AI-POWERED LAWYERING: AI REASONING MODELS, RETRIEVAL AUGMENTED GENERATION, AND THE FUTURE OF LEGAL PRACTICE

Daniel Schwarcz, Sam Manning, Patrick Barry, David R. Cleveland, JJ Prescott, & Beverly Rich*

Abstract: Generative AI is set to transform the legal profession, but its full impact remains uncertain. While AI models like GPT-4 improve the efficiency with which legal work can be completed, they can at times make up cases and "hallucinate" facts, thereby undermining legal judgment, particularly in complex tasks handled by skilled lawyers. This article examines two emerging AI innovations that may mitigate these lingering issues: Retrieval Augmented Generation (RAG), which grounds AI-powered analysis in legal sources, and AI reasoning models, which structure complex reasoning before generating output. We conducted the first randomized controlled trial assessing these technologies, assigning upper-level law students to complete six legal tasks using a RAG-powered legal AI tool (Vincent AI), an AI reasoning model (OpenAI's o1-preview), or no AI. We find that both AI tools significantly enhanced legal work quality, a marked contrast with previous research examining older large

* Schwarcz and Manning are co-first authors. Schwarcz is Fredrikson & Byron Professor of Law at the University of Minnesota Law School. Manning is Senior Research Fellow at the Centre for the Governance of AI. Barry is Clinical Assistant Professor and Director of Digital Academic Initiatives at the University of Michigan Law School. Cleveland is Clinical Professor of Law and Director of Legal Research & Writing at University of Minnesota Law School. Prescott is Henry King Ransom Professor of Law at the University of Michigan Law School. Rich is Practice Innovation Counsel at Ogletree Deakins. For generous financial support of this project, we thank Fredrikson & Byron PA, Robins Kaplan LLC, University of Minnesota Law School, and University of Michigan Law School. Neither OpenAI nor VLex (the company that owns Vincent AI) provided financial support for this project, but they both did make their AI platforms freely available to study participants. Given the nature of this Article, it should not be surprising that the authors used certain generative AI tools to assist them with drafting and analysis. For helpful comments and guidance, we thank Pablo Arredondo, Victor Bennett, Jonathan Choi, Miryam Gorelashvili, Dan Ho, Bryan Mechell, Amy Monahan, Daniel Rock, James Snelson, Tim Sullivan and Peter Wills. Tomás Aguirre provided excellent research assistance. Anonymized data and code used for the experiment and analysis made available upon request.

language models like GPT-4. Moreover, we find that these models maintain the efficiency benefits associated with use of older AI technologies. Our findings show that AI assistance significantly boosts productivity in five out of six tested legal tasks, with Vincent yielding statistically significant gains of approximately 38% to 115% and o1-preview increasing productivity by 34% to 140%, with particularly strong effects in complex tasks like drafting persuasive letters and analyzing complaints. Notably, o1-preview improved the analytical depth of participants' work product but resulted in some hallucinations, whereas Vincent AI-aided participants produced roughly the same amount of hallucinations as participants who did not use AI at all. These findings suggest that integrating domain-specific RAG capabilities with reasoning models could yield synergistic improvements, shaping the next generation of AI-powered legal tools and the future of lawyering more generally.

TABLE OF CONTENTS

INTRODUCTION

Generative AI is poised to transform the legal profession in the coming years.[1] Yet the scope and nature of this transformation remain uncertain. Some legal technology enthusiasts foresee a fundamental restructuring, where AI automates countless legal tasks and even replaces certain types of lawyers entirely.[2] Skeptics, however, argue that while AI may streamline aspects of legal work, it is unlikely to alter the core nature of lawyering.[3]

The stakes of this debate are enormous. Billions of dollars are pouring into AI-driven legal startups,[4] and industry giants like Westlaw and LexisNexis are racing to integrate AI into their own already-existing platforms.[5] Across the profession, lawyers—from Big Law partners[6] to legal aid attorneys[7]—are grappling with how best to incorporate AI into their work. Even judges are exploring ways AI may help them adjudicate cases and draft opinions.[8] Meanwhile, law schools and students face growing uncertainty about how to prepare for the profession's

---

[1] Jonathan H. Choi, Amy Monahan, & Daniel Schwarcz, *Lawyering in the Age of Artificial Intelligence*, 109 Minn. L. Rev. 147, 150 (2024); AKSH GARG & MEGAN MA, OPPORTUNITIES AND CHALLENGES IN LEGAL AI, STANFORD LAW SCHOOL (Jan. 2025); MICROSOFT, GENERATIVE AI FOR LAWYERS (2024).

[2] *See, e.g.*, RICHARD SUSSKIND & RICHARD E. SUSSKIND, TOMORROW'S LAWYERS: AN INTRODUCTION TO YOUR FUTURE (2023); Raymond H. Brescia, *What's a Lawyer For?: Artificial Intelligence and Third-Wave Lawyering*, 51 FL. ST. U. L. REV. 542 ((2024).

[3] *See, e.g.*, John Armour, Richard Parnham, & Mari Sako, *Augmented Lawyering*, 2022 U. ILL. L. REV. 71, 72.

[4] *See, e.g.*, Press Release, Harvey, Harvey Raises $100M Series C from Google Ventures, OpenAI, Kleiner Perkins, Sequoia Capital, Elad Gil, and SV Angel at a $1.5B valuation (July 23, 2024).

[5] *See* Press Release, LexisNexis, LexisNexis Introduces Protégé Personalized AI Assistant with Agentic AI, Making it Easier to Power Complex Legal Task Completion (Jan. 27, 2025); Press Release, Thompson Reuters, Get to Know Thomson Reuters: Our Technology Journey and What's Next (Jan. 23, 2025).

[6] *See* Roy Strom, *Big Law Is Questioning the 'Magical Thinking' of AI as Savior*, BLOOMBERG LAW (Aug. 8, 2024).

[7] *See* Miriam Kim & Colleen V. Chien, *Generative AI and Legal Aid: Results from a Field Study and 100 Use Cases to Bridge the Access to Justice Gap*, 57 LOYOLA LA LAW REV. 903, 904 (2025); Colleen V. Chien, Miriam Kim, Akhil Raj, & Rohit Rathish, *How Generative AI Can Help Address the Access to Justice Gap Through the Courts*, 57 LOYOLA LA LAW REV. 850 (2025).

[8] *See* John G. Roberts, Jr., *2023 Year-End Report on the Federal Judiciary,* SUP. CT. OF THE U.S. 5 (2023); Richard M. Re, *Artificial Authorship and Judicial Opinions*, 92 GEO. WASH. L. REV. 1558, 1559 (2024); John Zhuang Liu & Xueyao Li, *How Do Judges Use Large Language Models? Evidence From Shenzhen*, 16 J LEGAL ANALYSIS 235, 236 (2024); Yonathan Arbel & David A. Hoffman, *Generative Interpretation*, 99 NYU L. REV. 451 (2024).

increasingly uncertain future. How do you train for a legal landscape that could be radically transformed by the time you graduate?[9]

As this conversation unfolds, empirical research has begun to shed light on AI's actual impact on legal practice, though the findings have been mixed. Early studies showed that generative AI tools perform reasonably well on law school and bar exams, but the real-world implications of these results remain unclear.[10] The practice of law requires human judgment, and significant differences exist between producing a plausible-sounding answer on an exam and actually performing valuable legal work.[11]

Research more relevant to human lawyers' use of generative AI suggests that these tools can produce important efficiency benefits by reducing the time needed to perform certain legal tasks.[12] But this research offers limited evidence that AI tools can consistently improve the quality of legal work product,[13] and it highlights the risk that AI may

[9] See John Bliss, *Teaching Law In The Age Of Generative AI*, 64 JURIMETRICS 111 (2024); Amanda Head & Sonya Willis, *Assessing Law Students In A GenAI World To Create Knowledgeable Future Lawyers*, 31 INT'L J LEGAL PROF. 293 (2024).

[10] *See* Jonathan H. Choi, Kristin E. Hickman, Amy B. Monahan & Daniel Schwarcz, *ChatGPT Goes to Law School*, 71 J. LEGAL EDUC. 387 (2022); Daniel Martin Katz, Michael James Bommarito, Shang Gao & Pablo Arredondo, *GPT-4 Passes the Bar Exam*, PHIL. TRANS. R. SOC. A, Apr. 15, 2024, at 1, 3–5. *cf.* Eric Martínez, *Re-Evaluating GPT-4's Bar Exam Performance*, 1 Artificial Intelligence and Law 1 (2024) (finding that GPT-4 performance on the bar exam did not surpass 90% of human test takers, as had been touted by OpenAI and reported in numerous major media outlets).

[11] *See* Jonathan H. Choi & Daniel Schwarcz, *AI Assistance in Legal Analysis: An Empirical Study*, 73 J. LEGAL EDUC. (forthcoming 2025); Nicole Yamane, *Artificial Intelligence in the Legal Field and the Indispensable Human Element Legal Ethics Demands*, 33 GEO. J. LEGAL ETHICS 877, 882 (2020); W. Bradley Wendel, *The Promise and Limitations of Artificial Intelligence in the Practice of Law*, 72 OKLA. L. REV. 21, 24–26 (2019); John G. Browning, *Robot Lawyers Don't Have Disciplinary Hearings-Real Lawyers Do: The Ethical Risks and Responses in Using Generative Artificial Intelligence*, 40 GA. ST. U. L. REV. 917 (2023).

[12] *See* Choi, Monahan, & Schwarcz, *supra* note 1; Choi & Schwarcz, *supra* note 11; Aileen Nielsen, Stavroula Skylaki, Milda Norkute, & Alexander Stremitzer, *Building A Better Lawyer: Experimental Evidence That Artificial Intelligence Can Increase Legal Work Efficiency*, 21 J. EMP. LEGAL STUDS. 979 (2024).

[13] *See* Choi, Monahan, & Schwarcz, *supra* note 1; Choi & Schwarcz, *supra* note 11.

"hallucinate" fake source material and crowd out lawyers' independent judgment.[14]

To date, a key limitation of this research on AI and lawyering is its focus on older AI models, such as ChatGPT-3.5 and GPT-4. Because these models have limited ability to break down analytically complex tasks or draw from the most relevant legal source materials, their usefulness to lawyers is limited.[15] By contrast, two emerging technologies have the potential to significantly advance AI's role in law by improving reasoning capabilities and grounding outputs in authoritative legal sources.

The first is a new class of generative AI language models known as "reasoning models."[16] Unlike earlier AI chatbots, these models are explicitly designed to use additional computational resources at the point of use, planning responses before generating them—much like a human taking longer to think and outline thoughts before answering a complex question.[17] This shift is significant enough that, to highlight the distinction, OpenAI introduced an entirely new naming convention for its first reasoning model: "o1."[18] Early evidence suggests these reasoning models dramatically outperform their predecessors in complex tasks across fields such as mathematics, coding, and medical diagnosis.[19]

The second major advance relevant to the legal profession is Retrieval-Augmented Generation (RAG), a technique that integrates generative AI with legal source materials.[20] Unlike traditional models

---

[14] *See* Matthew Dahl, Varun Magesh, Mirac Suzgun, & Daniel E. Ho, *Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models,* 16 J. LEGAL ANALYSIS 64, 66 (2024) (finding a "widespread occurrence of legal hallucinations" in legal analysis of large language models); Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, Daniel E. Ho, *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools* (May 30, 2024), at https://arxiv.org/abs/2405.20362

[15] *See, e.g.*, Armin Alimardani, *Generative Artificial Intelligence Vs. Law Students: An Empirical Study On Criminal Law Exam Performance*, 16 LAW, INNOVATION & TECH. 777 (2024).

[16] *See* Press Release, OpenAI, Introducing OpenAI o1-preview (9/12/24), at https://openai.com/index/introducing-openai-o1-preview/.

[17] *See* Peter G. Brodeur et al, *Superhuman Performance Of A Large Language Model On The Reasoning Tasks Of A Physician* (Dec. 14, 2024), at https://arxiv.org/abs/2412.10849

[18] *See* Open AI, *supra* note 16

[19] *See* Brodeur et al, *supra* note 17; Anita Kirkovska & Akash Sharma, *Analysis: OpenAI o1 vs GPT-4o vs Claude 3.5* SONNET (Dec 17, 2024), at https://www.vellum.ai/blog/analysis-openai-o1-vs-gpt-4o; OpenAI, OpenAI o1 System Card (12/5/24), at https://openai.com/index/openai-o1-system-card/

[20] Patrick Lewis et al., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, 33 ADVANCES NEURAL INFO. PROCESSING SYS. 9459 (2020).

that rely solely on their training data to answer prompts, legal AI systems with RAG capabilities can retrieve relevant legal texts—such as case law, statutes, and regulations—before generating responses.[21] By grounding outputs in authoritative sources, RAG aims to minimize hallucinations and enhance the accuracy of AI-assisted legal analysis.[22] Perhaps even more importantly, RAG makes it easier for humans to check an AI's output by consulting the underlying sources on which it relied to generate an answer.[23] According to its proponents, RAG-enabled tools thus enable lawyers to accurately, confidently, and efficiently synthesize and analyze vast stores of legal source material.[24]

To better understand AI's impact on the future of lawyering, we conducted the first randomized controlled trial of these two emerging legal AI technologies: RAG and reasoning models. Our study involved 127 law students from the University of Minnesota and the University of Michigan law schools. Each of them completed six realistic legal assignments developed in collaboration with practicing lawyers.[25] For two tasks, participants received no AI assistance; for two, they used an AI reasoning model (o1-preview); and for the remaining two, they had access to Vincent AI, a leading tool that integrates RAG and automated prompting assistance.[26] The assignment of AI tools and control conditions was randomized to ensure a balanced distribution across participants.

Before beginning the assignments, participants received training on the effective use of both AI tools. This included both (1) general training on the use of AI models for legal work and (2) training specifically tailored to Vincent AI.[27] All assignments were blindly graded by team members who were lawyers with practice experience and who were uninvolved in data collection or analysis. Graders used standardized rubrics that assessed key attributes of quality legal work, including clarity, accuracy, and analytical depth.[28]

---

[21] *See id.*

[22] *See* Magesh et al, *supra* note 14.

[23] *See id.*

[24] *See, e.g.*, Niko Grupen & Julio Pereyra *BigLaw Bench – Retrieval* (11/13/24), at https://www.harvey.ai/blog/biglaw-bench-retrieval;

[25] The study design mirrored the basic research design of a prior study that was co-authored by one of the co-authors of this Article. *See* Choi, Monahan, & Schwarcz, *supra* note 1.

[26] *See* Press Release, VLex, AI that Knows the Law, at https://vlex.com/vincent-ai.

[27] The general AI training drew from prior work of one of the co-authors. See Daniel Schwarcz & Jonathan H. Choi, *AI Tools for Lawyers: A Practical Guide,*108 MINN. L. REV. HEADNOTES 1 (2023). See Part II, infra, for further details.

[28] The rubrics that were used for grading are contained in the Appendix.

Our findings reveal that access to both o1-preview and Vincent AI led to substantial, statistically significant improvements in the speed with which participants completed legal tasks. Our findings also reveal that, for four of the six tested assignments, the quality of the work product submitted by participants with access to those AI tools was considerably better than the work product submitted by participants without AI access.[29] While the speed-related improvements were comparable in magnitude to those observed in prior research examining the impact of GPT-4 on lawyering, the quality enhancements marked a significant departure from earlier studies. Those studies generally reported limited quality gains in realistic lawyering tasks.[30] In fact, this new study provides the first empirical evidence, to our knowledge, that AI tools can consistently and significantly enhance the quality of human lawyers' work across various realistic legal assignments.

We also observed variation in how—and the extent to which—these two AI tools enhanced quality.[31] For Vincent AI, quality improvements were primarily seen in the clarity, organization, and professionalism of submitted work.[32] The tool's impact on accuracy, however, was mixed. On the one hand, overall accuracy scores—which depended on whether an answer included and properly characterized the most relevant legal authorities and facts[33]—did not improve significantly in any assignment where participants used Vincent AI; for one task, the tool even appeared to *reduce* the accuracy of submitted work.[34] On the other hand, assignments completed with Vincent AI contained fewer hallucinations (3 total) than those produced using o1-preview (11 total). Assignments completed with Vincent also contained slightly fewer hallucinations than those completed without any AI assistance at all (4 total).[35]

We found that o1-preview led to stronger and more widespread improvements in the quality of legal work compared to Vincent.[36] Most

[29] See Part III, *infra*.

[30] *Compare* Choi, Monahan, & Schwarcz, *supra* note 1 (finding providing access to GPT-4 consistently and substantially increased speed of performance on lawyering tasks, but had limited and inconsistent effects on quality of work product); Choi & Schwarcz, *supra* note 11 (finding that providing students with access to GPT-4 produced mixed quality improvements on comparatively simple legal exams, but more marginal and inconsistent improvements on harder exams geared towards upper level law students).

[31] *See* Part III, *infra*.

[32] *See id.*

[33] *See* Appendix C, *infra*.

[34] *See* Part III, *infra*.

[35] *See id.*

[36] *See id.*

notably, in addition to enhancing clarity, organization, and professionalism, o1-preview produced statistically significant and substantial improvements in the quality of the legal analysis contained in three of the six assignments tested.[37] We assessed this metric based on the logical coherence and nuanced reasoning in the submitted assignments. This finding suggests that, when it comes to their potential to improve legal work, AI reasoning models represent a difference in kind—not just degree – relative to earlier LLMs like GPT-4.

Our results further illuminate a range of important questions for the future of lawyering. For instance, we find that improvements in quality from the two AI tools were concentrated in the litigation-oriented tasks that we tested; they did not appear to extend to the one transactionally oriented task we tested, which involved drafting a contract rather than tasks focused on potential or actual litigation.[38] We also found through post-experiment surveys that most participants felt their experience with the two AI tools in the study increased their likelihood of using similar tools in the future.[39] Many also reported gaining proficiency in using these tools over the course of the experiment. This positive subjective experience from using the two tools was particularly pronounced for Vincent AI relative to o1-preview.[40]

The implications of our findings for Vincent AI and o1-preview are each independently significant. Considered together, however, they are even more noteworthy, as these two AI technologies enhance legal work in distinct yet complementary ways. Vincent does so principally through retrieval-augmented generation. It supplements this capability with automated prompting, which supplies pre-crafted prompts based on factors such as the documents users upload, and the tasks they indicate they would like to complete. Notably, however, Vincent used an ensemble of non-reasoning OpenAI models—such as GPT-4 and GPT-4o—as its core foundation models at the time of the experiment.[41] By contrast, o1-preview enhances legal work through technological improvements to foundation models, which can be integrated into legal AI tools like Vincent.[42] Moreover, the reasoning model we tested— o1-preview —was

[37] *See id.*

[38] *See id.*

[39] *See id.*

[40] *See id.*

[41] *See* Email from Damien Riehl (9/27/24). In fact, OpenAI first provided API access to their reasoning models in December 2024, which was after the completion of this experiment. *See* Justin Sullivan, *OpenAI Brings Its o1 Reasoning Model To Its API — For Certain Developers*, TECH CRUNCH (12/7/24).

[42] *See*, *e.g.*, Gabe Pereyra & Winston Weinberg, *Harvey: Is Building Legal Agents And Workflows With OpenAI o1* (Sep 12, 2024).

the first publicly available reasoning model.[43] Since then, multiple new generations have been released, each improving upon its predecessors.[44] Early advances in new model types often yield particularly significant gains, suggesting that improvements in AI reasoning models are likely to continue, particularly if the returns to continued scaling up of inference-time compute are large.

This Article is structured in four Parts. Part I reviews current evidence on the impact of generative AI on the legal profession, emphasizing the emergence of retrieval-augmented generation (RAG) and reasoning-based foundation models. Part II details our methodology, which employs a randomized controlled trial to enable strong causal inferences about the effects of the AI tools we tested. Part III presents our findings. It shows that the latest generation of reasoning models enhances the quality of legal work product in two-thirds of the assignments we tested, while also delivering significant and consistent efficiency gains. Finally, Part IV explores the broader implications of our analysis. Given that the technologies we tested, when combined, offer complementary benefits, we suggest that our results understate the current potential of AI-powered lawyering. If each tool, on its own, can help lawyers in different ways, using them in tandem could provide even bigger advantages—like providing a hiker with both a map and a compass to navigate unknown terrain more effectively.

