STONE CENTER | INEQUALITY & SHAPING THE FUTURE OF WORK

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

Stone Center on Inequality and Shaping the Future of Work
70 Memorial Drive, Building E51-355 - Cambridge, MA 02142

shapingwork.mit.edu

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

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." The registration can be found at https://www.socialscienceregistry.org/trials/15823 We acknowledge generous input from Mauricio Carneiro, James Manyika, Ruth Porat, Kiera Russell, and Tom Shane (all of Google), as well as the Google InFlow 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

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
JEL No. I24, I29, J0

ABSTRACT

Whether AI assistance builds or erodes professional expertise is unsettled. 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 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. Paralleling findings from other white-collar domains, 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 requiring 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. The largest gains from AI thus accrued to the lawyers who retained the least. 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 (among others) master foundational skills while performing core job tasks under the supervision of senior practitioners. Such expertise has 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. But AI has the potential to disrupt this skill acquisition process (Beane, 2024; Acemoglu et al., 2026; Afrouzi et al., 2026; Asriyan et al., 2026; Ide, 2026): on the positive side, AI could facilitate the acquisition of expert judgment by providing high-quality examples of professional work and tailored, on-the-spot feedback that managers otherwise might not supply or that junior workers might hesitate to seek; alternatively, AI might stunt skill acquisition if junior workers bypass learning by cognitively offloading tasks to algorithms—crowding out the development of the expert judgment that effective AI use requires.[1] Evaluating the productivity consequences of 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, likely because it requires regular AI usage over an extended time interval, credible assessment of skills acquisition during AI usage, and a robust research design, a combination that is rarely available in existing studies of AI implementations. We meet these three 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 tactical scoping of claims), completeness and clarity.[3]

[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.

[3] Strategic ambiguity refers to the intentional use of broad, flexible language that maximizes a patent's future enforcement

The trial comprised three assessments at two points in time. At 10 days, lawyers completed a patent-drafting task; at 90 days, they completed both a second drafting task and an unassisted redlining task. The drafting and redlining tasks serve different intellectual objectives. The drafting tasks measure the immediate impact of AI on realized machine-assisted performance through a treatment-control comparison. Moreover, because treatment and control face the same task in each wave, comparing the gap between AI-assisted and control lawyers at 10 days against the same gap at 90 days reveals whether performance differences widen or narrow with extended use.

The redlining task instead evaluates the development of professional judgment. Conducted 'offline', i.e., without AI, it asks lawyers to critically evaluate the quality of existing professional work, a task patent lawyers routinely perform unaided and one for which AI is both less applicable and less commonly used than in drafting. Redlining consists mainly of reading and marking up existing work rather than producing new text, and hence foregrounds expert judgment. By comparing the offline performance of lawyers immersed in AI-assisted drafting for three months against that of lawyers without AI access, the redlining task lets us separate the skill lawyers have internalized from their AI-assisted performance.[4]

Our experimental treatment yields countervailing impacts of AI assistance on task performance versus skill acquisition. A first finding is that AI assistance meaningfully improved performance in patent drafting, with treated lawyers outperforming controls by 0.34 SD at 10 days ($p = 0.03$) and 0.38 SD at 90 days ($p = 0.01$). Treatment effects appeared across all five domains on which drafts were scored: enforceability, accuracy, ambiguity, completeness and clarity. Treated lawyers earned higher average scores not by generating a larger fraction of high scores, but by earning fewer low scores and more mid-range scores. We detect a comparable pattern among senior lawyers: gains in average performance were driven by a reduction in low scores. Consistent with a growing literature documenting AI-driven compression of realized performance differentials between workers with lesser versus greater skill and experience, junior lawyers realized larger gains from AI assistance than seniors.

The performance gains in output quality realized by AI-assisted lawyers did not, however, consistently translate into improved professional judgment in non-AI-assisted work. On the offline redlining task designed to evaluate professional judgment, the treatment group outperformed the control group (0.32 SD, $p = 0.04$). But these learning gains were concentrated *entirely* among senior lawyers; junior lawyers, the group that captured the largest contemporaneous gains in drafting tasks with AI access, showed no comparable carryover. As in AI-assisted drafting, the higher redlining performance of treated senior lawyers stemmed from fewer mediocre and poor scores rather than from any gain at the top of the distribution. Conversely, treated juniors showed no average gain in

scope while respecting the USPTO's "Broadest Reasonable Interpretation" standard.

[4] 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.)

performance and substantially greater dispersion: sharply fewer mediocre (fourth-quintile) scores, significantly more poor (bottom-quintile) scores, and an offsetting rise in good (second-quintile) scores. This dispersion is consistent with AI accelerating skill acquisition among some juniors and substituting for it among others. Our experiment does not answer why these treatment effects appear positive for some lawyers and negative for others.

Extensive robustness tests confirm the main findings. The aggregate effects of AI assistance on realized performance in AI-assisted tasks, and in offline performance in unassisted tasks, are broadly affirmed across all subscales of our performance metrics (enforceability, accuracy, ambiguity, completeness and clarity). The effects of three months of AI use on subsequent offline performance are similarly consistent. AI assistance improves senior performance in the redlining task near-uniformly across domains (though, of course, precision is diminished when focusing on subscales). It fails to improve junior performance across any domain. Leveraging the fact that each task is scored by at least two independent raters, we verify that the results are not driven by rater idiosyncrasies. Recognizing the risk of inconsistent standard error coverage in small samples, we also report a nonparametric permutation test that strongly affirms the primary findings.

We additionally deploy a Large Language Model to rate lawyers' outputs in both drafting and redlining tasks. These ratings support the conclusion that both juniors and seniors produce better patent drafts with AI assistance, with larger gains among junior lawyers. LLM evaluations also detect some performance improvement on the redlining task among treated juniors, though these effects are not as large as among treated seniors.

We extensively probe the reliability of our most intriguing finding, which is that senior lawyers realize significant performance gains in offline work following three months of AI assistance. We first evaluate this result's sensitivity to the classification of junior versus senior lawyers. Using seniority thresholds of as low as three and as high as nine years, we find no evidence that extended use of AI affected performance of junior lawyers in the offline redlining task. Moreover, when stratifying lawyers into three seniority tiers (junior, mid-level, and senior), we again find that performance gains in assisted tasks occur at all seniority levels but that performance gains in unassisted work is limited to senior lawyers (see Figure A4).

A second potential concern is that senior lawyers in the treatment group might have used AI for the redlining exercise itself, thus skewing the results towards detecting spillovers from AI-assisted to unassisted tasks. To evaluate this threat, we used a combination of LLM evaluations and manual audits to identify 'tells' of AI usage, including full-paragraph rewrites, artificial tonal transitions (e.g., "compliment sandwiches"), structural context loss and excessively stylized feedback. Additionally, we evaluated junior-level submissions for distinct behavioral tells, specifically flagging unnatural leaps in macro-level competence and a uniform edit density that lacks the sequential time-collapse and fatigue typical of human reviewers. Erring on the side of over-classifying non-adherence, we identify 15 of 91 sample-included redlining submissions where AI usage appears possible. We then re-ran our main analyses, controlling for possible non-adherence in one set of models and

excluding cases where non-adherence appears possible in another set of models. In models that control for non-adherence, we find that suspected non-adherents perform *insignificantly* better than other treated subjects in the redlining task. Controlling for suspected non-adherence does not affect our main inference, however. In a more conservative test, we remove these subjects from the sample entirely. This step reduces precision but does not overturn our main result. When we include LLM ratings in this restricted sample, our results are stronger still.

Additional supporting findings contextualize the main results. First, lawyers drafting patents using AI assistance worked modestly faster while producing higher-quality patent drafts. In the 10-day task, the treatment group saved 10 minutes ($p = 0.05$) against a baseline of 112 minutes, with directionally consistent, though not statistically significant, savings of 10 minutes ($p = 0.27$) against a 125-minute base at 90 days. Senior lawyers did not work faster with AI. Although in the control group, senior lawyers completed the patent drafting assignment in 13 (for the 90-day task) to 24 (for the 10-day task) percent faster than junior lawyers, AI assistance did not speed their work in either the 10 or 90-day drafting tasks. As a result, juniors with AI assistance worked as rapidly as seniors with or without AI assistance (and delivered equivalent quality drafts, as above).

Results for the redlining task, in which neither the treatment nor control group is assisted by AI, reveal the reverse seniority pattern for quality and time-savings. In both the treatment and control groups, senior lawyers spend about 12 (treatment) to 19 (control) percent *longer* than juniors marking up patent drafts, averaging 63 (senior) versus 53 (junior) minutes in the control group. And while AI assistance in patent drafting appears to yield higher (non-AI) redlining performance among senior lawyers, it does not reduce their time on task: treated and untreated seniors spend more time than juniors on the redlining task but roughly the same amount of time as each other.

Our findings point to a potential trade-off between immediate productivity gains and durable skill acquisition in this expertise-intensive professional domain. Although AI does not erode professional skill on average, junior lawyers capture the greatest immediate productivity gains during AI use, paralleling evidence from domains including writing (Noy and Zhang, 2023), customer service (Brynjolfsson et al., 2024) and software development (Ju and Aral, 2025). However juniors show no consistent durable skill gains, and in fact demonstrate increased skill dispersion: more good and more poor scores with no gain at the top of the distribution. Over a longer treatment period, it is plausible—though not certain—that this dispersion would yield a fanning out of performance levels. The finding that the greatest performance gains accrue to those who develop the least lasting expertise suggests that foundational expertise may be a prerequisite for extracting durable learning from AI assistance. This trade-off matters to the extent that professional work continues to demand judgment exercised without AI. We expect that it will: even if AI comes to assist tasks like patent redlining, it is unlikely to be available to lawyers in real time during client meetings and courtroom appearances, where stakes are high and judgment must be exercised extemporaneously.

2 AI in professional services: Productivity and skill development

AI adoption in the workplace is already high and rising rapidly. Bick et al. (2026) report that 43% of US workers used generative AI for work in the first quarter of 2026, with average reported time savings of 5.3 hours among AI users. Adoption is concentrated in knowledge-work occupations: close to 60% in business, management, and finance, compared with about 30% in personal service, blue-collar, healthcare, and sales (Bick et al., 2026).

Legal practice nonetheless poses distinctive challenges for AI integration: 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). Nevertheless, legal services have followed the trend of rapid AI adoption: the share of legal organizations actively using generative AI nearly doubled from 14% in 2024 to 26% in 2025, with 45% of law firms either currently using or planning to make AI central to their workflow within one year (Thomson Reuters Institute, 2025).

Legal practice is a consequential setting for studying AI's effects on professional work: the profession combines high-stakes output, domain-specific expertise that takes years to acquire, and an apprenticeship structure through which junior practitioners develop expert judgment (Remus and Levy, 2017). Intellectual property law, the domain we study, intensifies these features: the precise language of a patent claim determines its legal force, drafting defensible claims requires years of supervised practice, and evaluating patent quality is itself an expert domain.

2.1 Productivity impacts of AI use

Many recent studies evaluate the causal effects of AI usage on performance in professional services settings, with particular emphasis on writing tasks. Following the release of ChatGPT in late 2022, early empirical work found robust support for improvements in productivity across white-collar professions 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 use narrows the performance gap across both tenure and mastery, that is, 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). 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.

Results from studies of knowledge workers (or knowledge worker trainees) are heterogeneous: some find speed gains (Dillon et al., 2025), others find quality gains (Ju and Aral, 2025), and a few find both (Haslberger et al., 2025; Usdan et al., 2024). One explanation for these heterogeneous findings is that workers may reallocate time savings into additional effort on the same task, so an underlying productivity gain surfaces as either speed or quality or both, depending on circumstances (Humlum and Vestergaard, 2025a). Separately, several studies document improvements in job satisfaction

and emotional response to AI introduction, with mixed evidence on whether these gains accrue disproportionately to less-experienced or less-expert workers (Dell'Acqua et al., 2026; Jia et al., 2024; Cruces et al., 2026). (Some emerging evidence suggests that satisfaction gains from AI exposure may vary with the extent of exposure; Chiong and Xie, 2024)

A growing body of evidence qualifies these productivity gains, documenting cases in which AI fails to improve performance or actively 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) similarly 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 preregistered 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. Together, these findings suggest that gains from AI use are conditional on users' ability to evaluate, integrate, or override its outputs, a capacity that itself requires substantial domain expertise.

Our experiment builds most directly on studies of AI use in legal writing. 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, nearly all conducted with law students, find that these 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. Effects vary across the ability distribution: Choi and Schwarcz (2025) report an "equalizing effect on the legal profession," with sharp gains for students at the bottom of the class and declines for those at the top. Schwarcz et al. (2026), using two "reasoning" models that generate intermediate steps before answering, find consistent and significant quality improvements on five of six assignments completed by upper-level students. Closest to our 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. 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 of AI use in law to measure *skill acquisition* rather than contemporaneous productivity alone, and the first to do so with *practicing* lawyers rather than students.

2.2 Professional skill development

The literature on productivity impacts of AI use has rapidly advanced, but much less is known about whether AI use in judgment-critical settings fosters or inhibits the acquisition of that judgment. 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 development will diverge from its short-term effects on output.

Recent evidence supports the concern that AI access may hinder skill acquisition. 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 abilities. In professional services, Wiles et al. (2024) demonstrate in an experiment with management consultants that while generative AI acts as a "temporary exoskeleton" for tackling unfamiliar data science tasks, these enhanced technical abilities do not persist when the tool is removed. This lack of durable learning aligns with findings from Dell'Acqua et al. (2026), who document a "falling asleep at the wheel" effect where highly skilled professionals over-rely on AI, shutting down their own critical judgment and verification processes. In a non-experimental setting, Budzyń et al. (2025) find that routine AI assistance may lead to diagnostic skill deterioration among endoscopists performing routine colonoscopies.

