NBER WORKING PAPER SERIES

DOES AI ASSISTANCE ENHANCE OR ERODE EXPERTISE? EVIDENCE FROM A THREE-MONTH FIELD EXPERIMENT IN PATENT DRAFTING

David Autor
Tanya Rodchenko
Josh Martin
Zanna Iscenko
Scott Strand
David Pearl
Melissa Ferere

Working Paper 35720
http://www.nber.org/papers/w35720

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2026, Revised October 2026

This study was pre-registered in the AEA RCT Registry (AEARCTR-0015823) on May 21, 2025, under the title "AI & Expertise research with patent lawyers" (Autor et al., 2025). The registration can be found at https://www. socialscienceregistry.org/trials/15823. The MIT Committee on the Use of Humans as Experimental Subjects determined that MIT was not engaged in this research, so MIT human-subjects review was not required. The study was conducted by Google, and participants provided informed consent under Google's research procedures. We acknowledge generous input from Mauricio Carneiro, James Manyika, Ruth Porat, Kiera Russell, and Tom Shane (all of Google), as well as the Google Labs team. Autor gratefully acknowledges support from the Google Technology and Society Visiting Fellows Program, the Hewlett Foundation, the NOMIS Foundation, the Schmidt Sciences AI2050 Fellowship, the Smith Richardson Foundation, and the James and Cathleen D. Stone Foundation. Funding for the direct costs of this experiment was provided by Google. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed additional relationships of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w35720

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2026 by David Autor, Tanya Rodchenko, Josh Martin, Zanna Iscenko, Scott Strand, David Pearl, and Melissa Ferere. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Does AI Assistance Enhance or Erode Expertise? Evidence from a Three-Month Field Experiment in Patent Drafting

David Autor, Tanya Rodchenko, Josh Martin, Zanna Iscenko, Scott Strand, David Pearl, and Melissa Ferere

NBER Working Paper No. 35720
September 2026, Revised October 2026
JEL No. I24, I29, J0

ABSTRACT

AI assistance in expertise-intensive work presents a tension between performance and learning: AI tools may raise the quality of the work a professional delivers while stunting the expertise they would otherwise develop in producing it. In a pre-registered three-month randomized controlled trial, we gave 133 practicing patent lawyers at eleven U.S. intellectual property law firms access to a custom AI patent drafting assistant and measured both their performance while using AI and their professional judgment afterward without it. All work was scored by blinded expert patent attorneys. AI access raised the quality of work delivered on benchmark patent drafting tasks at 10 days (0.34 SD, $p = 0.03$) and 90 days (0.38 SD, $p = 0.01$), with larger gains among junior lawyers. After three months, all subjects redlined an existing patent application without AI, a core task of patent practice that requires expert judgment. Treated lawyers outperformed controls by 0.32 SD ($p = 0.04$), but this advantage was concentrated entirely among senior lawyers (0.45 SD, $p = 0.02$). Junior lawyers showed no average gain. Their scores instead bifurcated, with sharply fewer mediocre scores offset by more poor and more good ones, suggesting that AI assistance served as a springboard for some juniors and a cushion for others. Foundational expertise may be a prerequisite for extracting durable skill from AI-assisted practice.

David Autor
Massachusetts Institute of Technology
Department of Economics
and NBER
dautor@mit.edu

Tanya Rodchenko
Google
tanyarodchenko@google.com

Josh Martin
Google
joshuabmartin@google.com

Zanna Iscenko
Google
zannai@google.com

Scott Strand
Google
sstrand@google.com

David Pearl
Google
pearld@google.com

Melissa Ferere
Google
melissaferere@google.com

A randomized controlled trials registry entry is available at
https://www.socialscienceregistry.org/trials/15823

1 Introduction

Expertise in many professional domains is acquired through learning-by-doing: junior staff in law, medicine, science, engineering, and consulting master foundational skills while performing core job tasks under the supervision of senior practitioners. Such expertise has arguably become more, not less, necessary in the era of artificial intelligence (AI). Modern AI systems are easy to operate but hard to evaluate, so assessing their outputs demands the expert judgment that learning-by-doing instills. AI could strengthen this process or undermine it (Beane, 2024; Acemoglu et al., 2026; Afrouzi et al., 2026; Asriyan et al., 2026; Ide, 2026). It could accelerate the acquisition of expert judgment by supplying high-quality examples of professional work and on-the-spot feedback that managers might not provide and junior workers might hesitate to seek. Or it could stunt skill acquisition if junior workers offload tasks to algorithms and bypass the learning those tasks would otherwise generate, crowding out the judgment that effective AI use requires.[1] Evaluating AI tools in judgment-dependent tasks therefore requires more than measuring their impacts on short-term performance; it also requires quantifying how they affect skill acquisition.

A growing body of evidence addresses the first question, demonstrating that AI boosts immediate performance and often narrows the gap between high and low performers (Noy and Zhang, 2023; Peng et al., 2023; Brynjolfsson et al., 2024; Cruces et al., 2026; Ju and Aral, 2025; Dell'Acqua et al., 2026). The skill acquisition question, however, remains underexplored. Answering it requires sustained AI use over an extended period, a measure of skill that is distinct from AI-assisted output, and experimental control over AI access. To our knowledge, no existing study combines all three in a high-stakes professional setting. We meet these requirements by evaluating the consequences of AI use among intellectual property lawyers, whose legal drafting work requires both technical precision and strategic judgment, and whose outputs are amenable to blinded expert evaluation.

We partnered with eleven U.S. intellectual property law firms to conduct a three-month randomized controlled trial among 133 patent lawyers, 91 of whom completed the full protocol. All firms in the experiment have ongoing patent-drafting relationships with Google LLC, which facilitated recruitment, meaning that our sample best represents law firms that are sophisticated in patent law practice. Lawyers were randomly assigned in a 2:1 ratio to receive access to a custom AI drafting assistant or to a no-AI control condition, with randomization stratified by firm and experience level.[2] All work products were scored by blinded patent attorneys at an independent law firm on a standardized rubric covering enforceability, accuracy, strategic ambiguity (the use of broad, flexible claim language that maximizes future enforcement scope within the USPTO's "Broadest Reasonable Interpretation" standard), completeness, and clarity.

[1] See e.g. Brynjolfsson et al. (2025); Hosseini and Lichtinger (2025); Modestino et al. (2025); Liu et al. (2025) for evidence that AI diffusion is already leading to a decrease in junior-level hiring across sectors. Emanuel et al. (2026); Lambert and Schindler (2026) argue instead that the decline in junior hiring is better explained by the rise of remote work.

[2] The 2:1 ratio balanced the research team's preference for an even split against patent firms' preference for giving more of their lawyers AI access sooner.

The trial comprised three assessments at two points in time. At 10 days, lawyers completed a patent-drafting task; at 90 days, they completed a second drafting task and an unassisted redlining task. The drafting tasks measure the immediate effect of AI on assisted performance. Because treatment and control face the same task in each wave, comparing the treatment-control gap at 10 days with the same gap at 90 days reveals whether performance differences widen or narrow with extended use. The redlining task instead measures professional judgment. Conducted 'offline', without AI, it asks lawyers to evaluate and mark up existing professional work, a task patent lawyers routinely perform unaided and one that foregrounds expert judgment rather than text production. Comparing the offline performance of lawyers immersed in AI-assisted drafting for three months against that of lawyers without AI access separates the skill lawyers have internalized from their AI-assisted performance.[3]

Our central finding is that AI's effects on output and on expertise diverge. Measured by output, access to AI improved the quality of lawyers' patent drafts. Lawyers provided with AI access outperformed controls by 0.34 SD at 10 days ($p = 0.03$) and 0.38 SD at 90 days ($p = 0.01$). These gains are visible across all five scored domains, with larger point estimates for juniors than for seniors. Higher average scores reflect lower prevalence of weak work, not higher prevalence of excellent work.

Measured by unassisted judgment, only senior lawyers emerged from three months of AI use with improved capability. In the unassisted redlining task, the treatment group outperformed the control group by 0.32 SD ($p = 0.04$). But this effect stems entirely from better redlining by senior lawyers. Junior lawyers in the treatment group show no average gain in redlining quality relative to controls. Moreover, their scores were more dispersed, with significantly fewer mediocre scores, significantly more poor scores, and an offsetting rise in good but not excellent scores. This pattern is consistent with AI accelerating skill acquisition among some juniors and substituting for it among others. Our experiment does not answer why the effects appear positive for some lawyers and negative for others.

AI access also sped drafting. Treated lawyers spent about 10 fewer minutes on the 10-day drafting task than controls ($p = 0.05$), whose mean was 110 minutes, with similar but less precise savings at 90 days. Point estimates for time savings are substantially larger for junior than senior lawyers, though the contrast between these estimates is not statistically significant. Distinct from the patent drafting tasks, the treatment group realized no time savings relative to the control group in the unassisted redlining task. In both treatment and control groups, senior lawyers spend longer on average than juniors performing the redlining task, perhaps reflecting the expertise-intensive nature of this task.

[3] Our interpretation of differences in endline performance in redlining as reflecting differences in skill acquisition relies on the assumption of treatment-control skill balance at baseline. The evidence reported below supports this assumption. (Our preferred research design had subjects performing both a baseline and an endline skills assessment, but this feature proved an impediment to recruiting firms to the study.)

Several analyses confirm the robustness of these results to measurement and inference choices. Treatment effects on assisted drafting and on unassisted redlining hold across all five scoring subscales, with senior redlining gains near-uniform across domains and junior gains absent in every one, though precision falls at the subscale level. Because each task is scored by at least two independent raters, we verify that the results are not driven by rater idiosyncrasies. Given the risk of poor standard error coverage in small samples, we also report a nonparametric permutation test that affirms the primary findings. Ratings by a Large Language Model corroborate the human raters on the overall drafting gains, though not consistently on whether juniors or seniors gain more. On redlining, LLM ratings yield a positive but insignificant estimate for treated juniors and a larger, significant one for treated seniors.

We explore two concerns that apply to the central finding that senior but not junior lawyers gain in offline work. The first is the classification of seniority. Using thresholds from three to eleven years of experience, we find the same pattern: positive point estimates for assisted drafting in both groups at every threshold, and significant unassisted gains only among seniors. The second is that treated senior lawyers might have used AI on the redlining exercise itself. Using LLM evaluations and manual audits to flag textual tells of AI use, and erring toward over-classification, we identify 15 of 91 redlining submissions where AI use appears possible. Flagged submissions score *insignificantly* higher than unflagged submissions, and neither controlling for nor excluding them changes the inference for seniors. Excluding them lowers the full-sample estimate below conventional significance, while the senior effect remains significant at the 10 percent level. With LLM ratings, both estimates are larger and significant at the 1 percent level.

Our findings point to a trade-off between immediate productivity gains and durable skill acquisition. Junior lawyers capture the largest immediate gains from AI use, paralleling evidence from writing (Noy and Zhang, 2023), customer service (Brynjolfsson et al., 2024), and software development (Ju and Aral, 2025). But on the unassisted redlining task, The skill gains from AI exposure accrue entirely to senior lawyers. Treated juniors perform no better on average than control juniors, and their scores bifurcate: more good and more poor scores with no gain at the top of the distribution. That the group gaining most in output gains nothing in measured expertise suggests that foundational expertise may be a prerequisite for converting AI assistance into durable learning. This tradeoff matters so long as professional work demands judgment exercised without AI. In the domain of law, we expect that to remain true indefinitely. AI may increasingly assist with tasks such as redlining, but it is far less able to help lawyers as they advise clients, negotiate with counterparties, or respond to a patent examiner in real time.

2 AI in professional services: Productivity and skill development

Generative AI has diffused rapidly into knowledge work. Bick et al. (2026) report that 43% of US workers used generative AI for work in the first quarter of 2026, with adoption close to 60%

in business, management, and finance. Adoption in law is rising as well: the share of legal organizations actively using generative AI nearly doubled from 14% in 2024 to 26% in 2025, and 45% of law firms report either currently using or planning to make AI central to their workflow within a year (Thomson Reuters Institute, 2025). This uptake has proceeded despite distinctive obstacles: privileged client information limits what AI systems can access, high-stakes documents amplify the cost of AI-introduced errors (Charlotin, 2026; Robert, 2026), and AI's time savings threaten the time-based revenue model on which most law firms depend (Rapoport and Tiano, 2025).

Legal practice is a consequential setting for studying AI's effects on productivity and learning. The profession combines high-stakes output, domain-specific expertise that typically takes many years to acquire, and an apprenticeship structure through which junior practitioners develop expert judgment (Remus and Levy, 2017). These features are especially pronounced in intellectual property law, the domain we study. The precise language of a patent claim determines its legal force, drafting defensible claims requires extensive supervised practice, and evaluating patent quality is itself an expert domain. The two subsections that follow review the evidence on contemporaneous productivity and then on skill development.

2.1 Productivity impacts of AI use

Following the release of ChatGPT in late 2022, experimental studies found consistent productivity gains across white-collar occupations, from management consultants to call-center employees to freelancers (Noy and Zhang, 2023; Hui et al., 2023; Brynjolfsson et al., 2024; Dell'Acqua et al., 2026). A common finding is that AI narrows performance gaps between new and experienced workers, between average and high performers, and between generalists and domain experts (Noy and Zhang, 2023; Cruces et al., 2026; Shen and Tamkin, 2026). In law, Choi and Schwarcz (2025) report an "equalizing effect" among law students, with sharp gains for those at the bottom of the class and declines for those at the top. This leveling is not universal: in a field experiment with Kenyan entrepreneurs, Otis et al. (2026) find that an AI business mentor raised the performance of high performers while lowering that of low performers.

The form of the gain varies. Among knowledge workers and trainees, some studies find speed gains (Dillon et al., 2025), others find quality gains (Ju and Aral, 2025), and some find both (Haslberger et al., 2025; Usdan et al., 2024). One explanation is that workers reallocate time savings into additional effort on the same task, so an underlying productivity gain surfaces as speed, quality, or both, depending on circumstances (Humlum and Vestergaard, 2025a).

In some cases, AI fails to improve performance or erodes it. In an experiment with 227 professional radiologists, Agarwal et al. (2023) find that access to AI predictions does not raise diagnostic accuracy on average: clinicians underweight correct AI signals and treat their own judgment and AI output as statistically independent. Goh et al. (2024) report that physicians using GPT-4 perform no better than physicians using conventional diagnostic tools, even though GPT-4 alone

substantially outperforms both groups. Outside of medicine, Dell'Acqua et al. (2026) document that consultants working beyond the "jagged frontier" of AI capabilities produce worse outputs with AI than without it, as users fail to recognize when AI outputs are wrong. In a meta-analysis of 106 experiments and 370 effect sizes, Vaccaro et al. (2024) find that human-AI combinations on average perform worse than the better of human or AI alone, with shortfalls largest on decision tasks and when AI outperforms unaided humans. These findings suggest that gains from AI use may be conditional on users' ability to evaluate, integrate, or override its outputs, a capacity that may demand domain expertise.

Studies of AI in legal writing, the setting of our experiment, are mostly conducted with law students and measure contemporaneous output. Comparisons of human- and AI-generated legal text find AI output approaching and in many cases exceeding human quality (Wrzesniowska, 2023; Martin et al., 2024; Vals AI, 2025). Experiments that measure performance find that gains accrue mainly to speed. Choi et al. (2024) find that GPT-4 access produces large and consistent time savings across baseline performance levels but only slight and inconsistent quality improvements in legal analysis. Nielsen et al. (2024) find that AI-generated highlighting of relevant text cuts completion time by 30% with no measured quality loss, while AI summaries alone show no improvement. Schwarcz et al. (2026), using two "reasoning" models that generate intermediate steps before answering, find consistent quality improvements on five of six assignments completed by upper-level students. We are aware of one prior RCT with employed lawyers: Chien and Kim (2025) gave 91 legal aid lawyers AI access and randomly assigned 40% of them a broader set of tools and "concierge" support; in follow-up surveys, that subgroup self-reported higher productivity. Ours is the first study to measure skill acquisition, rather than contemporaneous productivity alone, and the first to do so with practicing lawyers rather than law students.

2.2 Professional skill development

If AI compresses the distribution of performance during assisted work but leaves underlying capabilities unchanged or degraded, its long-term effects on professional human capital will diverge from its short-term effects on output.

Evidence on this question is mixed. Shen and Tamkin (2026), in a randomized experiment with software engineers learning a new programming library, find that AI assistance impaired conceptual understanding, code reading, and debugging. Wiles et al. (2024) find that generative AI acts as a "temporary exoskeleton" for management consultants on unfamiliar data science tasks: these gains do not persist when the tool is removed. Among practicing endoscopists, Budzyń et al. (2025) find that routine AI assistance is associated with deterioration of unassisted diagnostic skill. Other studies find the opposite. He et al. (2026) document that novice radiologists who train with predictive AI subsequently make more accurate independent diagnoses. Cruces et al. (2026) find that less-educated participants who used generative AI on business problem-solving tasks improved modestly on unassisted follow-up tasks, and Lira et al. (2025) find that job-seekers who

practiced writing cover letters with AI improved their independent writing more than those who practiced without it. In the study closest to ours in design, Bednar et al. (2026) find that law students exposed to AI in an initial phase outperformed controls on subsequent tasks where AI access was removed.