I. BACKGROUND

Over the last several years, generative AI technology has advanced at an unprecedented pace.[45] New legal technologies incorporating generative AI have emerged just as rapidly.[46] These developments have elicited a wide range of reactions from the legal community—enthusiasm, caution, and even outrage.[47] Amid this

[43] *See* Karl Freund, *Will Open AI's o1 Reasoning Model Really Change The World?*, FORBES (Dec. 10, 2024).

[44] *See* Kyle Wiggers, *OpenAI Launches o3-mini, its Latest 'Reasoning' Model*, TECHCRUNCH (Jan. 31, 2025).

[45] *See The Great Acceleration: CIO Perspectives On Generative AI*, MIT TECH REV. (July 18, 2023), https://www.technologyreview.com/2023/07/18/1076423/the-great-acceleration-cio-perspectives-on-generative-ai.

[46] CASETEXT, *Meet Your New AI Legal Assistant*, https://casetext.com [https://perma.cc/5SDR-PG3S]; *Thomson Reuters to Acquire Legal AI Firm Casetext for $650 Million*, REUTERS (June 27, 2023), https://www.reuters.com/markets/deals/thomson-reuters-acquire-legal-techprovider-casetext-650-mln-2023-06-27.

[47] *See, e.g.,* Erin Mulvaney & Laura Webber, *End of the Billable Hour? Law Firms Get on Board with Artificial Intelligence*, WALL ST. J. (May 11, 2023,

uncertainty, empirical research has tempered both undue hype and unwarranted doubt, showing that while generative AI can enhance lawyers' efficiency across various tasks, it neither replaces them nor fundamentally transforms the nature of legal work—at least so far.

Section A of this Part traces that trajectory. It focuses on the evolution of generative AI models like ChatGPT and Claude, which are large language models with general purpose capabilities. Section B then examines a new wave of AI tools, including reasoning models and those leveraging Retrieval-Augmented Generation (RAG). These emerging technologies are driving another cycle of hype and speculation. Empirical evidence of their impact, however, has largely been absent.

A. The First Wave of Legal AI (Late 2022 to Mid 2024)

When OpenAI released the Large Language Model (LLM) ChatGPT for public use in late 2022, the effect on the world was immediate and profound.[48] While ChatGPT's core design was not entirely novel—like earlier chatbots, it generated text by predicting the next word in a sequence[49]—the technology reshaped the AI landscape with its

---

11:00 AM), https://www.wsj.com/articles/end-of-the-billable-hour-law-firms-get-on-board-with-artificial-intelligence-17ebd3f8; ST. BAR CAL. STANDING COMM. ON PRO. RESP. AND CONDUCT, PRACTICAL GUIDANCE FOR THE USE OF GENERATIVE A.I. IN THE PRACTICE OF LAW 3 (2023), https://www.calbar.ca.gov/Portals/0/documents/ethics/Generative-AI-Practical-Guidance.pdf; Stephanie Wilkins, *ChatGPT Is Impressive, But Can (and Should) It Be Used in Legal?*, LEGALTECH NEWS (Dec. 15, 2022), https://www.law.com/legaltechnews/2022/12/15/chatgpt-is-impressive-but-can-and-should-it-be-used-in-legal/?slreturn=20230223101453 [https://perma.cc/5QQM-Q6UT]; Roger E. Barton, *How Will Leveraging AI Change the Future of Legal Services?,* REUTERS (Aug. 23, 2023, 9:06 AM), https://www.reuters.com/legal/legalindustry/how-will-leveraging-ai-change-future-legal-services-2023-08-23; Daniel Farrar, *To Future-Proof Their Firms, Attorneys Must Embrace AI*, FORBES (July 13, 2023, 9:00 AM), https://www.forbes.com/sites/forbesbusinesscouncil/2023/07/13/to-future-proof-their-firms-attorneys-must-embrace-ai/?sh=6282438b245b; Steve Lohr, *A.I. is Coming for Lawyers, Again*, N.Y. TIMES (April 10, 2023), https://www.nytimes.com/2023/04/10/technology/ai-is-coming-for-lawyers-again.html; John Villasenor, *How AI Will Revolutionize the Practice of Law,* BROOKINGS INST. (March 20, 2023), https://www.brookings.edu/articles/how-ai-will-revolutionize-the-practice-of-law; Pierce, Natalie A. Pierce & Stephanie L. Goutos, *Why Lawyers Must Responsibly Embrace Generative AI,* 21 BERKELEY BUS. LJ 469 (2024).

[48] *See* Bernard Marr, *A Short History Of ChatGPT: How We Got To Where We Are Today*, FORBES (May 19, 2023).

[49] Priya Shree, *The Journey of Open AI GPT models*, MEDIUM (Nov. 9, 2020), https://medium.com/ walmartglobaltech/the-journey-of-open-ai-gpt-

remarkable ability to produce high-quality responses across diverse queries.[50]

ChatGPT's capabilities stemmed from several key innovations. First, the model's size expanded dramatically from prior LLMs, growing from 117 million parameters in early iterations of the chatbot to 175 billion in later versions.[51] Second, ChatGPT's training included Reinforcement Learning from Human Feedback (RLHF), a technique that fine-tuned responses based on human-provided evaluations.[52] This method was used to tune the model to follow instructions and respond to queries.[53]

Almost immediately after ChatGPT's public release, lawyers and commentators worldwide began speculating about whether the underlying technology could revolutionize legal practice.[54] This excitement grew as studies showed that ChatGPT could achieve passing—albeit low—grades on a range of law school exams simply by processing the exam text.[55] Equally significant, emerging research

---

models-32d95b7b7fb2; Harry Surden, *Machine Learning and Law*, 89 WASH. L. REV. 87, 87 (2014).

[50] *Introducing ChatGPT*, OpenAI (Nov. 30, 2022), https://openai.com/blog/chatgpt. For an excellent overview that is accessible to lawyers, see Katherine Lee, A. Feder Cooper, & James Grimmelmann, *Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain (The Short Version)*, PROCEEDINGS OF THE SYMPOSIUM ON COMPUTER SCIENCE AND LAW (2024).

[51] OpenAI, *supra* note 50.

[52] Paul F. Christiano et al., *Deep Reinforcement Learning From Human Preferences, In* PROCEEDINGS OF THE 31ST CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS 4302 (2017) (discussing RLHF).

[53] See Long Ouyang et al., Training Language Models To Follow Instructions With Human Feedback, (March, 2022), https://arxiv.org/abs/2203.02155

[54] JOSEPH BRIGGS ET AL., GOLDMAN SACHS, THE POTENTIALLY LARGE EFFECTS OF ARTIFICIAL INTELLIGENCE ON ECONOMIC GROWTH (2023); Kate Beioley & Cristina Criddle, *Allen & Overy Introduces AI Chatbot to Lawyers in Search of Efficiencies*, FIN. TIMES (Feb. 15, 2023), https://www.ft.com/content/baf68476-5b7e-4078-9b3e-ddfce710a6e2; Emily Hinkley, *Mishcon de Reya Is Hiring an "Engineer" to Explore How Its Lawyers Can Use ChatGPT*, LEGAL CHEEK (Feb. 16, 2023, 8:35:00 AM), https://www.legalcheek.com/2023/02/mishcon-de-reya-is-hiring-an-engineer-to-explore-how-its-lawyers-can-use-chatgpt; *Geoffrey Vance, AI + Human: A Bright Future For Legal Co-Pilots*, JDSUPRA (Apr. 20, 2023), https://www.jdsupra.com/legalnews/ai-human-a-bright-future-for-legal-co-7452383.

[55] *See* Choi, Monahan, Hickman, & Schwarcz, *ChatGPT Goes to Law School*, supra note 10. Subsequent studies confirmed this finding. *See* Margaret Ryznar, *Exams in the Time of ChatGPT*, 80 WASH. & LEE L. REV. ONLINE 305

suggested that giving humans access to ChatGPT could enhance their performance on basic non-legal writing tasks, further fueling interest in AI's role in legal work.[56]

At the same time, this early enthusiasm was tempered by widespread skepticism about the use of this technology in legal practice. There are several key concerns. First and foremost, ChatGPT had a well-documented tendency to "hallucinate" facts and legal sources. How could it be trusted as a research tool if the cases and details it generated weren't always real? The second concern was that its performance on law school exams fell well below the standard expected of most students. (It averaged about a C+.)[57] A third concern was that lawyering has traditionally been seen as requiring human judgment, and many feared that generative AI could undermine or even displace this fundamental aspect of the profession.[58] Finally, lawyers rightly worried that any privileged information input into tools like ChatGPT could be used to train future models, potentially compromising the confidentiality of sensitive client information.[59]

(2023); Andrew Blair-Stanek et al., GPT-4's Law School Grades: Con Law C, Crim C-, Law & Econ C, Partnership Tax B, Property B-, Tax B (May 24, 2023) (unpublished manuscript) (on file with authors; John Ney et al., *Large Language Models as Tax Attorneys: A Case Study in Legal Capabilities Emergence*, 381 PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A: MATHEMATICAL, PHYSICAL AND ENGINEERING SCIENCES (2023). Similar studies emerged across a range of white-collar professions. See, e.g., Tiffany H. Kung et al., *Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models*, PLOS DIGITAL HEALTH, Feb. 2023, at 1 (finding that ChatGPT performed "at or near the passing threshold" on the United States Medical Licensing Exam).

[56] *See* Shakked Noy & Whitney Zhang, *Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence*, 381 SCI. 187, 190 (2023) (finding that access to ChatGPT improved writing speed and quality, particularly for lower-performing participants, while high-performing participants saw only speed gains).

[57] *See* Choi, Monahan, Hickman, & Schwarcz, *ChatGPT Goes to Law School*, *supra* note 10.

[58] W. Bradley Wendel, *The Promise and Limitations of Artificial Intelligence in the Practice of Law*, 72 OKLA. L. REV. 21, 24-26 (2019); Nicole Yamane, *Artificial Intelligence in the Legal Field and the Indispensable Human Element Legal Ethics Demands*, 33 GEO. J. LEGAL ETHICS 877, 889-90 (2020). *See generally* Rebecca Crootof, Margot E. Kaminski & W. Nicholson Price II, *Humans in the Loop*, 76 VAND. L. REV. 429 (2023) (exploring how the law already regulates the interactions of humans and AI, and how it should evolve to perform this role more effectively).

[59] Jon Garon, *"Ethics 3.0—Attorney Responsibility in the Age of Generative AI,"* THE BUSINESS LAWYER, AM. BAR ASSOC (2024); Cooper, A. Feder,

Several of these concerns gained significant traction among lawyers and judges in May 2023 when an otherwise routine legal dispute made global headlines. In what would be the first of many similar incidents,[60] a New York lawyer submitted a court filing containing references to entirely fictitious cases.[61] When questioned in court, the lawyer admitted to using ChatGPT to write his brief. He further explained that, after the tool initially provided the citations, he had explicitly asked whether they were real. ChatGPT affirmed they were.[62] The judge publicly reprimanded the lawyer, sparking widespread media coverage and cementing the incident as a cautionary tale among legal professionals.[63]

Even as the story of the inept New York lawyer circulated widely within the legal community, multiple technology companies—including OpenAI, Google, Microsoft, Meta, and Anthropic—were unveiling new, more advanced LLM models that began shifting the narrative among legal professionals yet again. These newer models, such as GPT-4, featured expanded context windows, more efficient tokenization, and greater parameter counts.[64]

Yet what truly captured lawyers' attention was not these technical upgrades, but a widely reported study finding that GPT-4 passed the bar exam.[65] An OpenAI press release even claimed that the model not only passed the exam but also ranked among the top 10% of human test-takers. However, further analysis revealed that the top 10% figure was an overstatement—largely because it was based on a comparison with February exam takers, who historically performed

---

et al., *Report of the 1st Workshop on Generative AI and Law*, arXiv preprint arXiv:2311.06477 (2023).

[60] *See, e.g.*, Lydia Fontes, *Another Lawyer Faces ChatGPT Trouble* (Feb 4 2025), https://www.legalcheek.com/2025/02/another-lawyer-faces-chatgpt-trouble/; Jason Proctor, B.C. Lawyer Reprimanded For Citing Fake Cases Invented By ChatGPT, CBC NEWS (2/26/24).

[61] Benjamin Weiser, *Here's What Happens when Your Lawyer Uses ChatGPT*, N.Y. TIMES (May 27, 2023), https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html.

[62] *See id.*

[63] *See* Sara Merken, *New York Lawyers Sanctioned For Using Fake ChatGPT Cases In Legal Brief*, REUTERS, June 26, 2023; Larry Neumeister, *Lawyers Submitted Bogus Case Law Created By ChatGPT. A Judge Fined Them $5,000*, AP NEWS, June 22, 2023.

[64] OpenAI, Press Release, Hello GPT4o (2024). GPT-4o also leverages an advanced transformer architecture with improved self-attention mechanisms, enabling it to generate nuanced, contextually relevant responses. Its scalability benefits from larger training sessions, allowing it to handle up to 10 million tokens per minute.

[65] Katz et al., *supra* note 10, at 3–5.

below average.[66] But even after correcting for this detail, GPT-4's performance remained significantly above the passing threshold.

Although media discussions focused on GPT-4's bar exam performance, the more pressing empirical question was how access to tools like GPT-4 affected human attorneys' work.[67] After all, there continued to be broad consensus among lawyers, judges, and commentators that ethical and practical considerations necessitate human involvement in legal services.[68] Throughout late 2023 and 2024, several studies began to explore this issue.

---

[66] Martínez, *supra* note 10, at 1.

[67] Outside of the legal field, empirical evidence pointed in the same directions as the law-specific research: suggesting that while GPT-4 and comparable models can often enhance efficiency, their impact on the quality of human professionals' work product was decidedly mixed. A key theme in this research is the "jagged frontier" of AI's capabilities—while tools like GPT-4 enhanced human performance and efficiency in some tasks, they diminished accuracy and quality in others. *See* Fabrizio Dell'Acqua et al., *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality* 2 (Harvard Bus. Sch., Working Paper 24-013, 2023) (finding that for many tasks consultants using GPT-4 completed 12.2% more tasks, worked 25.1% faster, and produced 40% higher-quality results, but that these benefits did not extend to all tasks). In certain cases, this decline appeared to stem from over-reliance on AI, as humans exerted less mental effort when assisted by AIs, effectively allowing the AI to take the lead and causing them to "fall asleep at the wheel." *See* Fabrizio Dell'Acqua, *Falling Asleep at the Wheel: Human/AI Collaboration in a Field Experiment on HR Recruiters* 1 (Dec. 2, 2021) (unpublished manuscript) ("As AI quality increases, humans have fewer incentives to exert effort and remain attentive, allowing the AI to substitute, rather than augment their performance."). There were also numerous studies demonstrating that GPT-4 outperformed ChatGPT on various professional exams outside of law. *See* Peter Lee, Sebastien Bubeck & Joseph Petro, *Benefits, Limits, and Risks of GPT4 as an AI Chatbot for Medicine*, 388 NEW ENG. J. MED. 1233, 1238 (2023) (finding mixed results with respect to GPT-4's performance on various medical applications); Lakshmi Varanasi, *ChatGPT Can Ace the Bar, but It Only Has a Decent Chance of Passing the CFA Exams. Here's a List of Difficult Exams the ChatGPT and GPT-4 Have Passed*, BUS. INSIDER (Nov. 5, 2023), https://www.businessinsider.com/list-here-are-the-exams-chatgpt-has -passed-so-far-2023-1.

[68] *See, e.g.*, W. Bradley Wendel, *The Promise and Limitations of Artificial Intelligence in the Practice of Law*, 72 OKLA. L. REV. 21, 24-26 (2019); Nicole Yamane, *Artificial Intelligence in the Legal Field and the Indispensable Human Element Legal Ethics Demands*, 33 GEO. J. LEGAL ETHICS 877, 889-90 (2020); Cat Casey, *Why Human-Centered AI Is the Future of Legal: The Future Is More Ironman than Terminator, and That Is a Good Thing!*, LAW.COM (Jan. 30, 2023), https://www.law.com/legaltechnews/2023/01/30/why-human-centered-ai-is-the-future-of-legal-the-future-is-more-ironman-than-terminator-and-that-is-a-good-thing; *Geoffrey Vance, AI + Human: A Bright Future For Legal Co-Pilots*,

These studies suggested that GPT-4 could significantly enhance lawyers' speed and efficiency for certain legal tasks; but they supplied limited evidence that tools like GPT-4 could consistently improve the quality of the work lawyers produce. The most relevant study used a randomized controlled experiment similar to the one in this Article. It found that GPT-4 consistently increased speed across various legal tasks but had minimal and inconsistent effects on the quality of legal analysis.[69]

Another study examined whether GPT-4 could help law students perform better on exams. It found that the tool's benefits varied based on students' initial skill levels: lower-performing students experienced substantial performance gains with AI assistance, while top-performing students saw declines.[70] A third study provided GPT-4 and other legal technology tools to approximately 100 legal aid professionals. While 90% reported increased productivity and 75% planned to continue using AI, the study did not directly assess the quality of AI-assisted legal work.[71]

Empirical research also confirmed that models like GPT-4 remained vulnerable to the same types of hallucinations that had drawn widespread media attention in previous months. One high-profile study, for instance, found that general-purpose language models like GPT-4 and Meta's Llama 2 hallucinated between 58% and 82% of the time when responding to certain law-related queries.[72] Critics rightly noted that the study's queries were designed in ways that increased the likelihood of hallucinations and did not necessarily reflect how lawyers would use AI in practice. But the broader takeaway—that models like GPT-4 can easily generate incorrect legal information—reinforced skepticism among many lawyers who were already hesitant to adopt the technology.

B. The Second Wave of Gen AI and Legal Tech: Reasoning Models, RAG, and Automated Prompting

Throughout 2024 and early 2025, several key innovations in AI and legal technology have once again captured the legal community's attention, sparking a second wave of hype, skepticism, and uncertainty.

The first of these innovations is a new class of "reasoning models" that are specifically designed to tackle complex logical and analytical problems. OpenAI introduced the first such model in late 2024 with o1-

JDSUPRA (Apr. 20, 2023), https://www.jdsupra.com/legalnews/ai-human-a-bright-future-for-legal-co-7452383.

[69] *See* Choi, Monahan, & Schwarcz, *supra* note 1

[70] Choi & Schwarcz, *supra* note 11.

[71] Kim & Chien, *supra* note 7. Participants who received additional training and support reported even greater satisfaction and performance. *Id.*

[72] Dahl et al., *supra* note 14.

preview; since then, both OpenAI and other AI firms, including Google and DeepSeek, have released more advanced versions.[73]

These models mark a significant departure from earlier LLMs like ChatGPT-3.5 and GPT-4.[74] Unlike their predecessors, reasoning models allocate more compute at the time of inference, allowing them to process the prompt step-by-step in a way that earlier models could not.[75] By constructing an internal chain of reasoning, these models continuously reevaluate initial output to refine the answers they ultimately produce.[76] A large-scale reinforcement learning algorithm further enhances this ability during training, optimizing how the model evaluates and adjusts its reasoning.[77] This iterative approach enables the model to explore multiple solutions, analyze different components of a prompt separately, and strategically plan its response before finalizing its output. Figure One, sourced from OpenAI's own explanation, visually illustrates this reasoning process.[78]

[73] Wiggers, supra note 44.