Yet, empirical evidence also points to scenarios where AI successfully supports retention of knowledge acquired through working with AI tools. In radiology diagnostics, He et al. (2026) find that when predictive AI is incorporated into the workflows of novice radiologists, their subsequent independent diagnostic accuracy improves. In the context of general business problem-solving, Cruces et al. (2026) document that less-educated participants showed modest but significant improvements in unassisted follow-up tasks after using generative AI. Lira et al. (2025) also find that job-seekers who practiced writing cover letters with AI assistance improved their independent writing skills more than those who practiced without it.

These studies point to a common pattern: workflows that require users to interpret, evaluate, or verify AI outputs build skill, while those that permit uncritical acceptance bypass the friction needed for skill formation. Bastani et al. (2025), in a field experiment with nearly 1,000 high school mathematics students, find that unconstrained access to GPT-4 during practice improves in-session performance but harms subsequent unassisted exam performance; the harm disappears when the same model is reconfigured as a tutor that withholds direct answers and requires students to reason through problems. Lee et al. (2025), in a survey of 319 knowledge workers, find that higher confidence in AI is associated with less critical engagement during use, and that AI shifts cognitive effort away from task execution and toward output verification. He et al. (2026) document that skill acquisition is strongest when AI is deployed during both training and practice, yielding the largest unassisted gains in diagnostic precision and recall.

Additional research speaks to the longer-term effects of tool usage on expertise acquisition. 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. Emanuel et al. (2026) quantify this process, finding that software engineers who sit near their teammates receive more feedback and subsequently write better code, with the gains concentrated among junior engineers. A theoretical literature explores potential consequences, showing that automating entry-level tasks can raise current output while slowing the accumulation of expertise, though the long-run impacts turn on what each model assumes about how learning occurs (Afrouzi et al., 2026; Asriyan et al., 2026; Ide, 2026). These considerations imply that the consequences of AI usage on skill formation cannot be gauged from contemporaneous output; direct measures of learning are required.

The present study assesses how sustained AI use shapes quality, productivity, and, most centrally, the development of expert judgment. By tracking working professionals over three months in a high-stakes domain and pairing AI-assisted output with a separate unassisted task scored by independent expert raters, it combines a real-world setting, the time horizon needed to assess learning gains, and a design that separates AI's effects on performance from its effects 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 patent law firm, treated two-months apart, are treated as independent firms. Thus our analytic sample includes 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. Of the 156 intellectual property lawyers put forward by firms, 133 (85%) completed the onboarding requirements and were successfully randomized into the trial. Small post-randomization samples within each firm, as well as some later drop-out, led to slight deviation in the realized versus target ratios of treatment relative to control subjects; ultimately, 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. (These final two measures were collected as Likert items during an onboarding survey administered alongside the study consent form prior to InFlow access being granted.)

3.2 Experimental protocol

Subjects in the treatment group were provided access via a gated online platform to InFlow, an unreleased product developed by Google Labs. InFlow was designed to assist writing in professional services settings. InFlow integrates user-provided context, style guidance, and user-specified writing goals. Key tool attributes include:

- In-line modification: Prompt-based text modification and insertion for text selected by user;
- AI Typing: Contextual AI-generated text inserted at location specified by user;
- Chat with Editor / Expert / Audience: AI assistant feedback from multiple professional perspectives;
- Critique: Semantic inconsistencies within a document identified based on a user-defined or auto-created rule set.

As a pre-release product, InFlow underwent several feature changes during the study period of May 2025 through February 2026. The redlining feature was launched on July 24, 2025, after one firm had already completed the second test task and exited the study. (Incidentally, this firm was not exposed to the redlining task in any case, as the task had not been finalized and deployed until after this initial firm had finished their submissions) This feature aimed to enable standardization of document templates and automated editorial review against user-defined criteria.[5]

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,

[5] A full summary of InFlow feature changes is included in Appendix A2.1. InFlow was not ultimately released as a standalone product; its features were integrated into Google's other AI tools.

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 "Robustness of main results to choice of experience threshold" in Section 4.4 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.

covering: (1) commonly used AI terms; (2) an introduction to LLMs; and (3) prompt engineering. Participants assigned to the treatment group gained access to InFlow and to a second training module tailored to the intervention. This module covered: (1) recommendations for using InFlow; (2) instructions for optimizing InFlow 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 InFlow[6], all participants were assigned a patent drafting task (see Appendix A2.2 for full materials). 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* - claims 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 keep the write-up to 1,000 words. (35 subjects in the control group and 70 subjects in the treatment group successfully completed this exercise, representing 81% and 78% respectively of all randomized subjects in the group)

90-day assessments: Participants were assigned two tasks (see Appendix A2.3 for full materials):

- *Drafting:* A second drafting task, similar in structure to the first but with a different invention, figures, and notes (31 subjects in the control group and 67 subjects in the treatment group successfully completed this exercise, representing 72% and 74% respectively of all randomized subjects in the group); 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 4.4. (29 subjects in the control group and 62 subjects in the treatment group successfully completed this exercise, representing 67% and 69% respectively of all randomized subjects in the group)

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). Distinct from the two drafting tasks, the redlining task was intended to evaluate expert judgment without AI assistance, gauging whether

[6] In practice, 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 A5 for full chronology

performance gains attained while using AI translate into improved judgment when AI is unavailable.

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%, see Table 1).

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 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.[7] These estimates provide no evidence of differential attrition among control subjects; 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. None of these contrasts is statistically significant, however.

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. This pattern is not mechanically explained by the fact that one firm completed the study before the onboarding questions and redlining tasks were introduced, since that firm is excluded from the redlining attrition regression. We infer that participants who failed 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, blinded raters drawn at random from a pool of six raters at a contracted intellectual property law firm. Raters applied a rubric developed in collaboration with Google's in-house counsel that evaluated five core dimensions: enforceability, accuracy, ambiguity, completeness, and clarity (The full scoring rubric is detailed in Appendix A2.4). As shown in Figure A1, expert raters show substantial agreement on our main outcome variables, with Pearson correlations of 0.59 ($p = 0.00$), 0.47 ($p = 0.00$) and 0.55 ($p = 0.00$) for the 10-day drafting task, 90-day

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

drafting task and 90-day redlining task respectively.[8]. We also supplemented these professional evaluations with an LLM rater, discussed below.

Each drafting task had two separately scored parts, the dependent claims and the detailed description, and both raters scored each part. Given the similarity of the two parts, we average across parts but within raters and quality dimensions to obtain 10 individual scores per participant for each of the two drafting tasks (10-day and 90-day): 5 quality dimensions $\times$ 2 raters. The redlining task was a single exercise, yielding two scores per dimension.

InFlow telemetry did not record time spent, so we measure time-on-task in two ways. First, participants were instructed to run a stopwatch application while working and to record the elapsed 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.

The post-task surveys also recorded participants' perceptions of the task 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'). The 90-day survey added a module on subjects' on-the-job patent activity over the three-month study period: for each patent drafted in the course of routine work, excluding the test tasks, subjects reported completion time, rounds of internal feedback and review, and perceived difficulty of drafting. We average these per-patent reports to the subject level.

We also fielded monthly surveys, generally sent on the first of the month, as a monitoring tool rather than as a source of outcome measures. These surveys tracked perceived changes in routine-work difficulty, speed, job satisfaction, and internal reviewer scrutiny, along with subjects' use of generative AI tools outside of InFlow, allowing us to watch for problems with satisfaction, contamination, or spillover that might threaten our sample or our design. We do not use them for hypothesis testing.

3.4 Model specification

We estimate intent-to-treat (ITT) 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 standardized performance outcome for lawyer i at firm j on rubric dimension k, and T_i is the treatment indicator. Because our primary quality assessments span five rubric dimensions, we estimate two overall treatment effects. First, we regress on the unweighted average of the five

[8] We include a full exploration of human rater robustness as well as complementary LLM-provided ratings in Appendix A1

subscales for each participant, 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). Because the number of ratings per subject (n_i) varies somewhat due to miscommunications between our human raters, we estimate the pooled model at the rating level using weighted OLS with weights defined as $w_i = 1/n_i$ to avoid dropping any ratings while simultaneously ensuring each subject contributes equally to the pooled estimates.

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

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

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[9]:

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

To identify both subgroup baselines, we omit the global intercept and constrain the firm fixed effects appropriately. 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 key experimental outcomes: drafting quality at 10 and 90 days, and 90-day redlining quality, a task both groups were instructed to complete without AI assistance. Table 2 summarizes these outcomes overall and by experience subgroup, reporting the pooled mean across the five rubric dimensions alongside the individual subscale values. Because each subscale is standardized relative to the pooled control-group distribution (mean zero, standard deviation one), the overall treatment-group means can be read directly as effect sizes in standard deviations.[10] Within experience subgroups, the control means are not zero, so subgroup effects are given by the treatment-control difference rather than the treatment mean alone. many of our main findings are visible from these summary statistics; the regression tables that follow provide formal statistical tests.

[9] This threshold is based on industry standards as per advice from Google's in-house counsel and is tested for sensitivity (see Appendix Figure A4)

[10] Because participant scores covary across subscales, the variance of the overall averages 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 $N \times 2$.

4.1 Performance on patent drafting

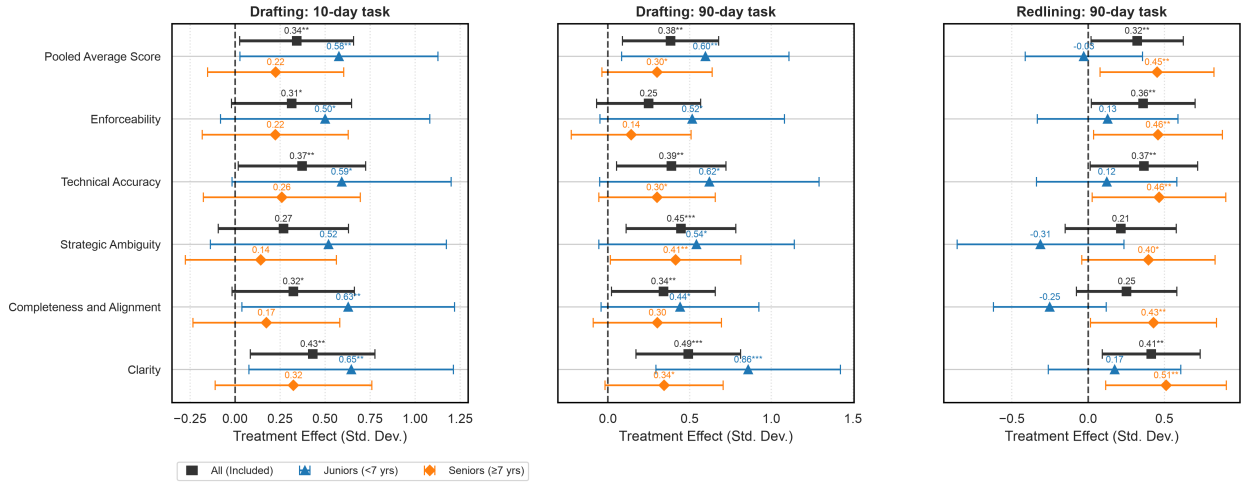
Our first finding is that lawyers assisted by AI draft stronger patents. Column 1 of Tables 3 and 4 reports estimated treatment effects on the quality of patent drafts at the 10-day and 90-day assessments, respectively, using pooled scores averaged across the five rubric subscales. Treated lawyers outperform 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 nearly unchanged in this more demanding specification. We use the estimates from column 2 as our preferred specification.

The robust effect of treatment on the quality of patent drafts is confirmed by the visual evidence in Figure 1. 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 reports a pooled version of column 2 that includes unconstrained treatment effects and intercepts for junior and senior lawyers (Eqn. 2). The effect of AI access on drafting quality is substantially larger for juniors than for seniors. Juniors gain 0.58 SD at 10 days ($p = 0.04$) and 0.60 SD ($p = 0.02$) at 90 days; seniors gain 0.22 SD at 10 days ($p = 0.24$) and 0.30 SD ($p = 0.08$) at 90 days.

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. (Average score also incorporates subindicator fixed effects in order to precisely match column 2 of the combined effects tables, i.e. Table 3, Table 4 and Table 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 a set of tests evaluating how AI access affects the relationship between experience and performance. The expected baseline ordering holds: untreated seniors outperform untreated juniors at both assessments. The junior–senior gap in treatment effects suggests a larger performance boost for juniors, though this contrast is not statistically distinguishable from zero ($p > 0.1$ at both assessments). Substantively, we do not reject the null that juniors using AI surpass the unassisted senior baseline at 10 days ($p = 0.95$) or 90 days ($p = 0.69$). But these tests are low-powered: we also cannot reject the null that that seniors using AI outperformed juniors using AI at both 10 days ($p = 0.19$) and 90 days ($p = 0.90$).