One reading of these divergent results is that workflows requiring users to interpret, evaluate, or verify AI output build skill, while those permitting uncritical acceptance bypass the effort on which learning depends. The most direct evidence for this interpretation comes from Bastani et al. (2025), who randomize nearly 1,000 high school mathematics students to practice with or without GPT-4. Unconstrained access improves in-session performance but lowers subsequent unassisted exam scores. When the same model is reconfigured as a tutor that withholds answers and requires students to reason through problems, the harm disappears. He et al. (2026) find that radiologists' unassisted gains are largest when AI is deployed during both training and practice.

A related concern is that AI may disrupt the apprenticeship through which expertise is transmitted. Drawing on field studies of robotic surgery and other technologically mediated work, Beane (2024) argues that tools which raise immediate productivity often sever the process by which skill passes from senior to junior practitioners. Theoretical work formalizes the tradeoff: automating entry-level tasks can raise current output while slowing the accumulation of expertise (Afrouzi et al., 2026; Asriyan et al., 2026; Ide, 2026). The consequences of AI use for skill formation therefore cannot be inferred from contemporaneous output. They must be measured directly.

The present study provides direct measurement. We assess how sustained AI use shapes quality, productivity, and, most centrally, the development of expert judgment, tracking practicing professionals over three months in a high-stakes domain and pairing their AI-assisted output with a separate unassisted task scored by independent expert raters. The design separates AI's effect on performance from its effect on skill acquisition.

3 Experimental design

3.1 Sample selection

We began with a sample of 24 intellectual property law firms that are engaged in regular, non-exclusive business with Google. Eleven firms agreed to participate in the experiment. Two cohorts from one firm, enrolled two months apart, are analyzed as separate firms, giving an analytic sample of 12 firm cohorts. Recruits ranged from 3 to 49 participants across firms. To ensure that the time demands of the experiment did not compete with other professional activities, participants were, with the consent of their employers, paid at their contractual hourly rate for the time spent on experimental tasks. Table 1 provides a summary of firm-level recruitment, randomization, and task completion statistics, as well as experience levels of participating lawyers.

Random assignment was conducted within each firm, with two-thirds of participants assigned to

treatment (AI assistance) and one-third assigned to control (no AI assistance). Assignment was stratified on the legal experience of participants, using the median level of years of experience within each firm to differentiate senior from junior intellectual property lawyers.[4] Of the 156 intellectual property lawyers put forward by firms, 133 (85%) completed the onboarding requirements and were successfully randomized into the trial. Because strata within firms were small, realized assignment ratios deviate slightly from the 2:1 target: 90 (68%) of the enrolled sample was placed in the treatment group and 43 (32%) in the control group.

Table A1 reports tests of experimental balance. We see no statistically significant differences between treatment and control subjects across the limited set of pre-treatment individual characteristics available, including years of experience in the legal profession, responsiveness to outreach emails (measured in days), self-reported confidence in ability to execute experimental tasks, and self-reported comprehension of study requirements.[5]

3.2 Experimental protocol

Subjects in the treatment group were provided access via a gated online platform to the AI tool, an unreleased product developed by Google Labs designed to assist writing in professional services settings. The tool integrated user-provided context, style guidance, and user-specified writing goals. Its main features are prompt-based text modification, feedback from multiple AI editorial personas, and rule-based review of documents for internal inconsistencies.[6]

Control subjects in the experiment accessed standard legal tools during the 90-day experimental treatment window but were not given access to generative AI writing tools. While it is widely observed that individuals access generative AI tools to perform their work even when this is discouraged by their employers (Microsoft and LinkedIn, 2024; Humlum and Vestergaard, 2025b), this is much less likely in legal services, where AI usage is tightly controlled to enforce accuracy and protect confidentiality.

The experiment had three phases:

Onboarding and training: All participants were offered optional, asynchronous AI training modules. Both the treatment and control groups received foundational instruction on AI basics, covering: (1) commonly used AI terms; (2) an introduction to LLMs; and (3) prompt engineering. Participants assigned to the treatment group gained access to the AI tool and to a second training

[4] The analysis below instead classifies lawyers with seven or more years of experience as senior, a cutoff that does not coincide with the within-firm median in every firm. Table A1 shows that years of experience are balanced across treatment and control, and Section 4.3 shows that the main results are insensitive to the seniority cutoff.

[5] The final two measures were collected as Likert items in an onboarding survey administered alongside the study consent form, before AI tool access was granted.

[6] As a pre-release product, the tool underwent several feature changes during the study period of May 2025 through February 2026, including the launch in July 2025 of a redlining feature that enabled automated editorial review against user-defined criteria. The tool was not ultimately released as a standalone product; its features were integrated into Google's other AI products.

Table 1: Summary Statistics: Firm Attributes, Within-Firm Randomization, Task Completion

		Sample Size		Experience		Completion rates among control subjects			Completion rates among treatment subjects		
Firm	N (Recruited)	N (Randomized)	N (Treated)	Average	N 'Seniors' (≥7 YoE)	10-day drafting	90-day drafting	90-day redlining	10-day drafting	90-day drafting	90-day redlining
Firm 1	49	40 (82%)	27 (68%)	10.6	24 (60%)	13 (100%)	12 (92%)	12 (92%)	25 (93%)	25 (93%)	25 (93%)
Firm 2	27	22 (81%)	14 (64%)	9.6	12 (55%)	6 (75%)	8 (100%)	7 (88%)	14 (100%)	13 (93%)	13 (93%)
Firm 3	12	10 (83%)	8 (80%)	3.5	2 (20%)	2 (100%)	1 (50%)	NA	7 (88%)	5 (62%)	NA
Firm 4	12	12 (100%)	8 (67%)	10.7	7 (58%)	2 (50%)	0 (0%)	0 (0%)	1 (12%)	0 (0%)	0 (0%)
Firm 5	10	10 (100%)	7 (70%)	17.0	10 (100%)	3 (100%)	3 (100%)	3 (100%)	6 (86%)	7 (100%)	7 (100%)
Firm 6	10	5 (50%)	3 (60%)	19.6	5 (100%)	2 (100%)	2 (100%)	2 (100%)	3 (100%)	3 (100%)	3 (100%)
Firm 7	9	9 (100%)	6 (67%)	14.2	7 (78%)	1 (33%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Firm 8	8	8 (100%)	5 (62%)	12.9	7 (88%)	2 (67%)	1 (33%)	1 (33%)	4 (80%)	4 (80%)	4 (80%)
Firm 9	7	7 (100%)	5 (71%)	18.4	6 (86%)	1 (50%)	1 (50%)	1 (50%)	3 (60%)	3 (60%)	3 (60%)
Firm 10	6	6 (100%)	4 (67%)	27.7	6 (100%)	2 (100%)	2 (100%)	2 (100%)	4 (100%)	4 (100%)	4 (100%)
Firm 11	3	1 (33%)	1 (100%)	17.0	1 (100%)	0 (0%)	0 (0%)	0 (0%)	1 (100%)	1 (100%)	1 (100%)
Firm 12	3	3 (100%)	2 (67%)	12.7	2 (67%)	1 (100%)	1 (100%)	1 (100%)	2 (100%)	2 (100%)	2 (100%)
All	156	133 (85%)	90 (68%)	12.4	89 (67%)	35 (81%)	31 (72%)	29 (67%)	70 (78%)	67 (74%)	62 (69%)

Notes: Firms arranged in descending order of original cohort size. 'Seniors' refers to lawyers with seven or more years of legal experience (see Section 4.3 for details pertaining to this choice). Treatment status, experience strata, seniors, and completion rates are conditional on being effectively randomized. Completion rates are split into control and treatment panels. Firm 3 completed the experimental protocol before the redlining exercise was finalized, thus is excluded from the final column of this table.

module tailored to the intervention. This module covered: (1) recommendations for using the AI tool; (2) instructions for optimizing their usage of the tool via custom settings and a patent-specific prompt library; and (3) common pitfalls in AI-augmented patent drafting and corresponding mitigation strategies. All materials remained accessible for reference throughout the study.

10-day assessment: Ten days after the treatment group gained access to the AI tool, all participants were assigned a patent drafting task.[7] Working from a patent document containing a *main claim* that describes the overall purpose of the invention, they were asked to produce five to seven *dependent claims*, which are subordinate to the main claim, along with a detailed description drawn from patent figures and mock-up notes attributed to the "inventor" and their "enterprise client team." This format closely simulates real-world settings in which intellectual property lawyers are engaged by their clients. Invention details were selected from a hold-out dataset not used in the AI model's training and reserved for model benchmarking. Participants were instructed by email to spend no more than two hours on the task and to limit the write-up to 1,000 words.

90-day assessments: Participants were assigned two tasks:

- *Drafting:* A second drafting task, similar in structure to the first but with a different invention, figures, and notes; and
- *Redlining:* A mark-up task in which AI usage was not allowed. All participants, regardless of group assignment, were required to review an AI-generated patent draft and to mark it up using "direct edit" and "comment." Because the research team could not monitor participants' use of AI, we could not directly enforce this restriction, instead relying on clear, highlighted instructions in both the email prompt and the task materials. We provide extensive evaluation of potential non-adherence in Section 4.3.

To avoid ceiling effects in measuring performance in representative job tasks, the 90-day drafting task was designed to be more challenging than the 10-day task, with a longer expected time on task, and a higher word count target (up to 2,000 words). The exact instructions given to experimental subjects are reported in Section 3 of the Supplementary Materials.[8]

Of the 133 subjects initially randomized at baseline, 105 (79%) successfully completed the 10-day task (an overall attrition rate of 21%), 98 completed at least the drafting section of the 90-day task (cumulative attrition of 26%); and 91 total subjects completed the additional redlining task (cumulative attrition of 32%). Table 1 reports these counts by firm.

Table A2 reports linear probability models for sample attrition, where the dependent variable equals one if a subject did not complete the task. Each model includes treatment assignment, three baseline covariates (senior attorney; self-reported understanding of the study, Likert; and confidence in

[7] Correspondence delays and public holidays meant that the actual average start date for the initial task was 13 days after onboarding, while the second task's average start date was 102 days after onboarding; see Figure A6 for full chronology.
[8] The Supplementary Materials are not intended for publication and accompany this submission.

ability to complete the study, Likert), and an indicator for whether the participant answered the two Likert onboarding questions. This last indicator is necessary because the onboarding questions and the redlining task were both introduced after the first firm had already completed the experiment.[9] These estimates provide no evidence of differential attrition by treatment status; the point estimate for the treatment main effect in the attrition regression is 0.04 ($p = 0.56$) on the 10-day drafting task, -0.02 ($p = 0.79$) on the 90-day drafting task, and -0.04 ($p = 0.63$) on the 90-day redlining task. Despite their higher opportunity cost of time, senior attorneys were somewhat less likely to attrit from the experiment across all three experimental tasks, though this contrast is not statistically significant.

The one robustly significant finding in the attrition table is that subjects who did not complete the onboarding survey were differentially likely to attrit from each of the three experimental tasks: -0.17 ($p = 0.08$) for the 10-day drafting task; -0.21 ($p = 0.03$) for the 90-day drafting task; and -0.24 ($p = 0.04$) for the 90-day redlining task.[10] We infer that participants who chose not to complete the onboarding survey were generally less inclined to spend the required time on the experiment.

3.3 Data and measurement

Our primary outcome variables are subjects' performance on the drafting and redlining tasks. All tasks were scored by two independent raters who were blind to treatment status and drawn at random from a pool of six raters at a contracted intellectual property law firm. Raters applied a rubric that was developed in collaboration with Google's in-house counsel and is reproduced in Appendix Table A7. The rubric evaluates five core dimensions of patent draft quality: enforceability, accuracy, ambiguity, completeness, and clarity. As shown in Figure A1, the scores given by the two raters randomly assigned to each task are moderately correlated, with Pearson correlations of 0.59, 0.47, and 0.55 for the 10-day drafting, 90-day drafting, and 90-day redlining tasks respectively.[11] We also supplemented these professional evaluations with an LLM rater, discussed in Section 4.3.

Each drafting task had two separately scored parts, the dependent claims and the detailed description, and two raters scored each part on the same five-dimension rubric. Averaging across parts within rater and dimension gives ten scores per participant for each drafting task (5 dimensions $\times$ 2 raters). The redlining task was a single exercise requiring no averaging, and it also yields ten scores per participant, two per dimension.

Due to privacy agreements with participating law firms, the time that participants spent on experimental tasks could not be directly observed. We indirectly measure it in two ways. First, participants were instructed to run a stopwatch application while working and to record the elapsed

[9] To avoid dropping these cases, we assign their two Likert values to their sample means and set the "Onboarding response" dummy to zero.

[10] For the redlining task, this pattern cannot be an artifact of the one firm that completed the study before the onboarding questions were introduced, since that firm is excluded from the redlining attrition regression.

[11] Figures A1 and A2 explore the robustness of the human ratings and report complementary LLM-provided ratings.

time at the top of their submission. Second, post-task surveys, sent immediately upon receipt of each submission, asked participants to report how long the task had taken. Where subjects supplied both measures, we average them. (All survey items used for analysis are reported in Section 3 of the Supplementary Materials.)

We had also planned to apply the same scoring rubric to patent drafts that participating lawyers produced for Google under contract during the study period. Two features of the setting made this infeasible: such drafts were produced infrequently across our sample; and lawyers worked in teams, which precluded attributing a draft to an individual's treatment assignment. In place of these scores, the 90-day survey asked subjects to report, for each patent they drafted in routine work for any client over the same period, its completion time, the rounds of internal feedback and review it required, and its perceived drafting difficulty.

The post-task surveys also recorded participants' perceptions of the tasks they had just completed. At both 10 and 90 days, Likert items asked how the task compared to their routine drafting in the time it required ('speed') and in the quality of the output ('quality'), and how satisfied they were with the drafting experience overall ('satisfaction'). We also fielded monthly surveys that tracked perceived changes in routine-work difficulty, speed, job satisfaction, and internal reviewer scrutiny, along with subjects' use of generative AI tools outside of the specific AI tool we tested. These surveys allowed us to detect potential problems with satisfaction, contamination, or spillovers.

To understand how experimentally induced variation in access to the AI tool interacted with subjects' regular use of generative AI, we included the following question in each of the three monthly surveys: "Outside of [the AI tool], how would you describe your use of generative AI tools (e.g., ChatGPT/Gemini/Claude, or specialized AI tools for lawyers such as Harvey AI, Ironclad)?" Respondents chose among four answers: 1) no use in personal or work contexts; 2) use only in personal contexts; 3) occasional use for work-related tasks; and 4) regular use in professional work. In all three months of the experiment, roughly four in five respondents report either no work-related use (responses 1 or 2) or occasional work-related use (response 3), as shown in Figure A3. However, the fraction of treated respondents reporting regular work-related use (response 4) exceeds that among control respondents in every wave, and this contrast is marginally significant in month two and significant at the 5 percent level in month three. The gap is 6.8 percentage points in month one ($p = 0.36$) on a control base of 13.5 percent; 12.9 points in month two ($p = 0.08$) on a control base of 7.1 percent; and 17.3 points in month three ($p = 0.05$) on a control base of 5.6 percent. The successive widening of the gap reflects a decline in regular use among control respondents and a corresponding rise in their occasional use; the share of treated lawyers who report regular work-related use is roughly 20 percent across all three waves. Pooling across the three survey waves, regular work-related use is significantly higher among treated than control lawyers in the full sample and the senior subsample, and marginally so among juniors (see lower panels of Figure A3). In summary, regular work-related use of generative AI was uncommon among patent lawyers during our study period (May 2025 to February 2026), but access to the experimental tool led to

greater use of other AI tools in work-related tasks.[12]

3.4 Model specification

We estimate intent-to-treat (ITT) effects using the following baseline specification:

$$Y_{ijk} = \alpha + \beta_1 T_i + \gamma_j + \delta_k + \epsilon_{ijk}. \tag{1}$$

Here Y_{ijk} is a performance outcome for lawyer i at firm j on rubric dimension k, and T_i is the treatment indicator. These outcomes are standardized relative to the control group's mean and standard deviation for each task: in column 1 the five subscales are averaged and the average is then standardized, and in the stacked models each subscale is standardized separately. Because our primary quality assessments span five rubric dimensions, we estimate two overall treatment effects. First, we use as the outcome each participant's unweighted average of the five subscales, which collapses dimension k and drops δ_k (column 1 of the quality effect tables). Second, we stack the subscales into a pooled model with both firm fixed effects (γ_j) and rubric dimension fixed effects (δ_k) (column 2).[13] We estimate the pooled model at the rating level, so that each rater's score on each dimension is a separate observation. Because assignment errors at the rating firm left some submissions with more or fewer than two raters, the number of ratings per subject (n_i) varies. We therefore weight each observation by $w_i = 1/n_i$, which retains every rating while giving each subject equal weight in the pooled estimates.