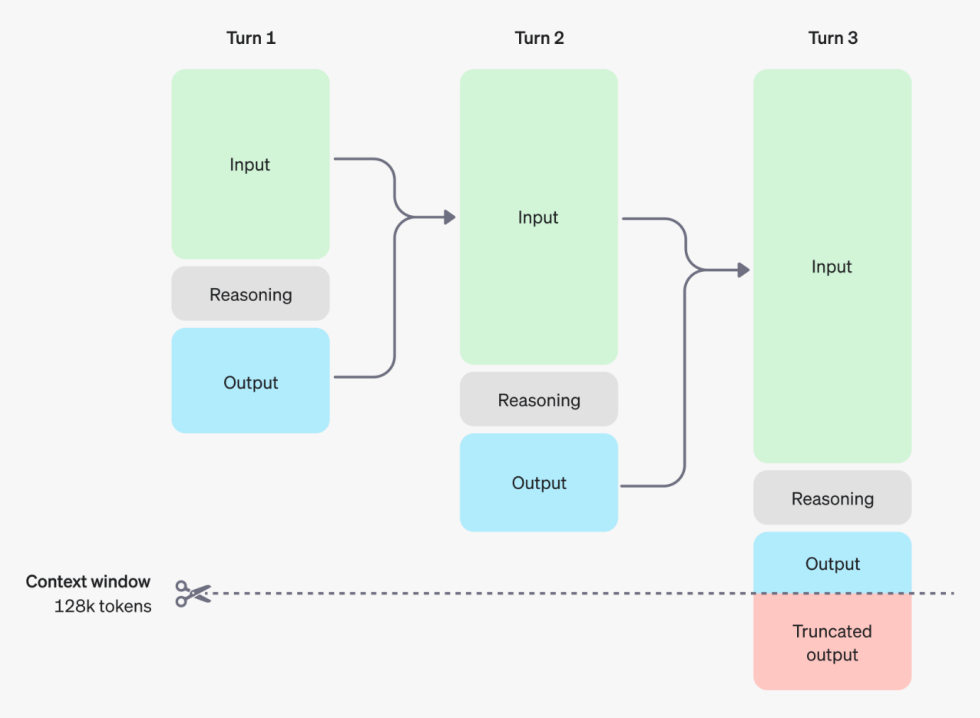
[74] *See* Press Release, OpenAI, Introducing OpenAI o1-preview (Sept. 12, 2024).

[75] *See id.*

[76] *See id.*

[77] *See id.*

[78] *See* OpenAI, *Reasoning models: Explore advanced reasoning and problem-solving models.* https://platform.openai.com/docs/guides/reasoning#how-reasoning-works



Compared to earlier models, this new class of reasoning models excels in multi-step problem-solving across domains such as mathematics, coding, and logic.[79] For example, OpenAI reports that its first reasoning model, o1, ranked in the 89th percentile on competitive programming questions, placed among the top 500 students in the USA Mathematical Olympiad qualifier, and exceeded PhD-level accuracy on a test covering physics, biology, and chemistry.[80]

Given these strengths, there has been widespread speculation that reasoning models will also excel at tackling complex legal questions, which often require problem-solving, planning, and analytical, step-by-step reasoning.[81] The problem is that no empirical evidence has been available regarding ability of these models to tackle legal questions—let alone the more critical question of how access to these models might impact human lawyers' work.

---

[79] *See id.*

[80] *See* Press Release, *OpenAI, Learning To Reason With LLMs* (Sept. 12, 2024), https://openai.com/index/learning-to-reason-with-llms/.

[81] See, e.g., OpenAI o1 Models Will Boost Legal GenAI + Agentic Flows, Artificial Lawyer, (Sept. 13, 2024), https://www.artificiallawyer.com/2024/09/13/openai-o1-models-will-boost-legal-genai-agentic-flows/; Joshua Dupuy, *Will AI Replace Lawyers? OpenAI's o1 And The Evolving Legal Landscape*, FORBES (Oct. 16, 2024).

The second major AI innovation shaping lawyering in recent years is Retrieval-Augmented Generation (RAG), which powers many of the latest AI-driven features introduced by leading legal technology companies.[82] RAG integrates LLMs with legal search engines and document retrieval systems. That combination enables AI-powered legal tools to respond to queries by retrieving relevant legal documents and generating answers based on them.[83] This approach has been widely touted for its potential to minimize—even eliminate—hallucinations.[84] Equally important, RAG enhances transparency by allowing users to verify the LLM's response based on the underlying source material on which it relied.[85]

For many legal tech companies, RAG is the primary mechanism by which they claim to deliver value beyond general-purpose models like ChatGPT. Nearly all major players—such as Westlaw, Lexis, and VLex—rely on foundational LLMs like ChatGPT to power their AI-driven legal services.[86] Their long-standing ownership of vast legal databases and advanced search tools positions them to enhance these models with RAG.[87] These companies widely tout that this technology is "hallucination-free"[88] and does "not make up facts."[89] Some legal tech

---

[82] *See* James Ju, *Retrieval-Augmented Generation In Legal Tech,* THOMPSON REUTERS (Dec. 4, 2024).

[83] *See id.*

[84] Research indicates that LLMs hallucinate far less when grounding their responses in specific source material. *See, e.g.*, Pu Xiao & Mingqi Gao, *Summarization is (Almost) Dead*, https://arxiv.org/pdf/2309.09558v1; Xiaojun Wan, Cezary Gesikowski, *AI vs Humans: Who is better at Summarizing Documents? Blind Proof of Concept Tests Reveal Clear Winner*, (Sept. 21, 2024).

[85] *See* Ju, *supra* note 82.

[86] See id.

[87] See id.

[88] LexisNexis, How Lexis+ AI Delivers Hallucination-Free Linked Legal Citations (2024), https://www.lexisnexis.com/community/insights/legal/b/product-features/posts/how-lexis-ai-delivers-hallucination-free-linked-legal-citations?srsltid=AfmBOoqfS5F5s2AWg9yaPEZi7SAlC_0FoDJ1xCdxCPUvoQcbXt6uZmRZ ("Unlike other vendors, however, Lexis+ AI delivers 100% hallucination-free linked legal citations connected to source documents, grounding those responses in authoritative resources that can be relied upon with confidence.").

[89] Press Release, Casetext, *GPT-4 alone is not a reliable legal solution—but it does enable one: CoCounsel harnesses GPT-4's power to deliver results that legal professionals can rely on (*2023), https://casetext.com/blog/cocounsel-harnesses-gpt-4s-power-to-deliver-results-that-legal-professionals-can-rely-on/. ("Unlike even the most advanced LLMs, CoCounsel does not make up facts, or 'hallucinate,' because we've implemented controls to limit CoCounsel to answering from known, reliable data sources—such as our comprehensive, up-

startups, like Harvey, take this approach further by offering to integrate law firms' proprietary databases into RAG systems, potentially allowing lawyers to more easily access and leverage their particular firm's extensive repository of past legal work.[90]

To date, however, limited empirical evidence exists on the impact of RAG-enabled legal technology on human lawyering. The most relevant study suggests that RAG-enabled AI can—and does—hallucinate.[91] While the hallucination rate for these tools was significantly lower than that of general-purpose models applied to legal queries, it still ranged from 17% to 33% in response to certain queries.[92] But the study has several key limitations that constrain its relevance to human lawyering. Most notably, it assessed only the capabilities of legal research tools in isolation, without human involvement. If human lawyers using these tools could have easily identified and corrected the hallucinations, the study's findings would be far less consequential. Additionally, the study relied on researcher-designed legal queries, many of which were explicitly crafted to induce hallucinations by embedding false premises.[93] The extent to which these queries reflect those typically made by real lawyers is questionable.

A third notable advance in AI-enabled legal technology has been the development of automated or embedded prompting.[94] In the months following the initial release of ChatGPT, users and researchers

---

to-date database of case law, statutes, regulations, and codes—or not to answer at all."). As one generative AI company operating in the legal space puts it, "many believe it's too soon for lawyers to rely on ChatGPT or GPT-4 for legal practice because they hallucinate, and because they don't access up-to-date, accurate legal data on their own." However, "it's not true that lawyers cannot trust generative AI for legal practice. It's only true that they cannot trust generative AI alone—a crucial distinction." *CoCounsel Harnesses GPT-4's Power to Deliver Results that Legal Professionals Can Rely on*, CASETEXT (May 5, 2023), https://casetext.com/blog/cocounsel-harnesses-gpt-4s-power-to-deliver-results-that-legal-professionals-can-rely-on.

[90] *See* Harvey, Professional Class AI, www.Harvey.AI.

[91] Varun Magesh et al., *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools*, https://arxiv.org/abs/2405.20362 (Working Paper, on file with Authors) (2024).

[92] See id.

[93] *See id.* This is particularly likely to result in hallucinations because LLMs do have a tendency to provide responses that affirm the user's embedded beliefs or assumptions. This is a byproduct of RLHF and similar tools, which prioritize results that humans rank as better. See Lars Malmqvist, *Sycophancy in Large Language Models: Causes and Mitigations*, (Nov. 22, 2024), arXiv:2411.15287v1.

[94] *See, e.g.,* Prompt-based Automation, Legartis, https://www.legartis.ai/prompt-based-automation

consistently discovered that certain prompting strategies led to significantly improved outputs.[95] For example, early experiments showed that instructing an AI to answer a query step-by-step often produced more accurate and reliable results.[96] Over time, these general insights into effective prompting evolved into a specialized skill known as "legal prompt engineering."[97] Such prompting might direct an AI to take on the persona of a lawyer, write in the style of a Supreme Court Justice, or emulate the voice of a renowned legal scholar. More substantively, it might instruct the AI to focus on legally pertinent issues, provide citations, apply specific legal rules, or tailor responses to particular legal documents, such as contracts or complaints.

Building on these insights, AI-enabled legal tech tools are increasingly automating the prompting process to help users ask more effective questions and improve AI-generated responses. Some tools, for example, analyze uploaded documents and generate a series of suggested questions tailored to the document type. Others present users with a menu of capabilities, each triggering a set of pre-formulated prompts. In some cases, legal technology companies embed prompts within their interfaces in ways that users do not see but that enhance outcomes. For instance, a company leveraging retrieval-augmented generation (RAG) may automatically prompt an AI model to provide citations for claims or ensure responses are based solely on relevant source material.

Unfortunately, the extent to which these automated prompting tools improve outcomes remains largely untested. One reason for skepticism is that foundation models are continuously improving at generating high-quality responses, even without specialized prompting. For instance, OpenAI advises users of its reasoning models to avoid explicitly requesting chain-of-reasoning generation, because this capability is already embedded within the model's reasoning process.[98] More broadly, LLMs are becoming increasingly adept at detecting context, which raises questions about the added value of certain prompting techniques. For example, it is unclear whether instructing an AI to respond to a legal query as an esteemed judge genuinely enhances

---

[95] See Schwarcz & Choi, supra note 27; Dils, *How to Use ChatGPT: Advanced Prompt Engineering*, WGMI Media (Jan. 31, 2023), https://wgmimedia.com/how-to-usechatgpt-advanced-prompt-engineering/;

[96] Bin Ji, et al., *Chain-of-Thought Improves Text Generation with Citations in Large Language Models*, PROCEEDINGS OF THE AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE. Vol. 38. No. 16. 2024.

[97] *See* Isabel Parker & Michal Morrison, *Introduction To Legal Prompt Engineering In Generative A*I, DELOITTE (Jan. 29, 2024); Catherine Reach, *Prompt Engineering 101 for Lawyers*, NC Bar (Aug. 21, 2024).

[98] See OpenAI, *supra* note 80.

the quality of its legal analysis compared to simply posing the question directly.[99]

II. METHODOLOGY

To gain deeper insights into the impact of emerging AI reasoning models and specialized legal AI platforms on the future of lawyering, we conducted a randomized controlled trial. The trial focused on two leading generative AI models as of late 2024. The first, o1-preview, is a general purpose AI reasoning model released by OpenAI in September 2024.[100] The second, VLex's Vincent AI, is a specialized AI tool for lawyers that uses RAG and automated prompting to facilitate the work of lawyers.[101] At the time of the study, Vincent AI used an ensemble of non-reasoning models, including GPT 4 and 4o, as its underlying foundation models.

The basic design of our study followed prior research examining AI's impact on lawyering.[102] Recruitment for the experiment began in

[99] Cf. Schwarcz & Choi, supra note 27.

[100] See OpenAI, Introducing OpenAI o1-preview (9/12/24), at https://openai.com/index/introducing-openai-o1-preview/. To help facilitate the experiment, OpenAI provided participants with free access to its Plus Accounts for the duration of the experiment. OpenAI did not otherwise provide any funding to support this project.

[101] See VLex, AI that Knows the Law, https://vlex.com/vincent-ai. Vincent AI use retrieval-augmented generation to address legal research questions by leveraging foundational generative AI models like ChatGPT to query a comprehensive range of legal source materials, including case law, statutory law, and secondary sources from all 50 states. Vincent AI also offers various workflows designed to support different legal tasks, such as answering objective legal research questions or constructing persuasive arguments. Notably, workflows like "analyze a contract" and "analyze a complaint" enable users to upload documents, providing suggested queries based on the uploaded materials to enhance the research process. To facilitate this experiment, VLex supplied University of Michigan participants with complimentary access to its Vincent AI tool for the duration of the experiment. University of Minnesota participants were able to access to the Vincent AI tool due to the law school's purchase of a law school subscription to the tool for the benefit of its students. Aside from the complimentary access to the Vincent AI tool for University of Michigan participants, Vlex did not provide any funding or other support for this project.

[102] We tested the impact of these two AI models on lawyering by following the core structure of the leading prior randomized controlled trial in the field, which focused on how well the GPT-4 model performed on various legal tasks. *See* Choi, Monahan, & Schwarcz, *supra* note 1. However, we enhanced our methodology in several important ways, incorporating lessons learned from the earlier study.

September 2024 and was led by two co-authors—a professor from the University of Minnesota Law School and a professor from the University of Michigan Law School.[103] They sent recruitment emails to all second- and third-year law students, as well as Master of Laws (LL.M.) students, at their respective institutions.[104] These emails had the subject line "U-M Research Opportunity: $300 to Experiment with AI Tools."[105] The response was substantial, with more than 250 students from the two schools expressing interest in participating. Of course, the subset of students expressing interest in participating may have been particularly enthusiastic about AI or interested in learning more about it than the overall student population that received recruiting emails.

Of these students, 153 formally enrolled in the study, and 127 successfully completed it.[106] During the enrollment process, we collected basic demographic and academic information about participants, including their law school, class year, first-year law school GPA (for second- and third-year law students), and their prior use of generative AI tools within the three months before enrollment. After participants formally enrolled in the study, they were randomly assigned to one of three evenly divided groups. Summary statistics for the participants who

[103] Because participants all affirmatively expressed interest in participating in the study, they are likely to not be perfectly reflective of the overall student body. For instance, it is possible that they may, on average, have greater familiarity than the overall student body with the use of AI tools. Table 1 does suggest, however, that there was significant variation across the participants in terms of their prior use of AI tools

[104] Both the University of Minnesota Law School and the University of Michigan Law School are top ranked law schools in the country. In 2024, The University of Minnesota Law School was ranked 16th in the country and The University of Michigan Law School was tanked 9th in the country, among approximately 200 law schools. U.S. News Ranking of Law Schools. *2023 Best Law Schools*, U.S. NEWS, https://www.usnews.com/best-graduate-schools/top-law-schools/law-rankings (last visited Aug. 5, 2023).

[105] The emails were sent at the same day and time to increase the comparability of responses.

[106] We were able to limit attrition by asking participants only to enroll if they could commit to completing the experiment; they were also informed that they would receive $300 upon full completion of the experiment but would not be compensated if they failed to complete it. Participants were advised that completing the experiment would require approximately 17 hours of work during October 2024. The informed consent, as well as the broader experimental design, was formally deemed exempt from IRB review by both the University of Minnesota IRB and the University of Michigan IRB. See IRB Exemption Determination, University of Minnesota IRB, (8/15/24).

completed the study, broken down by group assignments, are presented in Tables One and Two, below.

Table 1: Means and Proportions of Covariates by Group

Variable	Group A	Group B	Group C
GPA (Mean)	3.29	3.30	3.36
F-test p-value for GPA: 0.795			
Student Type (Proportions)			
2L Law Student	0.53	0.33	0.47
3L Law Student	0.37	0.55	0.39
LLM Student	0.10	0.12	0.14
chi-sq test p-value for Student Type: 0.304			
School (Proportions)			
University of Michigan	0.35	0.41	0.41
University of Minnesota	0.65	0.59	0.59
chi-sq test p-value for School: 0.781			
AI Use in Prev. 3 months (Proportions)			
0 times	0.17	0.19	0.20
1–5 times	0.35	0.38	0.44
6–10 times	0.23	0.27	0.13
More than 20 times	0.25	0.17	0.22
chi-sq test p-value for AI Use: 0.73			

Table 2: Counts by Group for Covariates

School	**University of Michigan**	**University of Minnesota**
A	18	33
B	21	30
C	21	30
Student Type		
	I'm currently a 2L law student	**I'm currently a 3L law student**
A	27	19
B	17	28
C	24	20
	I'm currently an LLM student	
A	5	
B	6	
C	7	
AI Use		
	0 times	**1-5 times**
A	8	17
B	9	18
C	9	20
	6-10 times	**More than 20 times**
A	11	12
B	13	8
C	6	10
GPA Bins		
	[0.0, 2.5)	**[2.5, 3.0)**
A	1	6
B	0	4
C	0	7
	[3.0, 3.5)	**[3.5, 4.0)**
A	27	10
B	28	12
C	19	17

As suggested in these tables, the three randomly constructed groups were approximately balanced in terms of participants' law school affiliation, year in law school, and first-year GPA.

Study participants completed the experiment remotely from October 1, 2024 to October 31, 2024, using a Canvas interface.[107] The study began with three online training modules that all participants in

[107] Participants were provided with suggested deadlines during October to encourage steady progress. However, they were allowed to complete the tasks at their own pace, provided all elements were finished by the study deadline of October 31, 2024. Among the 127 successful participants, 19 received extensions of between 1 day and 2 weeks to complete the experiment, depending on individual circumstances.

the study completed.[108] These were developed and delivered by a co-author, a representative from Vincent AI, and a research librarian.[109] Each module included a 20- to 30-minute video, with two modules also incorporating short exercises.[110] The first module focused on the use of general-purpose AI tools for legal research, highlighting the risks of AI "hallucinations" and the dangers of over-reliance on AI at the expense of independent legal reasoning.[111] Participants were encouraged to use AI as an aid to enhance their work rather than as a substitute for their own judgment. The second and third modules provided tailored instruction on Vincent AI, covering its various tools and workflows and offering guidance on distinguishing between AI-generated text and content from primary sources.

After completing the training, participants were tasked with six lawyering assignments, each accompanied by specific instructions on the use of generative AI. For instance, in Task One, participants in Group A were prohibited from using generative AI, Group B was required to use o1-preview, and Group C was required to use Vincent AI.[112] These instructions varied systematically across groups and tasks, ensuring that each participant completed two tasks without AI, two tasks using o1-

[108] All study participants used Vincent AI to compete two assignments and o1-preview to complete two assignments, as well as completing two assignments without AI. As such, the training was relevant to all participants irrespective of which group they were assigned to.

[109] The general AI training module was led by Daniel Schwarcz, and drew heavily on the co-author's prior work regarding the techniques for using AI in legal research and writing. *See* Schwarcz & Choi, *supra* note 27.The second training, which was focused on Vincent AI, was led by Damien Riehl, vice president and solutions champion at VLex, the company that designed Vincent AI. Riehl has extensive experience explaining and teaching lawyers about the use of Vincent AI. The final module was led by Andrew Martineau, who leads the University of Minnesota Law Library's instructional program, which includes coordinating research instruction for first-year law students, teaching a practice-ready legal research course, and giving presentations on specialized legal research topics in upper-division seminars.

[110] Two of the modules also required participants to complete brief training exercises related to the video content.

[111] *See* Schwarcz & Choi, *supra* note 27.