Improvements in drafting quality are not evenly distributed across quality levels, as shown in Table 5 and visualized in the histograms in Figure 2. Across the 10- and 90-day drafting tasks, AI access uniformly reduces the prevalence of performance at the bottom of the quality distribution, in most cases significantly so (probability of scores landing in the bottom quintile declines by -0.19, $p = 0.00$ at 10 days, with a more moderate and less-precisely estimated decline of -0.11, $p = 0.12$ at 90 days). These reductions are largely driven by junior participants, where the treatment–control contrast in the prevalence of bottom-quintile scores is -0.25 ($p = 0.01$) at 10 days and -0.13 ($p = 0.38$)

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)	
$H_0 : \alpha_{\text{junior}} < \alpha_{\text{senior}}$				0.26		0.19		0.19
$H_0 : \beta_{\text{junior}} > \beta_{\text{senior}}$				0.15		0.03**		0.05**
$H_0 : \alpha_{\text{junior}} + \beta_{\text{junior}} > \alpha_{\text{senior}}$				0.05**		0.00***		0.00***
$H_0 : \alpha_{\text{junior}} + \beta_{\text{junior}} < \alpha_{\text{senior}} + \beta_{\text{senior}}$				0.81		0.96		0.94
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-ind. 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. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

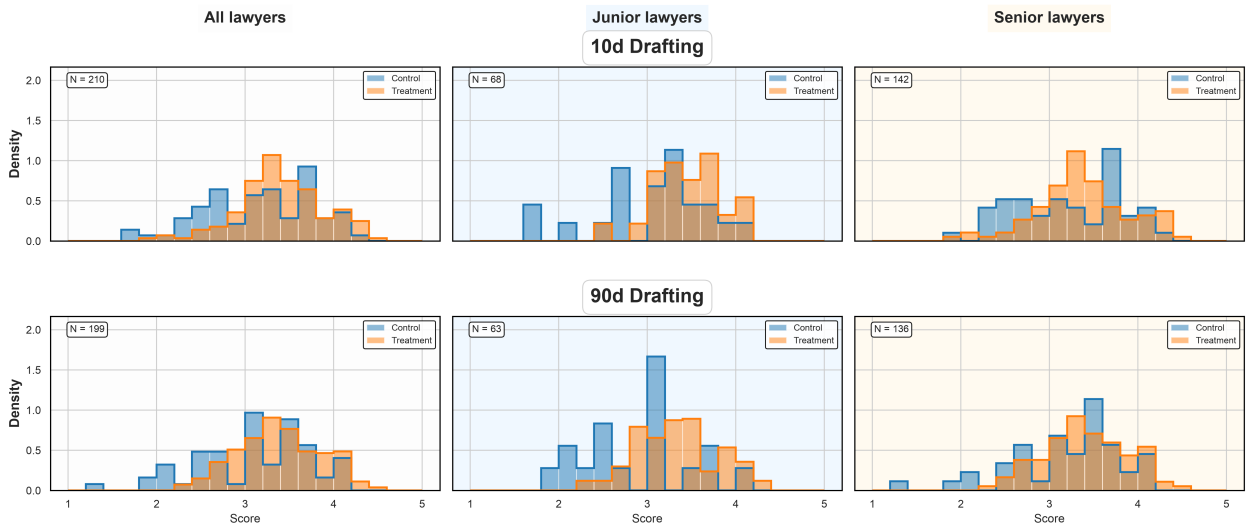
Table 4: Impact of AI Access on Performance in AI-Assisted Drafting Task at 90 Days

	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)	
$H_0 : \alpha_{\text{junior}} < \alpha_{\text{senior}}$				0.05**		0.96		0.29
$H_0 : \beta_{\text{junior}} > \beta_{\text{senior}}$				0.17		0.93		0.49
$H_0 : \alpha_{\text{junior}} + \beta_{\text{junior}} > \alpha_{\text{senior}}$				0.31		0.00***		0.00***
$H_0 : \alpha_{\text{junior}} + \beta_{\text{junior}} < \alpha_{\text{senior}} + \beta_{\text{senior}}$				0.10*		0.44		0.19
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-ind. 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. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

at 90 days relative to the control group. For seniors, the fall in bottom-quintile performance is less pronounced: -0.18 ($p = 0.03$) and -0.09 ($p = 0.24$) for the 10- and 90-day drafting tasks.

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.

In the case of junior lawyers, the displaced mass in the bottom segments of the distribution is largely mirrored by increased mass in the top quintile: the share of junior scores in the top 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. In the case of senior lawyers, the more modest reductions in the prevalence of bottom-quintile scores is offset by greater mass across most other segments of the distribution, particularly the middle quintile for the 10-day task (0.12, $p = 0.04$) and the second quintile at 90 days (0.11, $p = 0.14$), though the latter is less precisely estimated.

The drop in the frequency of bottom-quintile scores implies a reduction in performance variance among treated relative to control subjects—though this may be offset if mass shifts from the bottom to the top of the distribution rather than its center. A Levene test of the equality of variances finds that AI access significantly compresses performance variance on the 90-day drafting task, overall (Levene $p = 0.03$) and among senior lawyers (Levene $p = 0.08$). It does not significantly compress the variance of scores in the 10-day drafting task, though the p-values (in parentheses) are suggestive of an effect.

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

Our findings for patent drafting corroborate the expertise-leveling effects of AI access documented in other knowledge-intensive professional domains, such as writing, customer service, and software (Noy and Zhang, 2023; Cruces et al., 2026; Shen and Tamkin, 2026). This section addresses the key question our experiment was designed to answer: 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 attorneys is weakly negative though imprecisely estimated (-0.03 SD, $p = 0.89$).

This aggregate gain is prevalent across quality subscales for senior but not junior lawyers, as documented in the third panel of Figure 1. When scores are pooled across junior and senior lawyers, 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. These effects are driven entirely by senior lawyers, however. Treated seniors outperform senior controls in all five quality dimensions, four significantly at $p \leq 0.10$. By contrast, point estimates *negative*, though not significantly so, for juniors on two of five subscales.

Table 7 decomposes these distributional differences into their quintile components, with distributions visualized in the histograms in Figure 3. In the pooled comparison, treated lawyers have significantly greater mass in the second-to-top score quintile (0.22, $p = 0.00$) and significantly less mass in the fourth (next-to-bottom) quintile (-0.17, $p = 0.02$). This comparison averages across striking differences in treatment effects by seniority level, however. The redlining scores of treated junior lawyers appear dramatically more dispersed than those of controls, as seen in the middle panel of Figure 3 and confirmed by a Levene statistic of 3.66 ($p = 0.06$). This reflects a sizable fall of -0.42 in the share of fourth-quintile scores ($p = 0.01$), a significant rise 0.22 in the prevalence of bottom-quintile (*poor*) scores ($p = 0.04$), and corresponding gain in the prevalence of second-quintile (good) scores (0.19, $p = 0.20$).

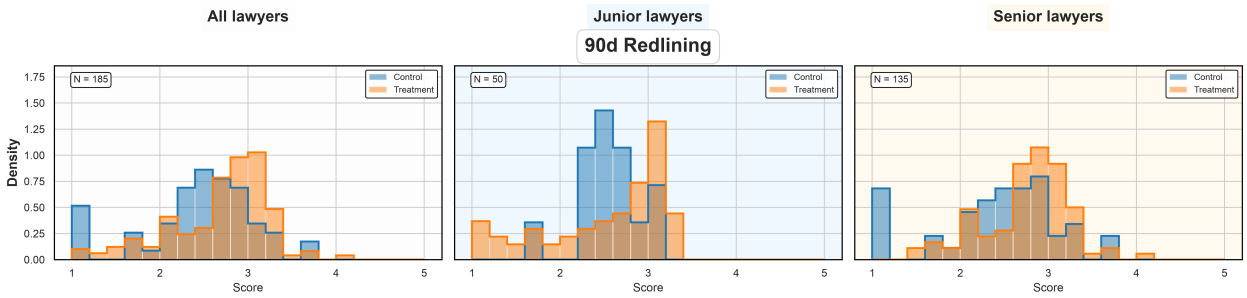
Senior lawyers also see an increase in score dispersion (Levene statistic 7.74, $p = 0.01$). But this comes primarily from a significant rise in the share of second-quintile (very good) scores (0.22, $p = 0.00$), with broad reductions in lower-tail scores. Specifically, the share of seniors scoring in the third, fourth, and fifth quintiles falls by -0.04, -0.08 and -0.14, respectively (all $p > 0.1$. Treated senior lawyers overperform the control group by achieving more strong scores and fewer mediocre and poor scores, while the performance of junior lawyers bifurcates, with significantly more poor (bottom quintile) scores, significantly fewer mediocre (fourth quintile) scores; and an offsetting increase in the prevalence of above-median (but not top) scores.

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)	
$H_0 : \alpha_{\text{junior}} < \alpha_{\text{senior}}$				0.74		0.60		0.70
$H_0 : \beta_{\text{junior}} > \beta_{\text{senior}}$				0.96		0.81		0.95
$H_0 : \alpha_{\text{junior}} + \beta_{\text{junior}} > \alpha_{\text{senior}}$				0.35		0.10		0.13
$H_0 : \alpha_{\text{junior}} + \beta_{\text{junior}} < \alpha_{\text{senior}} + \beta_{\text{senior}}$				0.05*		0.13		0.03**
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-ind. 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. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

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.

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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

In summary, we find that contemporaneous gains during AI-assisted drafting and subsequent gains on the unassisted evaluation task diverge: those who benefit most contemporaneously (junior lawyers) show no aggregate gain in offline performance; those who realize more modest contemporaneous gains from AI assistance (senior lawyers) show robust offline performance improvements.

4.3 Characterizing treatment effects on redlining performance with an LLM

Our experiment does not tell us *why* 90 days of AI access improved senior but not junior performance in the judgment-intensive (AI-unassisted) redlining task. To develop insight, we ran an LLM-based qualitative analysis of submissions, reported in full in Appendix A2.7.[11]

The behavioral patterns it identifies are concentrated among treated seniors, consistent with our finding that durable skill gains do not extend to juniors. Treated seniors more often restructured the patent draft than copy-edited existing text, addressing macro-level legal liabilities by rewriting whole sections. They concentrated on maximizing the patent's commercial scope, removing scope-limiting language, and paired their redlines with rationale-rich annotations naming the legal principles behind each edit. Lower-performing controls tended to make small surface edits and missed these macro-level vulnerabilities.

Junior lawyers did not show these patterns. Their effort went toward stylistic and administrative adjustments rather than broader commercial vulnerabilities, and their time allocation was rigid and sequential, often exhausting the allotted time on lower-stakes sections (such as the Background) before reaching the Claims. They frequently diagnosed high-level problems without executing the corresponding redlines, leaving short notes such as 'the claims use indefinite language' or 'there are antecedent basis issues' rather than performing the corrections themselves. While these notes read like instructions issued to an AI assistant, our qualitative review found this formalistic, issue-spotting posture was equally common among unassisted control juniors (see Appendix A2.7). Therefore, rather than AI actively instilling a 'prompting' habit, it appears this diagnose-without-execute posture is simply a baseline junior-level archetype that extended AI assistance failed to remediate.

These patterns admit at least two non-exclusive interpretations. The first concerns editing habits. Sustained AI use may weaken attachment to existing prose, licensing the wholesale rewrites we observe among seniors, and may train users to articulate the "why" and "how" of an edit, which surfaces as rationale-rich annotation. For juniors, relying on AI to execute edits during the drafting phase may fail to correct — or even reinforce — their baseline tendency to flag problems without solving them. The second interpretation concerns the locus of attention. Having AI perform the task of baseline text generation may free experienced lawyers to spend their attention on strategic decisions. Simultaneously, it may deprive juniors of the effortful practice through which strategic

[11] Conducted using NotebookLM without preregistered hypotheses, so we report patterns as qualitative observations rather than formal prevalence estimates.

capacity is acquired. An open question is whether these divergent impacts of AI use on junior versus senior lawyers would have attenuated or intensified with longer exposure.

4.4 Robustness of key results

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 a separate, independent firm using a standardized evaluation rubric covering enforceability, accuracy, ambiguity, completeness, and clarity. 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 periodic task that contained a brief task summary and the same scoring rubric given to the human raters. (Full prompts are reported in Appendix A2.5.)

The summary of expert ratings provided in Table A5 shows that LLM evaluations 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.

Among both expert and LLM evaluators in each of the three experimental tasks, the within-evaluator correlation among participant scores across all five dimensions of performance (enforceability, accuracy, ambiguity, completeness, and clarity) is high, as shown in Tables A3 and A4. However, the inter-rater correlation between LLMs and expert human raters is much lower. Taking averages across sub-scales and averaging the two expert evaluations for each submission, the LLM-expert correlation is $r = 0.47$ in the 10-day drafting task ($p = 0.00$), $r = 0.19$ in the 90-day drafting task ($p = 0.06$), and $r = 0.14$ in the 90-day redlining task ($p = 0.18$; see Figure A2).

We analyze these LLM evaluations by combining them with human evaluations in columns 5 through 8 of each of our three main results tables (Tables 3, 4, and 6). Columns 5 and 6 present results based only on LLM evaluations, while columns 7 and 8 pool human and LLM evaluations. Given the relatively low correlations between LLM and expert evaluations in the two 90-day tasks, we focus on models that pool expert and LLM evaluations (columns 7 and 8). Expert evaluations contribute two of every three observations in these pooled models—since two experts evaluated each submission—while our LLM evaluator contributes one evaluation per submission.

Combining human and LLM ratings in the 10-day drafting exercise, reported in Table 3, raises the main effect point estimate for the impact of AI access on patent draft quality from 0.34 standard deviations ($p = 0.03$) to 0.40 standard deviations ($p = 0.00$). In the column 8 specification that allows for separate treatment effects for junior and senior lawyers, the estimated treatment effect for juniors rises from 0.58 SD ($p = 0.04$) to 0.74 SD ($p = 0.00$). The corresponding point estimate for senior lawyers is unaffected: 0.22 SD ($p = 0.24$) in the human-evaluator estimate and 0.23 SD

($p = 0.14$) in the human-plus-LLM estimate.

Columns 5 through 8 of Table 4 report analogous estimates for 90-day drafting. Adding LLM evaluations raises the estimated main effect of AI access from 0.38 standard deviations ($p = 0.01$) to 0.55 standard deviations ($p = 0.00$). Recall that in the 90-day patent drafting exercise, we found a statistically significant impact of AI access on drafting quality among both junior and senior lawyers, at 0.60 SD ($p = 0.02$) and 0.30 SD ($p = 0.08$) respectively. Adding LLM raters increases the precision of both point estimates, while slightly lowering the estimated impact for junior lawyers and raising it for seniors. These point estimates are 0.56 SD ($p = 0.00$) and 0.55 SD ($p = 0.00$), respectively. This is the only instance where the addition of LLM raters changes the pattern of results, implying that seniors with AI access get a comparable performance boost to juniors with AI access at 90 days.