To allow for the possibility that firms may perform systematically better or worse on specific rubric dimensions, we also report a model with full interactions between firm and rubric dummies (column 3):

$$Y_{ijk} = \alpha + \beta_1 T_i + \gamma_{jk} + \epsilon_{ijk}.$$

To study whether treatment effects on quality and learning differ systematically with professional experience, all tables also report models with separate intercepts and treatment effects for junior (under 7 years of experience) and senior (7 or more years) patent lawyers (column 4):[14]

[12] While each firm cohort was treated over the course of three months, cohorts started as early as May 2025 and as late as September 2025.

[13] We include firm fixed effects since treatment was randomized within firms, but excluding them has little impact on the estimates: the column 2 main effect is 0.37 rather than 0.34 SD in 10-day drafting, 0.46 rather than 0.38 SD in 90-day drafting, and unchanged at 0.32 SD in 90-day redlining, where the column 4 senior effect is likewise unchanged at 0.45 SD ($p = 0.03$).

[14] This threshold follows industry convention, on the advice of Google's in-house counsel. Appendix Figure A5 tests its sensitivity.

$$Y_{ijk} = \alpha_{\text{junior}} \cdot \text{Junior}_i + \alpha_{\text{senior}} \cdot \text{Senior}_i + \beta_{\text{junior}} (T_i \times \text{Junior}_i) + \beta_{\text{senior}} (T_i \times \text{Senior}_i) + \gamma_j + \delta_k + \epsilon_{ijk}. \quad (2)$$

To identify both subgroup baselines, we omit the global intercept and one firm's fixed effect, so that α_{junior} and α_{senior} are the control-group baselines for the omitted firm. For models using the collapsed average score (column 1), we use robust standard errors, following Abadie et al. (2023). For the stacked pooled models (columns 2–4), we cluster standard errors at the individual level to account for within-subject correlation across rubric dimensions.

4 Primary results: AI-assisted drafting and AI-unassisted redlining

This section reports the impacts of AI assistance on our primary outcomes: drafting quality at 10 and 90 days, and 90-day redlining quality. The treatment group was instructed to use AI on the drafting tasks; both groups were instructed to complete the redlining task without it. Table 2 summarizes these outcomes for the full sample and by experience subgroup, reporting the mean pooled across the five rubric dimensions and each dimension separately. Each dimension is standardized relative to the pooled control-group distribution (mean zero, standard deviation one), so full-sample treatment-group means read directly as effect sizes in standard deviations.[15] Control means within experience subgroups are not by construction zero, so subgroup effects must be read as treatment–control differences. Most of our main findings are apparent in these summary statistics; the regression tables that follow provide formal tests.

4.1 Performance on patent drafting

Our first finding is that lawyers assisted by AI draft stronger patents. Column 1 of Tables 3 and 4 reports estimated treatment effects on the quality of patent drafts at the 10-day and 90-day assessments, respectively, using pooled scores averaged across the five rubric subscales. Treated lawyers outperform control lawyers by 0.34 standard deviations at the first assessment ($p = 0.05$) and 0.38 standard deviations at the second assessment ($p = 0.01$).

The estimated impacts on pooled average scores reported in column 1 discard information on within-subject variation across subscales. In subsequent columns of each table, we switch from pooled outcomes to a model that stacks the five subscales for each participant. In these specifications, firm effects (γ_j) capture baseline differences in firm-level performance, and subscale effects (δ_k) account for baseline differences in the difficulty of each subscale. The estimated treatment effects using the stacked specification, reported in column 2, show slight gains in precision: p-values fall to $p = 0.03$ and $p = 0.01$ for drafting performance at 10 and 90 days, respectively. Column 3 provides a further robustness test by introducing firm-by-subscale interactions. The standard errors are

[15] Because participant scores covary across dimensions, the variance of the pooled average need not equal one.

Table 2: Experimental Performance Scores, Overall and by Seniority, Using Expert Ratings

	All		Junior		Senior	
	Control	Treatment	Control	Treatment	Control	Treatment
A. Main Outcomes: 10-Day Drafting						
Pooled Score	0.00	0.37	−0.06	0.49	0.03	0.31
	(0.90)	(0.79)	(0.87)	(0.61)	(0.92)	(0.86)
Enforceability	0.00	0.33	−0.07	0.42	0.03	0.29
	(1.00)	(0.82)	(0.98)	(0.70)	(1.02)	(0.87)
Technical Accuracy	0.00	0.39	−0.08	0.49	0.04	0.34
	(1.00)	(0.97)	(0.87)	(0.87)	(1.06)	(1.01)
Strategic Ambiguity	0.00	0.30	−0.08	0.41	0.04	0.25
	(1.00)	(0.92)	(1.07)	(0.80)	(0.98)	(0.97)
Completeness & Alignment	0.00	0.36	−0.06	0.56	0.03	0.26
	(1.00)	(0.88)	(1.01)	(0.63)	(1.01)	(0.97)
Clarity	0.00	0.44	−0.02	0.55	0.01	0.38
	(1.00)	(0.91)	(0.96)	(0.74)	(1.03)	(0.97)
Observations	70	140	22	46	48	94
B. Main Outcomes: 90-Day Drafting						
Pooled Score	0.00	0.45	−0.32	0.36	0.13	0.50
	(0.91)	(0.71)	(0.85)	(0.78)	(0.92)	(0.67)
Enforceability	0.00	0.35	−0.32	0.32	0.13	0.36
	(1.00)	(0.88)	(0.99)	(1.01)	(0.99)	(0.82)
Technical Accuracy	0.00	0.46	−0.42	0.30	0.17	0.54
	(1.00)	(0.74)	(1.01)	(0.77)	(0.96)	(0.72)
Strategic Ambiguity	0.00	0.49	−0.26	0.34	0.11	0.56
	(1.00)	(0.85)	(0.99)	(0.91)	(1.00)	(0.82)
Completeness & Alignment	0.00	0.40	−0.19	0.29	0.08	0.46
	(1.00)	(0.78)	(0.79)	(0.88)	(1.07)	(0.73)
Clarity	0.00	0.56	−0.39	0.55	0.16	0.57
	(1.00)	(0.76)	(0.94)	(0.76)	(0.99)	(0.77)
Observations	62	137	18	45	44	92
C. Main Outcomes: 90-Day Redlining						
Pooled Score	0.00	0.26	0.02	−0.06	−0.01	0.38
	(0.88)	(0.78)	(0.52)	(1.02)	(0.97)	(0.63)
Enforceability	0.00	0.29	−0.10	−0.02	0.03	0.41
	(1.00)	(0.88)	(0.86)	(1.14)	(1.05)	(0.73)
Technical Accuracy	0.00	0.30	−0.07	−0.03	0.02	0.43
	(1.00)	(0.87)	(0.68)	(1.06)	(1.09)	(0.75)
Strategic Ambiguity	0.00	0.16	0.23	−0.14	−0.08	0.27
	(1.00)	(0.97)	(0.61)	(1.15)	(1.10)	(0.86)
Completeness & Alignment	0.00	0.17	0.18	−0.14	−0.06	0.29
	(1.00)	(0.83)	(0.53)	(0.97)	(1.11)	(0.75)
Clarity	0.00	0.37	−0.13	0.03	0.04	0.50
	(1.00)	(0.91)	(0.76)	(1.17)	(1.07)	(0.75)
Observations	56	129	14	36	42	93

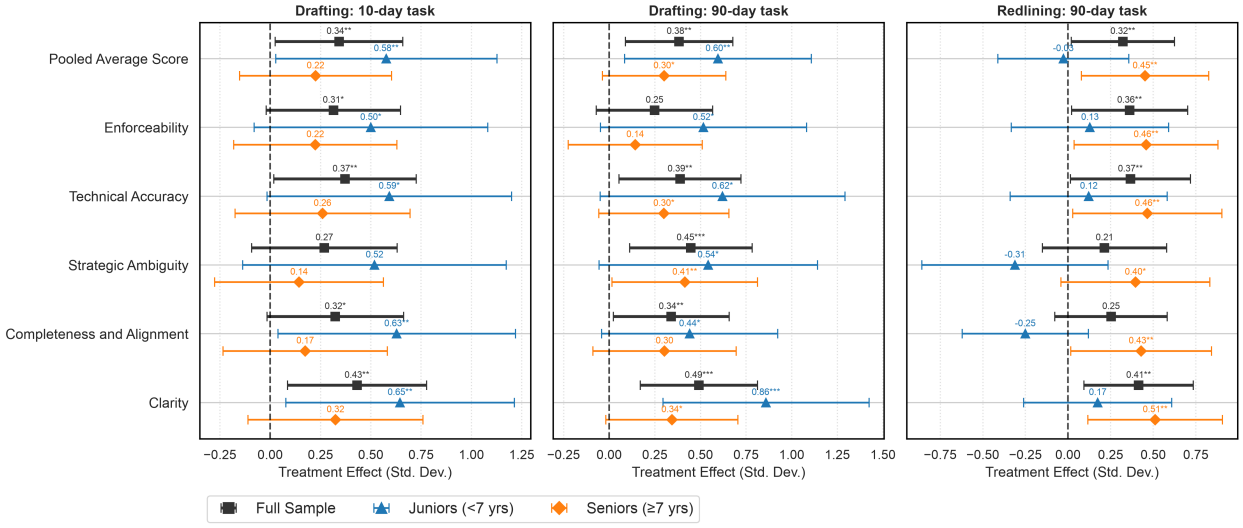
Notes: Table presents standard normal cell means with standard deviations reported in parentheses below. Sample restricts to successfully randomized subjects completing the study. Main outcome observations are at the individual expert rating level. Experience ('Senior') requires ≥ 7 years of practice. All dependent variables are standardized using the control group mean and variance. Pooled task scores represent the simple arithmetic average of standardized subcomponents. Observation counts reflect rater-level evaluations. Due to rater miscommunications detailed in Section 3.4, a small number of subjects received more or fewer than exactly two evaluations, resulting in observation counts that slightly diverge from $N \times 2$.

nearly unchanged in this more demanding specification. We use the estimates from column 2 as our preferred specification.

Figure 1 reports the treatment effect by rubric subscale. Relative to controls, treated subjects earned higher evaluation scores at both 10 and 90 days on all five components of the evaluation rubric: enforceability, accuracy, ambiguity, completeness, and clarity. Although not all individual effects are precisely estimated, their uniformity across subscales points to broad improvements in the work product of AI-assisted lawyers.

The visual breakdown in Figure 1 also previews the differential impact of AI across experience levels: across every subscale and on both drafting tasks, the point estimates for junior lawyers exceed those for senior lawyers. Column 4 of Tables 3 and 4 re-estimates column 2 with separate intercepts and treatment effects for junior and senior lawyers (Eqn. 2). The point estimates of the effect of AI access on drafting quality are substantially larger for juniors than for seniors. Juniors gain 0.58 SD at 10 days ($p = 0.04$) and 0.60 SD at 90 days ($p = 0.02$); the senior estimates are 0.22 SD at 10 days ($p = 0.24$) and 0.30 SD at 90 days ($p = 0.08$).

Figure 1: Impact of AI Access on Performance on Patent Drafting and Redlining Tasks: Pooled Scales and Component Subscales, Overall and By Seniority



Notes: This plot presents intent-to-treat estimates of treatment effects on the pooled quality score and on each of the five rubric subscales (enforceability, accuracy, ambiguity, completeness, and clarity) across the 10-day drafting, 90-day drafting, and 90-day redlining tasks. Estimates are reported for all subjects and separately for junior (under 7 years of experience) and senior (7 or more years) lawyers. Sample restricted to randomized participants with non-missing outcome data, using rating-level observations for human raters only. Dependent variables are standardized relative to the control group's mean and standard deviation. Models are estimated by OLS with weights $w_i = 1/n_i$, where n_i is the number of ratings individual i receives so that each individual carries equal total weight, incorporating firm fixed effects. The pooled average score also incorporates subscale fixed effects, matching column 2 of Tables 3, 4, and 6. Error bars represent 95% confidence intervals with standard errors clustered at the individual level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

The bottom rows of Tables 3 and 4 report one-sided tests of how AI access affects the relationship between experience and performance.[16] Broadly, AI access reduces junior–senior performance deficits. As expected, untreated seniors outperform untreated juniors at both assessments, significantly so at 90 days ($p = 0.26$ at 10 days and $p = 0.05$ at 90 days against the null that untreated juniors perform at least as well). The estimated treatment effect of AI access is larger for juniors. Though this differential effect is not itself significant ($p = 0.15$ and $p = 0.17$), it affects performance ordering. With AI access, juniors outperform *untreated* seniors at 10 days ($p = 0.05$), though not significantly at 90 days ($p = 0.31$). At 10 days, performance of treated juniors exceeds that of treated seniors, though not significantly ($p = 0.19$). At 90 days, treated seniors outperform treated juniors at the 10 percent level ($p = 0.1$ against the null that treated juniors perform at least as well).

These gains in drafting performance stem from a reduction in the prevalence of low scores. The reduction holds at both 10 and 90 days and in every experience subgroup, though it is not always precisely estimated. (Throughout, we number quintiles from the top of the score distribution, so that the first quintile holds the highest scores and the fifth the lowest, and we label them top, good, middle, mediocre, and poor.) As Table 5 and the histograms in Figure 2 show, AI access lowers the share of fifth-quintile (poor) scores by 0.19 ($p < 0.01$) at 10 days and by a smaller and less precisely estimated 0.11 ($p = 0.12$) at 90 days. The decline is larger among juniors, whose share of fifth-quintile scores falls by 0.25 ($p = 0.01$) at 10 days and 0.13 ($p = 0.38$) at 90 days, than among seniors, whose share falls by 0.18 ($p = 0.03$) and 0.09 ($p = 0.24$), respectively.

Among junior lawyers, the displaced mass in the bottom segments of the distribution is largely mirrored by increased mass in the first (top) quintile: the share of junior scores in the first quintile rises by 0.21 ($p = 0.01$) in the 10-day drafting task, and by a directionally consistent but statistically insignificant 0.12 ($p = 0.15$) in the 90-day drafting task. Among senior lawyers, the more modest reduction in the prevalence of fifth-quintile (poor) scores is offset by greater mass across most other segments of the distribution, particularly the third (middle) quintile for the 10-day task (0.12, $p = 0.04$) and the second (good) quintile at 90 days (0.11, $p = 0.14$), though the latter is less precisely estimated. This drop in the frequency of fifth-quintile (poor) scores tends to reduce performance variance among treated relative to control subjects. A Levene test finds that AI access significantly compresses performance variance on the 90-day drafting task, overall (Levene $p = 0.03$) and among senior lawyers (Levene $p = 0.08$). At 10 days, the test does not reject equal variances (Levene $p = 0.12$ overall).

These findings corroborate the expertise-leveling effects of AI access documented in other knowledge-intensive professional domains, such as writing, customer service, and software development (Noy and Zhang, 2023; Cruces et al., 2026; Shen and Tamkin, 2026).

[16] Each row reports the one-sided p-value against the null hypothesis in its label. Where the text reports a difference that runs in the direction of that null, it cites the p-value against the opposite null, which equals one minus the tabulated value.