[112] Group assignment was stable across the entire experiment; we did not randomize group assignment on each task. This approach ensured that each participant completed two tasks without AI assistance, two tasks with the assistance of GPT o1 preview, and two assignments with the assistance of Vincent AI. This structure makes it especially important that assignment to the three groups was effectively randomized.

preview, and two tasks using Vincent AI.[113] The deliberately equal division enabled a balanced evaluation of performance under different conditions.[114] Participants completed the assignments in the same order to avoid confounding the treatment effect of AI assistance with ordering effects.[115]

All six assignments were developed in collaboration with at least one co-author and a practicing attorney to ensure they reflected realistic scenarios typically assigned to first- or second-year law firm associates. These assignments, along with their respective time limits, were as follows:

1. **Assignment One**: Draft an email for a client (60-minute time limit).[116]

---

[113] We cannot eliminate the possibility that participants exerted more or less effort when completing tasks without the assistance of AI then when completing tasks with the assistance of AI. This could occur if, for instance, participants expected that they would perform better when provided with AI assistance, and adjusted their effort levels as a result, perhaps even subconsciously.

[114] Participants in the No AI treatment were instructed "You may use other online resources, including traditional (non-AI based) Westlaw or Lexis tools to help craft your answer. Please make sure not to use any AI powered tools, including Google AI, to complete this task (regular Google searches are permitted)." Participants in the o1-preview treatment were instructed: "You must use o1-preview to ASSIST you with producing an answer. You are also permitted to access other non AI online resources (like conventional Westlaw or Lexis) to the extent you feel doing so is necessary or desirable. You may NOT use Vincent AI. Remember to do your best to ensure that your answer you submit does not APPEAR to be drafted by an AI." Finally, participants in the Vincent AI treatment were instructed "You must use VINCENT AI to ASSIST you with drafting your answer. You are also permitted to access other non AI online resources (like conventional Westlaw or Lexis) to the extent you feel doing so is necessary or desirable. You may not use o1-preview or any AI tools other than Vincent AI. Remember to do your best to ensure that the answer you submit does not APPEAR to be drafted by an AI."

[115] It is possible that the order in which assignments were completed interacted with the treatments. It is conceivable, for instance, that AI tools were more useful for later tasks when participants were rushing to complete the experiment. We find little evidence of this effect in the data, however.

[116] This assignment required drafting a concise, research-backed email to a client explaining why a defamation claim cannot be based on statements made solely within litigation. The assignment specified that the email should reference authoritative case law or statutes— particularly within the jurisdiction of the Tenth Circuit—to provide a strong foundation for the explanation. The assignment had a word limit of 700 words.

2. **Assignment Two**: Draft a legal memo for a partner (240-minute time limit).[117]
3. **Assignment Three**: Analyze a complaint and draft a written analysis (120-minute time limit).[118]
4. **Assignment Four**: Draft a non-disclosure agreement (NDA) for a client (180-minute time limit).[119]
5. **Assignment Five**: Draft a motion to consolidate (150-minute time limit).[120]
6. **Assignment Six**: Draft a persuasive letter addressing the enforceability of a covenant not to compete (150-minute time limit).[121]

---

[117] This assignment required drafting an objective legal research memo analyzing whether, under Massachusetts and New Hampshire insurance laws, an insurer is obligated to pay $200,000 in attorneys' fees in addition to a $2 million liability coverage limit. The assignment provided participants with the relevant insurance policy language. It also instructed them that they should (1) focus solely on U.S. law, (2) disregard Ontario or Canadian law, (3) consider relevant case law and (4) distinguish precedents where applicable. The assignment had a word limit of 1,500 words.

[118] This assignment involved drafting a concise memo summarizing the key allegations and claims in a class action complaint, assessing the strength of these claims, and outlining potential defense strategies. The assignment included the complaint, which was taken from a real case filed in federal court but never resolved. The assignment had a word limit of 1,000 words.

[119] This assignment required drafting a concise, enforceable nondisclosure agreement (NDA) for a company that protected its proprietary trade secrets while complying with the legal limitations in Minnesota and neighboring states. It specified that the NDA should be (1) written in plain English, (2) favorable to the company, and (3) formatted to be no more than three single-spaced pages. It also included an overbroad sample NDA that participants were instructed to use as a starting point for drafting.

[120] This assignment involved drafting a persuasive brief in support of a motion to consolidate two cases in Minnesota state court that share overlapping facts and issues. The assignment specified that the brief should advocate for consolidation to ensure efficiency and consistency and that it should cite Minnesota case law and civil procedure rules. It included a word limit 1,000 words.

[121] This assignment required participants to draft a persuasive letter arguing that a covenant not to compete signed by a restaurant's former chef was reasonable and enforceable under Indiana law. It specified that the letter should focus on demonstrating that the scope, geographic restrictions, and duration of the covenant are justified to protect the restaurant's legitimate business interests, including its proprietary recipes, client relationships, and specialized training. It included a word limit of 1,250 words.

Four of the six assignments (Assignments one, two, five and six) were designed to focus on research-oriented tasks for which retrieval-augmented generation using legal source materials was expected to be especially beneficial. However, we also strove to vary the complexity of the assignments, the extent to which they were litigation or transaction oriented, and the extent to which they required an objective or persuasive analysis.[122]

In addition to completing the six assignments, participants were required to report the amount of time they spent on each task. To encourage participants to complete the assigned work efficiently and effectively, we instructed them as follows:

> As with all assignments completed in connection with this experiment, you should approach the assignment as if you are a junior attorney who has been asked to produce work for a fee-sensitive client. While you can take up to the maximum time allotment to complete the task, you should stop working at the point where you would feel comfortable submitting your work product to a supervising attorney, given that your client would prefer to minimize the amount they pay for your work product. If you reach the end of the maximum time allocation and have not finished, you should simply turn in the work product you were able to produce within the allotted time. Do not spend any more than the maximum time on any assignment. **As a reminder, your study compensation is not based on the actual time spent completing these assignments.** Timekeeping is only used to gather data on the efficiency of both methods of completion.[123]

After participants completed the assigned tasks, their work product was evaluated by three of the co-authors. The grading was conducted anonymously, with the graders unaware of the participants'

[122] Section A of the Appendix contains each assignment in full.

[123] Although we designed these instructions to replicate real-world conditions facing lawyers, there are of course differences between the incentives our participants faced and those facing practicing lawyers. For instance, many lawyers have an incentive to spend more time than necessary to complete tasks so as to maximize billable hours. So the speed-related benefits we find may not always translate well to practicing lawyers being paid on an hourly basis.

identities, GPAs, use of AI, or time spent on each task. Each of the three grading co-authors graded two of the assignments that aligned most closely with their expertise. To ensure anonymity in the grading process, the three co-authors responsible for grading were different from the co-authors who coordinated the experiment and handled the data.

Before grading the assignments, the co-authors developed a general grading rubric centered on five core criteria:

1. **Accuracy**: The precision and usefulness of the research.
2. **Analysis**: The depth and insightfulness of the analysis.
3. **Organization**: The clarity and structure of the work product.
4. **Clarity**: The quality and persuasiveness of the writing.
5. **Professionalism**: The extent to which directions were followed effectively.

Each co-author responsible for grading then adapted this general rubric to create a tailored rubric for their specific assignment, as detailed in the Appendix. Additionally, each rubric included a separate binary metric to flag whether any sources cited in the assignment appeared to be hallucinated, either because the sources were non-existent or because their descriptions were entirely inaccurate.

To evaluate treatment effects, we used ordinary least squares (OLS) regression with two treatment indicator variables. Our base specification can be written as:

$$Y_i = \beta_0 + \beta_1 \text{Vincent}_i + \beta_2 \text{o1-preview}_i + \epsilon_i \quad (1)$$

where Y_i represents our outcome measures for participant i (scores on overall quality and on each individual quality criteria, time spent, or productivity), Vincent$_i$ and o1-preview$_i$ are dummy variables equal to 1 if the participant was assigned to use Vincent AI or o1-Preview respectively, and 0 otherwise. The omitted category is the no-AI control group, so β_1 and β_2 represent the average treatment effects of Vincent and o1-preview relative to completing tasks without any AI assistance. We initially ran this specification without controls, relying on the randomized assignment of participants to treatment conditions to limit confounding factors and because post-survey participation (in which some of the data on the control variables was collected) was optional.

As a robustness check, we expanded the specification to include additional controls for participants who provided the data:

$$Y_i = \beta_0 + \beta_1 \text{Vincent}_i + \beta_2 \text{o1-preview}_i + \gamma X_i + \epsilon_i \quad (2)$$

where *Xi* is a vector of control variables including GPA, indicators for law school class year (2L, 3L, LLM), and self-reported prior AI use. The results were largely unchanged with these controls included, as detailed in in the Appendix. We use heteroskedasticity-robust standard errors throughout our analysis.

At the conclusion of the experiment, all participants were invited to complete a post-experiment survey about their experiences. At the time of the survey, participants had not yet received grades or feedback on their submitted work. Of the 127 participants, 114 completed the survey, which asked them to evaluate o1-preview and Vincent AI across several dimensions. The survey focused on the perceived impact of the two AI tools on the quality, speed, and personal satisfaction of participants' work, as well as whether their ability to use the tools effectively improved over the course of the experiment. Participants were also asked how their experiences influenced their anticipated future use of similar tools for legal work.[124]

We pre-registered our methods and hypotheses prior to analyzing our results; the pre-analysis plan is archived with the American Economic Association's registry for randomized controlled trials.[125]

III. RESULTS

Our most significant finding is that access to both o1-preview and Vincent AI led to statistically significant and meaningful improvements in overall quality of work across four of the six assignments tested—with o1-preview producing larger and more statistically significant gains than Vincent AI. For both AI tools, these improvements were primarily reflected in enhanced clarity, organization, and professionalism of submitted work. Notably, o1-preview also significantly improved the logic and nuance of the legal argumentation in three of the six assignments. In contrast, we found mixed evidence that either tool improved accuracy, with only one exception: o1-preview did improve accuracy when the assigned task required participants to focus their analysis on a single document (a complaint) with which they were supplied as part of the assignment. To our knowledge, this data is the

[124] *See* Part III.C, *infra.*

[125] *See* The Impact of Specialized AI tools for Lawyering Tasks, AEARCTR-0014957 (December 20, 2024), at https://www.socialscienceregistry.org/trials/14957. *See generally* Jason M. Chin & Kathryn Zeiler, *Replicability in Empirical Legal Research*, 17 ANN. REV. L. & SOC. SCI. 239, 243 (2021) (discussing the benefits of pre-registering a data collection and analysis plan in the context of empirical legal research).

first evidence that providing humans with AI access can enhance the quality of written work involving complex legal reasoning.

Shifting from quality to speed and productivity, we found that participants generally completed legal tasks more quickly when using both o1-preview and Vincent AI than when working without AI. However, the magnitude and variability of these speed gains were comparable to those observed with GPT-4 alone in a prior study. By contrast, productivity gains—measured by changes in overall quality score points per minute of work—appeared to be greater for both o1-preview and Vincent AI than for GPT-4 alone. We find that Vincent yields statistically significant productivity boosts of approximately 38% to 115% and o1-preview increased productivity by between roughly 34% and 140%, with particularly strong effects in complex tasks like drafting persuasive letters and analyzing complaints. These core findings on quality, speed, and productivity are detailed in Section A. Section B then analyzes results by participant sub-type, to evaluate whether our results vary by participant characteristics. Finally, Section C reviews the post-study survey results, which reveal strong enthusiasm for both o1-preview and Vincent AI.

A. Quality, Speed and Productivity

1. Quality Results

Access to both o1-preview and Vincent AI led to statistically significant improvements in the overall quality of legal work product in four of the six assignments tested. Table 3, below, presents those results.

Table 3: Treatment Effects on Total Score Across Tasks

Task	Control Mean	Model	Effect	SE	% Change	N
Draft Client Email	19.341	Vincent	2.926*	(1.580)	+15.1%	136
		o1-preview	1.829	(1.435)	+9.5%	136
Draft Legal Memo	16.909	Vincent	2.277*	(1.212)	+13.5%	126
		o1-preview	3.988***	(1.194)	+23.6%	126
Analysis of Complaint	24.400	Vincent	1.941*	(1.098)	+8.0%	127
		o1-preview	2.484**	(1.207)	+10.2%	127
Draft NDA	26.395	Vincent	0.066	(0.824)	+0.3%	127
		o1-preview	−1.106	(0.862)	−4.2%	127
Draft Motion to Consolidate	17.489	Vincent	2.093*	(1.244)	+12.0%	127
		o1-preview	4.921***	(1.079)	+28.1%	127
Draft Persuasive Letter	19.564	Vincent	−1.746	(1.697)	−8.9%	126
		o1-preview	4.087***	(1.560)	+20.9%	126

Notes: Effects shown as absolute increase relative to No AI control group. Total score is calculated by summing the grades for each of the fiver quality critiera. So the possible range of total scores across tasks is 5-35. Percent changes are calculated relative to control group mean. Robust standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Sample size (N) represents the number of observations used in the regression.

As shown in Table 3, the improvements in quality varied in magnitude, statistical significance, and task type for both Vincent AI and o1-preview. Notably, o1-preview yielded more statistically significant

quality improvements than Vincent AI, while also demonstrating a greater average effect size.[126] Both tools improved the quality of three common assignments—the legal memo, complaint analysis, and motion to consolidate—as well as one additional assignment specific to each tool: the persuasive letter for o1-preview and the client email for Vincent.[127] These results are a notable contrast to a previous randomized controlled trial of GPT-4, which found no statistically significant improvements in overall quality across all four of the assignments tested in that study.[128]

To explore how the two AI tools impacted participants' performance across the six different assignment types, Figures 1 through 6 illustrate the distribution of scores for each of these assignments. These Figures are density plots, meaning they present the share of participants (on the y-axis) who received each score (on the x-axis).[129] In each figure, the scores of the three different groups are separately represented: those who completed each assignment without AI assistance, those assisted by o1-preview, and those assisted by Vincent AI.

[126] For o1-preview, the quality improvements were significant at the 1% level for three assignments, and for the fourth, at the 5% level. In contrast, the quality improvements linked to Vincent AI were only significant at the 10% level across the four assignments with a statistically significant effect. o1-preview's statistically significant improvements in scores ranged from approximately 10% to 28% (10.2%, 20.9%, 23.6%, and 28.1%), while those for Vincent AI ranged from 8% to 15% (8%, 12%, 13.5%, and 15.1%)

[127] Both o1-preview and Vincent produced quality benefits on three legal memo, complaint analysis, and motion to consolidate assignments. However, only Vincent improved the client email assignment, while only o1-preview improved the persuasive letter.

[128] *Cf.* Choi, Monahan, & Schwarcz, *supra* note 1.

[129] *See generally* Adriano Z. Zambom & Ronaldo Dias, *A Review of Kernel Density Estimation with Applications to Econometrics*, 5 INT'L ECONOMETRIC REV. 20, 29–33 (2013).