The final three columns of Table 6 report estimates for the impact of AI access on our key outcome variable, AI-unassisted 90-day redlining performance, while pooling combining human and LLM evaluations. Including LLM evaluations raises the point estimate for the overall impact of AI access on redlining quality from 0.32 standard deviations ($p = 0.04$) to 0.42 standard deviations ($p = 0.00$). In the column 8 specification that allows for separate treatment effects for junior and senior lawyers, the estimated treatment effect for juniors rises from -0.03 SD ($p = 0.89$) to 0.12 SD ($p = 0.57$), thus confirming the null effect that we found using expert ratings alone. The corresponding point estimate for senior lawyers similarly confirms the verdict of the expert raters. Adding LLM evaluations raises this positive effect from 0.45 SD ($p = 0.02$) in the human-evaluator estimate to 0.54 SD ($p = 0.00$) in the human-plus-LLM estimate.

In net, supplementing human evaluator ratings with the LLM-as-evaluator ratings provides corroborating evidence for our key results above, with the caveat that the LLM rater is likely less reliable than the lawyers contracted to provide our expert ratings.

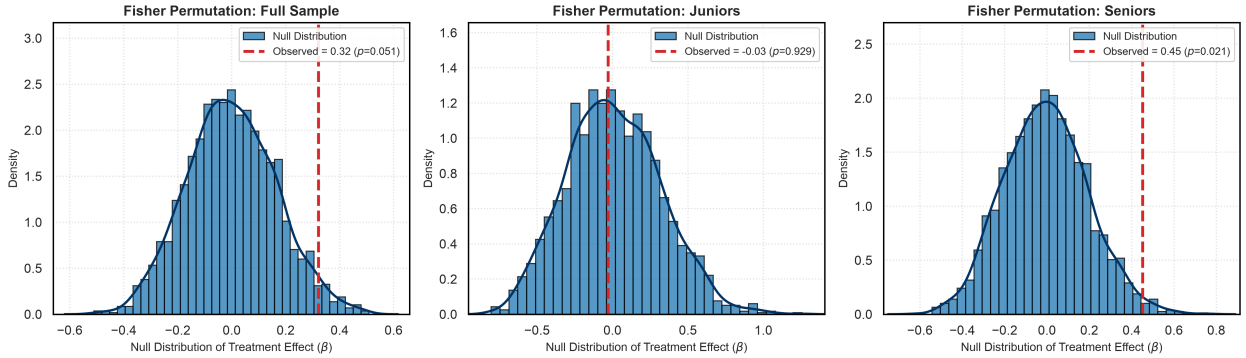
Nonparametric test of redlining results

The positive impact of AI access on senior but not junior performance in AI-unassisted redlining work is arguably 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). 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 spilled over into 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. For the main effect, the same OLS specification is applied as in column 2 of Table 6 (i.e. with fixed effects for sub-indicators and firms, weights $w_i = 1/n_i$ to correct for varying numbers of ratings per individual and standard errors clustered at the individual level, with only human ratings used) and the subgroup effects are estimated similarly to 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 firm fixed effects

Our main estimates include firm fixed effects to remove potentially confounding heterogeneity stemming, for example, from firm size, management structure, practice specialization, etc. Excluding firm fixed effects has minimal effects on our main results. In the 10-day drafting task, the full-sample main effect (Table 3, column 2) moves from 0.34 to 0.37 SD with no change in statistical significance. Neither the junior nor senior subgroup effects show changes in statistical significance despite slight changes to the point estimate. In the 90-day drafting task, the main effect likewise increases from 0.38 to 0.46 SD (Table 4, column 2), with p-values moving from 0.011 to 0.004. The senior effect in this task similarly rises from 0.30 to 0.37 SD, with p-values going from 0.080 to 0.047. On the redlining task (Table 6), the main effect remains stable (0.32 SD), though we see a slight loss in precision with p-values rising from 0.037 to 0.058. Removing firm fixed effects has no effect on the statistical significance of either the junior or senior subgroup effects.

Excluding specific firms

Our experimental sample includes eleven firms, one of which contributes 40 of 133 participants, raising the concern that our key finding—specifically, that AI access improves offline redlining

performance—is driven by a non-representative subset of the sample.[12] To explore this concern, we re-estimate the treatment effect on redlining performance while iteratively excluding each firm from the sample. Figure A3 reports these results: the estimates are stable across subsamples. Removing the largest firm in the experiment (Firm 1, second row in the figure) has only a modest effect on the point estimates, and both the pooled junior-senior treatment effect and the senior-only treatment effect remain statistically significant at $p < 0.10$. The estimated junior-only treatment effect is close to zero in all subsamples.

Robustness of main results to choice of experience threshold

Our estimates classify junior lawyers as those with fewer than seven years of experience, while those with seven or more years of experience are classified as senior lawyers. Figure A4 explores whether the choice of thresholds meaningfully affects our conclusions. In this exercise, we split the sample into junior, "midlevel," and senior groups, and we vary the thresholds that define these groups over three different start and end years, as documented in the figure.

These perturbations to our seniority definitions have almost no substantive effect on the conclusions: AI access differentially increases 10-day and 90-day drafting performance of juniors relative to seniors. It fails to increase redlining performance of juniors but substantially boosts redlining performance of seniors. The mid-level group, variously defined as those with three to six, four to seven, and five to eight years of experience, generally shows modest performance gains on the two drafting tasks but no gains on the redlining task. These results corroborate the finding that AI access differentially boosts AI-assisted performance among juniors but only boosts AI-unassisted performance among seniors.

Potential contamination by unauthorized AI usage

Participants were instructed not to use AI tools in the redlining task, with this admonition highlighted in both the task assignment email and the assigned task materials (see Appendix A2.3 for full instructions). This instruction was not directly enforceable, however, since end-user privacy commitments preclude observing individual-level tool use. If treated subjects used AI while redlining, our estimate of the treatment effect of AI access on non AI-assisted performance could be contaminated by non-adherence. Three pieces of evidence suggest that non-adherence does not drive our results.

First, our data provide a modest amount of direct evidence on compliance. One firm completed the redlining task in a session separate from the drafting task, and we observe zero use of InFlow's redlining feature at that firm during the session. This check is narrow: it covers one feature at one firm, and it cannot detect use of other InFlow features, of other AI tools, or of AI outside the timed session. We note separately that firm-level usage intensity is roughly half as high in the long interval

[12] The completed redlining task sample encompasses nine firms and 91 participants, 37 of whom were from Firm 1.

between the two test tasks as during the task windows themselves (Figure A5), which indicates that subjects concentrated their tool use around assigned work rather than using InFlow at a constant rate throughout. This pattern carries no information about redlining adherence, however.

Second, we deployed NotebookLM (which runs on Gemini 3.0 Pro) to act as a forensic linguist and senior patent examiner, seeking evidence of AI use. We tasked the model with comparing each redlined submission against its unedited baseline draft to detect 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; high-density use of LLM-typical transitional adverbs ("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. We seek one additional marker specific to juniors: uniform edit density across all sections, which is inconsistent with the typically uneven time allocation of junior lawyers across the early versus the later sections of a document. Each flag requires a confidence score and a verbatim supporting quotation. Appendix A2.6 reports the full prompt. Erring toward over-classification, we flag 15 of 91 redlining submissions in our sample as possible non-adherence. The research team also manually audited the submissions for AI-generated artifacts, conservatively categorizing any anomalies as potential non-adherence.

We re-estimated the redlining results under two treatments of the flagged submissions, reported in Table A6. Controlling for suspected non-adherence leaves the results intact. Flagged submissions score 0.18 SD higher than other treated submissions, an insignificant difference ($p = 0.31$), and the senior treatment effect is 0.44 SD ($p = 0.02$) against 0.45 SD in the uncorrected specification. Dropping the flagged submissions is more conservative and costs precision. The overall treatment effect falls from 0.32 to 0.25 SD ($p = 0.14$), while the senior effect falls from 0.45 to 0.39 SD ($p = 0.06$). The junior estimate stays negative and imprecise throughout, at -0.11 SD ($p = 0.61$). Repeating both exercises with LLM rather than human ratings strengthens the results: excluding flagged submissions, the overall effect is 0.61 SD ($p = 0.00$) and the senior effect 0.76 SD ($p = 0.00$), as reported in Table A7.

Two corroborating patterns appear elsewhere in the paper. First, as reported in Section 5, AI access reduced subjects' time on task during the drafting task but yielded no time savings among treated subjects in the redlining task, which we would anticipate if subjects were using AI during redlining. Second, the LLM characterization in Section 4.3 finds that the qualitative pattern of treated juniors' redlining performance was opposite to what AI use would produce: they frequently diagnosed problems without executing the corresponding edits, leaving notes of the kind one would issue as a prompt rather than the executed edits an AI tool would return. In this sense, AI access spilled over to *reduced* performance in redlining: juniors seemingly attempted to prompt AI even when it was unavailable.

5 Supplementary findings

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 on task, patent drafting performed in the course of routine work during the study period, and post-study interviews. Because the field measures cover only part of the sample and the interviews only a handful of participants, we treat these results as context for the experimental findings rather than as tests of our hypotheses.

5.1 Drafting speed

Using subjects' self-reported time spent on each experimental task as outcome variables (10-day drafting, 90-day drafting, and 90-day redlining), Table 8 shows that AI access reduces time on both drafting tasks, with time savings concentrated among juniors. AI access did not reduce time spent on the redlining task, consistent with the protocol's stipulation that subjects not use AI assistance for this task.

In the 10-day drafting task, treated participants finished an estimated 10 minutes faster than controls against a baseline of 112 minutes ($p = 0.05$); the 90-day task shows directionally consistent but smaller and statistically uncertain time savings of 10 minutes ($p = 0.27$). Time savings are concentrated among junior lawyers: at 10 days, treated juniors finished 18 minutes faster than control juniors ($p = 0.05$), whereas treated seniors finished only 5 minutes faster, a contrast that is not distinguishable from zero ($p = 0.35$). This pattern persists directionally at 90 days, with juniors showing larger time savings than seniors, but neither contrast is statistically distinguishable from zero.

Untreated seniors spent *longer* on the redlining task than untreated juniors, an average of 63 versus 53 minutes. This gap is not statistically significant, but its direction is what we would expect if seniors bring greater care and deliberation to their unassisted work. The same ordering holds within the treated group. Both treated juniors and treated seniors spent slightly longer on redlining than their respective control groups (+4.49 minutes among junior lawyers, $p = 0.76$, and +1.23 minutes among senior lawyers, $p = 0.89$), and the total redlining time of treated seniors again exceeded that of treated juniors.

5.2 On-the-job patent drafting performance

To assess whether measured productivity gains in experimental tasks also accrue to participants' non-experimental work tasks, we asked participants to report on their own on-the-job patent drafting during the 90-day study period. These measures, reported in Table 9, are exploratory in three respects. First, we introduced them after recruitment had begun, so they cover only 68 participants (51.1% of the sample). Second, they lack the control of the benchmark tasks, since subjects drafted patents of differing complexity under differing supervision. Finally, they are

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)	
$H_0 : \alpha_{\text{junior}} < \alpha_{\text{senior}}$		1.00		0.76		0.23
$H_0 : \beta_{\text{junior}} > \beta_{\text{senior}}$		0.89		0.80		0.42
$H_0 : \alpha_{\text{junior}} + \beta_{\text{junior}} > \alpha_{\text{senior}}$		0.07*		0.74		0.70
$H_0 : \alpha_{\text{junior}} + \beta_{\text{junior}} < \alpha_{\text{senior}} + \beta_{\text{senior}}$		0.99		0.38		0.26
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
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. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

under-powered for the effect sizes at issue: on time to completion, the closest field analogue to our experimental outcome, only reductions exceeding roughly 31% of the control baseline of 5.1 weeks would register as statistically significant, far larger than the 9% time saving we measure on the 10-day drafting task.[13]

Treated participants completed patents 0.76 weeks faster than controls ($p = 0.35$), reported less reviewer input on their applications (-0.39 SD, $p = 0.17$), and reported fewer edits, corrections, and comments from reviewers on their drafts (-0.15 SD, $p = 0.58$). All three point estimates point toward better performance among subjects given AI access, though none is statistically distinguishable from zero. The seniority pattern of time savings in non-experimental tasks echoes the pattern detected in experimental tasks. Juniors show larger self-reported per-patent time savings of 1.65 weeks, against 0.3 weeks for seniors, though neither estimate is significant ($p = 0.33$ and $p = 0.74$). Power is weaker still in this subgroup split. The standard error on the junior estimate is 1.71 weeks, so only reductions exceeding 3.4 weeks, against a junior control baseline of 4.4 weeks, would attain significance.

Because firms were not blinded to treatment assignment, they could in principle have shifted workloads in a way that favored or penalized the treated group. We see no significant change in the number of patents worked on during the study period (-1.06, $p = 0.18$), but this estimate is too imprecise to support strong conclusions. Treated participants also report no difference relative to controls in their level of formal responsibility for the patents they drafted (0.01, $p = 0.98$). The confidence interval on this estimate spans roughly ±0.5 SD, however, so it excludes only sizable compositional shifts. The pooled estimate also averages over subgroup estimates of opposite sign, of 0.41 SD for juniors ($p = 0.18$) against -0.16 SD for seniors ($p = 0.63$). We therefore cannot rule out that treated juniors were assigned more responsible drafting work.

The field estimates are directionally aligned with our experimental findings, but they are too imprecise to establish whether the experimental gains translate into on-the-job efficiency.

[13] The standard error on the time-to-completion estimate is 0.81 weeks, implying a minimum detectable effect of roughly 1.6 weeks at conventional significance levels. We therefore read what follows as suggestive.