Table 3: Impact of AI Access on Performance in AI-Assisted Drafting Task at 10 Days

	Expert Ratings				LLM Ratings		Pooled Ratings	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	0.34**	0.34**	0.34**		0.51***		0.40***	
	(0.17)	(0.16)	(0.17)		(0.19)		(0.14)	
Treat × Junior				0.58**		1.01***		0.74***
				(0.28)		(0.32)		(0.26)
Treat × Senior				0.22		0.26		0.23
				(0.19)		(0.22)		(0.16)
Junior				−0.18		−0.22		−0.38
				(0.27)		(0.33)		(0.25)
Senior				0.03		0.13		−0.12
				(0.15)		(0.16)		(0.12)
LLM Rater							0.68***	0.68***
							(0.10)	(0.10)
Constant	0.05	0.10	0.19		0.02		−0.16	
	(0.18)	(0.17)	(0.18)		(0.21)		(0.14)	
$H_0 : \alpha_{\text{junior}} < \alpha_{\text{senior}}$				0.74		0.81		0.81
$H_0 : \beta_{\text{junior}} < \beta_{\text{senior}}$				0.15		0.03**		0.05**
$H_0 : \alpha_{\text{junior}} + \beta_{\text{junior}} < \alpha_{\text{senior}}$				0.05**		0.00***		0.00***
$H_0 : \alpha_{\text{junior}} + \beta_{\text{junior}} < \alpha_{\text{senior}} + \beta_{\text{senior}}$				0.19		0.04**		0.06*
Subject-level average	Yes	No	No	No	No	No	No	No
Rating-level disaggregated	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
Sub-indicator FE	No	Yes	No	Yes	Yes	Yes	Yes	Yes
Firm × sub-indicator FE	No	No	Yes	No	No	No	No	No
Observations	105	1,050	1,050	1,050	525	525	1,575	1,575
R^2	0.23	0.14	0.15	0.14	0.10	0.14	0.19	0.21

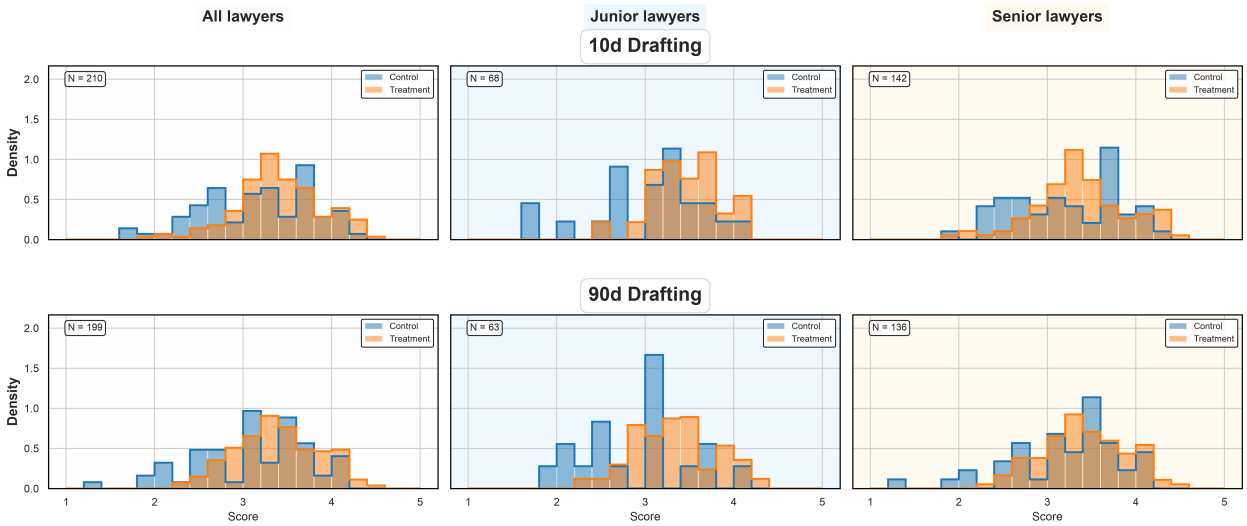
Notes: Table reports intent-to-treat estimates using OLS regressions. Model (1) reports effects on the subject-level average of subcomponents using robust (HC1) standard errors. Models (2)-(8) report rating-level disaggregated regressions using standard errors clustered at the individual level. Pooled models (7)–(8) control for rater type. All outcome subcomponents are standardized to mean zero and unit variance using the control group calculated locally within the full active sample. Models (1), (2), and (4)-(8) incorporate firm fixed effects, with Models (2) and (4)-(8) also including sub-indicator fixed effects. Model (3) incorporates firm × sub-indicator fixed effects. Models are weighted by $w_i = 1/n_i$, where n_i is the number of ratings individual i receives so that each individual carries equal total weight. The hypothesis tests in the bottom panel report one-sided p-values. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 4: Impact of AI Access on Performance in AI-Assisted Drafting Task at 90 Days

	Expert Ratings				LLM Ratings		Pooled Ratings	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	0.38**	0.38**	0.38**		0.90***		0.55***	
	(0.16)	(0.15)	(0.15)		(0.20)		(0.11)	
Treat × Junior				0.60**		0.48		0.56***
				(0.26)		(0.32)		(0.20)
Treat × Senior				0.30*		1.06***		0.55***
				(0.17)		(0.23)		(0.13)
Junior				−0.12		0.31		−0.23
				(0.23)		(0.23)		(0.17)
Senior				0.37		−0.23		−0.11
				(0.14)		(0.20)		(0.11)
LLM Rater							0.90***	0.90***
							(0.11)	(0.11)
Constant	−0.16	−0.09	0.00		−0.04		−0.36	
	(0.17)	(0.17)	(0.17)		(0.22)		(0.13)	
$H_0 : \alpha_{\text{junior}} < \alpha_{\text{senior}}$				0.95		0.04**		0.71
$H_0 : \beta_{\text{junior}} < \beta_{\text{senior}}$				0.17		0.93		0.49
$H_0 : \alpha_{\text{junior}} + \beta_{\text{junior}} < \alpha_{\text{senior}}$				0.31		0.00***		0.00***
$H_0 : \alpha_{\text{junior}} + \beta_{\text{junior}} < \alpha_{\text{senior}} + \beta_{\text{senior}}$				0.90		0.56		0.81
Subject-level average	Yes	No	No	No	No	No	No	No
Rating-level disaggregated	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
Sub-indicator FE	No	Yes	No	Yes	Yes	Yes	Yes	Yes
Firm × sub-indicator FE	No	No	Yes	No	No	No	No	No
Observations	98	995	995	995	490	490	1,485	1,485
R^2	0.29	0.17	0.18	0.20	0.21	0.22	0.29	0.29

Notes: Table reports intent-to-treat estimates using OLS regressions. Model (1) reports effects on the subject-level average of subcomponents using robust (HC1) standard errors. Models (2)-(8) report rating-level disaggregated regressions using standard errors clustered at the individual level. Pooled models (7)–(8) control for rater type. All outcome subcomponents are standardized to mean zero and unit variance using the control group calculated locally within the full active sample. Models (1), (2), and (4)-(8) incorporate firm fixed effects, with Models (2) and (4)-(8) also including sub-indicator fixed effects. Model (3) incorporates firm × sub-indicator fixed effects. Models are weighted by $w_i = 1/n_i$, where n_i is the number of ratings individual i receives so that each individual carries equal total weight. The hypothesis tests in the bottom panel report one-sided p-values. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Figure 2: Performance Distributions of Treated and Untreated Subjects on Drafting Tasks: Overall and by Experience Level



Notes: The figure plots empirical distributions of the average drafting scores for treated and control subjects, for the full sample and separately by experience. Observations are individual ratings; 'Junior' denotes fewer than 7 years of experience in intellectual property law. Scores are averaged Likert ratings (1–5), adjusted for firm fixed effects. Control distributions are indicated with blue bars and treatment with orange bars; bin size is set to 0.2 Likert increments in all cases.

Table 5: Impact of AI Access on Distributional Probability of Expert Ratings of the Drafting Tasks by Subgroup

	10 Day Drafting			90 Day Drafting		
	All (N=210)	Juniors (N=68)	Seniors (N=142)	All (N=199)	Juniors (N=63)	Seniors (N=136)
Levene Statistic	2.48	2.47	0.93	4.60**	0.05	3.19*
	(0.12)	(0.12)	(0.34)	(0.03)	(0.83)	(0.08)
Pr(Top (First) Quintile)	0.11**	0.21***	0.04	0.06	0.12	0.02
	(0.06)	(0.08)	(0.07)	(0.06)	(0.08)	(0.07)
Pr(Second Quintile)	−0.01	0.04	−0.01	0.08	0.04	0.11
	(0.06)	(0.12)	(0.08)	(0.06)	(0.12)	(0.07)
Pr(Third Quintile)	0.11**	0.09	0.12**	0.05	0.09	0.01
	(0.05)	(0.10)	(0.06)	(0.06)	(0.08)	(0.08)
Pr(Fourth Quintile)	−0.02	−0.10	0.03	−0.07	−0.12	−0.05
	(0.06)	(0.12)	(0.08)	(0.06)	(0.12)	(0.07)
Pr(Fifth (Bottom) Quintile)	−0.19***	−0.25**	−0.18**	−0.11	−0.13	−0.09
	(0.06)	(0.10)	(0.08)	(0.07)	(0.15)	(0.08)

Notes: This table presents intent-to-treat estimates on distributional changes and variance equality using rating-level observations, weighted by the inverse of the number of ratings per subject. Summary outcome variables are standardized to mean zero and unit variance using the control group. Cutoffs for quintiles are determined using the full valid sample for each respective outcome. Coefficients indicate the change in probability of falling into a specific quintile. All models include firm fixed effects and report HC1 robust standard errors in parentheses. The Levene statistic tests the null hypothesis of equal variances between treatment and control groups (centered at the median), with corresponding p-values in parentheses. The full sample includes all participants properly randomized. Juniors are defined as having < 7 years of experience. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

4.2 Performance in AI-unassisted patent redlining

This section addresses the key question that motivates our experiment: does extended access (90 days) to AI assistance affect performance in professional work executed without AI? We answer this question by evaluating work quality in the 90-day AI-unassisted redlining task introduced above. Table 6 reports the main results.

The first three columns of the table document an overall improvement in the quality of AI-unassisted redlining performed by lawyers in the treatment relative to the control group. Treated subjects outperform controls on the redlining task by 0.32 SD overall ($p = 0.04$, column 2). As column 4 of the table reveals, however, this aggregate gain is driven entirely by senior lawyers, whose work quality improved by 0.45 SD ($p = 0.02$). By contrast, the point estimate for junior lawyers is essentially zero (-0.03 SD, $p = 0.89$).

The aggregate performance gain in AI-unassisted redlining holds across quality subscales for senior but not junior lawyers, as documented in the third panel of Figure 1. In the full sample, the treatment effects of AI access on redlining quality are positive across all five quality dimensions (enforceability, accuracy, ambiguity, completeness, and clarity) and statistically significant at $p \leq 0.10$ in three of five cases. Here too the effects come from senior lawyers: treated seniors outperform senior controls in all five quality dimensions, all significantly at $p \leq 0.10$. By contrast, point estimates are *negative*, though not significantly so, for juniors on two of five subscales.

The bottom rows of Table 6 repeat the hypothesis tests above, here applied to unassisted redlining. Untreated juniors and seniors perform comparably: the point estimate favors juniors, but not significantly (one-sided $p = 0.26$). But the effect of treatment on junior–senior relative performance is the mirror image of the drafting result. The treatment effect is significantly larger for seniors than for juniors ($p = 0.04$ against the null that the junior effect is at least as large). Treated juniors do not significantly outperform untreated seniors ($p = 0.35$), and treated seniors significantly outperform treated juniors ($p = 0.05$ against the null that treated juniors perform at least as well). Thus, three months of AI-assisted drafting opened a junior–senior gap in unassisted judgment that is absent among controls.

As in drafting, the redlining gains are not evenly distributed across quality levels. Table 7 decomposes the treatment effects into quintile components, with the underlying distributions depicted in the histograms in Figure 3. In the full sample, treated lawyers have significantly greater mass in the second (good) quintile (0.22, $p = 0.00$) and significantly less mass in the fourth (mediocre) quintile (-0.17, $p = 0.02$). This comparison averages across striking differences in treatment effects by seniority level. The redlining scores of treated junior lawyers are substantially more dispersed than those of controls, as seen in the middle panel of Figure 3 and as a Levene statistic of 3.66 ($p = 0.06$) indicates. This reflects a significant rise of 0.22 in the share of fifth-quintile (poor) scores ($p = 0.04$), a sizable fall of 0.42 in the share of fourth-quintile (mediocre) scores ($p = 0.01$), and an imprecisely estimated rise of 0.19 in the share of second-quintile (good) scores ($p = 0.20$).

Table 6: Impact of AI Access on Performance in Non-AI-Assisted Redlining Task

	Expert Ratings				LLM Ratings		Pooled Ratings	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	0.32**	0.32**	0.32**		0.56***		0.42***	
	(0.16)	(0.15)	(0.16)		(0.19)		(0.13)	
Treat × Junior				−0.03		0.29		0.12
				(0.20)		(0.36)		(0.20)
Treat × Senior				0.45**		0.66***		0.54***
				(0.19)		(0.22)		(0.16)
Junior				0.21		0.07		−0.05
				(0.12)		(0.30)		(0.15)
Senior				0.09		−0.02		−0.16
				(0.15)		(0.17)		(0.13)
LLM Rater							0.64***	0.64***
							(0.14)	(0.14)
Constant	−0.28	−0.20	−0.19		0.29		−0.19	
	(0.17)	(0.17)	(0.17)		(0.20)		(0.14)	
$H_0 : \alpha_{\text{junior}} < \alpha_{\text{senior}}$				0.26		0.40		0.30
$H_0 : \beta_{\text{junior}} < \beta_{\text{senior}}$				0.96		0.81		0.95
$H_0 : \alpha_{\text{junior}} + \beta_{\text{junior}} < \alpha_{\text{senior}}$				0.35		0.10		0.13
$H_0 : \alpha_{\text{junior}} + \beta_{\text{junior}} < \alpha_{\text{senior}} + \beta_{\text{senior}}$				0.95		0.87		0.97
Subject-level average	Yes	No	No	No	No	No	No	No
Rating-level disaggregated	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
Sub-indicator FE	No	Yes	No	Yes	Yes	Yes	Yes	Yes
Firm × sub-indicator FE	No	No	Yes	No	No	No	No	No
Observations	91	925	925	925	455	455	1,380	1,380
R^2	0.16	0.10	0.11	0.12	0.26	0.27	0.17	0.18

Notes: Table reports intent-to-treat estimates using OLS regressions. Model (1) reports effects on the subject-level average of subcomponents using robust (HC1) standard errors. Models (2)-(8) report rating-level disaggregated regressions using standard errors clustered at the individual level. Pooled models (7)–(8) control for rater type. All outcome subcomponents are standardized to mean zero and unit variance using the control group calculated locally within the full active sample. Models (1), (2), and (4)-(8) incorporate firm fixed effects, with Models (2) and (4)-(8) also including sub-indicator fixed effects. Model (3) incorporates firm × sub-indicator fixed effects. Models are weighted by $w_i = 1/n_i$, where n_i is the number of ratings individual i receives so that each individual carries equal total weight. The hypothesis tests in the bottom panel report one-sided p-values. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Senior lawyers' scores also differ in dispersion between arms (Levene statistic 7.74, $p = 0.01$), but in the opposite direction: treated seniors' scores are less dispersed than controls'. This reflects a significant rise in the share of second-quintile (good) scores (0.22, $p = 0.00$) and a fall in the share of fifth-quintile (poor) scores (0.14, $p = 0.07$). The third (middle) and fourth (mediocre) shares fall by smaller and insignificant amounts (0.04 and 0.08).

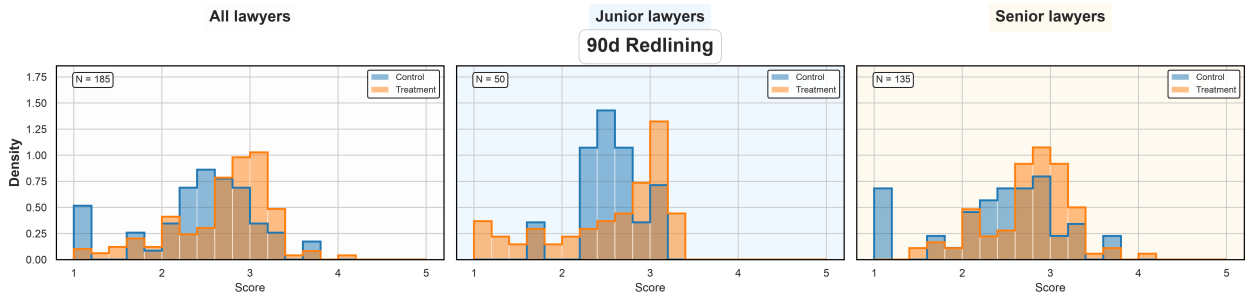
In summary, contemporaneous gains from AI-assisted drafting and subsequent gains on the unassisted redlining task diverge. Junior lawyers, who gain most while assisted by AI, show no aggregate gain in unassisted work, and their unassisted performance becomes more dispersed. Senior lawyers, whose contemporaneous gains are smaller, show substantial gains in unassisted work, concentrated in fewer fifth-quintile (poor) and more second-quintile (good) scores, with no significant change in the prevalence of first-quintile (top) scores.

Table 7: Impact of AI Access on Distributional Probability of Expert Ratings of the Redlining Task by Subgroup

	90 Day Redlining		
	All (N=185)	Juniors (N=50)	Seniors (N=135)
Levene Statistic	0.88	3.66*	7.74***
	(0.35)	(0.06)	(0.01)
Pr(Top (First) Quintile)	0.05	0.06	0.03
	(0.06)	(0.04)	(0.07)
Pr(Second Quintile)	0.22***	0.19	0.22***
	(0.06)	(0.15)	(0.07)
Pr(Third Quintile)	−0.05	−0.05	−0.04
	(0.06)	(0.10)	(0.07)
Pr(Fourth Quintile)	−0.17**	−0.42***	−0.08
	(0.07)	(0.16)	(0.08)
Pr(Fifth (Bottom) Quintile)	−0.05	0.22**	−0.14*
	(0.06)	(0.11)	(0.08)

Notes: This table presents intent-to-treat estimates on distributional changes and variance equality using rating-level observations, weighted by the inverse of the number of ratings per subject. Summary outcome variables are standardized to mean zero and unit variance using the control group. Cutoffs for quintiles are determined using the full valid sample for each respective outcome. Coefficients indicate the change in probability of falling into a specific quintile. All models include firm fixed effects and report HC1 robust standard errors in parentheses. The Levene statistic tests the null hypothesis of equal variances between treatment and control groups (centered at the median), with corresponding p-values in parentheses. The full sample includes all participants properly randomized. Juniors are defined as having < 7 years of experience. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure 3: Performance Distributions of Treated and Untreated Subjects on Redlining Task: Overall and by Experience Level



Notes: The figure plots empirical distributions of the average redlining score for treated and control subjects, for the full sample and separately by experience. Observations are individual ratings; 'Junior' denotes fewer than 7 years of experience in intellectual property law. Scores are averaged Likert ratings (1–5), adjusted for firm fixed effects. Control distributions are indicated with blue bars and treatment with orange bars; bin size is set to 0.2 Likert increments in all cases.