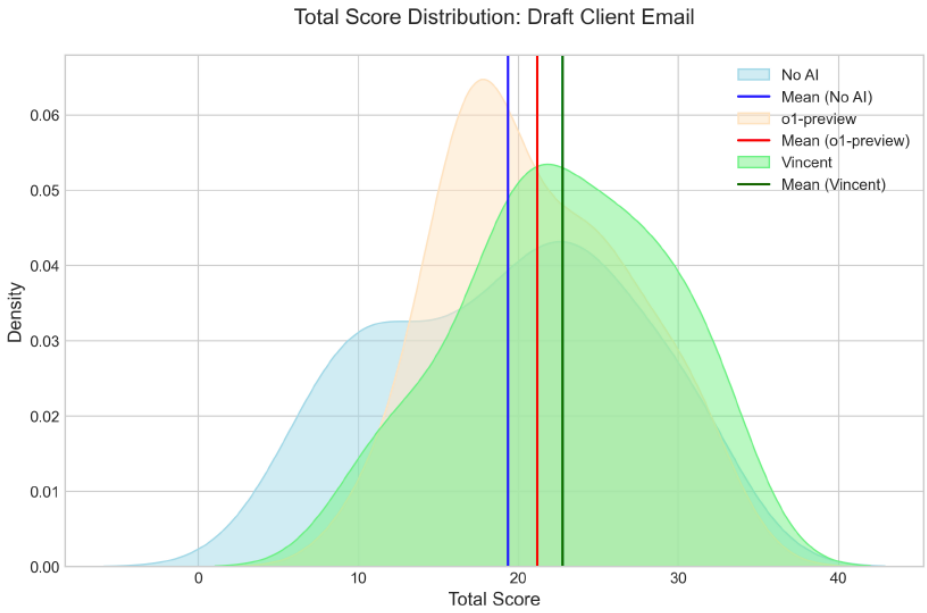


Figure 1: Task 1 Score Density

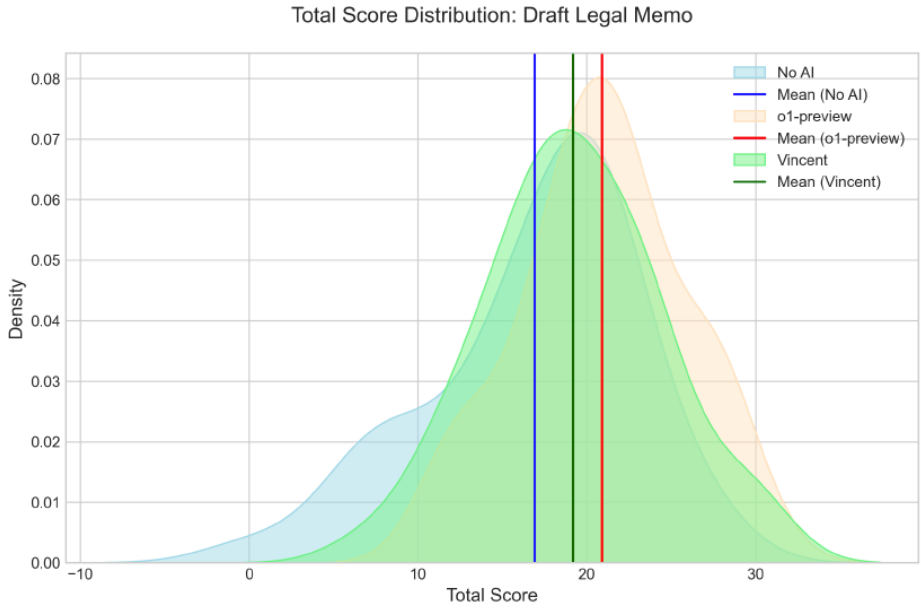


Figure 2: Task 2 Score Density

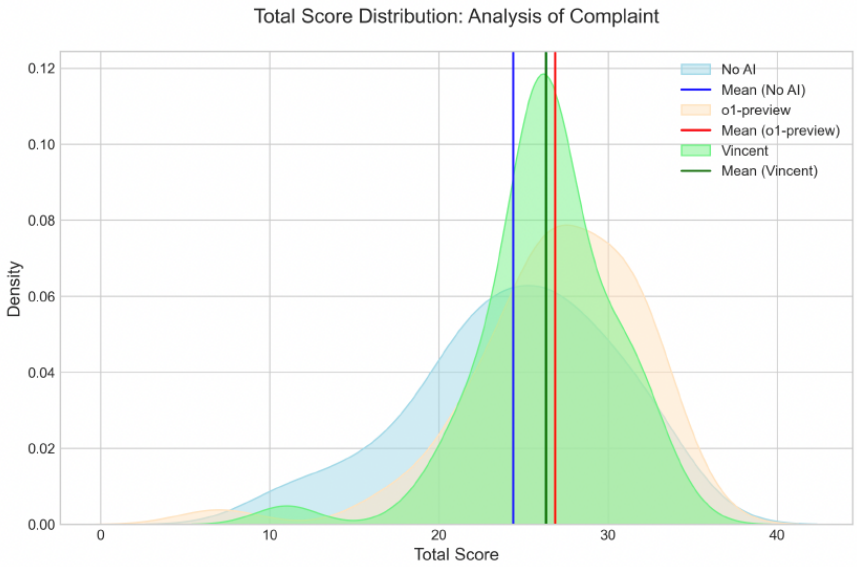


Figure 3: Task 3 Score Density

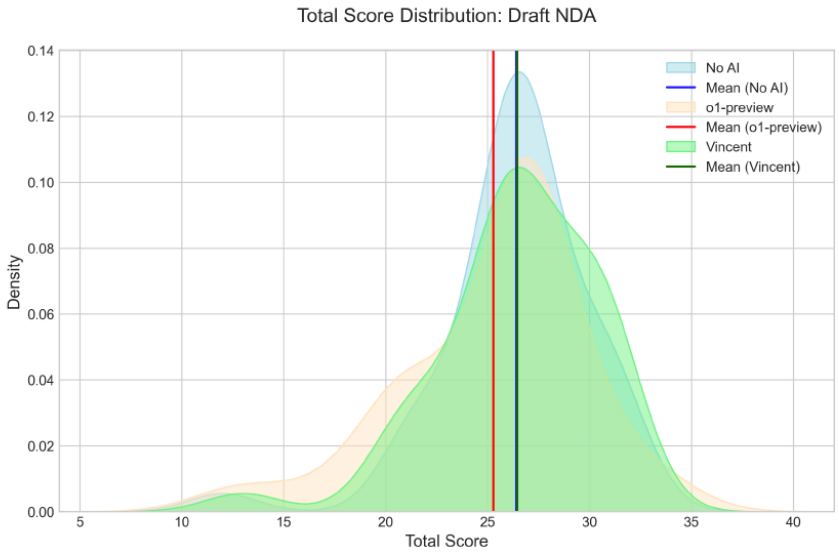


Figure 4: Task 4 Score Density

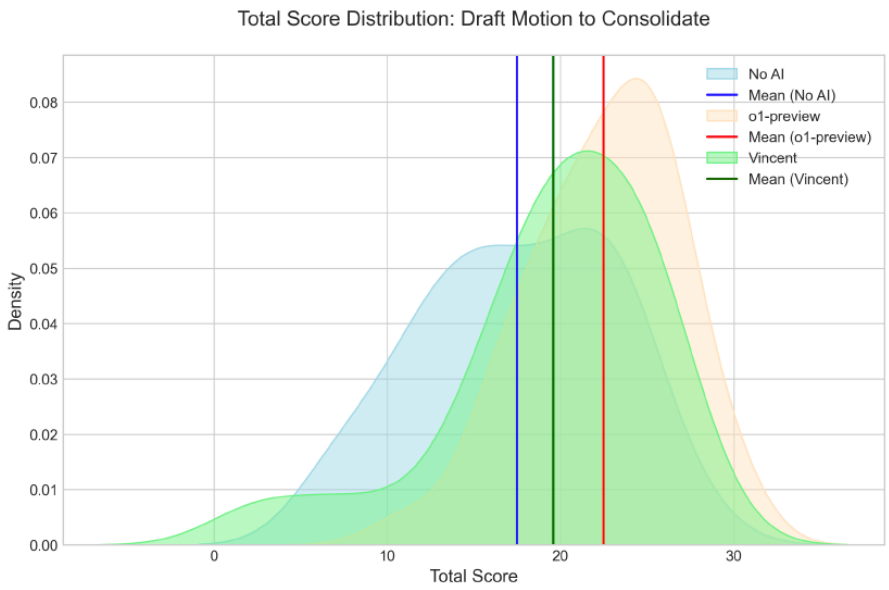


Figure 5: Task 5 Score Density

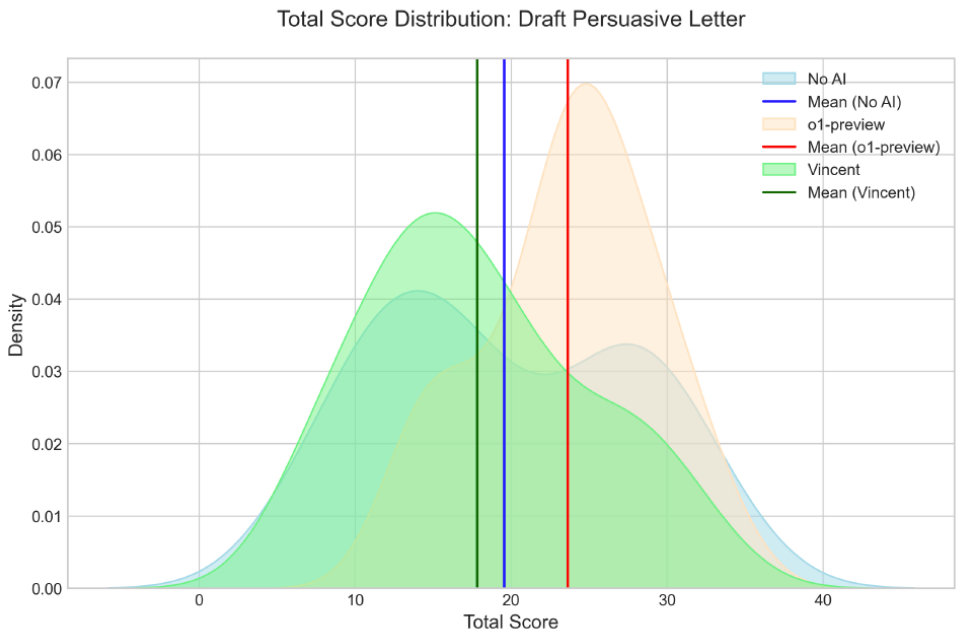


Figure 6: Task 6 Score Density

As shown in Table 3 and Figures 1 to 6, the only assignment on which neither AI tool led to a statistically significant improvement in quality was the Draft NDA, a result which is illustrated in Figure Four. This outcome may be explained by two key differences between the Draft NDA and the other five assignments. Unlike the other assignments, which were litigation-focused, the Draft NDA assignment was transactionally oriented. Additionally, it was the only assignment on which participants were provided with a general template to use in producing their response—a common practice in transactional work, but not in the litigation-related tasks analyzed.[130] These factors may have reduced the potential for AI-driven quality improvements in this assignment—particularly for Vincent AI, which is marketed more towards litigators than transactional attorneys.[131]

We also analyzed how access to the two AI technologies influenced five quality-related metrics—accuracy, analysis, organization, clarity, and professionalism—which we aggregated to produce the overall quality

[130] *See* Carol Goforth, *Transactional Skills Training Across the Curriculum*, 66 J. LEGAL ED. 904 (2017).

[131] *See* VLex, *supra* note 101.

scores reported above.[132] Tables 4–8 below report these results for each of the six assignments.[133] Each quality criteria is graded on a 1-7 scale.

Table 4: Treatment Effects on Clarity Across Tasks

Task	Control Mean	Model	Effect	SE	% Change	N
Draft Client Email	4.114	Vincent	1.068***	(0.281)	+26.0%	135
		o1-preview	1.078***	(0.271)	+26.2%	135
Draft Legal Memo	3.273	Vincent	0.774***	(0.259)	+23.6%	126
		o1-preview	0.932***	(0.268)	+28.5%	126
Analysis of Complaint	5.050	Vincent	0.155	(0.232)	+3.1%	127
		o1-preview	0.183	(0.244)	+3.6%	127
Draft NDA	5.140	Vincent	0.066	(0.162)	+1.3%	127
		o1-preview	−0.206	(0.191)	−4.0%	127
Draft Motion to Consolidate	3.422	Vincent	0.461**	(0.189)	+13.5%	126
		o1-preview	0.815***	(0.183)	+23.8%	126
Draft Persuasive Letter	4.436	Vincent	0.178	(0.299)	+4.0%	126
		o1-preview	0.750***	(0.281)	+16.9%	126

Notes: Effects shown as absolute increase relative to No AI control group. Percent changes calculated relative to control group mean. Robust standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Sample size (N) represents the number of observations used in the regression.

Table 5: Treatment Effects on Accuracy Across Tasks

Task	Control Mean	Model	Effect	SE	% Change	N
Draft Client Email	3.250	Vincent	0.432	(0.359)	+13.3%	135
		o1-preview	−0.399	(0.354)	−12.3%	135
Draft Legal Memo	2.955	Vincent	0.301	(0.269)	+10.2%	126
		o1-preview	0.097	(0.278)	+3.3%	126
Analysis of Complaint	5.100	Vincent	0.264	(0.246)	+5.2%	127
		o1-preview	0.458*	(0.264)	+9.0%	127
Draft NDA	5.698	Vincent	−0.082	(0.183)	−1.4%	127
		o1-preview	−0.187	(0.181)	−3.3%	127
Draft Motion to Consolidate	3.778	Vincent	0.083	(0.302)	+2.2%	127
		o1-preview	0.402	(0.275)	+10.6%	127
Draft Persuasive Letter	3.436	Vincent	−0.754*	(0.396)	−21.9%	126
		o1-preview	0.308	(0.385)	+9.0%	126

Notes: Effects shown as absolute increase relative to No AI control group. Percent changes calculated relative to control group mean. Robust standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Sample size (N) represents the number of observations used in the regression.

[132] *See* Part II, *supra*, for additional information on how these metrics were identified and evaluated.

[133] Tables 11-16 in Part B of the Appendix provide the same information, organized by assignment rather than by quality-related metric.

Table 6: Treatment Effects on Analysis Across Tasks

Task	Control Mean	Model	Effect	SE	% Change	N
Draft Client Email	3.636	Vincent	0.523	(0.322)	+14.4%	135
		o1-preview	−0.211	(0.322)	−5.8%	135
Draft Legal Memo	3.091	Vincent	0.374	(0.253)	+12.1%	126
		o1-preview	0.524**	(0.248)	+17.0%	126
Analysis of Complaint	4.625	Vincent	0.330	(0.261)	+7.1%	127
		o1-preview	0.445	(0.296)	+9.6%	127
Draft NDA	4.930	Vincent	−0.135	(0.192)	−2.7%	127
		o1-preview	−0.152	(0.201)	−3.1%	127
Draft Motion to Consolidate	3.422	Vincent	0.299	(0.268)	+8.7%	127
		o1-preview	0.834***	(0.257)	+24.4%	127
Draft Persuasive Letter	3.462	Vincent	−0.348	(0.379)	−10.1%	126
		o1-preview	0.841**	(0.350)	+24.3%	126

Notes: Effects shown as absolute increase relative to No AI control group. Percent changes calculated relative to control group mean. Robust standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Sample size (N) represents the number of observations used in the regression.

Table 7: Treatment Effects on Organization Across Tasks

Task	Control Mean	Model	Effect	SE	% Change	N
Draft Client Email	4.045	Vincent	0.568*	(0.314)	+14.0%	135
		o1-preview	0.295	(0.297)	+7.3%	135
Draft Legal Memo	4.045	Vincent	0.164	(0.344)	+4.1%	126
		o1-preview	0.980***	(0.315)	+24.2%	126
Analysis of Complaint	4.800	Vincent	0.473*	(0.252)	+9.8%	127
		o1-preview	0.526**	(0.259)	+10.9%	127
Draft NDA	5.047	Vincent	0.030	(0.194)	+0.6%	127
		o1-preview	−0.247	(0.176)	−4.9%	127
Draft Motion to Consolidate	3.467	Vincent	0.766**	(0.310)	+22.1%	127
		o1-preview	1.482***	(0.251)	+42.8%	127
Draft Persuasive Letter	4.000	Vincent	−0.523	(0.401)	−13.1%	126
		o1-preview	0.977**	(0.380)	+24.4%	126

Notes: Effects shown as absolute increase relative to No AI control group. Percent changes calculated relative to control group mean. Robust standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Sample size (N) represents the number of observations used in the regression.

Table 8: Treatment Effects on Professionalism Across Tasks

Task	Control Mean	Model	Effect	SE	% Change	N
Draft Client Email	4.295	Vincent	0.841**	(0.393)	+19.6%	135
		o1-preview	1.066***	(0.377)	+24.8%	135
Draft Legal Memo	3.545	Vincent	0.664**	(0.327)	+18.7%	126
		o1-preview	1.455***	(0.318)	+41.0%	126
Analysis of Complaint	4.825	Vincent	0.720***	(0.270)	+14.9%	127
		o1-preview	0.873***	(0.287)	+18.1%	127
Draft NDA	5.581	Vincent	0.188	(0.222)	+3.4%	127
		o1-preview	−0.315	(0.249)	−5.6%	127
Draft Motion to Consolidate	3.400	Vincent	0.484	(0.347)	+14.2%	127
		o1-preview	1.497***	(0.300)	+44.0%	127
Draft Persuasive Letter	4.231	Vincent	−0.299	(0.438)	−7.1%	126
		o1-preview	1.211***	(0.395)	+28.6%	126

Notes: Effects shown as absolute increase relative to No AI control group. Percent changes calculated relative to control group mean. Robust standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Sample size (N) represents the number of observations used in the regression.

Several key findings emerge from these tables. First, both AI tools significantly enhanced the clarity, organization, and professionalism of submitted work across multiple assignments.[134] However, o1-preview consistently outperformed Vincent AI in terms of the frequency, magnitude, and statistical significance of these improvements.[135]

Second, as shown in Table 6, only o1-preview produced statistically significant improvements in respondents' legal analysis, doing so for three of the six assignments: the legal memo, the motion to consolidate, and the persuasive letter. By contrast, Vincent AI did not yield statistically significant improvements in legal analysis for any of the six assignments. This finding is particularly noteworthy because the

[134] For clarity, both tools led to statistically significant improvements in three assignments—the draft email, legal memo, and motion to consolidate—with o1-preview also enhancing clarity in the persuasive letter. Similarly, both tools improved organization in two assignments—the complaint analysis and motion to consolidate. Additionally, Vincent AI enhanced organization in the client email, while o1-preview did so for the legal memo and persuasive letter. Both tools also contributed to statistically significant improvements in professionalism across three assignments—the draft email, legal memo, and complaint analysis. o1-preview further improved professionalism in the motion to consolidate and persuasive letter.

[135] For clarity, o1-preview (26.2% and 28.5%) and Vincent AI (26% and 23.6%) produced similar improvements for the client email and legal memo. However, o1-preview (23.8% and 16.9%) yielded substantially greater clarity gains for the motion to consolidate and the persuasive letter compared to Vincent AI (4% and 13.5%). Similarly, in terms of organization, o1-preview outperformed Vincent AI in every instance where a statistically significant effect was observed, except for the client email, where o1-preview had no significant impact. Finally, for professionalism, o1-preview consistently produced greater improvements across all assignments with statistically significant effects.

nuance and depth of legal analysis is arguably the most critical factor in assessing the quality of legal work, even though our score aggregation approach did not weight it differently.[136]

Third, neither AI tool consistently led to statistically significant improvements in the accuracy of respondents' assignments. The only instance of a statistically significant accuracy gain occurred in the complaint-drafting task, where access to o1-preview resulted in improvement—likely because participants were provided with the key document (the complaint) as part of the assignment. In contrast, Vincent AI resulted in a decrease in accuracy for the persuasive-letter assignment. This finding is surprising, given that RAG is designed to enhance the accuracy of legal research, such as that required for this task.[137] However, because this effect was only statistically significant at the 10% level and observed in a single assignment, its broader implications remain limited.

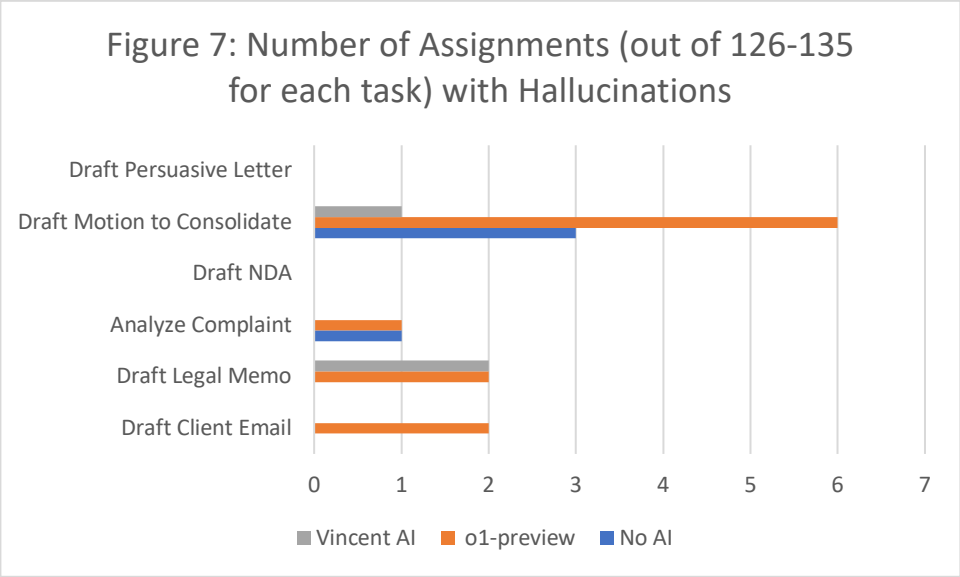
In addition to evaluating assignments based on the five quality-related metrics, we separately tracked instances of hallucinations, which we defined as citations to entirely fabricated sources.[138] The results, presented in Figure 7, indicate that while hallucinations were rare, they did occur. Although the small sample size of reported hallucinations limits our ability to draw definitive conclusions about their comparative likelihood, the data suggest that RAG technology, such as that used by Vincent AI, does indeed reduce hallucinations. In fact, we identified fewer hallucinations in assignments completed with Vincent AI (3 total) than in those completed without any AI assistance at all (4). By contrast, assignments completed with o1-preview exhibited a substantially higher number of hallucinations (11).

---

[136] While clarity, organization, and professionalism can often be improved with additional time, effort, and attention, the quality of legal analysis is much harder to enhance. It is often what separates highly skilled attorneys from less proficient ones.

[137] *See* Part I, *supra*.

[138] Although certain graders noted instances in which participants mis-cited cases (mixing up jurisdictions, for instance), or cited cased that did not appear to stand for the relevant proposition, we did not count these as hallucinations.



2. Speed Results

Access to both AI tools also resulted in statistically significant and meaningfully large improvements in the speed with which participants completed five of the six tasks. This result is reflected in Table 9.

Table 9: Treatment Effects on Time Spent Across Tasks (Minutes)

Task	Control Mean	Model	Effect	SE	% Change	N
Draft Client Email	50.302	Vincent	−7.147***	(2.544)	−14.2%	135
		o1-preview	−6.111***	(2.277)	−12.1%	135
Draft Legal Memo	183.909	Vincent	−30.374***	(11.302)	−16.5%	127
		o1-preview	−25.984**	(12.097)	−14.1%	127
Analysis of Complaint	107.103	Vincent	−39.512***	(5.226)	−36.9%	126
		o1-preview	−29.917***	(5.715)	−27.9%	126
Draft NDA	96.302	Vincent	−5.456	(10.590)	−5.7%	127
		o1-preview	−13.302	(9.297)	−13.8%	127
Draft Motion to Consolidate	95.800	Vincent	−17.405**	(8.135)	−18.2%	128
		o1-preview	−17.500**	(8.476)	−18.3%	128
Draft Persuasive Letter	110.950	Vincent	−38.677***	(8.128)	−34.9%	127
		o1-preview	−29.136***	(7.710)	−26.3%	127

Notes: Effects shown as absolute increase relative to No AI control group. Percent changes calculated relative to control group mean. Robust standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Sample size (N) represents the number of observations used in the regression.

Table 9 also indicates that the magnitude of speed improvements associated with Vincent AI and o1-preview was generally comparable, with some variation across assignments. Specifically, both tools produced nearly identical reductions in completion time for three assignments, while o1-preview led to greater speed gains in one assignment, and Vincent AI outperformed o1-preview in two assignments. On the legal

tasks where speed improvements were identified, they ranged from 14-37% for Vincent and 12-28% for o1-preview.

Figures 7 through 12, which are density plots with the number of participants (on the y-axis) and time spent (on the x-axis), provide additional information about the distribution of time spent to complete each of the six assignments.