Table 9: 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)	
$H_0 : \alpha_{\text{junior}} < \alpha_{\text{senior}}$		0.33		0.75		0.75		0.72		0.89
$H_0 : \beta_{\text{junior}} > \beta_{\text{senior}}$		0.34		0.11		0.75		0.44		0.46
$H_0 : \alpha_{\text{junior}} + \beta_{\text{junior}} > \alpha_{\text{senior}}$		0.84		0.02**		0.62		0.53		0.13
$H_0 : \alpha_{\text{junior}} + \beta_{\text{junior}} < \alpha_{\text{senior}} + \beta_{\text{senior}}$		0.52		1.00		0.47		0.79		0.93
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. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

5.3 Qualitative interviews

To better understand subjects' perception of how AI affected their work, we conducted post-experiment qualitative interviews with four treated participants, two juniors and two seniors.

One senior interviewee described the AI as providing a "formidable first draft," comparable to the baseline output of a junior associate. Neither reported accepting that text as given. One described using it as a "logic auditor" to pressure-test strategic reasoning and evaluate rejections from new perspectives, and both reported that working from a flawed but structurally complete draft led them to interrogate logical gaps they might otherwise have passed over. Both also reported giving junior associates more granular feedback on structural and technical inconsistencies, though we have no measure of the feedback juniors actually received. These accounts are consistent with the redlining behavior documented in Section 4.3, where treated seniors paired their redlines with rationale-rich annotations naming the legal principles behind each edit.

Junior lawyers described a "drafting jumpstart effect," wherein the tool eliminated the friction of producing a first draft and getting past the "blank page problem." Post-task survey results (see Section 3.3 for details) corroborate that juniors especially *believed* that the tool accelerated their work, with treatment subjects showing a 0.97 ($p = 0.04$) SD increase in their self-reported perception of their drafting speed at 10 days and a 1.47 ($p = 0.00$) SD increase at 90 days compared to increases of 0.58 ($p = 0.04$) and 0.99 ($p = 0.00$) SD for seniors at these time intervals (see Table A8 for full results). Both junior lawyers interview reported that this let them move earlier into a reviewing posture, checking technical accuracy and the linguistic precision that is critical for patent drafting. Treated juniors report greater satisfaction in their own drafting during the experimental drafting task. Self-reported satisfaction is 1.48 SD higher among treated relative to control juniors at 90 days ($p = 0.00$), the largest treatment effect we estimate on any self-reported outcome.[14]

Juniors' accounts of their own development are not reflected in their performance absent AI. We detect no gain in juniors' unassisted redlining performance, and the behavioral evidence suggests weaker skill accumulation, as in the case of juniors flagging antecedent basis problems without correcting them. Our unassisted measure gives no corroboration of the skill development these accounts describe.

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 the larger gains going to junior lawyers. Juniors also captured nearly all of the time savings. On the unassisted redlining

[14] The specific survey question was: "How satisfied are you with the overall experience drafting the patent sections in this task?"

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, markedly 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 one key respect. 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). We instead observe experienced practitioners performing a core task of their own specialty, following 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 in the shape of the performance distribution, not merely its mean. Treated seniors improved by earning more good scores and fewer mediocre and poor ones. Treated juniors show no detectable average improvement. Rather, their scores bifurcate: significantly more bottom-quintile (poor) scores, significantly fewer fourth-quintile (mediocre) ones, and an offsetting increase in second-quintile (good) scores, with no gain at the top of the distribution. This pattern is consistent with AI use having abetted learning for some juniors and impaired it for others.

Text analysis of the redlining submissions hints at the sources of divergence. As documented in Section 4.3, treated seniors tasked with redlining restructured the flawed patent draft to neutralize its macro-level legal liabilities, and they annotated their edits with formal explanations. Treated juniors worked sequentially through stylistic and administrative corrections and repeatedly diagnosed serious defects without repairing them, in some cases, leaving notes that read as instructions issued to an AI assistant that was not present. Because our qualitative analysis found this formalistic, issue-spotting posture was also prevalent in the control group, it appears to represent a baseline junior-level deficit rather than a habit actively acquired from AI use.

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. Neither the analysis of AI-assisted performance nor subjects' own assessment of their work would forecast this result. Treated juniors attained mean drafting scores at or above those of untreated seniors at both assessments, reaching 0.36 SD against 0.13 SD at 90 days, while their mean redlining score of -0.06 SD sat at the control baseline. Juniors also believed the tool was improving their performance, a belief borne out in AI-assisted work that did not carry over to unassisted work. For seniors, three months of AI-assisted drafting raised near-term output and unassisted judgment alike.

This divergence between lawyers' performance on assisted and unassisted tasks is consequential because unassisted judgment remains a routine 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 go-to 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 identifies a similar tension and also points to one margin along which it may operate. 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. The losses accrue to students whose study time falls after adoption; students who maintain their study time perform similarly to non-adopters, who are not themselves positively selected on pre-AI achievement.

The seniority contrast in our setting is also consistent with a time-investment dynamic. Juniors using AI worked faster but showed no unassisted gains. Seniors drafted better patents with AI but with almost no time acceleration. They worked more slowly than juniors on the unassisted redlining task, whether or not they had spent three months drafting with AI, yet treated seniors supplied more substantive and comprehensive redlines. 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. Establishing time investment as the mediating channel in either setting would require manipulating time on task directly, something that neither our study nor Strömberg et al. (2026) does.

We caveat that our sample is relatively small, reflecting the high hourly billing rate of top-tier lawyers. It is drawn from a set of firms contracted to perform ongoing patent work for a sophisticated, demanding 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 across which patent expertise is built. Finally, we cannot observe individual tool use directly to verify full compliance with the stipulation that participants not use AI during the offline redlining task. However, the analysis reported in Section 4.4 suggests that potential non-compliance does not skew our results.

These limitations underscore opportunities for further experimentation. 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 trends toward convergence or further divergence. Randomizing workflow, and time on task in particular, rather than access alone would test the

mechanism directly; Bastani et al. (2025) show how the detailed design of user-AI interactions shapes learning outcomes. 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.

Chiong, K. and Xie, Y. (2024). Generative AI and job satisfaction: Evidence from glassdoor employee reviews. *Working Paper, Available at SSRN*.

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.

Jia, F. et al. (2024). Effect of genAI dependency on university students' academic achievement. *BMC Psychology*.

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.

Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., and Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM.

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.

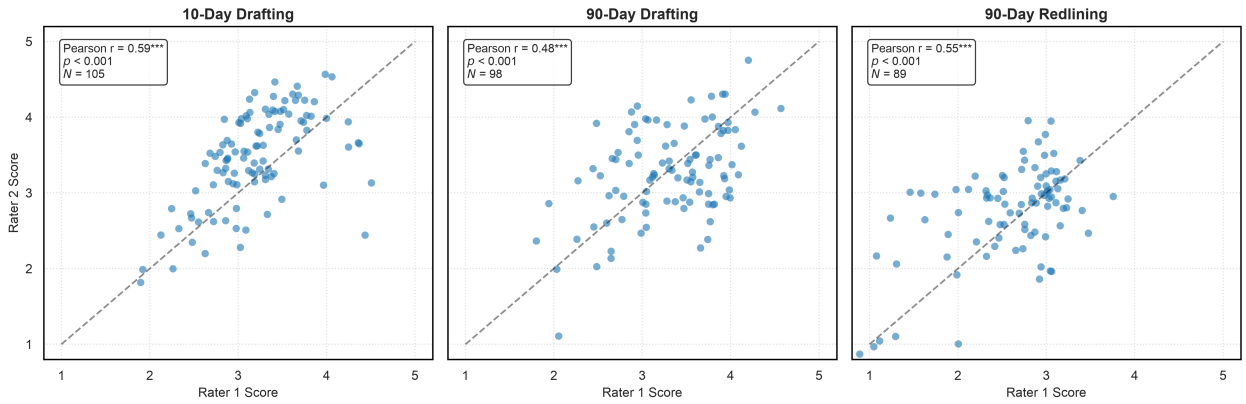
Wiles, E., Krayer, L., Abbadi, M., Awasthi, U., Kennedy, R., Mishkin, P., Sack, D., and Candelon, F. (2024). GenAI as an exoskeleton: Experimental evidence on knowledge workers using GenAI on new skills. Working paper, Boston University. Available at SSRN: https://ssrn.com/abstract=4944588.

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

Appendices

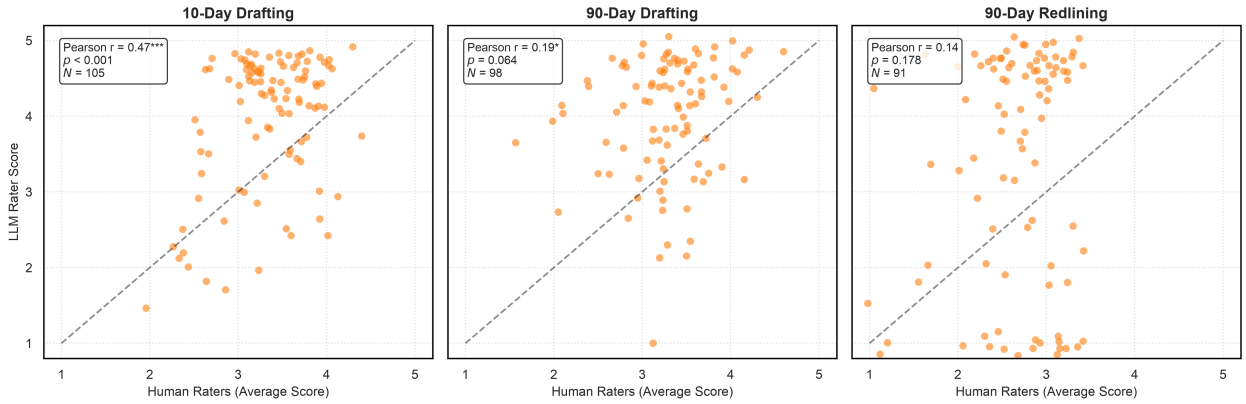
A1. Appendix Figures and Tables

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



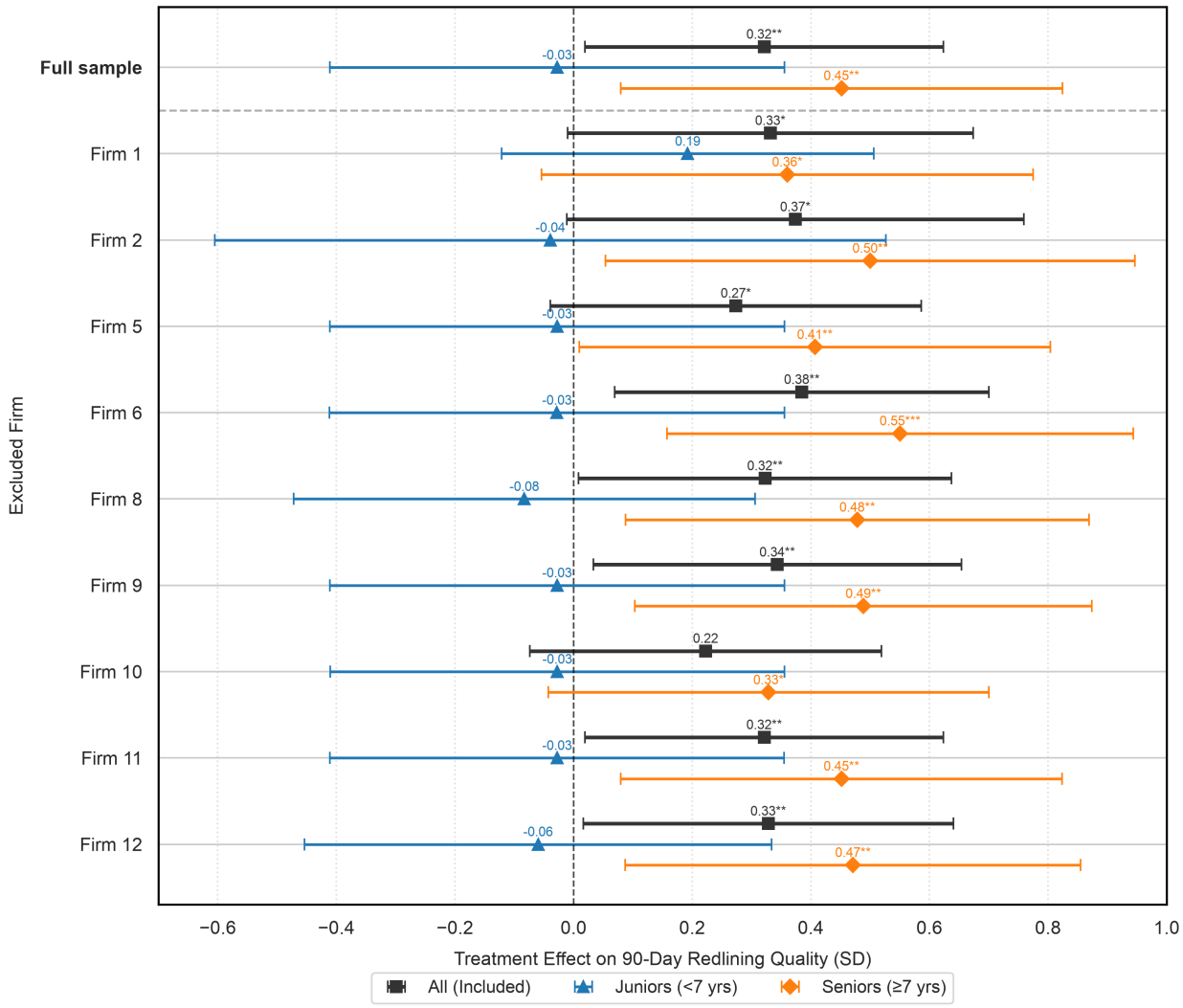
Notes: This figure evaluates inter-rater reliability by plotting pairwise human rater scores for the 10-day drafting, 90-day drafting, and 90-day redlining tasks. (For the redlining task, two participants have only one expert rating where two were expected, leading to a slightly lower N that we account 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 0.48 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