4.3 Robustness of key results

We probe the main results in five ways: supplementing expert ratings with LLM ratings, a nonparametric permutation test, dropping individual firms, varying the seniority threshold, and auditing for unauthorized AI use in the redlining task.

Using LLM evaluations to supplement human evaluations

Our main analysis compares the impact of AI access on the quality of patent drafting and redlining at eleven law firms, as assessed by blinded patent attorneys at an independent firm using the rubric described above. To supplement these expert assessments, we re-scored every drafting and redlining submission with a large language model (Gemini 2.5-flash), supplying one prompt per task that contained a brief task summary and the same scoring rubric given to the human raters. (Section 3 of the Supplementary Materials provides the full prompts.)

The summary of LLM ratings in Table A3 shows that they recover the same broad pattern of results as the human evaluations, but the treatment-control gaps they imply are systematically larger. In pooled scores, the LLM ratings place treated subjects 0.50 standard deviations above controls in 10-day drafting against 0.37 in the human ratings, 0.89 against 0.45 in 90-day drafting, and 0.60 against 0.26 in 90-day redlining.

Within each evaluator, scores across the five rubric dimensions are highly correlated. Across all tasks and evaluator types, including pooled expert and LLM ratings, the minimum pairwise Pearson correlation between dimensions is 0.75. (Section 2 of the Supplementary Materials reports within-evaluator correlations among quality dimensions.) The correlation between LLM and expert ratings is much lower. Averaging across subscales and across the two expert evaluations for each

submission, the LLM-expert correlation is $r = 0.47$ in the 10-day drafting task, $r = 0.19$ in the 90-day drafting task, and $r = 0.14$ in the 90-day redlining task, as Figure A2 shows.

Columns 5 through 8 of each main results table (Tables 3, 4, and 6) incorporate the LLM evaluations. Columns 5 and 6 use LLM evaluations alone, while columns 7 and 8 pool human and LLM evaluations. Given the low LLM-expert correlations in the two 90-day tasks, we treat the LLM as a third, less reliable rater and focus on the pooled models, in which expert evaluations contribute two of every three observations while the LLM evaluation contributes the third.

Pooling expert and LLM ratings leaves the pattern of results intact and generally reinforces earlier conclusions. In 10-day drafting, the main effect rises from 0.34 SD ($p = 0.03$) to 0.40 SD ($p = 0.00$), the junior effect from 0.58 to 0.74 SD, and the senior effect is unchanged at 0.23 SD. In 90-day drafting, the main effect rises from 0.38 to 0.55 SD. The junior estimate is little changed at 0.56 SD, while the senior estimate rises from 0.30 SD ($p = 0.08$) to 0.55 SD ($p = 0.00$). This is the one place where adding LLM ratings alters the pattern: in the pooled ratings, seniors and juniors receive a comparable boost at 90 days. In 90-day redlining, our key outcome, the main effect rises from 0.32 to 0.42 SD and the senior effect from 0.45 to 0.54 SD, while the junior estimate moves from -0.03 SD ($p = 0.89$) to 0.12 SD ($p = 0.57$) and remains indistinguishable from zero.

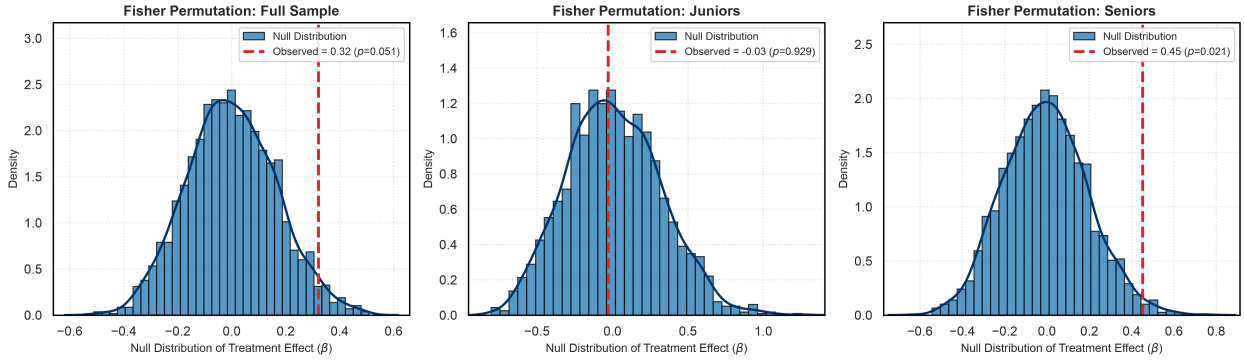
On net, supplementing expert ratings with LLM ratings corroborates our key results, with the caveat that the LLM rater is likely less reliable than the lawyers contracted to provide the expert ratings.

Nonparametric test of redlining results

The positive impact of AI access on senior but not junior performance in AI-unassisted redlining work is the most important and least expected finding in our study. Given the small senior and junior subsamples that support these findings, we further probe the robustness of our main inference by applying a nonparametric test of treatment-control differences. Following Holt and Sullivan (2023), we re-estimate the specifications of columns 2 and 4 of Table 6 on the disaggregated subscale ratings, then draw 2,000 Monte Carlo permutations of the treatment assignment to construct null distributions for the intent-to-treat effect. Figure 4 plots the resulting null distributions. The permutation exercise leaves the point estimates unchanged and alters only the inference; all reported permutation p-values are two-sided.

For the full sample, the estimated treatment effect of 0.32 falls at the 97.1 percentile of the null distribution ($p = 0.05$). Among seniors, the estimated effect of 0.45 falls at the 98.7 percentile ($p = 0.02$). An effect of this magnitude is unlikely under the sharp null of no treatment effect, supporting the inference that AI access carried over to improved performance in the judgment-intensive, unassisted patent redlining task. The junior estimate is near zero, at -0.03, and sits at the 46.8 percentile ($p = 0.93$), near the center of its null distribution, so we cannot distinguish it from zero.

Figure 4: Nonparametric Estimation of Effects of AI Access on AI-Unassisted Redlining Performance



Notes: Results of a nonparametric permutation test (2,000 Monte Carlo permutations) of the treatment effect on 90-day redlining quality scores, plotted against the null distribution under random treatment assignment. The main effect uses the specification of column 2 of Table 6, with subscale and firm fixed effects, weights $w_i = 1/n_i$, standard errors clustered at the individual level, and human ratings only; the subgroup effects use the specification of column 4. All reported permutation p-values are two-sided. Panels show results for the full sample and separately for junior and senior cohorts. See Figure 3 for cohort definitions.

Excluding individual firms

One firm contributes 40 of the 133 randomized participants and 37 of the 91 who completed the redlining task, raising the concern that a single firm drives the redlining results.[17] Figure A4 re-estimates the redlining treatment effects excluding one firm at a time. Across the nine subsamples, the full-sample estimate ranges from 0.22 to 0.38 SD, compared with 0.32 SD with all firms included, and the senior estimate ranges from 0.33 to 0.55 SD, compared with 0.45 SD. Excluding Firm 1, the largest, leaves the full-sample estimate at 0.33 SD and lowers the senior estimate to 0.36 SD, with both significant ($p = 0.06$ and $p = 0.09$ respectively). The most influential firm is Firm 10, whose six participants are all senior. Excluding it lowers the full-sample estimate to 0.22 SD, which is not statistically significant, while the senior estimate of 0.33 SD remains significant ($p = 0.08$). The effect of AI access on juniors' AI-unassisted redlining performance is statistically insignificant in every subsample.

Varying the experience threshold

Our main estimates classify lawyers with fewer than seven years of experience as junior and those with seven or more years as senior. Figure A5 re-estimates the junior and senior treatment effects from column 4 of each main results table using thresholds of three, five, seven, nine, and eleven years. The redlining results are insensitive to this choice. The effect on AI-unassisted redlining by

[17] The redlining sample spans nine of the twelve firm cohorts in Table 1. Firm 3 completed the protocol before the redlining task was finalized, and Firms 4 and 7 registered no submissions.

seniors ranges from 0.38 to 0.46 SD across thresholds and is statistically significant at the 5 percent level or better at every threshold. The effect on AI-unassisted redlining by juniors is statistically insignificant at every threshold. Its point estimate is near zero at thresholds of up to nine years and rises to 0.20 SD at eleven years (as the junior group absorbs more experienced lawyers). In both drafting tasks, the junior point estimate exceeds the senior point estimate at thresholds of five years and above. A three-year threshold leaves few lawyers in the junior group, and its estimates are imprecise.

Potential contamination by unauthorized AI use

Participants were instructed not to use AI tools in the redlining task, and this instruction was highlighted in both the task assignment email and the task materials. We could not enforce it directly, however, since end-user privacy commitments preclude observing individual tool use. If treated subjects used AI while redlining, our estimate of the effect of AI access on AI-unassisted performance would be contaminated. Three pieces of evidence suggest that non-adherence does not drive our results.

As a starting point, we have a limited amount of direct evidence against non-adherence. One firm completed the redlining task in a session separate from the drafting task, and we observe no use of the AI tool's redlining feature at that firm during the session. This check is narrow, however.

Second, we audited the redlining submissions for textual signatures of AI use. We prompted NotebookLM, running Gemini 3.0 Pro, to compare each redlined submission with its unedited baseline draft and to flag signatures of machine authorship: full-paragraph rewrites in place of targeted line edits; criticism bracketed by strained praise, a softening pattern characteristic of instruction-tuned models; dense use of LLM-typical vocabulary ("furthermore," "additionally," "moreover," "delves"); loss of structural context across patent sections; change-logs supplied in lieu of executed inline edits; excessively nested outlines in the free-form feedback; and theatrical over-commitment to a reviewing-partner persona. For juniors, the prompt added one marker: uniform edit density across sections, which is inconsistent with junior lawyers' typically uneven allocation of time between the early and later sections of a document. Each flag required a confidence score and a verbatim supporting quotation. (Section 3 of the Supplementary Materials reproduces the prompt in full.) The research team also audited the submissions flagged by Gemini manually, confirming that flagged submissions indeed showed evidence of non-adherence.

Erring toward over-classification, these procedures flag 15 of 91 redlining submissions.[18] Table A4 re-estimates the redlining results under two treatments of the flagged submissions. Both condition on a characteristic of the submission itself, so we read them as diagnostics rather than as causal estimates. Columns 3 and 4 add an indicator for flagged submissions. These score 0.18 SD

[18] Flags are about as frequent among control submissions (4 of 29) as among treated submissions (11 of 62), which suggests that they reflect either false positives or AI use that did not differ by treatment arm.

higher than unflagged submissions, a statistically insignificant difference ($p = 0.31$), and the senior treatment effect is 0.44 SD ($p = 0.02$), compared with 0.45 SD at baseline. Columns 5 and 6 exclude flagged submissions. The full-sample effect falls from 0.32 to 0.25 SD and is no longer statistically significant ($p = 0.14$), while the senior effect falls from 0.45 to 0.39 SD, which remains significant at the 10 percent level ($p = 0.06$). The junior estimate is −0.11 SD ($p = 0.61$). Table A5 repeats both exercises with LLM ratings. Excluding flagged submissions, the full-sample effect is 0.61 SD and the senior effect is 0.76 SD, both with $p < 0.01$.

Third, AI use for redlining should save time. As Section 5 reports, AI access reduced time on task in drafting, where AI use was permitted, but not in redlining. Had treated subjects used AI for redlining, we would expect time savings on that task as well.

5 Self-reported outcomes

Our primary results rest on blinded expert ratings of performance on experimental tasks. This section reports three further analyses that rely instead on subjects' self-reports: time spent on the experimental tasks, perceptions of those tasks, and patent drafting performed in the course of routine work during the study period. We treat the on-the-job measures as context for the experimental findings rather than as tests of our hypotheses.

Table 8 reports the effect of AI access on subjects' self-reported time on each experimental task. AI access reduced time on both drafting tasks, though only the 10-day estimate is statistically significant. It did not reduce time on the redlining task, consistent with the protocol's stipulation that subjects not use AI for redlining.

In the 10-day drafting task, treated participants finished an estimated 10 minutes faster than controls ($p = 0.05$), against a control-group mean of 110 minutes. The 90-day estimate is similar in size, at 10 minutes, but less precise ($p = 0.27$). Untreated seniors drafted faster than untreated juniors, averaging 104 versus 124 minutes at 10 days and 114 versus 123 minutes at 90 days. The point estimates indicate larger time savings for juniors. At 10 days, treated juniors finished 18 minutes faster than control juniors ($p =$) and treated seniors 5 minutes faster than control seniors ($p = 0.35$), though the difference between the two estimates is not statistically significant (two-sided $p = 0.22$). At 90 days, the junior and senior estimates are 22 and 5 minutes, and neither is statistically significant.

Untreated seniors spent *longer* on the redlining task than untreated juniors, 76 versus 69 minutes, though this difference is not statistically significant. AI access did not reduce redlining time for either group. Treated juniors spent 4 minutes longer than control juniors ($p = 0.76$) and treated seniors 1 minute longer than control seniors ($p = 0.89$).

Table 8: Impact of AI Access on Time Spent on Experimental Tasks

	10-day drafting; time on task (Minutes)		90-day drafting; time on task (Minutes)		90-day redlining; time on task (Minutes)	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	−9.97*		−9.61		1.69	
	(5.11)		(8.65)		(7.42)	
Treat × Junior		−17.91**		−21.63		4.49
		(9.13)		(18.24)		(14.95)
Treat × Senior		−4.91		−4.54		1.23
		(5.22)		(9.59)		(8.61)
Junior		114.14		107.27		53.00
		(7.25)		(16.60)		(11.98)
Senior		86.79		93.21		63.31
		(3.50)		(7.39)		(7.06)
Constant	111.97		124.56		71.18	
	(4.75)		(8.38)		(6.76)	
$H_0 : \alpha_{\text{junior}} = \alpha_{\text{senior}}$		0.00***		0.48		0.46
$H_0 : \beta_{\text{junior}} = \beta_{\text{senior}}$		0.22		0.40		0.85
$H_0 : \alpha_{\text{junior}} + \beta_{\text{junior}} = \alpha_{\text{senior}}$		0.13		0.51		0.60
$H_0 : \alpha_{\text{junior}} + \beta_{\text{junior}} = \alpha_{\text{senior}} + \beta_{\text{senior}}$		0.02**		0.76		0.52
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Unconditional control mean (all)	110.4		116.4		74.4	
Unconditional control mean (juniors)	123.6		123.3		69.0	
Unconditional control mean (seniors)	104.0		113.5		76.2	
Observations	102	102	96	96	91	91
R^2	0.18	0.28	0.17	0.18	0.04	0.05

Notes: Table reports intent-to-treat estimates using OLS regressions for time-on-task measures across both drafting and redlining tasks. All estimations conducted at the subject level with robust (HC1) standard errors. Columns (2), (4) and (6) report split specifications with no global intercept. The hypothesis tests in the bottom panel report two-sided p-values. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

The post-task surveys also recorded participants' perceptions of each drafting task: its speed and output quality relative to their routine drafting, and their satisfaction with the drafting experience, with details reported in Section 2 of the Supplementary Materials. Treated participants perceived large speed gains, of 0.72 SD at 10 days and 1.14 SD at 90 days (both $p < 0.01$), among both juniors (0.97 and 1.47 SD) and seniors (0.58 and 0.99 SD). Perceived quality gains are smaller and statistically insignificant, at 0.29 SD ($p = 0.23$) at 10 days and 0.34 SD ($p = 0.18$) at 90 days. Among juniors, perceived quality gains are 0.95 SD at 10 days ($p = 0.01$) and 0.59 SD at 90 days ($p = 0.12$). The treatment effect on satisfaction with the drafting experience is statistically insignificant at 10 days (0.30 SD, $p = 0.27$) and is 0.65 SD at 90 days ($p = 0.01$). The 90-day gain is concentrated among juniors, at 1.48 SD ($p < 0.01$), against 0.28 SD for seniors ($p = 0.34$), and the difference between the two is statistically significant (one-sided $p = 0.01$).