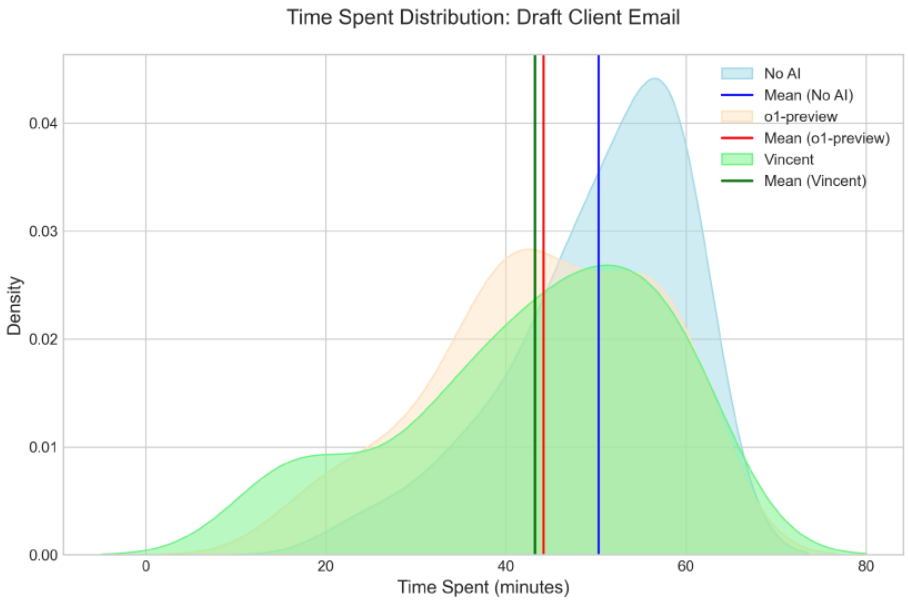


Figure 7: Task 1 Time Density

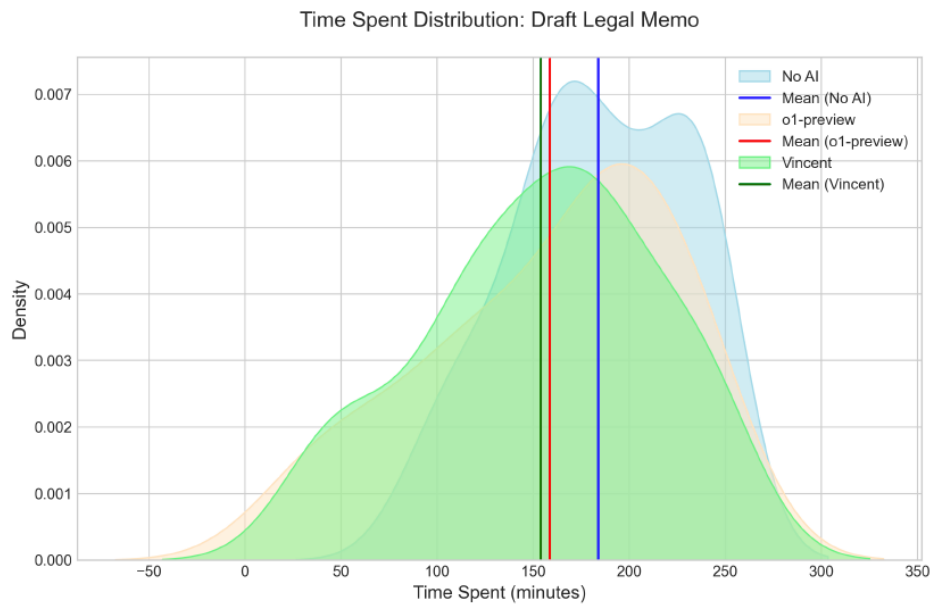


Figure 8: Task 2 Time Density

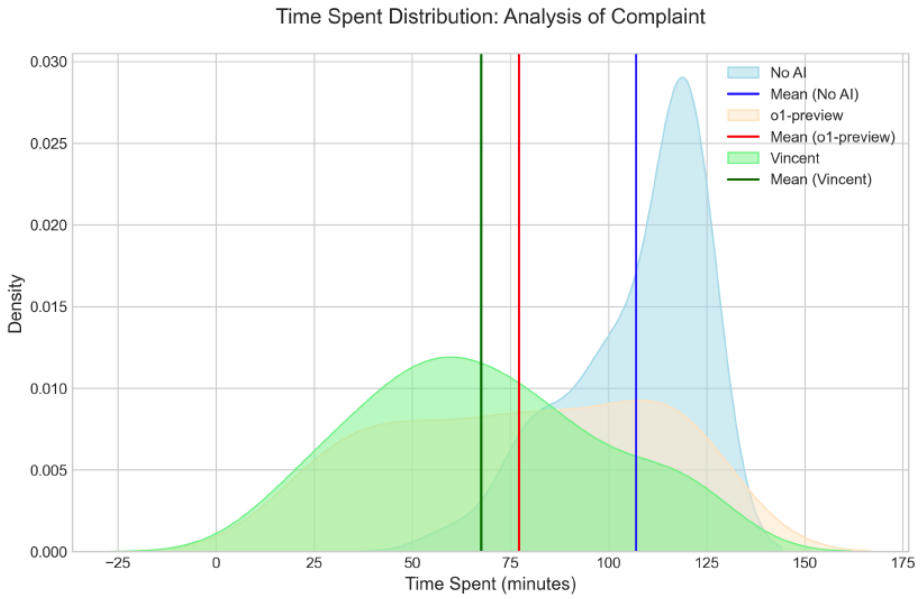


Figure 9: Task 3 Time Density

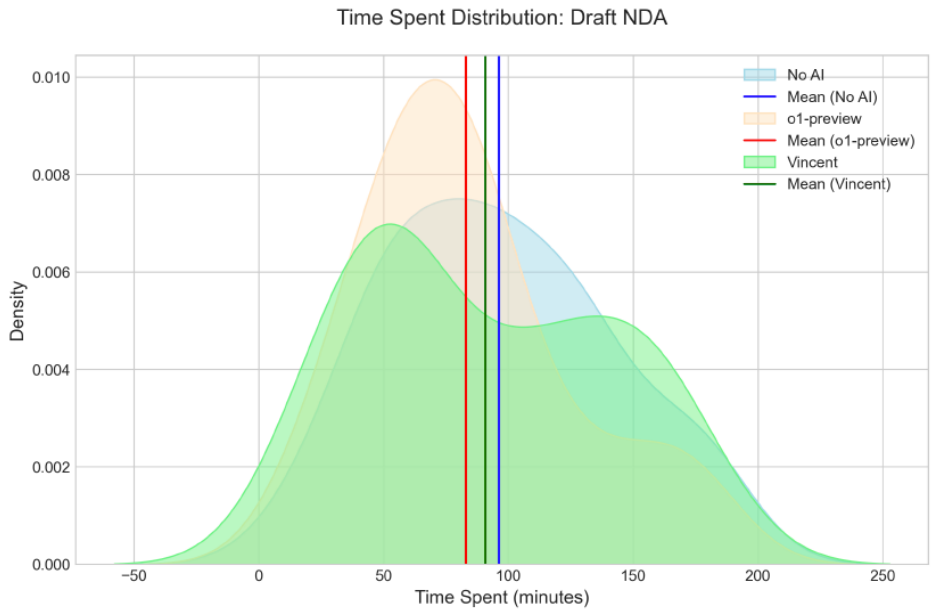


Figure 10: Task 4 Time Density

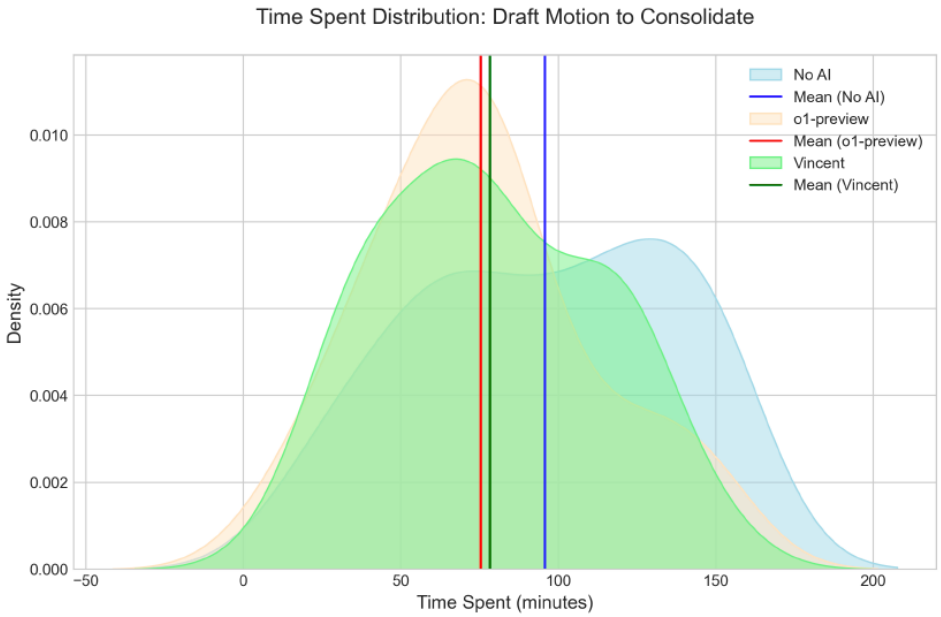


Figure 11: Task 5 Time Density

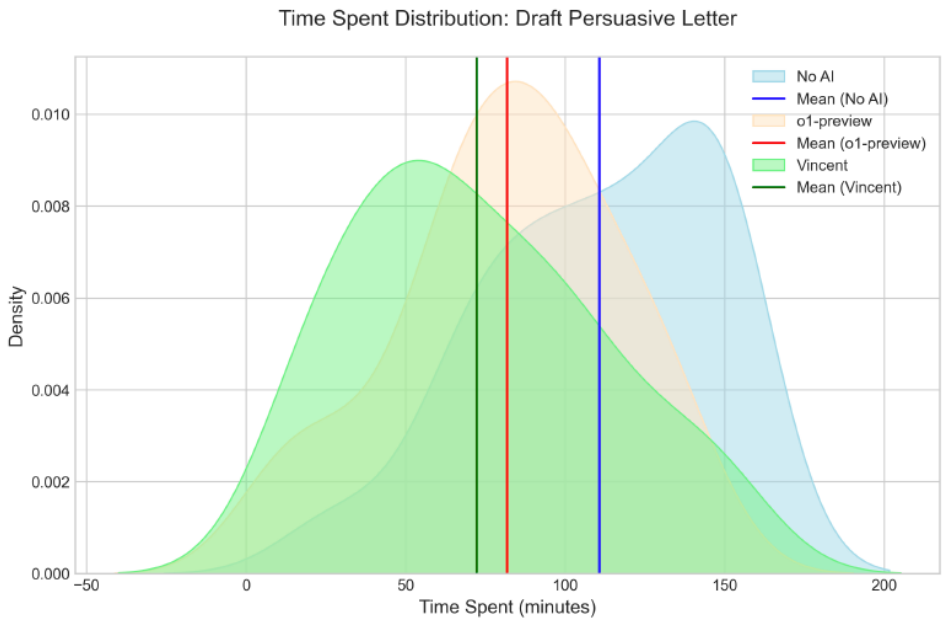


Figure 12: Task 6 Time Density

As evident from both Table 9 and Figure 10, the draft NDA assignment was the only task where the two AI tools did not lead to statistically significant reductions in completion time. Notably, this was also the assignment on which neither AI tool produced a statistically significant improvement in quality.[139] As with quality, we attribute this difference to the transactionally oriented nature of the assignment as well as the fact that all participants were provided with a template to use in crafting their answer.[140]

Unlike the findings on quality improvements, prior research using a nearly identical methodology found that access to GPT-4 consistently reduced the time participants took to complete

[139] *See* Part III.A.1, *supra*.

[140] Regarding speed, the key factor influencing completion time may be the availability of a template rather than the transactional nature of the assignment. A prior randomized controlled trial of GPT-4 found that access to that AI model decreased participants' time to complete a contract drafting task. While that assignment was also transactional, a crucial difference was that participants in the prior study did not receive a contract template, whereas they did in this study. Another potentially relevant distinction is that the assignment tested here more closely reflects a real-world legal task compared to the contract drafting assignment in the previous study.

assignments.[141] This raises an important question: Were the speed improvements observed with o1-preview and Vincent AI greater or more consistent than those found in the previous study with GPT-4? While a perfect comparison is not possible due to differences in the assignments tested in each study, we can gain insight into this question by looking at the results in Table 9 alongside the corresponding findings from Table 2 in the prior study, which reported the reductions in time associated with using GPT-4 for each of the four tasks tested in that study.

Table 2: Average Time Taken on Tasks with and Without GPT-4 (Minutes)

Task	No GPT-4 (Std. Dev.)	With GPT-4 (Std. Dev.)	Difference (95% CI)	% Diff.	*p*-value
Complaint Drafting	160.69	122.00	-38.77	24.1	0.0018
	(72.38)	(66.80)	(-64.00, -13.36)		
Contract Drafting	69.72	47.59	-22.40	32.1	0.0000
	(32.00)	(31.09)	(-33.71, -10.91)		
Employee Handbook	37.24	29.41	-7.84	21.1	0.0000
	(9.55)	(13.42)	(-12.03, -3.74)		
Client Memo	244.41	215.69	-28.75	11.8	0.0152
	(58.03)	(72.96)	(-52.59, -5.05)		

Comparing these results with those in Table 9, there is little evidence to suggest that the magnitude and variability of time-to-completion improvements differed significantly between GPT-4 and the two AI tools tested in this study: o1-preview and Vincent AI. To illustrate, in the prior study GPT-4 reduced completion time by an average of approximately 22% across four assignments, while Vincent AI achieved an average reduction of 21% across six assignments, and o1-preview resulted in a 19% improvement.

3. Productivity Results

Not surprisingly given the findings detailed above, we found consistent evidence that access to both o1-preview and Vincent AI led to

[141] Choi, Monahan, & Schwarcz, *supra* note 1.

statistically significant and substantial productivity gains in five of the six tested assignments, with the NDA assignment again being the exception. Productivity was measured by total points scored on the overall quality assessment per minute spent on the task. These results are presented in Table 10 below.

Table 10: Treatment Effects on Task Productivity (Total Quality Score Points per Minute)

Task	Control Mean	Model	Effect	SE	% Change	N
Draft Client Email	0.393	Vincent	0.216***	(0.055)	+55.0%	134
		o1-preview	0.135***	(0.047)	+34.3%	134
Draft Legal Memo	0.098	Vincent	0.060***	(0.018)	+61.0%	125
		o1-preview	0.076***	(0.021)	+77.5%	125
Analysis of Complaint	0.238	Vincent	0.273***	(0.058)	+114.6%	126
		o1-preview	0.206***	(0.047)	+86.7%	126
Draft NDA	0.373	Vincent	0.070	(0.070)	+18.7%	127
		o1-preview	0.038	(0.060)	+10.3%	127
Draft Motion to Consolidate	0.235	Vincent	0.090*	(0.048)	+38.3%	127
		o1-preview	0.172***	(0.062)	+73.3%	127
Draft Persuasive Letter	0.193	Vincent	0.158***	(0.043)	+82.1%	126
		o1-preview	0.271***	(0.087)	+140.5%	126

Notes: Effects shown as absolute increase in points per minute relative to No AI control group. Percent changes calculated relative to control group mean. Robust standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Sample size (N) represents the number of observations used in the regression.

As is evident from Table 10, the magnitude of these productivity improvements was quite large for both AI tools. We find that Vincent yields statistically significant productivity boosts of approximately 38% to 115% and o1-preview increased productivity by between roughly 34% and 140%, with particularly strong effects in complex tasks like drafting persuasive letters and analyzing complaints.

4. Qualitative Assessment of Results

To gain deeper insight into how AI usage affected the quality of submitted work, the three grading co-authors conducted a post-analysis qualitative review of the assignments they graded. During this review, they had access to information about which assignments had been completed without AI assistance or with one of the two tested AI tools.

A clear trend in this unblinded qualitative review was that participants who used an AI tool generally produced writing that was easier to read and more polished than those who did not. Their sentences were more concise, their paragraphs flowed more smoothly, and their overall structure presented information in a more coherent and user-friendly manner. Additionally, these submissions were largely free of typos, comma splices, and other distracting errors. Within the AI-

assisted group, the differences between participants using Vincent AI and those using o1-preview were less pronounced.

These benefits from AI influenced scores in the "Clarity," "Organization" and "Professionalism" categories. In all three categories, participants who had access to AI produced work that was consistently solid. By contrast, the quality of participants' writing varied widely when they did not have access to AI. As one grading co-author put it in the context of a bowling analogy, it was as if students with AI not only were playing with bumpers built into the gutters—to prevent huge mistakes—but also were told which ball to use, which shoes to use, and where to aim.

The stabilizing, "raise-the-floor" effect of AI was less pronounced in the "Analysis" and "Accuracy" categories compared to others. However, in several assignments, AI assistance helped participants focus on the most relevant issues and questions. For example, access to AI appeared to reduce the likelihood of participants veering off on tangents and ensured they spent less time struggling with the research phase in a way that might otherwise leave them with insufficient time to write. This suggests a connection between AI's speed and quality-related benefits: by streamlining the research process, AI allowed participants to dedicate more time to analyzing and refining their work.

However, the positive effects of AI on analysis and accuracy were not consistent, particularly with Vincent AI. AI assistance, and Vincent AI in particular, tended to be more beneficial when the assignment presented a narrower issue clearly outlined in the prompt. Conversely, its advantage diminished on broader tasks where participants needed to identify the key issue themselves. Vincent users, in particular, were more likely to struggle with task identification on such broader assignments.[142] They sometimes responded to a far broader question than the one actually asked and frequently included fewer relevant citations. In some cases, they provided case names without citations or relied on non-binding administrative or secondary sources. Additionally, AI sometimes

[142] For Assignment Six, Vincent-assisted responses were more likely to address other (or all) CNTC enforceability elements, minimizing the assigned "scope of restraint" issue, causing organizational issues (as well as accuracy and analysis issues discussed below). Addressing the entire CNTC enforceability test (consideration, protectible interests, scope of restraint, and consistent with public policy) created a structure that required the reader to dig through irrelevant issues to get to the assigned issue. It also meant that the writer spent more space on the irrelevant issues and were far less likely to delve into the organizational subparts of a reasonable scope of restraint (time, geography, and activity restrained.)

led participants to oversimplify legal questions or, in the case of o1-preview, omit legal authorities altogether.[143]

The qualitative review also highlighted o1-preview's disproportionate tendency to generate hallucinated sources. Although the overall number of hallucinated citations was very small (18 across all treatment groups on a total of 768 tasks), a clear pattern emerged regarding inaccuracies in the law. In particular, o1-preview users occasionally cited cases that were entirely fabricated—meaning they did not exist under the names or citations provided. A more subtle issue arose in the types of sources used. Vincent users, in particular, were more likely to include obscure—and often unnecessary—sources, setting them apart from other participants.

B. Variation Across Participants

In addition to assessing how access to o1-preview and Vincent influenced overall performance, we also examined the impact of these two AI tools on participants with varying baseline skill levels. Prior research suggested that when GPT-4 affected the quality of legal work, it did so unevenly—tending to benefit those with lower initial skill levels more than those with higher baseline proficiency.[144]

To examine whether similar patterns emerge with o1-preview and Vincent, Figure 13 plots average task productivity across all six assignments on the Y-axis, comparing performance when participants used o1-preview (red line and red dots) versus no AI (blue line and blue dots). These scores are mapped against participants' GPAs on the X-axis. If lower-GPA students experienced a greater productivity boost from access to o1-preview, we would expect a larger gap between the red and blue lines on the left side of the graph than on the right side of the graph, as GPAs increased. While Figure 13 shows some evidence of this pattern, the effect is not pronounced, suggesting that the productivity gains from o1-preview access are relatively consistent across skill levels.

[143] For instance, for Assignment #1, the underperformance of GPT users seems driven by a significant number of them treating the problem as a simple binary and providing an answer with only a little to no legal authority.