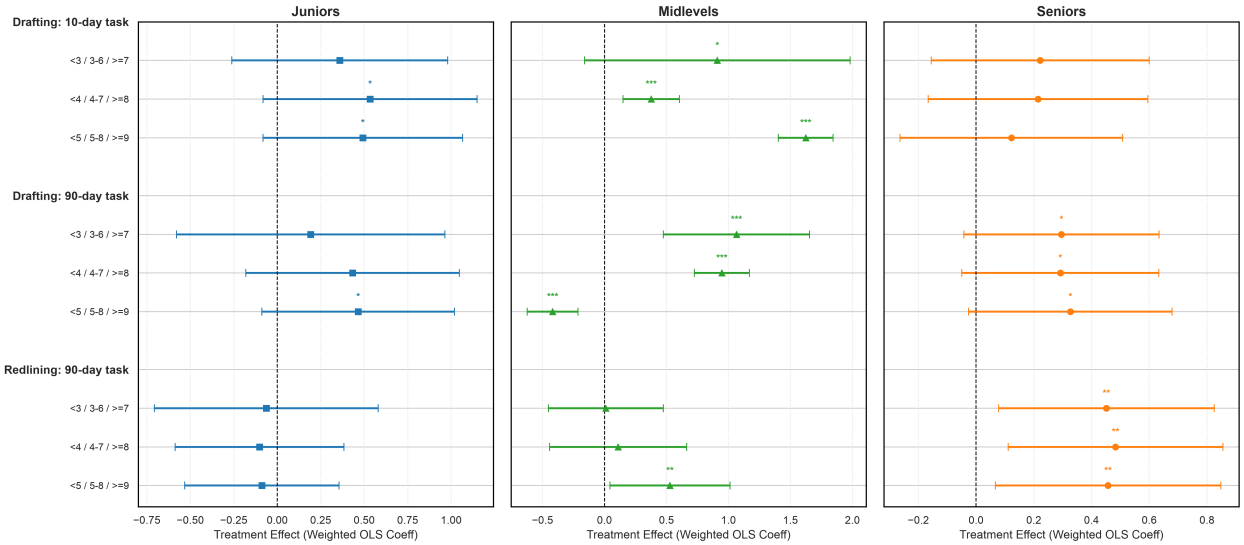
Notes: This plot compares the average score given by human raters to the score provided by the LLM rater for the same participants across the three main study tasks. Like the human rater plots, data points are lightly jittered to prevent overplotting, and Pearson correlations are calculated to measure 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: Impact of AI Access on 90-Day Redlining Quality, Excluding Individual Firms



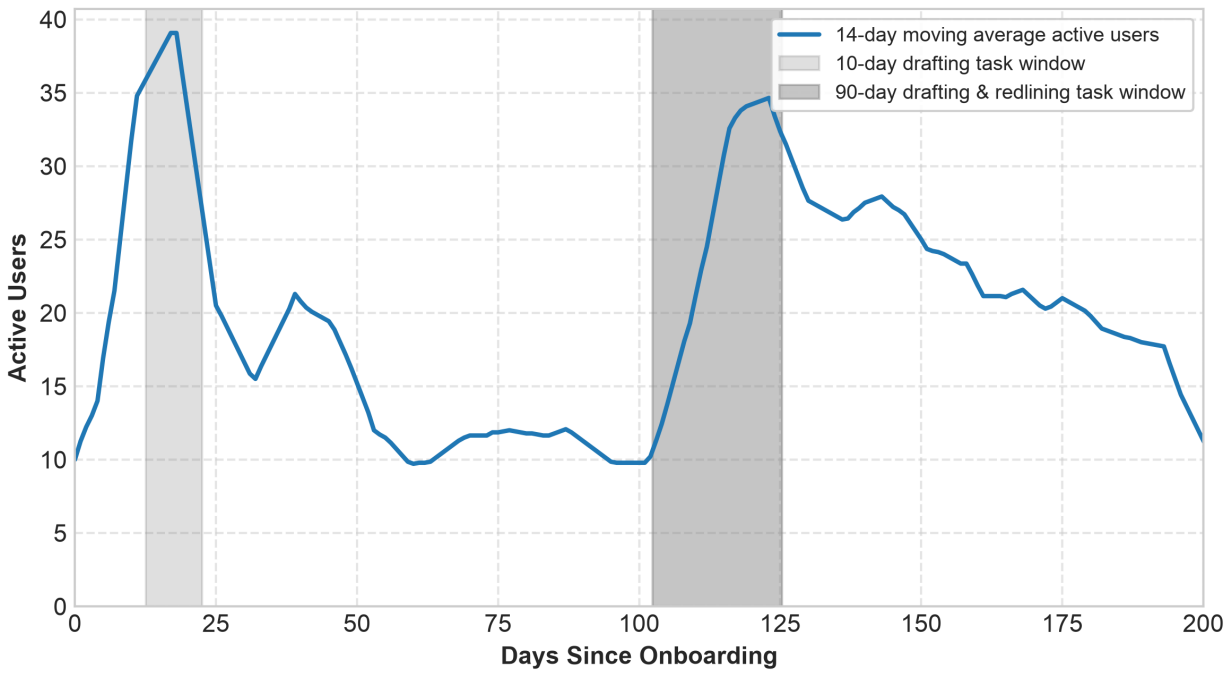
Notes: This forest plot presents intent-to-treat estimates of treatment effects on the 90-day redlining task when iteratively excluding each individual firm from the sample. Estimates are shown for all participants and separately for junior (< 7 years of experience) and senior (≥ 7 years of experience) lawyers. The top row ("Full sample") reports baseline estimates with all firms included. Subsequent rows report estimates after dropping the indicated firm, where firm numbering corresponds to Table 1. Dependent variables standardized relative to the control group mean and standard deviation across the full active sample prior to firm exclusions. Models are estimated by OLS with weights $w_i = 1/n_i$, where n_i is the number of human ratings individual i receives so that each individual contributes equally, incorporating firm and sub-indicator fixed effects with standard errors clustered at the individual level. We leave out three firms: Firm 3 completed the protocol prior to the finalization of the redlining task, while Firms 4 and 7 registered no submissions; these therefore show identical estimates to the full sample. Error bars represent 95% confidence intervals. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Figure A4: Sensitivity of Main Results to Choice of Experience Threshold for Junior vs. Senior Classification



Notes: This figure presents a sensitivity analysis of the treatment effect when partitioning the sample into three subgroups: Juniors, Midlevels, and Seniors. (We only report on Juniors and Seniors in the text above but here we include Midlevels as an additional check on the robustness of our results to category definitions) Each row on the y-axis specifies the experience boundaries used for the split ($< X$ / X to $Y-1$ / $\geq Y$), where X bounds Juniors from Midlevels and Y bounds Seniors from Midlevels. The markers represent the estimated subgroup-specific treatment effects for each task (10-day Drafting, 90-day Drafting, and 90-day Redlining. Models are estimated by OLS with weights $w_i = 1/n_i$, where n_i is the number of human ratings individual i receives so that each individual carries equal total weight. All models control for firm fixed effects and subcomponent fixed effects, omitting a global intercept. Standard errors are clustered at the participant level. Error bars represent 95% confidence intervals.. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Figure A5: 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 InFlow 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 InFlow. * 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 InFlow. 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: Relationships Between Quality Dimensions in Expert Ratings

A. 10-day Drafting Task (N = 105, Cronbach's α = 0.953)					
	(1)	(2)	(3)	(4)	(5)
(1) Enforceability	1.00				
(2) Accuracy	0.83	1.00			
(3) Ambiguity	0.82	0.79	1.00		
(4) Completeness/Alignment	0.81	0.82	0.80	1.00	
(5) Clarity	0.83	0.82	0.77	0.77	1.00
B. 90-day Drafting Task (N = 98, Cronbach's α = 0.956)					
	(1)	(2)	(3)	(4)	(5)
(1) Enforceability	1.00				
(2) Accuracy	0.75	1.00			
(3) Ambiguity	0.80	0.79	1.00		
(4) Completeness/Alignment	0.82	0.82	0.85	1.00	
(5) Clarity	0.83	0.85	0.83	0.86	1.00
C. 90-day Redlining Task (N = 91, Cronbach's α = 0.952)					
	(1)	(2)	(3)	(4)	(5)
(1) Enforceability	1.00				
(2) Accuracy	0.83	1.00			
(3) Ambiguity	0.80	0.78	1.00		
(4) Completeness/Alignment	0.83	0.79	0.78	1.00	
(5) Clarity	0.83	0.82	0.78	0.76	1.00

Notes: Pearson correlation coefficients between sub-dimensions of task quality. Cronbach's α is reported for each set of sub-dimensions.

Table A4: Relationships Between Quality Dimensions in LLM Ratings

A. 10-day Drafting Task (N = 105, Cronbach's α = 0.964)					
	(1)	(2)	(3)	(4)	(5)
(1) Enforceability	1.00				
(2) Accuracy	0.84	1.00			
(3) Ambiguity	0.81	0.79	1.00		
(4) Completeness/Alignment	0.89	0.92	0.82	1.00	
(5) Clarity	0.89	0.85	0.78	0.89	1.00
B. 90-day Drafting Task (N = 98, Cronbach's α = 0.949)					
	(1)	(2)	(3)	(4)	(5)
(1) Enforceability	1.00				
(2) Accuracy	0.77	1.00			
(3) Ambiguity	0.76	0.81	1.00		
(4) Completeness/Alignment	0.81	0.80	0.80	1.00	
(5) Clarity	0.88	0.85	0.76	0.78	1.00
C. 90-day Redlining Task (N = 91, Cronbach's α = 0.981)					
	(1)	(2)	(3)	(4)	(5)
(1) Enforceability	1.00				
(2) Accuracy	0.88	1.00			
(3) Ambiguity	0.93	0.87	1.00		
(4) Completeness/Alignment	0.93	0.88	0.89	1.00	
(5) Clarity	0.97	0.90	0.92	0.93	1.00

Notes: Pearson correlation coefficients between sub-dimensions of task quality evaluated by LLM raters. Cronbach's α is reported for each set of sub-dimensions.

Table A5: 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 A6: Impact of 90-Day AI Access on Non-AI-Assisted 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)
Reported 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)	
$H_0 : \alpha_{\text{junior}} < \alpha_{\text{senior}}$		0.74		0.74		0.75
$H_0 : \beta_{\text{junior}} > \beta_{\text{senior}}$		0.96		0.95		0.95
$H_0 : \alpha_{\text{junior}} + \beta_{\text{junior}} > \alpha_{\text{senior}}$		0.35		0.35		0.46
$H_0 : \alpha_{\text{junior}} + \beta_{\text{junior}} < \alpha_{\text{senior}} + \beta_{\text{senior}}$		0.05*		0.05*		0.07*
Rating-level Disaggregated	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Sub-indicator 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 Appendix A2.6. 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 sub-indicator 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. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A7: Impact of 90-Day AI Access on Non-AI-Assisted 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)
Reported 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)	
$H_0 : \alpha_{\text{junior}} < \alpha_{\text{senior}}$		0.60		0.59		0.52
$H_0 : \beta_{\text{junior}} > \beta_{\text{senior}}$		0.81		0.81		0.86
$H_0 : \alpha_{\text{junior}} + \beta_{\text{junior}} > \alpha_{\text{senior}}$		0.10		0.10		0.22
$H_0 : \alpha_{\text{junior}} + \beta_{\text{junior}} < \alpha_{\text{senior}} + \beta_{\text{senior}}$		0.13		0.11		0.03**
Rating-level Disaggregated	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Sub-indicator 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 Appendix A2.6. 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 sub-indicator 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. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A8: Impact of AI Access on Perceptions of Speed, Quality, and Satisfaction in Drafting Tasks at 10 and 90 Days

	10-Day Drafting						90-Day Drafting					
	Speed		Quality		Satisfaction		Speed		Quality		Satisfaction	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Treatment	0.72***		0.29		0.30		1.14***		0.34		0.65***	
	(0.25)		(0.24)		(0.27)		(0.23)		(0.25)		(0.24)	
Treat × Junior		0.97**		0.95***		0.20		1.47***		0.59		1.48***
		(0.46)		(0.36)		(0.46)		(0.43)		(0.39)		(0.40)
Treat × Senior		0.58**		−0.09		0.37		0.99***		0.22		0.28
		(0.29)		(0.29)		(0.33)		(0.27)		(0.32)		(0.30)
Junior		−0.28		−0.77		0.36		−0.11		0.03		−0.62
		(0.42)		(0.30)		(0.39)		(0.37)		(0.30)		(0.32)
Senior		−0.02		0.22		−0.12		−0.19		0.06		−0.14
		(0.19)		(0.19)		(0.22)		(0.19)		(0.26)		(0.21)
Constant	−0.12		−0.11		−0.02		−0.06		0.11		0.21	
	(0.27)		(0.25)		(0.22)		(0.22)		(0.23)		(0.21)	
$H_0 : \alpha_{\text{junior}} < \alpha_{\text{senior}}$		0.30		0.01***		0.85		0.57		0.47		0.11
$H_0 : \beta_{\text{junior}} > \beta_{\text{senior}}$		0.24		0.01**		0.62		0.17		0.23		0.01***
$H_0 : \alpha_{\text{junior}} + \beta_{\text{junior}} > \alpha_{\text{senior}}$		0.02**		0.54		0.02**		0.00***		0.08*		0.00***
$H_0 : \alpha_{\text{junior}} + \beta_{\text{junior}} < \alpha_{\text{senior}} + \beta_{\text{senior}}$		0.66		0.57		0.84		0.97		0.84		0.98
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	101	101	101	101	101	101	95	95	95	95	95	95
R^2	0.20	0.21	0.09	0.15	0.12	0.13	0.34	0.37	0.13	0.14	0.20	0.25

Notes: Table reports intent-to-treat estimates using OLS regressions for self-reported speed, quality, and satisfaction outcomes from the 10-day and 90-day post-task surveys. All estimations conducted at the subject level with robust (HC1) standard errors. Columns (2), (4), (6), (8), (10), and (12) 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. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

A2. Prompts and Other Reference Materials

A2.1. InFlow Feature Update Timeline

Date	Feature Changes
May 8, 2025	• Model Contextualization: Upgraded AI to use all source files as default context for "AI Typing." • Workflow Simplification: Automated the "Summary for AI" step to streamline project initialization. • Hierarchical Organization: Introduced sub-folders and direct file creation within the project menu.
June 11, 2025	• Grounded Chat (Beta): Launched a chat interface for document-based queries and brainstorming. • Patent Research: Integrated public patent search capabilities directly into the right-side panel. • Inline Command Expansion: Enabled prompt box nesting and AI prediction triggers via keyboard shortcuts ($Ctrl + J/K$).
July 24, 2025	• Blueprints: Released a templating system to standardize AI instructions and context across documents. • Critique System: Introduced automated editorial review against user-defined criteria with AI-suggested resolutions. • Model Upgrades: Transitioned default chat model to Gemini Pro 2.5 and added "Personas" for varied tonal responses.
Sept 2, 2025	• Instruction Bifurcation: Separated AI instructions into "Style" (voice capture) and "Purpose" (task-specific context). • Source Management: Added the ability to convert chat logs into persistent source documents. • Critique Automation: Introduced a "Generate Criteria" button to automatically create review standards based on the Blueprint.
Oct 13, 2025	• Prompt Library: Launched a searchable library with Slash Command (/) integration for rapid prompt retrieval. • Custom Personas: Enabled users to create and save experimental custom personas to Blueprints. • UI Standardization: Introduced visual indicators (purple borders) to distinguish Blueprint editing from standard document drafting.

