1,920
Total annual cost (USD)	**\$123,230**	**\$100,504**
Monthly cases resolved	1,224	1,287
Annual saving (USD)	—	**\$22,726**

Note: Salary cost computed using the average gross annual salary of PKR 3,429,678 (\$12,323 at 1 USD = 278.30 PKR, June 2026) per judge (SalaryExpert). *JudgeGPT* infrastructure cost computed at \$20 per active user per month, multiplied by 8 judges and 12 months. This per-user cost excludes labor costs for development, knowledge-base construction, maintenance, and user support. Monthly cases resolved in Scenario B reflect the per-judge monthly productivity of 160.9 cases.

TABLE C9: Net Savings per Dollar Invested in *JudgeGPT* (10-Judge Bench, $20 per User)

Step	Quantity	Formula	Result
1	Salary cost, no AI (10 judges)	$10 \times \$12{,}323$	\$123,230/yr
2	Salary cost, with AI (8 judges)	$8 \times \$12{,}323$	\$98,584/yr
3	*JudgeGPT* cost	$8 \times \$20 \times 12$	\$1,920/yr
4	Total cost, with AI	$\$98{,}584 + \$1{,}920$	\$100,504/yr
5	Annual net saving	$\$123{,}230 - \$100{,}504$	\$22,726/yr
6	**Return per dollar invested**	$\frac{\$22{,}726}{\$1{,}920}$	**\$11.8**

Note: The calculation reuses the parameters and worked example of Table C8: baseline productivity of 122.4 cases per judge per month, 160.9 cases per judge per month with *JudgeGPT*, an annual judge salary of \$12,323 (PKR 3,429,678 at 1 USD = 278.30 PKR, June 2026; SalaryExpert), and an infrastructure cost of \$20 per active user per month. Step 6 divides the annual net saving (Step 5) by the annual *JudgeGPT* cost (Step 3). Even at more than triple the observed per-user cost, the return remains an order of magnitude above unity.

Summary

These calculations suggest that providing district judges with access to *JudgeGPT* is inexpensive relative to expanding judicial capacity through additional hiring. At an average infrastructure cost of \$5.99 per active user per month, access costs approximately \$72 per judge per year. The annual salary of an additional judge, by comparison, is approximately \$12,323, implying a ratio of roughly **171 to 1**.[19]

This comparison should be viewed as a salary-equivalent capacity benchmark rather than as an estimate of the intervention's realised benefit-to-cost ratio. In particular, it does not establish that providing access to *JudgeGPT* is equivalent to hiring an additional judge, either in terms of output or in the range of tasks performed, nor does the experiment directly observe reduced headcount or budgets; realising the saving would require corresponding staffing decisions. The calculation also excludes the fixed costs of developing, adapting, deploying, and administering the platform and abstracts from variation in usage and treatment effects across judges. Subject to these limitations, it suggests that the recurring per-user infrastructure cost is modest when compared with conventional judicial personnel expenditures.

[19]This ratio is calculated by dividing the annual salary cost of an additional judge, \$12,323, by the annual infrastructure cost of providing one judge with access to *JudgeGPT*, \$72 ($\$5.99 \times 12$).