[144] See Part I, supra.

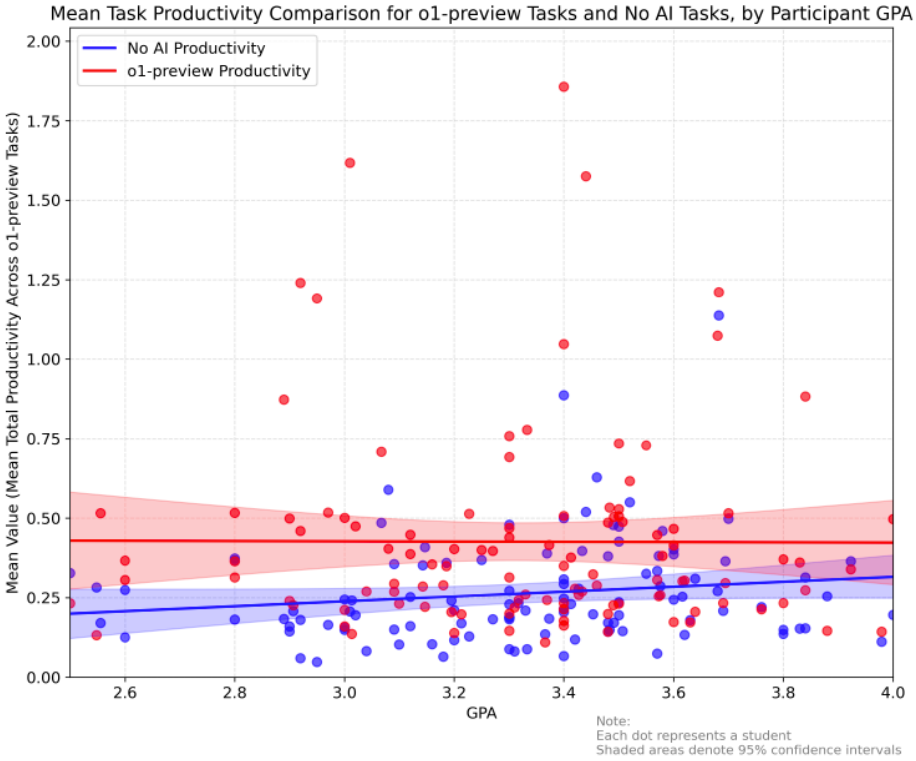


Figure 13: Productivity Comparison: No AI vs o1-preview

We conducted a similar analysis to assess how access to o1-preview influenced work quality, rather than productivity, across participants with varying baseline skill levels. Figure 14 plots participants' average scores across all six assignments on the Y-axis against their GPAs on the X-axis. As in the previous figure, the red line and dots represent scores with access to o1-preview, while the blue line and dots represent scores without AI.

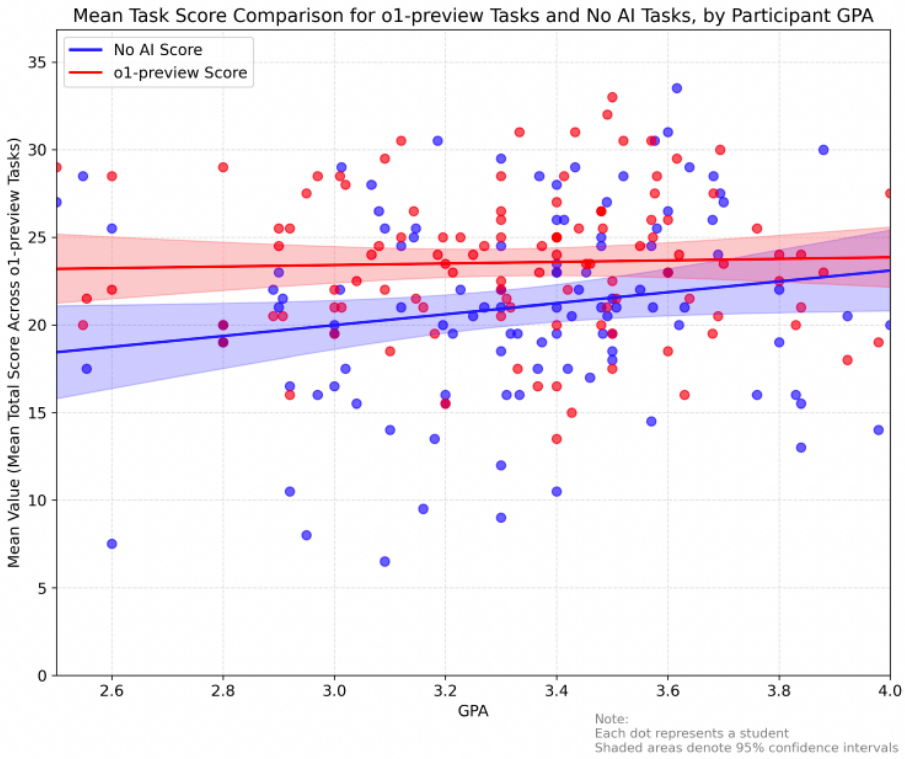


Figure 14: Score Comparison: No AI vs o1-preview

In contrast to the findings on productivity, the differential effect by ability—as measured by GPA—is more pronounced when focusing on scores. Specifically, the near convergence in Figure 14 of the blue and red lines as GPAs increase suggests that o1-preview provides a greater boost in quality for participants with lower baseline skill levels compared to those with higher baseline skills. However, even among the highest skill levels, there is no indication that access to o1-preview reduces the overall quality of work, which is a notable contrast with earlier research on GPT-4.[145]

Figures 15-16 below repeat this same analysis for Vincent AI. Interestingly, these figures reveal a pattern opposite to that observed with o1-preview. In terms of productivity, a differential effect based on baseline skill level (measured by GPA) is evident from the convergence of lines in Figure 15. In contrast, Vincent's impact on overall scores appears relatively uniform across baseline skill levels, as shown in Figure 16.

[145] *See* Schwarcz & Choi, *supra* note 27.

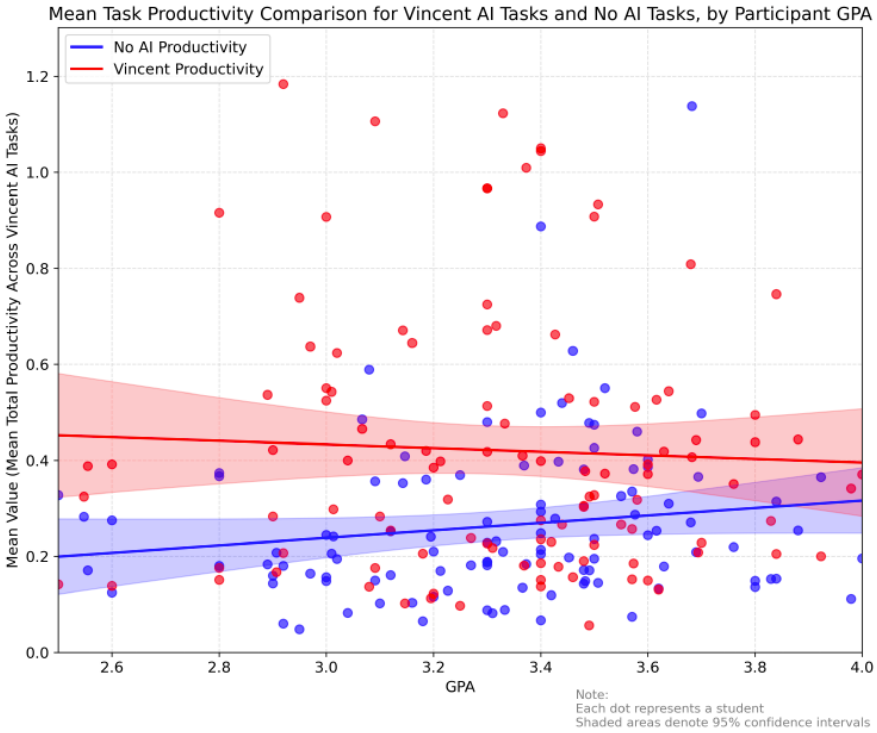


Figure 15: Productivity Comparison: No AI vs Vincent

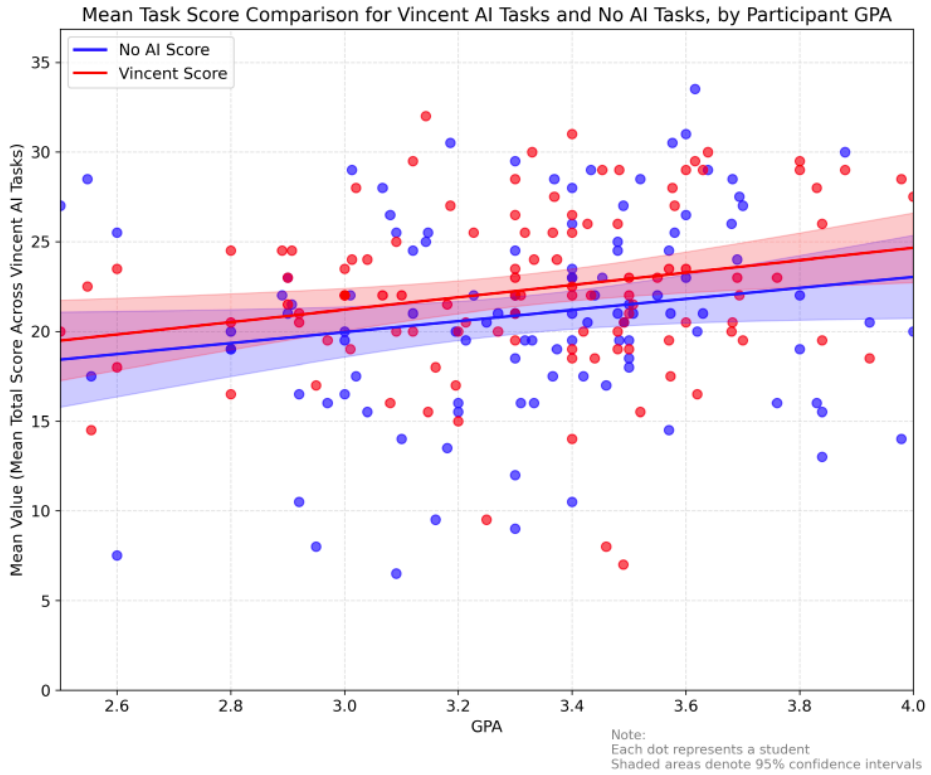


Figure 16: Score Comparison: No AI vs Vincent

Another way to assess the relative impact of AI access across participants with different baseline skill levels is to measure baseline skill not by GPA, but by the scores participants received on assignments completed without AI. Figures 17 and 18 illustrate this approach for o1-preview and Vincent, respectively.

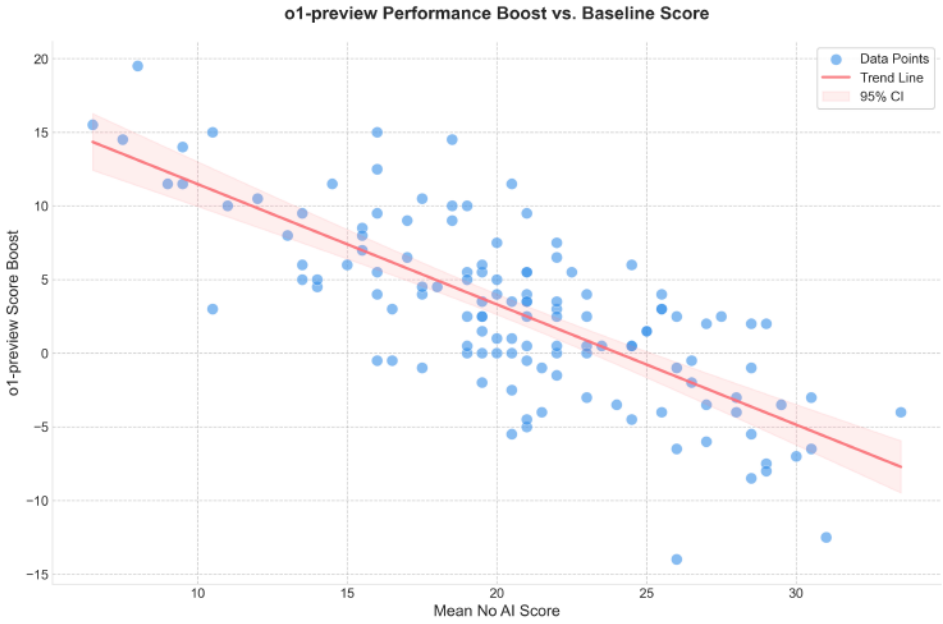


Figure 17: o1-preview Performance Boost vs No AI Performance

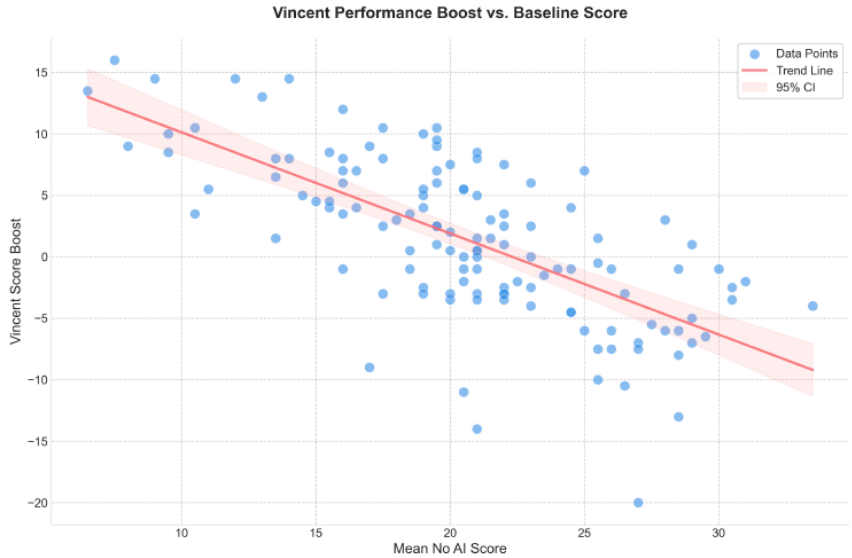


Figure 18: Vincent Performance Boost vs No AI Performance

This approach to measuring the relative impact of AI access on overall quality reveals more pronounced differences in the boost provided by the two AI tools across baseline skill levels. In both cases, the fact that the trend line dips below zero on the Y-axis for a significant number of participants suggests that, for those with the highest baseline skill levels

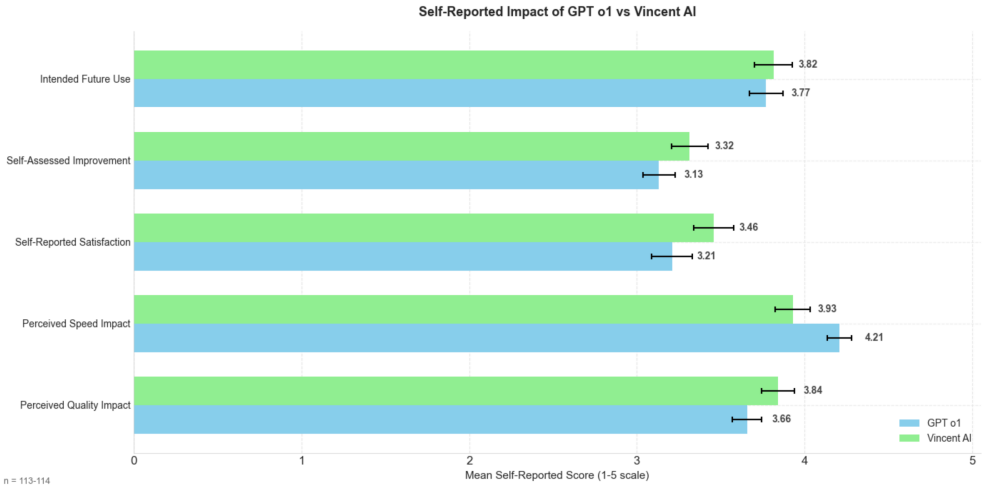
(as measured by performance on tasks completed without AI), access to the AI tool can actually reduce overall performance quality.[146]

Overall, these findings align with prior research on GPT-4's impact on legal analysis, as well as broader studies on AI-driven productivity. In summary, individuals with lower baseline ability tend to benefit more from AI tools than those with higher baseline ability.

C. Post-Experiment Survey Results

The post-experiment survey results were generally consistent with the overall findings on quality, speed, and productivity described above, though some discrepancies emerged. Figure 19 displays the average responses to several survey questions: how access to the two AI tools influenced participants' intended future use of these tools, the perceived improvement in their proficiency with these tools over the experiment, the impact of these tools on their overall satisfaction during task completion, and their perceptions of how the tools affected both the quality of their work and their speed of completion.

Figure 19



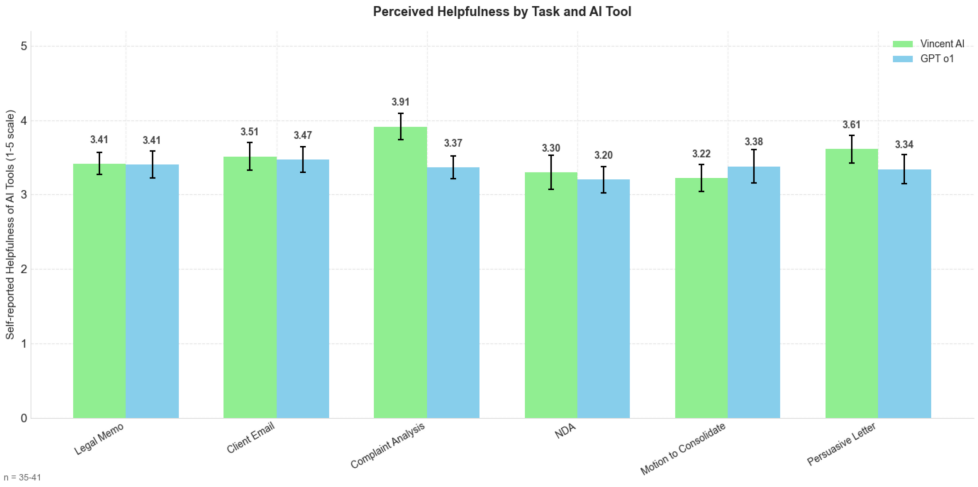
These data indicate that participants generally believed that both AI tools enhanced the quality of their work and increased their speed in completing the work. Interestingly, participants perceived o1-preview as more effective for improving speed and Vincent AI as more helpful for enhancing quality. These subjective impressions diverge somewhat from the actual results discussed earlier, which showed that both tools had comparable effects on speed, while o1-preview yielded broader and more

[146] *See* Schwarcz & Choi, *supra* note 27.

significant improvements in quality.[147] Additionally, the survey results in Figure 19 suggest that participants had a largely positive experience using both AI tools during the experiment, with particularly strong approval for Vincent AI.

Figure 20 displays participants' average ratings of the overall helpfulness of the two AI tools across each of the six assigned tasks.

Figure 20



These results reveal mixed perceptions about which of the six assignments benefited most and least from the two AI tools. Notably, these perceptions do not fully align with the actual performance data, which indicate that both tools were less effective in enhancing speed or quality for the draft NDA assignment compared to the other five tasks. In contrast, the survey responses show only small to moderate differences in how participants rated the AI tools' usefulness for the NDA assignment relative to the others.[148]

[147] *See* Part III.A, *supra*.

[148] For example, while approximately 30% of respondents found Vincent AI either not helpful or only slightly helpful for the NDA task (the highest percentage across all assignments), nearly 60% rated it as very or extremely helpful—outperforming three other assignments. Similarly, although o1-preview received the lowest percentage of "extremely helpful" ratings for the NDA assignment, its overall positive ratings (more than slightly helpful) were comparable to two other assignments.