Date	**Feature Changes** (Continued)
Oct 23, 2025	• Visual Overhaul: Redesigned the interface for a "Light & Clean" aesthetic to reduce fatigue. • Drive Interoperability: Implemented full Google Drive import/export functionality for Documents, Slides, and PDFs. • Conversational Setup: Replaced manual setup with a natural language interface for document creation.
Jan 16, 2026	• Workflow Finalization: Refined the critique and review interface to reduce latency and steps in the feedback cycle. • Navigation Improvements: Added a centralized modal for rapid switching between Document, Purpose, Style, and Sources.

A2.2. 10-Day Task (Drafting Only)

AI & Expertise: Designing AI for Productivity and Skills Development

Benchmark task 1

I. INSTRUCTIONS

Please follow the instructions below carefully. Should you have any questions, don't hesitate to reach out to Joshua Martin (xWF) on the Economic Research team.

1. Confirmation of prior activities

By now you should have completed the onboarding form and a 30 minute training exercise as a pair. If not, please make sure you do these first.

2. Submission format

Once you are done with the test task, using the tools & methods specific to your Group, please copy the full content of the exercise to a separate Google doc. Please label each section clearly, i.e.:

I. DRAFTING

a. DEPENDENT CLAIMS
b. DETAILED DESCRIPTION

At the top of the document please state how many hours & minutes the exercise took to complete by stating "Test task time - drafting exercise: X hrs Y mins." To track time on the task, please go to https://www.tickcounter.com/stopwatch and turn on the timer. Click "Lap" after finishing the drafting exercise, then click "Stop" after finishing the critique exercise.

Please save the Google document with the title of **Your Firm_First 4 letters of email address_Date**. Please upload the submission into a Secure Shared Drive created for your Firm (the onboarding email has the link in it). We need to link your email address to your submission so that we can match it to your survey responses. Rest assured that beyond that all analysis is done anonymously.

Reminder: The exercise should be completed within 2 hrs - if you are not done by the end of 2 hrs, please pause where you are and submit as is, indicating time on task as instructed above.

3. Evaluation criteria

The evaluation of the finished test tasks' output quality will be conducted by 2 independent third-party raters (experienced patent lawyers) - unaware of your group assignment or tenure - using a predefined rubric focused on both patent draft quality and robustness.

On top of that the research team will collect time on tasks, perceived quality & time gains or losses, increased or decreased satisfaction with patent drafting experience through a survey.

4. What's next

We will send:

- Post test task survey shortly (estimated time required <2 mins)
- Monthly rolling feedback survey in the next few weeks (estimated time required 10 mins)
- More about planned assessments of real patents drafted for Google shortly.

II. DRAFTING EXERCISE

Reminder: please go to https://www.tickcounter.com/stopwatch and turn on the timer!

Note: You are encouraged to use InFlow for these tasks, but not required to do so.

1. Handout

On the following tabs (also linked below) you will see a set of materials for you to draft a portion of a mock patent application:

- **Main Claim**
- **Inventor Notes**
- **Background of the Invention** – Excerpts detailing the technical field, relevant prior art, and the problem the invention is solving.
- **Relevant Figures/ Drawings**

2. Task

Your task is to use these materials to draft:

1. **Five (5) to seven (7) dependent claims** that build upon (i.e., depend from) the main claim provided. Your claims need to include:
    a. At least one (1) dependent claim that relies on a prior dependent claim (i.e., nested dependency)
    b. At least one (1) element drafted purely in functional language
    c. At least one (1) element drafted purely in structural language
    d. No two dependent claims may recite the same additional feature set.

    Your dependent claims can go beyond what was disclosed in the inventor notes, but have to make logical and technical sense.

2. **A detailed description** of the invention (capped at **1000 words**), focusing on the **primary embodiment** of the invention.

A detailed description should explain how the invention works, its configuration, and how it addresses the problem stated in the background. This should be drafted purely on the information in the provided materials.

A2.3. 90-Day Task (Drafting and Redlining)

AI & Expertise: Designing AI for Productivity and Skills Development

Benchmark task 2

I. INSTRUCTIONS

Please follow the instructions below carefully. Should you have any questions, don't hesitate to reach out to Joshua Martin (xWF) on the Economic Research team.

1. Confirmation of prior activities

By now it should be about three months since you joined the InFlow expertise study. You have already completed the first Test Task, towards the beginning of your participation. *This test task has two parts. The first part – the 'drafting' exercise – is formatted similarly but substantively different to the first test task. The second – the 'critique' exercise – is wholly unique.*

2. Submission format

Once you have read through the test task, please copy the full content of the exercise to a separate Google doc. Please label each section clearly, i.e.:

I. DRAFTING

a. DEPENDENT CLAIMS
b. DETAILED DESCRIPTION

II. CRITIQUE

a. REDLINE EDITS
b. FEEDBACK

At the top of your submission document, please copy-paste the following table. To track time, please go to https://www.tickcounter.com/stopwatch and turn on the timer. Click "Lap" after finishing the drafting exercise, then click "Stop" after finishing the critique exercise in order to get time lengths for each task.

Drafting exercise:	X hrs, Y min
Critique exercise:	X hrs, Y min

Please save the Google document with the title of **Your Firm_First 4 letters of email address_Date**. Please upload the submission into a Secure Shared Drive created for your Firm (the email you received from Josh has the link in it). We need to link your email address to your

submission so that we can match it to your survey responses. Rest assured that beyond that all analysis is done anonymously.

The full exercise - both subtasks - should be completed alone and within 4 hrs - if you are not done by the end of 4 hrs, please pause where you are and submit as is, indicating time on task as instructed above.

3. Evaluation criteria

The evaluation of the finished test tasks' output quality will be conducted by 2 independent third-party raters (experienced patent lawyers) - unaware of your group assignment or tenure - using a predefined rubric focused on the quality and robustness of your work.

On top of that the research team will collect time on tasks, perceived quality & time gains or losses, increased or decreased satisfaction with patent drafting experience through a survey.

On top of that the research team will collect time on tasks, perceived quality & time gains or losses, increased or decreased satisfaction with patent drafting experience through a survey.

6. What's next

We will send:

- Post test task survey shortly (estimated time required <2 mins)
- Monthly rolling feedback survey in the next few weeks (estimated time required 10 mins)
- More about planned assessments of real patents drafted for Google shortly.

II. DRAFTING EXERCISE

Reminder: please go to https://www.tickcounter.com/stopwatch and turn on the timer!

1. Handout

On the following tabs (also linked below) you will see a set of materials for you to draft a portion of a mock patent application:

- **Main Claim**
- **Inventor Notes**
- **Background of the Invention** – Excerpts detailing the technical field, relevant prior art, and the problem the invention is solving.
- **Relevant Figures/ Drawings**

2. Task

Your task is to use these materials to draft:

1. **Five (5) to seven (7) dependent claims** that build upon (i.e., depend from) the main claim provided. Your claims need to include:

a. At least one (1) dependent claim that relies on a prior dependent claim (i.e., nested dependency)
b. At least one (1) element drafted purely in functional language
c. At least one (1) element drafted purely in structural language
d. No two dependent claims may recite the same additional feature set.

Your dependent claims can go beyond what was disclosed in the inventor notes, but have to make logical and technical sense.

2. **A detailed description** of the invention (capped at **1000 words**), focusing on the **primary embodiment** of the invention.
A detailed description should explain how the invention works, its configuration, and how it addresses the problem stated in the background. This should be drafted purely on the information in the provided materials.

III. CRITIQUE EXERCISE

***** Important note: Please DO NOT use any form of AI for this task.** Submissions that use InFlow or any other AI tool are at risk of being rejected.

1. Handout

On the following tab (also linked below) you will see a patent draft for you to critique / redline:

- **Patent:** ADAPTIVE ATTACHMENT COMPONENTS FOR AN ACCESSORY DEVICE

2. Task

Your task is to:

1. **Make "redline" edits ONLY to the following sections of the patent to improve quality**
    - Title
    - Field
    - Background
    - Summary
    - Claims

2. Share **5-10 bullets of feedback** about the patent (summarizing its strengths & vulnerabilities, best and worst written section, as well as proposing revision strategies),

Reminder 1: Please go to https://www.tickcounter.com/stopwatch and click "Lap" on the timer!

Reminder 2: When you finish the task, please go to https://www.tickcounter.com/stopwatch and click “Stop” on the timer, then please record your time on task at the top of your submission.

Reminder 3: Do not use InFlow or any other form of AI for this task.

A2.4. 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.

Category	1	2	3	4	5
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.

A2.5. LLM prompts for evaluating patent draft quality

10-day Task Scoring Prompt

Task: Score the provided patent draft against the rubrics below. The draft contains an Invention Description and Dependent Claims. Give each dimension a score from 1 (lowest) to 5 (highest) based strictly on the rubric descriptions.

Rubric for "Invention Description":
- *Score 1:* The description is highly inadequate. It lacks sufficient detail for a person of ordinary skill in the art (POSITA) to understand or implement the invention. It may be confusing, contradictory, or missing critical elements entirely.
- *Score 2:* The description has significant gaps. It provides a basic overview but lacks the necessary depth or clarity for full comprehension or implementation without undue experimentation. Important components or their interactions are poorly described.
- *Score 3:* The description is adequate. It covers the essential elements and their interactions, allowing a POSITA to understand the invention. However, it might lack detail on alternative embodiments, edge cases, or broader applications.
- *Score 4:* The description is good. It is clear, detailed, and comprehensive. It covers the main embodiment thoroughly and touches upon relevant alternatives or variations, demonstrating a solid understanding of the invention's scope.
- *Score 5:* The description is excellent. It is exceptionally clear, detailed, and robust. It not only covers the main embodiment exhaustively but also anticipates and details numerous variations, alternatives, and potential future applications, maximizing the patent's potential value and defensibility.

Rubric for "Dependent Claims":
- *Score 1:* Claims are very poorly drafted. They may be legally invalid (e.g., lacking antecedent basis, indefinite), overly narrow, or completely fail to capture the invention's core value.
- *Score 2:* Claims have significant weaknesses. They might be unnecessarily limiting, miss key features, or have structural flaws that could invite easy challenges or design-arounds.
- *Score 3:* Claims are adequate. They are legally sound and cover the basic features of the invention. However, they might lack strategic foresight, leaving potential gaps in protection or failing to capture broader commercial applications.

- *Score 4:* Claims are strong. They are well-structured, legally robust, and strategically drafted to cover the invention effectively. They anticipate potential design-arounds and provide solid protection for the core concepts.
- *Score 5:* Claims are excellent. They are masterful in their construction, offering broad and robust protection. They strategically employ varying levels of specificity, anticipate future technological shifts, and are highly defensible against challenges.

Input:
{TEXT_PLACEHOLDER}

90-day Task Scoring Prompt

Task: Score the provided patent drafts and redlines against the rubrics below. Give each dimension a score from 1 (lowest) to 5 (highest) based strictly on the rubric descriptions.

Drafting Rubric:
Same as 10-day rubric above.

Redlining Rubric:
- *Score 1 (Poor):* The redlines address only trivial or cosmetic issues (e.g., typos, formatting) and completely miss substantive legal or technical flaws in the original draft.
- *Score 2 (Fair):* The redlines address some substantive issues but miss major vulnerabilities (e.g., indefinite language, lack of antecedent basis). The reasoning provided (if any) is weak or inaccurate.
- *Score 3 (Adequate):* The redlines correct the most obvious legal and technical errors, making the claims legally sound. However, the edits lack strategic foresight and do not significantly improve the patent's overall strength or scope.
- *Score 4 (Good):* The redlines significantly improve the draft, addressing both glaring errors and more subtle vulnerabilities. The edits demonstrate a strong understanding of patent law and technical accuracy, with sound reasoning provided.
- *Score 5 (Excellent):* The redlines transform the draft. They aggressively address all vulnerabilities, optimize the language for maximum scope and enforceability, and demonstrate exceptional strategic foresight (e.g., anticipating design-arounds, removing "patent profanity"). The accompanying reasoning is highly persuasive and legally rigorous.

Input:
{TEXT_PLACEHOLDER}

Prompt for LLM-based analysis of redlining task

Turn 1: The Anchor Prompt (Objective Extraction & Persona Setup)

I have uploaded two sets of sources:

1. Experiment Submissions: 90+ unique mock patent critique submissions from patent lawyers. These documents contain the participants' redlines, edits, and written feedback on a flawed mock patent draft, including summary feedback at the end. 2. Matching Sheet: A spreadsheet containing metadata and scores for each participant. Every submission has been rated by 2

raters, so each participant has 2 separate rows in this document.