Finally, we asked participants to report on their own on-the-job patent drafting during the study

period. These measures cover only 68 participants (51% of those randomized), since we introduced them after recruitment had begun, and they are underpowered: the minimum detectable effect on weeks to complete a patent draft is roughly 2.3 weeks at 80 percent power. None of the five outcomes in Appendix Table A6 shows a statistically significant treatment effect. Treated participants report completing drafts 0.76 weeks faster than controls ($p = 0.35$). Because firms were not blinded to treatment assignment, they could have shifted workloads toward or away from treated lawyers, but neither the number of patents worked on (-1.06, $p = 0.18$) nor reported responsibility for them (0.01 SD, $p = 0.98$) differs significantly by treatment arm. (Section 1 of the Supplementary Materials reports post-experiment interviews with treated participants and an LLM-based qualitative analysis of redlining submissions.)

6 Discussion

This field experiment provided patent lawyers at eleven firms access to a custom AI drafting assistant for three months, benchmarked their drafting performance at 10 and 90 days, and finally evaluated all subjects' performance in marking up ('redlining') a flawed patent draft without the use of AI. AI assistance improved drafting quality at both assessments, with larger point estimates for junior lawyers. Point estimates of time savings were also larger for juniors. On the unassisted redlining task, the experience ordering reversed. Senior lawyers who had spent three months drafting with AI marked up patents more capably than their untreated peers. Junior lawyers showed no average gain, more dispersed performance, and a higher incidence of poor work.

Our design differs from most existing evidence on the skill consequences of AI assistance in two respects. Studies that measure unassisted performance following AI use have largely examined students and novices (Bastani et al., 2025; Lira et al., 2025; He et al., 2026), or professionals working outside their established expertise: management consultants attempting data science (Wiles et al., 2024), software engineers learning an unfamiliar library (Shen and Tamkin, 2026). Most also observe a single session or a short practice period. We instead observe experienced practitioners performing a core task of their own specialty, after three months of AI use embedded in their ordinary work. When experimental subjects are using AI, our results reproduce the leveling pattern documented in writing, customer service, and software work (Noy and Zhang, 2023; Brynjolfsson et al., 2024; Cruces et al., 2026; Ju and Aral, 2025). Our subsequent results indicate that this leveling does not carry over to unassisted performance.

This contrast is visible across the performance distribution, not merely at its mean. Treated seniors improved by earning more good scores and fewer mediocre and poor ones. Treated juniors show no detectable average improvement. Instead, their scores bifurcate: significantly more fifth-quintile (poor) scores, significantly fewer fourth-quintile (mediocre) ones, and an offsetting increase in second-quintile (good) scores, with no gain in the first (top) quintile. This pattern is consistent with AI use having abetted learning for some juniors and impaired it for others.

Raising measured performance and building expertise are distinct objectives. A tentative interpretation of our results is that these objectives are in tension when workers have limited experience but complementary for those with seasoned judgment. Expertise may thus partly determine whether AI assistance builds skill or substitutes for it. This pattern is not evident in either subjects' AI-assisted performance or their own assessments of their work. With AI access, juniors' mean drafting scores equaled or exceeded those of untreated seniors at both assessments, at 0.36 versus 0.13 SD at 90 days. Yet their mean redlining score, at −0.06 SD, was close to the control mean. The reverse holds for seniors: their gains in AI-assisted drafting were modest, while their gains in unassisted work were substantial. Juniors believed the tool was improving their performance (Section 5). This belief was borne out in AI-assisted work but not in unassisted work.

This divergence between lawyers' performance on assisted and unassisted tasks is consequential because unassisted judgment remains a regular demand of legal practice. We adopted the redlining task because it draws intensively on individual judgment and is therefore less likely to be delegated to AI than drafting, which requires producing new text within a well-defined scope. Even if AI eventually becomes a standard tool for patent redlining, it remains unlikely that AI will substitute for lawyers' expert judgment in real time during client and courtroom interactions.

Recent evidence from education shows a similar pattern and highlights a possible mechanism. Studying Chinese secondary students who adopt generative AI, Strömberg et al. (2026) find that time spent on homework falls, homework scores *rise*, and scores on the subsequent high-stakes examinations that determine high school and university placement fall by 1.3 to 1.5 standard deviations, with the largest losses among high-achieving students. These losses accrue to students whose study time falls after adoption; students who maintain their study time perform similarly to non-adopters.

The seniority contrast in our setting is consistent with the same mechanism. Juniors using AI worked faster but showed no unassisted gains. Seniors drafted better patents with AI but barely faster. On the unassisted redlining task, seniors worked at least as long as juniors whether or not they had spent three months drafting with AI, while treated seniors' redlining work was rated substantially higher than those of untreated seniors. Two differences limit the comparison with Strömberg et al. (2026): our subjects are paid professionals working in their own specialty rather than students completing assignments, and our horizon is three months against two years. Testing whether time spent is the causal channel would require manipulating time on task directly, which neither our study nor Strömberg et al. (2026) does.

Several limitations of our work bear mention. The sample is relatively small, reflecting the high hourly billing rate of top-tier lawyers. The junior subgroup is smaller still, so our estimates for junior lawyers are less precise than those for seniors. The sample is drawn from a set of firms contracted to perform ongoing patent work for a sophisticated client, and so is not representative of professional services generally. We were also unable to perform a baseline skills assessment (something our

preferred design included but partner firms rejected), so interpreting endline differences as skill acquisition rests on treatment-control balance at randomization. Additionally, we observe the effect of access over only three months, which is brief relative to the years over which patent expertise is built. Finally, we cannot observe individual tool use directly to verify full adherence to the stipulation that participants not use AI during the redlining task. However, the analysis reported in Section 4.3 suggests that potential non-adherence does not drive our results.

These limitations suggest directions for further experiments. A baseline unassisted assessment would enable direct measurement of skill acquisition rather than relying on randomization to infer it. Repeated unassisted assessments over a longer horizon would also reveal whether the dispersion we detect among juniors ultimately narrows or widens. One unambiguous takeaway from our work is that assessing the effect of AI assistance on skill development requires evaluations that separate the contribution of the software from the expertise of its user.

References

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. M. (2023). When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*, 138(1):1–35.

Acemoglu, D., Kong, D., and Ozdaglar, A. E. (2026). AI, human cognition and knowledge collapse. NBER Working Paper 34910, National Bureau of Economic Research.

Afrouzi, H., Blanco, A., Drenik, A., and Hurst, E. (2026). Automation, learning, and career dynamics. NBER Working Paper 35157, National Bureau of Economic Research.

Agarwal, N., Moehring, A., Rajpurkar, P., and Salz, T. (2023). Combining human expertise with artificial intelligence: Experimental evidence from radiology. NBER Working Paper 31422, National Bureau of Economic Research.

Asriyan, V., Reichardt, H., and Shelegia, S. (2026). Immiserizing automation. CEPR Discussion Paper DP21872, Centre for Economic Policy Research.

Autor, D., Rodchenko, T., Iscenko, Z., Strand, S., and Pearl, D. (2025). Ai & expertise research with patent lawyers. AEA RCT Registry.

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., and Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122.

Beane, M. (2024). *The Skill Code: How to Save Human Ability in an Age of Intelligent Machines*. Harper Business, New York.

Bednar, N., Cleveland, D. R., Erbsen, A., and Schwarcz, D. (2026). Artificial intelligence and human legal reasoning. *Minnesota Legal Studies Research Paper*, (2026-21).

Bick, A., Blandin, A., Deming, D. J., Fuchs-Schündeln, N., and Jessen, J. (2026). Mind the gap: AI adoption in Europe and the US. Brookings papers on economic activity conference draft, Brookings Papers on Economic Activity.

Brynjolfsson, E., Chandar, B., and Chen, R. (2025). Canaries in the coal mine? Six facts about the recent employment effects of artificial intelligence. Working paper, Stanford Digital Economy Lab.

Brynjolfsson, E., Li, D., and Raymond, L. R. (2024). Generative AI at work. *The Quarterly Journal of Economics*, 139(4):1721–1784.

Budzyń, K., Romańczyk, M., Kitala, D., Kołodziej, P., Bugajski, M., Adami, H. O., Blom, J., Buszkiewicz, M., Halvorsen, N., Hassan, C., et al. (2025). Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. *The Lancet Gastroenterology & Hepatology*, 10(10):896–903.

Charlotin, D. (2026). AI hallucination cases. Database, https://www.damiencharlotin.com/hallucinations/. Retrieved April 18, 2026.

Chien, C. V. and Kim, M. (2025). Generative AI and legal aid: Results from a field study and 100 use cases to bridge the access to justice gap. *Loyola of Los Angeles Law Review*, 57(4):903.

Choi, J. H., Monahan, A. B., and Schwarcz, D. (2024). Lawyering in the age of artificial intelligence. *Minnesota Law Review*, 109(1):147–218.

Choi, J. H. and Schwarcz, D. (2025). AI assistance in legal analysis: An empirical study. *Journal of Legal Education*, 73. Available at SSRN.

Cruces, G., Fernández Meijide, D., Galiani, S., Gálvez, R. H., and Lombardi, M. (2026). Does generative AI narrow education-based productivity gaps? Evidence from a randomized experiment. NBER Working Paper 34851, National Bureau of Economic Research.

Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. (2026). Navigating the jagged technological frontier: Field experimental evidence of the effects of Artificial Intelligence on knowledge worker productivity and quality. *Organization Science*, 37(2):403–423.

Dillon, E. W., Jaffe, S., Immorlica, N., and Stanton, C. T. (2025). Shifting work patterns with generative AI. NBER Working Paper 33795, National Bureau of Economic Research.

Emanuel, N., Harrington, E., and Pallais, A. (2026). The power of proximity to coworkers: Training for tomorrow or productivity today? *The Quarterly Journal of Economics*, 141(3):1825–1870.

Goh, E., Gallo, R., Hom, J., Strong, E., et al. (2024). Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Network Open*, 7(10):e2440969.

Haslberger, M. et al. (2025). Rage against the machine? generative AI exposure, subjective risk, and policy preferences. *Journal of European Public Policy*.

He, V. F., Li, S., Puranam, P., and Lin, F. (2026). Predictive AI can support human learning while preserving error diversity. Working Paper 2026/09/STR, INSEAD.

Holt, C. A. and Sullivan, S. P. (2023). Permutation tests for experimental data. *Experimental Economics*, pages 1–38. Epub ahead of print. PMID: 37363160; PMCID: PMC10066020.

Hosseini, S. M. M. and Lichtinger, G. (2025). Generative AI as seniority-biased technological change: Evidence from U.S. résumé and job posting data. Working Paper 5425555, SSRN.

Hui, X., Reshef, O., and Zhou, L. (2023). The short-term effects of generative artificial intelligence on employment: Evidence from an online labor market. CESifo Working Paper 10601, Center for Economic Studies and ifo Institute (CESifo), Munich.

Humlum, A. and Vestergaard, E. (2025a). Still waters, rapid currents: Early labor market transformation under generative AI. NBER Working Paper 33777, National Bureau of Economic Research.

Humlum, A. and Vestergaard, E. (2025b). The unequal adoption of ChatGPT exacerbates existing inequalities among workers. *Proceedings of the National Academy of Sciences*.

Ide, E. (2026). Automation, AI, and the intergenerational transmission of knowledge. Working paper, IESE Business School.

Ju, H. and Aral, S. (2025). Collaborating with AI agents: Field experiments on teamwork, productivity, and performance.

Lambert, P. J. and Schindler, Y. (2026). The broken ladder: AI, remote work, and early-career hiring. Working Paper 6787638, SSRN.

Lira, B. et al. (2025). Coach not crutch: Evidence that AI can improve writing skill despite reducing effort.

Liu, Y., Wang, H., and Yu, S. (2025). Labor demand in the age of generative AI: Early evidence from the u.s. job posting data. Policy Research Working Paper Series 11263, The World Bank.

Martin, L., Whitehouse, N., Yiu, S., Catterson, L., and Perera, R. (2024). Better call gpt, comparing large language models against lawyers. *arXiv preprint arXiv:2401.16212*.

Microsoft and LinkedIn (2024). AI at work is here. Now comes the hard part. 2024 work trend index annual report, Microsoft and LinkedIn.

Modestino, A. S. et al. (2025). The impact of generative AI on job opportunities for junior software developers. Working paper, Northeastern University.

Nielsen, P. A. et al. (2024). Building a better lawyer: Experimental evidence that artificial intelligence can increase legal work efficiency. *Journal of Empirical Legal Studies*, 21(4):979–1022.

Noy, S. and Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192.

Otis, N. G., Clarke, R., Delecourt, S., Holtz, D., and Koning, R. (2026). The uneven impact of generative artificial intelligence on entrepreneurial performance: Evidence from a field experiment in Kenya. *Management Science*.

Peng, S., Kalliamvakou, E., Cihon, P., and Demirer, M. (2023). The impact of AI on developer productivity: Evidence from github copilot. *arXiv:2302.06590*.

Rapoport, N. B. and Tiano, Jr., J. R. (2025). Fighting the hypothetical: Why law firms should rethink the billable hour in the generative AI era. *Washington Journal of Law, Technology & Arts*, 20(2):41.

Remus, D. and Levy, F. (2017). Can robots be lawyers? computers, lawyers, and the practice of law. *Georgetown Journal of Legal Ethics*, 30(3):501–558.

Robert, A. (2026). Sanctions ramping up in cases involving AI hallucinations. ABA Journal.

Schwarcz, D., Manning, S., Barry, P. J., Cleveland, D. R., Prescott, J. J., and Rich, B. (2026). AI-powered lawyering: AI reasoning models, retrieval augmented generation, and the future of legal practice. *Journal of Law and Empirical Analysis*. Forthcoming. Minnesota Legal Studies Research Paper No. 25-16; U of Michigan Public Law Research Paper No. 24-058.

Shen, J. H. and Tamkin, A. (2026). How AI impacts skill formation.

Strömberg, D., Lei, V., and Wu, Y. (2026). The generative AI learning penalty: Evidence from Chinese secondary education. Working paper, Stockholm University and University of Hong Kong.

Thomson Reuters Institute (2025). 2025 generative AI in professional services report. Report, Thomson Reuters.

Usdan, M., Connell Pensky, A. E., and Chang, R. (2024). Generative AI's impact on graduate student writing productivity and quality. Working paper, SSRN.

Vaccaro, M., Almaatouq, A., and Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8:2293–2303.

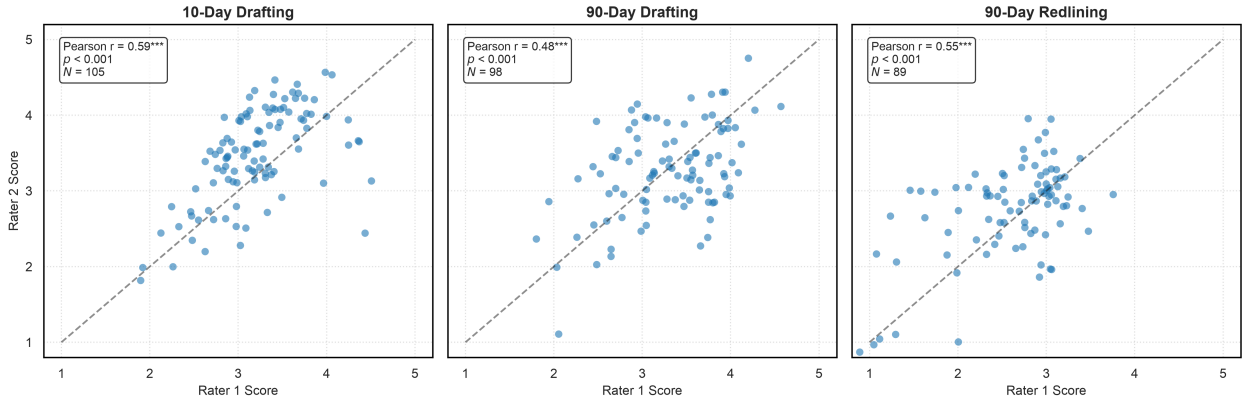
Vals AI (2025). Vals legal AI report (vlair): Benchmarking large language models against human lawyer performance. Technical report, Vals AI in Partnership with Legaltech Hub.

Wiles, E., Krayer, L., Abbadi, M., Awasthi, U., Kennedy, R., Mishkin, P., Sack, D., and Candelon, F. (2024). GenAI as an exoskeleton: Experimental evidence on knowledge workers using GenAI on new skills. Working paper, Boston University. Available at SSRN: https://ssrn.com/abstract=4944588.

Wrzesniowska, L. (2023). Can AI make a case? AI vs. lawyer in the dutch legal context. *International Journal of Law, Ethics, and Technology*.

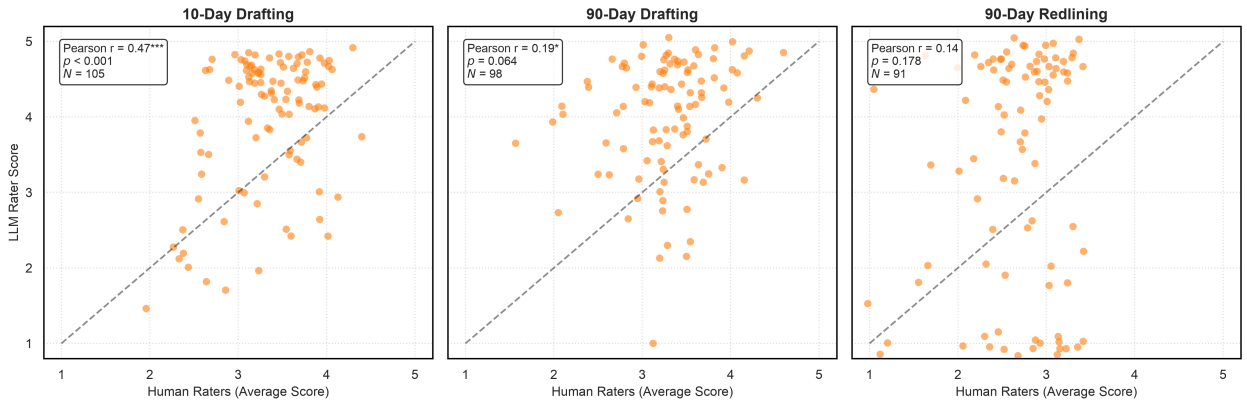
Appendices

Figure A1: Interrater Reliability among Expert Evaluations of Participant Performance in Primary Tasks



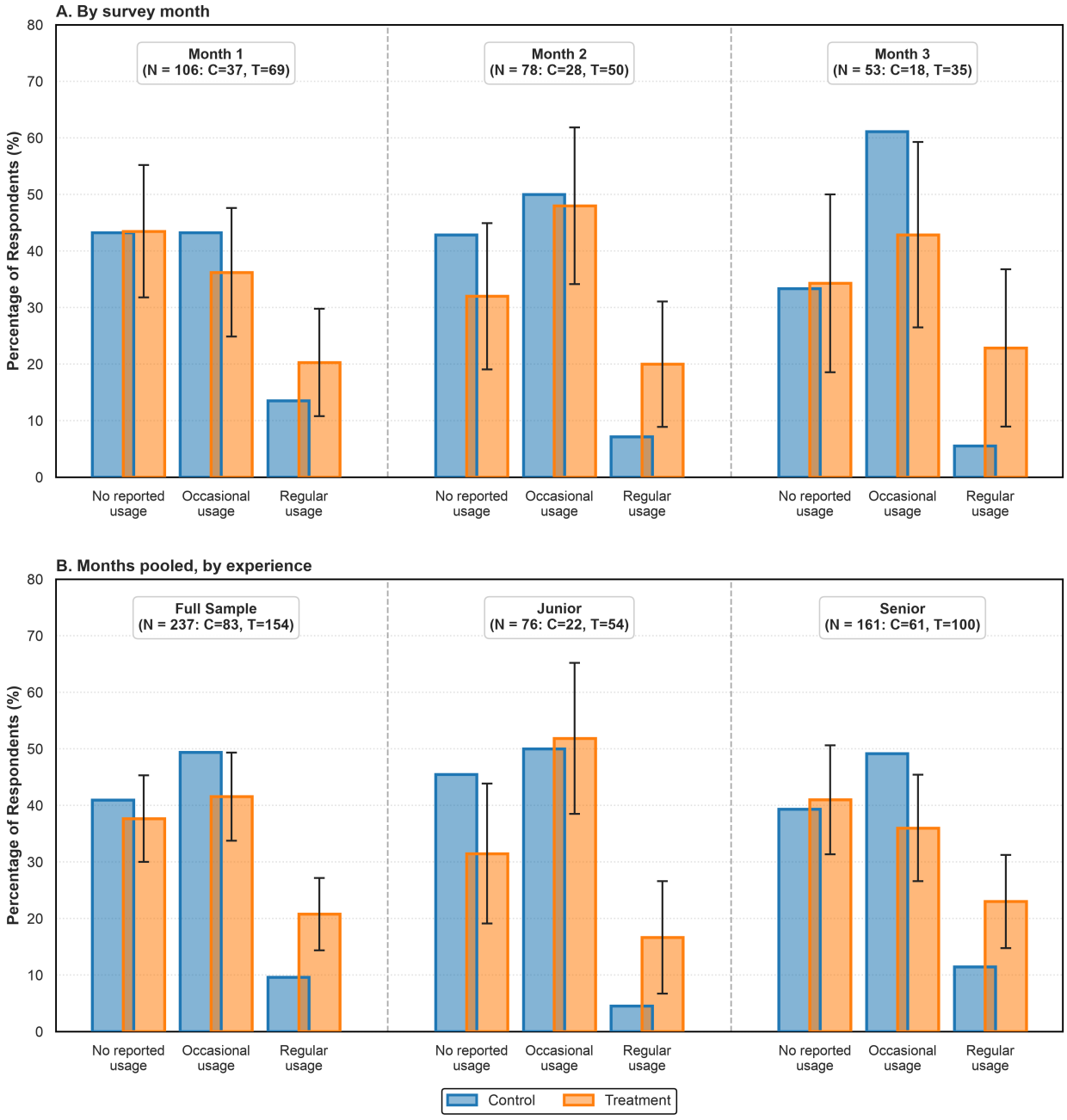
Notes: This figure evaluates inter-rater reliability by plotting pairwise human rater scores for the 10-day drafting, 90-day drafting, and 90-day redlining tasks. (For the redlining task, two participants have only one expert rating where two were expected, leading to a slightly lower N that we account for through weighting in our regression estimates) A small amount of random jitter is added to discrete scores to make overlapping data points visible, and Pearson correlation coefficients are calculated alongside a 45-degree reference line. The scatter plots reveal a moderate to strong positive correlation (Pearson r ranging from 0.47 to 0.59) among human raters across the tasks. These statistically significant correlations indicate acceptable levels of agreement in how legal professionals evaluate the quality of the patent drafting and redlining tasks.

Figure A2: Comparison of Expert and LLM Ratings of Participant Performance in Primary Tasks



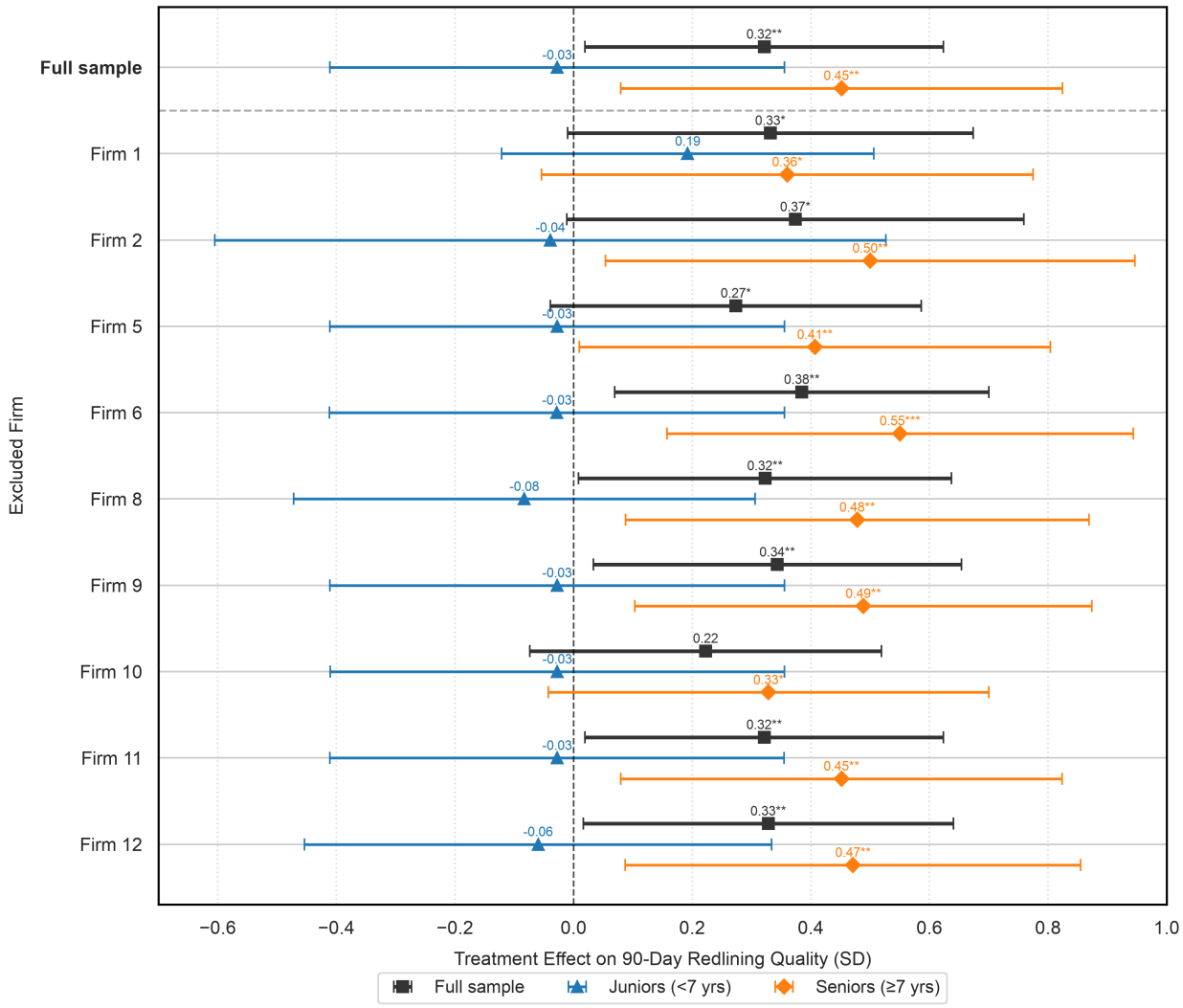
Notes: This plot compares the average score given by human raters to the score provided by the LLM rater for the same participants across the three main study tasks. Like the human rater plots, data points are lightly jittered to prevent overplotting, and Pearson correlations are calculated to measure the alignment between human rater average scores and machine evaluation. The LLM scores show a moderate, statistically significant correlation with human averages in the initial 10-day drafting task (r = 0.47). However, the correlations are notably weaker and less statistically significant for the 90-day tasks, suggesting that the LLM may diverge from human consensus on more difficult tasks.

Figure A3: Background (i.e. Non-Study-Related) AI Usage During Study Period, by Month, Seniority and Experimental Group



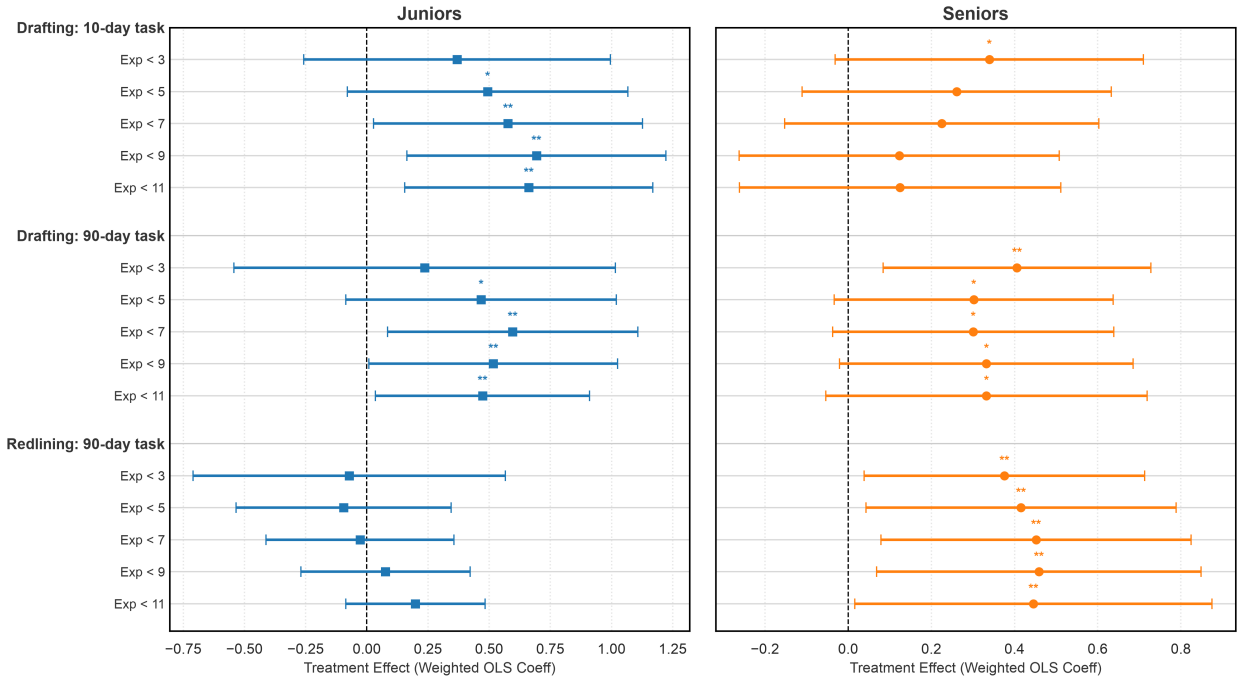
Notes: This plot compares non-study-related generative AI tool usage in both the treatment and control groups. The top panel (A) presents means by group for each month (1, 2 and 3) of the study period, while the bottom panel (B) presents responses pooled over months by sample subgroup (Full, Junior and Senior, where Senior means at least 7 years of experience consistent with our usage throughout the paper). Error bars represent 95% confidence intervals for treatment group responses.

Figure A4: Impact of AI Access on 90-Day Redlining Quality, Excluding Individual Firms



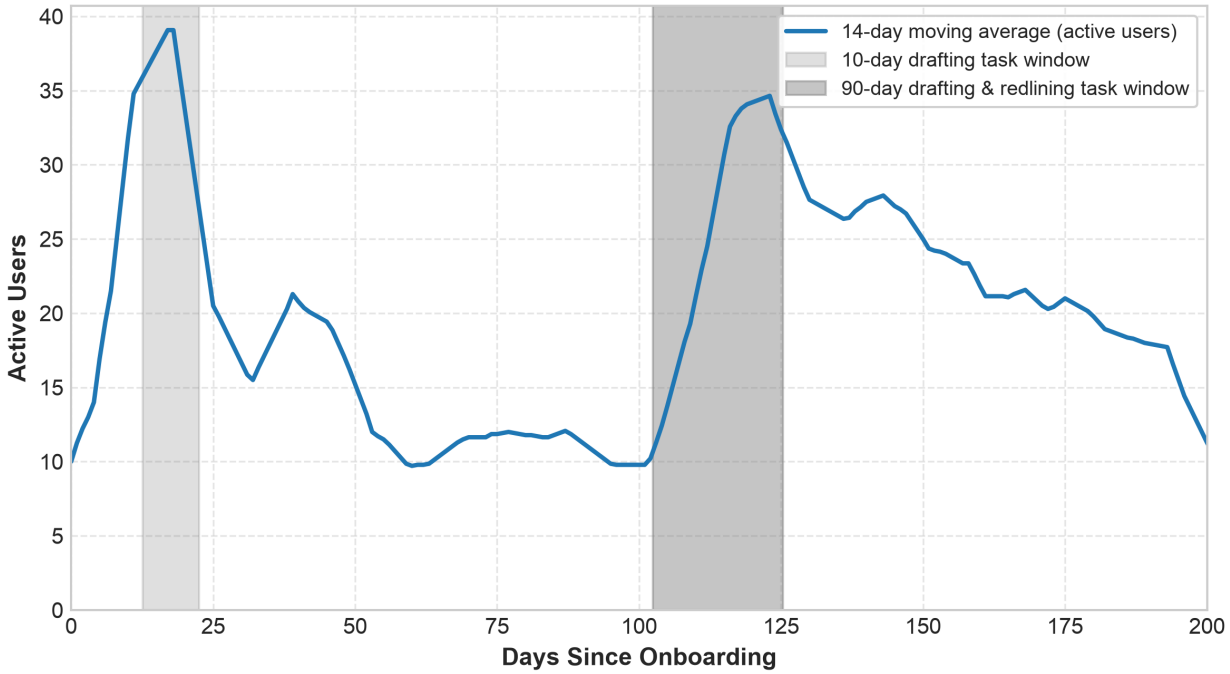
Notes: This figure plots intent-to-treat estimates of the effect of AI access on 90-day redlining quality, excluding one firm at a time. The top row reports estimates with all firms included, and each subsequent row excludes the indicated firm, with firm numbering following Table 1. Three firm cohorts contribute no redlining submissions and are omitted: Firm 3 completed the protocol before the redlining task was finalized, and Firms 4 and 7 registered no submissions. Where the excluded firm has no junior redlining completers, the junior estimate is identical to the full-sample estimate. Full-sample estimates use the specification of column 2 of Table 6 and junior and senior estimates use the specification of column 4, with human ratings only; see Figure 4 for details. Outcomes are standardized using the control group mean and standard deviation in the full redlining sample, before any firm is excluded. Error bars represent 95% confidence intervals. See Figure 3 for cohort definitions. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Figure A5: Sensitivity of Main Results to Choice of Experience Threshold for Junior vs. Senior Classification



Notes: This figure plots intent-to-treat estimates of the effect of AI access on junior and senior lawyers under alternative experience thresholds. Each row classifies lawyers with fewer than the indicated years of experience as junior and all others as senior. The seven-year threshold is the classification used in the main analysis. Estimates use the specification of column 4 of Tables 3, 4, and 6, with human ratings only; see Figure 4 for details. Error bars represent 95% confidence intervals. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Figure A6: AI Tool Usage by Period



Notes: This figure summarizes daily usage intensity (calculated as a 14-day moving average) across five phases of the experiment: the period preceding the first test task, the first (10-day) task window, the period between the two tasks, the second (90-day) task window, and the 30 days following the second submission. Usage intensity is lowest before the first task and in the interval between the two tasks, and it is roughly twice as high during each task window. (While we use "10-day" and "90-day" task labels throughout, the actual temporal window for each task varied slightly across firms due to operational technical hurdles, holiday periods and firm response rates; the actual post-onboarding day count averages are included in the legend at upper-right) Higher usage after the second task may be partially explained by the control group gaining access to the AI tool at the conclusion of their study obligations.