Act as an expert qualitative researcher. I need you to perform a strictly objective analysis comparing 'Top Performers' against 'Low Performers'.

Crucial Constraint: You must use the pre-calculated 'Performance Bucket' column in the Matching Sheet to filter and group participants into these two cohorts:

- Top Performers are participants explicitly labeled as 'Top Performer' in the Performance Bucket column (whose average score across the 5 rubric dimensions is strictly greater than 3).
- Low Performers are participants explicitly labeled as 'Low Performer' in the Performance Bucket column (whose average score across the 5 rubric dimensions is less than or equal to 2). Ignore any participants labeled 'Excluded'.

Next, analyze how these two groups engaged with the document. Do not assume any assigned cohort or seniority level automatically performed better. What specific legal liabilities did Top Performers consistently catch that Low Performers missed (e.g., patent profanity, admitted prior art, antecedent basis issues, functional language)? How did their overall engagement, editing strategy, depth of redlining, and time management differ? Extract the most prominent substantive patterns that objectively differentiate a Top Performer's submission from a Low Performer's submission.

Turn 2: Structuring the Core Patterns

Now, formalize these findings into the 5-6 most prominent patterns that differentiate the score groups. For each pattern, you must provide:

1. Title: A descriptive 'X vs. Y' title (e.g., 'Constructive vs. Dismissive Engagement'). 2. Cohort Focus: Cross-reference the Matching Sheet (specifically the 'Junior/expert' and 'Group' columns) and tell me if this pattern skews towards Seniors, Juniors, or applies Universally across levels. 3. Prevalence & Density: Assign a qualitative density tier to this behavior in Top vs. Low performers (Pervasive [>75% of group], Frequent [50% of group], or Occasional [<25% of group]). To substantiate this, list the specific Participant IDs exhibiting this pattern in a verification footnote. 4. Detailed Explanation: Contrast how Top Performers handled the specific legal liabilities versus Low Performers. Ensure every claim you make is grounded in the actual edits and comments found in the submissions, noted in exact quotes "".

Turn 3: Generating the Exhaustive Markdown Table

Create an 'Exhaustive Pattern Table: Top vs. Low Performers'. The table must include exactly these columns: - Pattern Category - Main Difference - Detailed Clarification - Density Tier & Participant IDs - Top Performer Example (Avg Score >3) - Low Performer Example (Avg Score

<=2)

Constraint: For the 'Top Performer Example' and 'Low Performer Example' columns, you must objectively search the documents, cite specific participant IDs (from the 'File Name' column), and provide exact quotes "" of their actual redlines, edits, or written comments to substantiate the pattern. Do not paraphrase examples.

Below the table, write a dedicated paragraph highlighting the single most critical differentiator in mindset between high and low scorers.

Turn 4: Mapping to the Grading Rubric

Now, map these objectively identified patterns against our 5 evaluation dimensions (Columns U through Y in the Matching Sheet). Create a 'Deep Dive: Patterns by Rubric Dimension' section covering exactly these 5 dimensions: 1. Enforceability 2. Technical Accuracy 3. Strategic Ambiguity 4. Completeness and Alignment 5. Clarity

For each dimension, contrast the Top Performer approach versus the Low Performer approach. You must find and cite specific participant file names from the documents to show exactly who succeeded or failed in that specific dimension, and quote "" their exact edits or comments (e.g., cite who specifically purged business logic vs. who left it in; cite who specifically corrected antecedent basis errors vs. who missed them).

Turn 5: Isolating the Feedback Section (Section II.b)

Next, I want to look strictly at the written 'Feedback' section (Section II.b at the very end of the submitted document) authored by the participants. Create a detailed summary table comparing the feedback style of Top Performers (who acted as 'Fixers') versus Low Performers (who acted as 'Critics').

Compare them across these 5 specific features: 1. Attitude (e.g., collaborative/remedial vs. dismissive/punitive) 2. Vocabulary / Terminology used 3. Legal Basis cited 4. Assessment of the Background section 5. Best Section Identified

Support your table with short verbatim illustrative quotes "" from Section II.b for both groups.

A2.6. Prompt for LLM-based assessment of possible AI use

System Persona & Task Overview

Act as an expert Forensic Linguist and Senior Patent Law Partner. Your task is to analyze a set of manually completed patent redlining and commenting exercises (also known as the Critique exercise) to detect unauthorized Large Language Model (LLM) usage.

Source Context

- The source document titled TT2 contains the original exercise instructions and the unedited baseline text.
- All other source documents are individual participant submissions.
- Participants were instructed to complete inline edits on specific sections of a patent draft and then share top-down feedback on the patent draft in free form.
- Participants were instructed to complete the redlining, commenting, and top-down feedback manually, without using any AI tools.

Strict Execution Rules

- You must evaluate EVERY submitted document in the source list. Do not summarize or group submissions.
- Before assessing a submission, ensure you are seeing comments and inline edits including both added and removed text.
- You must cross-reference each submission against TT2 to determine exactly what the participant changed.
- Before assessing a submission, ensure you are not attributing the original rich text formatting/structure of TT2 to the participant.

Evaluation Criteria

Analyze every submission against TT2. Look specifically for the following forensic markers of LLM usage:

General Tells:

1. **Full Paragraph Rewrites:** Replacing entire paragraphs rather than making targeted, surgical line edits typical of human patent attorneys.
2. **RLHF Compliment Sandwich:** An unnatural, desperate attempt to find strengths or soften critiques before/after delivering corrections.
3. **Lexical Tells:** Overly enthusiastic tone, or high-density use of LLM-typical transitional adverbs (e.g., "furthermore," "additionally," "moreover," "delves"). *Guardrail: Do not penalize standard, dry patent terminology; flag only words that create an inappropriately enthusiastic, conversational, or distinctly AI-like tone.*
4. **Loss of Structural Context:** Mixing up distinct patent sections (e.g., confusing claims with the specification or background).

5. **Action Summarization/Changelog:** Providing a summary list of changes made in lieu of making actual inline edits, or appending a robotic changelog to the text.
6. **Overly Structured Feedback:** Submitting the "top-down feedback" as a highly stylized, excessively nested outline (e.g., multi-level bullet points summarizing the review) that looks machine-generated rather than a human's free-form thoughts.
7. **Persona Over-commitment:** Assuming a "reviewing partner" persona to an absurd degree, resulting in theatrical or unrealistic scolding that does not match the actual context of a timed experiment task.

Junior-Specific Tells (Apply strictly if the participant is a Junior):

1. **Perfect Pacing / Absence of Time Collapse:** Human juniors exhibit "sequential time collapse" (dense, meticulous edits in the first few paragraphs, degrading to sparse, rushed edits at the end due to fatigue/time limits). LLMs maintain uniform, perfect edit density across all sections.

Output Format

Generate a Markdown table analyzing all participant submissions. Do not include any introductory or concluding text outside the table. The table must contain exactly these columns:

1. **Submission Name:** [Extract exact source document title]
2. **Seniority:** [Identify as Junior or Senior based on the document title/contents]
3. **LLM Suspected in Critique (Yes/No):** [Yes if any tells are present, No otherwise]
4. **Reason(s):** [List the specific tell(s) from the 9 criteria above. If none, write "N/A"]
5. **Confidence:** [Score from 1 to 10]
6. **Example Text:** [Extract a direct, verbatim quote from the submission that proves the specified tell. If none, write "N/A"]

A2.7. LLM-based analysis of redlining task

We used Gemini 3.0 Pro, accessed through the NotebookLM interface, to analyze how subjects engaged with the 90-day redlining task. Redlining is the only task for which we hold both the initial flawed draft and each participant's edits and comments on it, so it is the only task in which we observe the process of engagement rather than only its product. The model received the full text of all 91 sample-included redline submissions, comprising both text edits and margin comments, along with each participant's experience cohort, treatment assignment, and quality score given by each rater (calculated as an average of 5 rubric dimension scores). We instructed the model to isolate submissions into two groups: "Top Performers" (average rater score > 3) and "Low

Performers" (average rater score ≤ 2), and to identify emergent qualitative patterns differentiating how these two groups diagnosed and corrected the draft. We supplied no a priori hypotheses about behavioral archetypes, instructing the model only to "identify emergent patterns in how participants diagnosed and corrected the draft, and evaluate whether these patterns map onto the provided cohort or treatment group metadata." Appendix A2.5 reproduces the prompt in full. We use this analysis to organize and describe the qualitative features of the submissions that drove the human expert ratings, not to derive quantitative point estimates.

The analysis identified six behavioral patterns separating high- from low-performing submissions. Because our quantitative results demonstrate that treated seniors captured the largest performance gains, they make up the bulk of the 'Top Performer' cohort. Consequently, the high-performing behavioral variants described below are predominantly observed among treated seniors.

1. **Constructive versus dismissive engagement.** *Senior-specific; roughly 35 high performers, predominantly treated, against 6 low performers, all of them controls.*

   Constructive engagement is an active willingness to salvage the flawed draft through comprehensive overhaul. High performers resolved logical gaps, repaired broken dependency chains, and standardized nomenclature between the specification and the claims.

   Dismissive engagement is a passive refusal to engage with the flawed draft. Low performers often used the feedback section to criticize the hypothetical drafter's competence instead of suggesting improvements. They left the text itself unedited while writing notes such as "These claims are pretty unsalvageable," "I did not review the dependent claims in any detail," and "I would have given it back to the associate."

2. **Restructuring versus passive acceptance.** *Universal across experience cohorts; roughly 30 high performers, predominantly treated, against 20 low performers drawn from both cohorts.*

   Restructuring neutralizes macro-level legal liabilities even when doing so means going beyond redlines and comments. High performers deleted blatant enablement traps such as "All alternatives and hypothetical futures are hereby claimed," stripped admitted prior art from the Background, and rewrote the Summary so that it tracked the corrected claim language.

   Passive acceptance treats the initial draft as a rigid framework and confines engagement to localized redlines and commentary. Low performers retained flawed structures, narratives, and legal traps on the (seeming) assumption that the document's macro-structure had to be preserved.

3. **Prioritized versus sequential time use.** *Junior-skewed on the sequential side; roughly 10 high performers, predominantly treated, against 8 low performers, predominantly controls.*

   Prioritized time use allocates effort by legal value rather than by document order. High performers deliberately bypassed the Title, Field, and Background sections in order to spend

time rebuilding the Claims and conforming the Summary to those new claims, returning to the remaining sections only as time allowed.

Sequential time use follows a rigid top-to-bottom progression. Low performers spent heavily on the prose of the Title, Field, and Background, then ran out of time before reaching the Claims, submitting no redlines at all to the section carrying the most legal weight and leaving notes such as "N/A (out of time)."

4. **Strategic focus versus undifferentiated effort.** *Senior-specific; roughly 25 high performers, predominantly treated, against 15 low performers drawn from both experience cohorts.*

   Strategic focus concentrates analytical rigor on the commercial boundaries of the patent. High performers diagnosed the original claims as overly specific "picture claims" of little commercial value. They purged "patent profanity" such as "must" and "essential" that needlessly narrowed scope, and they removed business-focused language such as "market demands" and "SKUs" from the specification.

   Undifferentiated effort spreads attention uniformly across the document, treating all text as carrying equal legal weight. These participants misallocated effort rather than exhausting it, which distinguishes them from the sequential group above. They tweaked stylistic flow, formatting, and boilerplate while leaving commercially limiting constraints embedded in the patent's scope, retaining claim limitations keyed to "assembly conditions" or "material cost."

5. **Rationale-heavy versus proofreading inputs.** *Universal across experience cohorts; roughly 25 high performers, predominantly treated, against 20 low performers, predominantly controls.*

   Rationale-heavy inputs state the legal doctrine underpinning each edit. High performers explained why a change was made, noting for instance that vague terms such as "generally" must go because they forfeit doctrine-of-equivalents arguments, or that characterizing prior art in the Background invites admitted prior art rejections.

   Proofreading inputs treat the task as copyediting and mechanical compliance. Low performers executed small, localized fixes that did nothing for enforceability: globally swapping "may" or "might" for "can," correcting passive voice, and repairing typographic omissions and numeral formatting. Confining themselves to these adjustments, they left the statutory vulnerabilities and scope limitations in the claim language untouched.

6. **Robust versus high-level inputs.** *Universal across experience cohorts; roughly 40 high performers, predominantly treated, against 15 low performers drawn from both cohorts.*

   Robust inputs are granular and precise, eliminating severe flaws such as indefiniteness vulnerabilities, antecedent basis errors, and subjective or functional limitations. One participant replaced "having a size and shape generally sufficient to cover a broad range of foreseeable form-factors" with "configured to receive a plurality of different form-factors."

High-level inputs offer sweeping commentary without acting on it. Unlike the dismissive engagement described above, these participants diagnosed real problems, noting in feedback bullets that "the claims use indefinite language" or that "there are antecedent basis issues," but they made no corresponding corrections in the document.

A2.7.1. Observational note on junior cohort behaviors

Due to the relatively small number of junior lawyers in the study, it is difficult to make robust comparisons of qualitative differences between treated and control juniors. But telltale editing archetypes appear among treated and control juniors alike. Junior work is marked by rigid formalism and cosmetic issue-spotting rather than strategic problem-solving. Asked to evaluate the patent's flaws, juniors hunted for rule violations and administrative non-compliance, reporting that "apparatus claims are written using method terminology" or that "some terminology does not comply with Google's guidelines" rather than assessing the commercial boundaries of the text.

Many edits made by juniors focused on surface-level synonym swapping and left structural liabilities intact. Even those juniors who correctly diagnosed a macro-level vulnerability, such as noting that the "background section includes detailed discussion of prior art," typically recorded the observation as a comment rather than redlining the unsound text. This formalistic posture separates junior submissions from the strategic rewriting seen among seniors. It is as common among juniors in the treatment and control groups.