Table A1: Tests of Treatment-Control Covariate Balance

	Control		Treatment		
Variable	N	Mean	N	Mean	T-C
Time (days) for response after onboarding email	31	8.72	70	9.05	0.33
		(1.40)		(1.23)	(1.86)
Understanding of study (Likert)	31	4.65	69	4.75	0.11
		(0.10)		(0.06)	(0.12)
Confidence in ability to complete tasks (Likert)	31	4.52	69	4.59	0.08
		(0.11)		(0.07)	(0.13)
Years of experience	43	13.15	90	11.99	-1.15
		(1.32)		(0.91)	(1.60)
Seniority (Exp >= 7)	43	0.72	90	0.64	-0.08
		(0.07)		(0.05)	(0.09)

Notes: Comparison of treatment and control means using two-sided Welch's t-tests. Standard errors are reported in parentheses below the means and differences. "Understanding of study" and "Confidence in ability" were both collected during the onboarding process, prior to granting access to the study-provided AI tool. * p<0.10, ** p<0.05, *** p<0.01.

Table A2: Linear Probability Models for Sample Attrition from Primary Experimental Tasks

	(1) 10-Day Drafting	(2) 90-Day Drafting	(3) 90-Day Redlining
Treatment	0.04	−0.02	−0.04
	(0.07)	(0.08)	(0.09)
Senior	−0.03	−0.08	−0.14
	(0.08)	(0.09)	(0.10)
Understanding of study (Likert)	0.03	0.05	0.02
	(0.12)	(0.11)	(0.10)
Confidence in ability to complete tasks (Likert)	−0.07	−0.03	0.02
	(0.10)	(0.10)	(0.09)
Onboarding response	−0.17*	−0.21**	−0.24**
	(0.09)	(0.10)	(0.12)
Constant	0.51	0.41	0.37
	(0.36)	(0.37)	(0.38)
Observations	133	133	123
F-Statistic	1.05	1.20	1.27
Prob > F	0.39	0.31	0.28

Notes: Robust standard errors in parentheses. "Understanding of study" and "Confidence in ability" were both collected during the onboarding process, prior to granting access to the study-provided AI tool. To include the full sample of randomized participants (133), missing responses for these two covariates were imputed using their sample means. An "Onboarding response" indicator variable is included which equals 1 if the participant provided responses to these questions, and 0 otherwise. Firm 3 completed the experimental protocol before the redlining exercise was finalized, thus is excluded from the final column of this table. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A3: Experimental Performance Scores, Overall and by Seniority, Using LLM Ratings

	All		Junior		Senior	
	Control	Treatment	Control	Treatment	Control	Treatment
A. Main Outcomes: 10-Day Drafting						
Pooled Score	0.00	0.50	−0.27	0.76	0.12	0.37
	(0.95)	(0.79)	(1.04)	(0.29)	(0.90)	(0.92)
Enforceability	0.00	0.54	−0.32	0.82	0.15	0.40
	(1.00)	(0.84)	(1.06)	(0.39)	(0.96)	(0.96)
Technical Accuracy	0.00	0.46	−0.32	0.70	0.15	0.35
	(1.00)	(0.87)	(1.27)	(0.45)	(0.84)	(1.00)
Strategic Ambiguity	0.00	0.40	−0.28	0.64	0.13	0.28
	(1.00)	(0.87)	(0.83)	(0.33)	(1.06)	(1.02)
Completeness & Alignment	0.00	0.52	−0.19	0.85	0.09	0.36
	(1.00)	(0.93)	(1.10)	(0.39)	(0.96)	(1.06)
Clarity	0.00	0.57	−0.23	0.79	0.11	0.46
	(1.00)	(0.78)	(1.17)	(0.38)	(0.92)	(0.90)
Observations	35	70	11	23	24	47
B. Main Outcomes: 90-Day Drafting						
Pooled Score	0.00	0.89	0.41	0.89	−0.17	0.89
	(0.88)	(0.92)	(0.66)	(1.07)	(0.91)	(0.85)
Enforceability	0.00	0.93	0.39	0.96	−0.16	0.92
	(1.00)	(1.25)	(0.72)	(1.27)	(1.07)	(1.26)
Technical Accuracy	0.00	0.67	0.34	0.71	−0.14	0.64
	(1.00)	(1.03)	(0.94)	(1.20)	(1.01)	(0.95)
Strategic Ambiguity	0.00	0.89	0.46	0.77	−0.19	0.94
	(1.00)	(0.85)	(0.80)	(1.07)	(1.03)	(0.74)
Completeness & Alignment	0.00	1.03	0.59	0.98	−0.24	1.06
	(1.00)	(0.95)	(0.86)	(1.12)	(0.97)	(0.88)
Clarity	0.00	0.92	0.30	1.00	−0.12	0.89
	(1.00)	(0.93)	(0.82)	(1.04)	(1.06)	(0.89)
Observations	31	67	9	21	22	46
C. Main Outcomes: 90-Day Redlining						
Pooled Score	0.00	0.60	0.15	0.52	−0.05	0.63
	(0.96)	(0.88)	(1.05)	(0.91)	(0.96)	(0.88)
Enforceability	0.00	0.62	0.06	0.48	−0.02	0.68
	(1.00)	(0.89)	(1.02)	(0.93)	(1.02)	(0.88)
Technical Accuracy	0.00	0.51	0.11	0.44	−0.04	0.54
	(1.00)	(0.92)	(1.17)	(0.90)	(0.97)	(0.94)
Strategic Ambiguity	0.00	0.63	0.11	0.63	−0.04	0.63
	(1.00)	(0.96)	(1.10)	(0.98)	(0.99)	(0.97)
Completeness & Alignment	0.00	0.58	0.29	0.49	−0.09	0.62
	(1.00)	(0.92)	(1.12)	(0.92)	(0.97)	(0.93)
Clarity	0.00	0.64	0.19	0.55	−0.06	0.67
	(1.00)	(0.89)	(1.01)	(0.94)	(1.01)	(0.88)
Observations	29	62	7	17	22	45

Notes: Table presents standard normal cell means with standard deviations reported in parentheses below. Sample restricts to successfully randomized subjects completing the study. Main outcome observations are at the individual LLM rating level. Experience ('Senior') requires ≥ 7 years of practice. All dependent variables are standardized using the control group mean and variance. Pooled task scores represent the simple arithmetic average of standardized subcomponents.

Table A4: Impact of 90-Day AI Access on AI-Unassisted Redlining Performance: Controlling for Suspected Non-Adherence Using Expert Ratings

	Full Sample		Controlling for Non-Adherence		Excluding Non-Adherence	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.32**		0.32**		0.25	
	(0.15)		(0.16)		(0.17)	
Treat × Junior		−0.03		−0.03		−0.11
		(0.20)		(0.20)		(0.21)
Treat × Senior		0.45**		0.44**		0.39*
		(0.19)		(0.19)		(0.20)
Flagged for possible AI use			0.18	0.16		
			(0.18)	(0.18)		
Junior		0.21		0.18		0.19
		(0.12)		(0.13)		(0.12)
Senior		0.09		0.06		0.06
		(0.15)		(0.15)		(0.15)
Constant	−0.20		−0.21		−0.16	
	(0.17)		(0.17)		(0.18)	
$H_0 : \alpha_{\text{junior}} < \alpha_{\text{senior}}$		0.26		0.26		0.25
$H_0 : \beta_{\text{junior}} < \beta_{\text{senior}}$		0.96		0.95		0.95
$H_0 : \alpha_{\text{junior}} + \beta_{\text{junior}} < \alpha_{\text{senior}}$		0.35		0.35		0.46
$H_0 : \alpha_{\text{junior}} + \beta_{\text{junior}} < \alpha_{\text{senior}} + \beta_{\text{senior}}$		0.95		0.95		0.93
Rating-level Disaggregated	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Subscale FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	925	925	925	925	775	775
R^2	0.10	0.12	0.10	0.12	0.09	0.11

Notes: Table reports intent-to-treat estimates using OLS regressions for 90-day redlining quality evaluated by expert raters. All models report rating-level disaggregated regressions using standard errors clustered at the individual level. Models (1) and (2) report estimates on the full active sample. Models (3) and (4) control for suspected non-adherence, while Models (5) and (6) exclude participants with suspected non-adherence. Non-adherence is flagged by Gemini on the basis of the prompt we provide in the Supplementary Materials. All outcome subcomponents are standardized to mean zero and unit variance using the control group calculated locally within the full active sample. All models incorporate firm fixed effects and subscale fixed effects, and are weighted by $w_i = 1/n_i$, where n_i is the number of ratings individual i receives so that each individual carries equal total weight. The hypothesis tests in the bottom panel report one-sided p-values. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A5: Impact of 90-Day AI Access on AI-Unassisted Redlining Performance: Controlling for Suspected Non-Adherence Using LLM Ratings

	Full Sample		Controlling for Non-Adherence		Excluding Non-Adherence	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.56***		0.58***		0.61***	
	(0.19)		(0.19)		(0.21)	
Treat × Junior		0.29		0.29		0.24
		(0.36)		(0.38)		(0.42)
Treat × Senior		0.66***		0.68***		0.76***
		(0.22)		(0.22)		(0.25)
Flagged for possible AI use			−0.41*	−0.43*		
			(0.25)	(0.26)		
Junior		0.07		0.14		−0.03
		(0.30)		(0.32)		(0.34)
Senior		−0.02		0.06		−0.04
		(0.17)		(0.17)		(0.18)
Constant	0.29		0.31		0.32	
	(0.20)		(0.20)		(0.23)	
$H_0 : \alpha_{\text{junior}} < \alpha_{\text{senior}}$		0.40		0.41		0.48
$H_0 : \beta_{\text{junior}} < \beta_{\text{senior}}$		0.81		0.81		0.86
$H_0 : \alpha_{\text{junior}} + \beta_{\text{junior}} < \alpha_{\text{senior}}$		0.10		0.10		0.22
$H_0 : \alpha_{\text{junior}} + \beta_{\text{junior}} < \alpha_{\text{senior}} + \beta_{\text{senior}}$		0.87		0.89		0.97
Rating-level Disaggregated	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Subscale FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	455	455	455	455	380	380
R^2	0.26	0.27	0.28	0.29	0.27	0.30

Notes: Table reports intent-to-treat estimates using OLS regressions for 90-day redlining quality evaluated by LLM raters. All models report rating-level disaggregated regressions using standard errors clustered at the individual level. Models (1) and (2) report estimates on the full active sample. Models (3) and (4) control for suspected non-adherence, while Models (5) and (6) exclude participants with suspected non-adherence. Non-adherence is flagged by Gemini on the basis of the prompt we provide in the Supplementary Materials. All outcome subcomponents are standardized to mean zero and unit variance using the control group calculated locally within the full active sample. All models incorporate firm fixed effects and subscale fixed effects, and are weighted by $w_i = 1/n_i$, where n_i is the number of ratings individual i receives so that each individual carries equal total weight. The hypothesis tests in the bottom panel report one-sided p-values. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A6: Impact of AI Access on Patent Drafting Performance at 90 Days

	Patents worked on during study period (Number)		Level of responsibility for drafting of patents worked on		Length of time to completion for patents (Weeks)		Extent of reviewer input on patent applications		Extent of edits, corrections or comments from reviewers on drafts	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Treatment	−1.06		0.01		−0.76		−0.39		−0.15	
	(0.79)		(0.25)		(0.81)		(0.28)		(0.28)	
Treat × Junior		−0.58		0.41		−1.65		−0.31		−0.09
		(1.33)		(0.31)		(1.71)		(0.49)		(0.53)
Treat × Senior		−1.29		−0.16		−0.30		−0.41		−0.15
		(1.01)		(0.34)		(0.92)		(0.37)		(0.32)
Junior		4.39		0.24		4.39		0.30		0.40
		(1.17)		(0.31)		(1.51)		(0.32)		(0.38)
Senior		5.03		−0.06		3.11		0.03		−0.20
		(0.84)		(0.24)		(0.86)		(0.27)		(0.24)
Constant	5.27		−0.05		5.14		0.04		−0.07	
	(0.74)		(0.27)		(0.97)		(0.26)		(0.27)	
$H_0 : \alpha_{\text{junior}} < \alpha_{\text{senior}}$		0.67		0.25		[0.75]		0.28		0.11
$H_0 : \beta_{\text{junior}} < \beta_{\text{senior}}$		0.34		0.11		[0.25]		0.44		0.46
$H_0 : \alpha_{\text{junior}} + \beta_{\text{junior}} < \alpha_{\text{senior}}$		0.84		0.02**		[0.38]		0.53		0.13
$H_0 : \alpha_{\text{junior}} + \beta_{\text{junior}} < \alpha_{\text{senior}} + \beta_{\text{senior}}$		0.48		0.00***		[0.47]		0.21		0.07*
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	68	68	66	66	67	67	66	66	66	66
R^2	0.11	0.12	0.21	0.32	0.23	0.24	0.10	0.12	0.19	0.25

Notes: Table reports intent-to-treat estimates using OLS regressions for on-the-job patent metrics collected during the 90-day post-task survey. All estimations conducted at the subject level with robust (HC1) standard errors. Columns (2), (4), (6), (8), and (10) report split specifications with no global intercept. Unless otherwise indicated in column titles, all outcomes are standardized with mean 0 and standard deviation of 1 using the control group distribution. The hypothesis tests in the bottom panel report one-sided p-values; values in square brackets in column (6) indicate that the inequality direction in each null hypothesis is reversed for time to completion. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A7: Evaluation Rubric for Drafting and Redlining

Category	1	2	3	4	5
Enforceability & Legal Strength	Claims are vague/ contradictory; disclaiming language present; major PR vulnerabilities; fails global standards.	Claims are imprecise; potential disclaimers or PR vulnerabilities; meets only basic global norms.	Claims are sufficiently precise with minimal "patent profanity"; minimal PR risks; meets most global norms.	Claims are precise and compliant with most global standards; effectively avoids disclaimers/PR pitfalls.	Claims drafted with robust precision, free of disclaimers, expertly manage PR exposure, fully satisfy global requirements.
Technical Accuracy	Technical details are inaccurate; major errors suggest poor understanding of the invention.	Several technical inaccuracies undermine credibility and clarity of the draft.	Details generally accurate with minor errors that do not undermine the description. Conveys core technical aspects.	Details are accurate and sufficiently detailed. Demonstrates solid grasp of the invention's complexities.	Details are exemplary and meticulously detailed. Demonstrates masterful understanding of relevant technical details.
Strategic Ambiguity	Ambiguity is haphazard with no rationale; confusion outweighs protective benefit.	Some ambiguity present but does not seem deliberate or beneficial.	Ambiguity occasionally employed to broaden claims but not always optimized; overall meaning intact.	Ambiguity thoughtfully applied to expand claim coverage without sacrificing essential clarity.	Ambiguity masterfully balanced to maximize protective scope while retaining clear, effective disclosures.
Completeness & Alignment	Missing essential elements/nuance; does not meet fundamental patent requirements; generic.	Some key elements present, but notably incomplete/lacks nuance; partially fulfills objectives.	Covers essential elements needed; meets standard patent requirements.	Thoroughly covers all necessary elements; strong alignment with patent's objectives.	Exhaustively addresses every relevant element, reflects appropriate nuance, surpasses all requirements/objectives.
Clarity	Disorganized, ambiguous, lacks detail. Inconsistent terminology creates confusion, severely impedes understanding.	Unclear in several parts; terminology usage is not fully consistent.	Generally clear and consistent; minor confusion does not prevent overall understanding.	Very clear and cohesive throughout; terminology is consistently applied and well-explained.	Exceptionally clear, organized, easy to follow; terminology used precisely and enhances coherence.