IV. IMPLICATIONS

Our findings demonstrate that Vincent AI and o1-preview each independently enhance legal productivity and the quality of certain types of legal work. Because they accomplish these results in distinct ways, the combined impact of these technologies is virtually certain to be greater than our results suggest. Section A elaborates on this point. Next, Section B underscores the need for continued empirical evaluation of AI's legal capabilities. It argues that randomized-controlled trials, like those used in this Article, are a particularly reliable method for accurately measuring and projecting AI's impact on lawyering. Given the strong likelihood that AI will fundamentally reshape legal work in the near-to-medium term, greater attention should be devoted to assessing the impact of this technology across various legal tasks, as well as its potential to transform legal education and training.

A. The Combined Power of RAG and Reasoning Models

The implications of our separate findings for Vincent AI and o1-preview are each independently significant. Viewed together, however, they are even more noteworthy. That is because each AI system appears to enhance legal work through distinct mechanisms, which can be and already are being combined with one another in updated legal technology tools.[149] Although this integration may result in additive benefits, it may also produce multiplicative benefits.

Consider the primary mechanism through which Vincent AI impacts legal work beyond facilitating access to a foundation model: retrieval-augmented generation (RAG).[150] Perhaps the most significant limitation of this technology is its ability to correctly identify and leverage the most relevant sources among millions of potentially relevant cases, statutes, regulations, and secondary materials. This challenge is particularly acute in legal analysis, where one of the key difficulties is determining which materials are most pertinent and how best to use them in constructing an argument.[151] This inherent difficulty helps explain why Vincent AI did not improve the accuracy or analysis scores across any of the six assignments, despite yielding fewer hallucinations relative to o1-preview and even study participants who were not using

[149] See, e.g., Gabe Pereyra & Winston Weinberg, Harvey: is building legal agents and workflows with OpenAI o1 (Sep 12, 2024).

[150] *See* Part I, *supra*.

[151] In some legal contexts, the most useful sources may not be immediately obvious—they may involve analogous legal arguments, decisions from jurisdictions with ideologically aligned judges, or precedents from different deal structures that illuminate similar contractual issues.

AI.[152] Legal accuracy and analytical quality depend not only on correctly summarizing legal sources but also on strategically selecting and persuasively leveraging the most compelling authorities to support an argument. Our results suggest that reasoning models like o1-preview have the potential to excel in precisely these areas relative to older AI models like GPT-4.

Just as o1-preview and increasingly advanced reasoning models can enhance RAG-based tools by addressing their primary weakness, the reverse is also true. When combined with extensive legal databases, RAG technology can mitigate a key shortcoming of general purpose foundation models in legal analysis: their lack of direct access to legal source materials.[153] This limitation was evident in our results, which showed a higher rate of hallucinations in o1-preview-assisted assignments compared to those using Vincent AI or no AI at all.[154] Additionally, this limitation was evident in the lack of statistically significant improvement in accuracy scores across the six assignments—except for the complaint analysis assignment, where the key source material (the complaint) was directly provided to all participants.[155]

A further reason that our results are likely to understate the potential impact of AI on lawyering is more familiar: AI technology is continuing to improve at a blistering pace: even the next-generation technology we tested in this experiment is already outdated.[156] But this point has special salience in the context of our Article, because the reasoning model we tested (o1-preview)—and found to improve the quality of human legal reasoning in ways that differed from any other previously tested model—was the very first reasoning model publicly available.[157] The pace of innovation in AI, or any other field, is typically greatest when new types of approaches are first released.[158] Indeed, since OpenAI publicly released o1-preview—an event that immediately preceded the start of our experiment in October, 2024—the company has

---

[152] *See* Part II, *supra*.

[153] *See* Part I, *supra*.

[154] *See* Part II, *supra*.

[155] *See id.*

[156] See Choi, Monahan, & Schwarcz, *supra* note 1.

[157] See Kylie Robison, OpenAI releases o1, its first model with 'reasoning' abilities, The Verge, (Sept. 12, 2024).

[158] See generally Everett M. Rogers, Arvind Singhal & Margaret M. Quinlan, *Diffusion of Innovations*, in AN INTEGRATED APPROACH TO COMMUNICATION THEORY AND RESEARCH 432, 432–48 (Don W. Stacks & Michael B. Salwen eds., 3d ed. 2014); Xuli Tang et al., *The Pace of Artificial Intelligence Innovations: Speed, Talent, and Trial-and-Error*, 15 J. INFORMETRICS 101147 (2021).

released several new generations of reasoning models.[159] Not surprisingly, the company's most recently announced reasoning model, o3, substantially outperforms the previous o1 model on numerous objective benchmarks.[160]

Ultimately, our findings suggest that legal AI may be moving toward an inflection point. AI tools are increasingly capable of handling two key aspects of legal work: (1) information retrieval and (2) reasoning. The convergence of these capabilities points to a future in which AI—by enhancing both the efficiency and the quality of the work that attorneys can produce—becomes more than just a helpful accessory. It becomes an integral part of the profession.

B. Empirically Testing the Power of AI in Law

As AI continues to improve, the legal community must thoughtfully consider how best to integrate these tools. Legal institutions will need to develop systematic, empirical methods for evaluating AI's capabilities on a broad range of legal tasks. Yet remarkably few evaluation tools currently exist—largely because assessing legal work remains inherently challenging. Unlike fields such as coding, mathematics, or science, where automated benchmarks can precisely measure performance,[161] no standardized metrics exist for assessing AI's effectiveness on complex lawyering tasks like drafting persuasive briefs or crafting airtight contracts.

Some firms and commentators have attempted to address this gap by developing legal benchmarks to objectively evaluate AI tools.[162] But

---

[159] *See* Maxwell Zeff & Kyle Wiggers, OpenAI announces new o3 models, TECHCRUNCH (December 20, 2024).

[160] *See id.*

[161] Tidor-Vlad Pricope, HardML: A Benchmark for Evaluating Data Science and Machine Learning Knowledge and Reasoning in AI, arXiv preprint arXiv:2501.15627 (2025), https://arxiv.org/abs/2501.15627; M. Tian et al., Scicode: A Research Coding Benchmark Curated by Scientists, arXiv preprint arXiv:2407.13168 (2024), https://arxiv.org/abs/2407.13168; Swaroop Mishra et al., Lila: A Unified Benchmark for Mathematical Reasoning, arXiv preprint arXiv:2210.17517 (2022), https://arxiv.org/abs/2210.17517.

[162] See, e.g., Fei, Zhiwei, et al*., Lawbench: Benchmarking Legal Knowledge Of Large Language Models*, arXiv preprint arXiv:2309.16289 (2023) (proposing "a comprehensive legal benchmark for AI that aims to provide a precise assessment of the LLMs' legal capabilities from three cognitive levels: (1) Legal knowledge memorization . . . (2) Legal knowledge understanding, . . . [and] (3) Legal knowledge applying."); Harvey, Introducing BigLaw Bench (Aug 29, 2024); Niko Grupen & Julio Pereyra, Harvey, BigLaw Bench – Retrieval, Nov 13, 2024, https://www.harvey.ai/blog/biglaw-bench-retrieval (using a series of objective metrics to conclude that "Harvey's retrieval system outperforms

these benchmarks are becoming increasingly inadequate for measuring AI's legal capabilities. First, many are saturated. AI models have already achieved near-maximum—or even superhuman—performance, leaving little room for meaningful improvement.[163] Second, because valuable lawyering tasks cannot be easily measured formulaically, the real-world relevance of AI performance on these tests is limited. Third, the practice of law is fundamentally human. It I both a technical skill and a value-laden activity. The key question is not how well AI performs in isolation but how its capabilities can be effectively leveraged by human lawyers. Existing AI benchmarks are simply not designed to measure this critical dimension.

The approach employed in this Article—randomized controlled trials focused on realistic lawyering tasks—offers the key to better understanding the role AI will play in the work that lawyers do.[164] Unlike formulaic benchmarks, randomized controlled trials allow researchers to reliably evaluate AI's impact on humans' ability to perform virtually any realistic lawyering task. Given the transformative potential of AI on the profession, it is important that clients, law firms, and law schools start to embed periodic trials into their operations and then adapt accordingly.[165]

Consider a pair of ways in which way law schools can and should be actively testing the impact of AI on legal education. The first involves a well-known weakness in legal training: the lack of individualized feedback for students and young lawyers.[166] Generative AI has the potential to revolutionize this feature of legal training by providing frequent, detailed, and personalized feedback on legal work. But whether AI can fulfill this promise is ultimately an empirical question that depends on both the AI models used and the techniques for implementing them. Early testing suggests that advanced reasoning models can generate highly accurate and specific feedback on complex legal exams when supplied with the exam question and answer, a blank grading rubric, and several instructor-completed rubric examples. Law schools

---

commonly used embedding-based and reranking methods, identifying up to 30% more relevant content than alternative retrieval methods across a diverse range of legal document types.").

[163] See, e.g., Artificial Lawyer, Paxton Hits 94% Accuracy On Stanford GenAI Benchmark (30th July 2024), at https://www.artificiallawyer.com/2024/07/30/paxton-hits-94-accuracy-on-stanford-genai-benchmark/.

[164] See also Choi, Monahan, & Schwarcz, *supra* note 1.

[165] See id.

[166] *See* Daniel Schwarcz & Dion Farganis, *The Impact of Individualized Feedback on Law Student Performance*, 67 J. LEGAL EDUC. 139 (2017).

should therefore invest significant resources into evaluating this approach using randomized controlled trials.

A second key empirical question for law schools is whether allowing students to use AI in their legal training may hinder their development of critical legal skills. Some recent research on knowledge workers suggests that repeated AI use can undermine analytical abilities,[167] raising concerns that similar effects could emerge in legal education. Yet these concerns have yet to be evaluated empirically and will likely depend on factors such as how and when students are encouraged to use AI. Only through systematic testing of different approaches can law schools determine the most effective strategies for training the next generation of lawyers.

CONCLUSION

This Article presents the first rigorous empirical evidence that advanced AI tools—specifically Retrieval-Augmented Generation (RAG) and reasoning models—can significantly enhance the quality of legal work in realistic lawyering tasks, while preserving the efficiency gains observed with earlier generations of generative AI. Our findings demonstrate that reasoning models improve not only the clarity, organization, and professionalism of legal work but also the depth and rigor of legal analysis itself.

Additionally, we provide evidence that RAG-enabled legal AI tools may be able to reduce hallucinations in human legal work to levels comparable to those found in work completed without AI assistance. The distinct yet complementary strengths of these technologies suggest that their integration could yield even greater benefits, a development already taking shape in emerging legal tech.

The rapid advancement of reasoning models also indicates that the improvements observed in this study may only be the beginning of AI's transformative potential for legal practice. As law schools, practitioners, and policymakers navigate AI's evolving role, our findings highlight the critical importance of empirical research in shaping informed, forward-looking strategies for the future of the legal profession.

[167] Hao-Ping (Hank) Lee et al, The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers, https://doi.org/10.1145/3706598.3713778.

APPENDIX

A. Assignments

The six Assignments we gave to participants were as follows:

(1) Assignment One: Client Email, Time Limit: One Hour

A Client is annoyed that our opponent has filed a counterclaim that included several lies about the Client. I know we can look at the Rule 11 path and try to prove that Opponent is lying through discovery. But the Client has also asked us to add a defamation claim to the suit and any other claim we can think of.

I'm pretty sure that you're not allowed to base a defamation or other tort claim on statements made solely inside a litigation—but we need to draft a short email to the client explaining this prohibition. The email needs to be supported by sound research, because the Client is very insistent that we pursue a defamation claim and won't be satisfied with an explanation that is not rock solid.

For additional context, the case is in a district court located in the Tenth Circuit.

Please draft an email no longer than 700 words to the Client and then send it me to review.

(2) Assignment Two: Legal Research Memo, Time Limit: Four Hours

ABC Trucking (ABC) is large trucking company doing business in Massachusetts and New Hampshire. It purchases a commercial auto liability insurance policy. The policy explicitly covers ABC's liability for trucking accidents in the United States and Canada. It contains a $2 million coverage limit for liability insurance. It also contains the following language, which is relatively common in auto insurance policies:

2. Coverage Extensions

a. Supplementary Payments

We will pay for the "insured":

* * *

(5) All costs taxed against the "insured" in any "suit" against the "insured" we defend.

These payments will not reduce the Limit of Insurance.

Other portions of the insurance policy make clear that the "coverage extensions" detailed above are additional insurance provided by the policy beyond the $2 million coverage limit for liability insurance.

One of the company's trucks was involved in an accident in Ontario, Canada. Several passengers in the other vehicle were severely injured. The insurer agrees to pay its $2 million limit to settle litigation brought by the injured passengers. But lawyers representing these victims insist in the settlement negotiations with the insurer that, according to the coverage extensions excerpted above, the insurer also needs to pay the victims' attorneys' fees. The amount of those fees is $200,000. Unlike in the United States—where the normal rule is that both parties pay for their own attorneys' fees—the rule in Ontario is that a losing party in litigation generally pays the attorneys' fees of the prevailing party. This rule is contained in the Ontario Courts of Justice Act, which establishes general procedural rules applicable to all civil disputes in Ontario. See Ontario Courts of Justice Act, R.S.O. 1990, c. C.43. The victims' attorneys therefore argue that their expenses constitute "costs taxed against the 'insured' in any 'suit' against the 'insured'" that the insurer defends.

We represent the insurer. Another lawyer is analyzing whether Massachusetts or New Hampshire law will govern the dispute. Your task is to draft an objective research memo analyzing whether, under Massachusetts and New Hampshire laws governing insurance disputes, the insurer is obligated to pay the $200,000 in attorneys' fees in addition to the $2 million coverage limit. Don't spend any time looking into Ontario law or Canadian law more generally. (To the extent that you need any more information about those jurisdictions, simply note that in your memo. We'll have our Canadian counterparts look into the issue.)

Your memo should be no longer than 1,500 words. Make sure to consider the extent to which existing Massachusetts and New Hampshire caselaw squarely addresses the issue presented here, or can plausibly be distinguished from the facts at issue in this case. Also be sure to articulate both the best arguments for and against coverage, as well as the answer that a court is most likely to reach.

(3) **Assignment Three: Complaint Analysis, Time Limit: Two Hours**

Our client, Foundever Operating Corporation, provides customer service operations for a variety of different companies. Yesterday, it was served with the attached complaint, which appears to be a putative class action lawsuit brought by a number of current and former employees of Foundever.

Please draft an initial memo no longer than 1,000 words that accomplishes the following:

- Briefly summarize the key allegations and claims in the complaint.
- Provide a short assessment of the strength of the claims based on the allegations.
- Outline potential next steps or a preliminary defense strategy.

Don't worry about jurisdiction or venue. Focus instead on the claims under (1) the Fair Labor Standards Act and (2) Nevada state law.

Note: The complaint was provided as part of the assignment, but is omitted here due to space constraints. It was 17 pages long, and taken from a real case that was resolved without subsequent filings from the parties.

(4) **Assignment Four: NDA** , **Time Limit: Three Hours**

As a condition of their job offer, Sue Scientist will be asked to sign a nondisclosure agreement. Eddie Employer has hired you to draft the agreement for Acme Co.

Sue is an experienced employee in this field and is expected to work on confidential products involving Acme Co.'s proprietary trade secrets. If she leaves her employment with Acme Co., it is reasonable to expect it would be for an opportunity at a competitor—or some other company.

It will be important to Acme Co.'s continued viability that those trade secrets are not made public in either its state of business (Minnesota) or neighboring states for a reasonable period of time after her departure. You will need to confirm each state's specific limitations before drafting that part of the agreement.

Please use the drafting standards of plain English to create an enforceable nondisclosure agreement favorable to Acme Co. I've attached a sample nondisclosure agreement that is overbroad, but it should provide a starting point for you to work from. The contract should be no more than three pages. Please also make sure it is single-spaced, uses 12-point type, and is formatted with 1" margins.

Note: A sample NDA was provided as part of the assignment, but is omitted here due to space constraints. It was 10 pages long, and taken from a model used by a law firm.

(5) **Assignment Five: Motion to Consolidate**, **Time Limit: 2.5 Hours**

We are going to bring a motion to consolidate two cases in Minnesota state court. They are already assigned to the same judge. I won't get into the facts, but the basics are that Acme originally sued Beta in Minnesota state court seeking declaratory judgment that a contract between Acme and Beta has terminated based on Beta's insolvency.

Near the same time, the owners of Beta in their individual capacities sued Acme in a different jurisdiction for fraud, but that case has now been moved to Minnesota state court. The second case revolves around the same core set of facts and events as the first case.

We represent Beta in the first case and Beta's owners in the second case. We think the Court should consolidate the two cases so that all issues can be tried together given the overlap.

Can you put together a basic draft brief in support of a motion to consolidate?

Please cite to Minnesota case law and Minnesota civil procedure. And make it as persuasive as you can!

(6) **Assignment Six: Covenant not to Compete**, **Time Limit: 2.5 Hours**

Our client, Yes Chef, provides private chef services throughout the greater Indianapolis area by deploying one of its three chef employees for daily or weekly engagements. In addition to the chefs, Yes Chef employs several part-time workers to manage the Yes Chef central prep kitchen and offices in downtown Indianapolis, which stores commonly used ingredients and specialized cooking equipment. The prep kitchen also

has a small event space where clients can host dinner parties catered by the company.

Yes Chef sends a chef to work with the client for the given day or week to do menu planning, then it provides complete grocery shopping, food preparation, meal service, and clean-up services. Yes Chef was founded by Bear Grills, a retired chef and culinary school instructor in 2015. It initially served just downtown Indianapolis but has expanded outward since then, doing at least some business in the surrounding counties. Bear hopes to expand it further.

Yes Chef has a number of proprietary recipes, advanced chef techniques, and client development methods that Bear has developed over the years. Yes Chef has a significant and extended training program for newly hired chefs that lasts for the first few years of employment. The employee chefs each play an important role in customer relations and customer service, developing relationships on Yes Chef's behalf and acquiring specialized information about a client's requirements, preferences, allergies, meal price requirements, and other information only shared with the in-home chef. Yes Chef collects much of this information into a password-protected database accessible only by the three chefs and Bear. Chefs are expected to develop relationships with both the customers but also potential customers that attend their dinners as guests. So far, Yes Chef has been extremely successful and retains numerous recurring clients who have used the services for several years and some since its inception.

Yes Chef's longest-serving chef, Antonio Brand, left six months ago. Yes Chef believed he was moving out of state, but it just learned that he has started his own competing business in the area, On Brand Eats (OBE). Brand signed a covenant not to compete with the company when he started back in 2018. The covenant prohibits him from "performing private chef services, for himself as an employee of another, in Marion County, any adjacent county, or any other Indiana county for a period of three years after his employment with Yes Chef ends." The covenant also prohibits him from directly soliciting any Yes Chef customers, wherever located, during that period.

When confronted with the covenant and Brand's obvious violation OBE's counsel claimed the scope of the restraint is unreasonable and therefore unenforceable. The parties have agreed to private judging on this single issue. Please draft a letter for my review making the case that the covenant is reasonable in its restraint and arguing for enforceability to the maximum extent allowed by Indiana law. The letter should be no longer than 1,250 words long.

B. Additional Tables and Figures

Table 11: Treatment Effects for Draft Client Email

Outcome	Control Mean	Model	Effect	SE	% Change	N
Accuracy	3.25	Vincent	0.43	(0.36)	+13.3%	135
		o1-preview	−0.40	(0.35)	−12.3%	135
Analysis	3.64	Vincent	0.52	(0.32)	+14.4%	135
		o1-preview	−0.21	(0.32)	−5.8%	135
Organization	4.05	Vincent	0.57*	(0.31)	+14.0%	135
		o1-preview	0.29	(0.30)	+7.3%	135
Clarity	4.11	Vincent	1.07***	(0.28)	+26.0%	135
		o1-preview	1.08***	(0.27)	+26.2%	135
Professionalism	4.30	Vincent	0.84**	(0.39)	+19.6%	135
		o1-preview	1.07***	(0.38)	+24.8%	135
Total Score	19.34	Vincent	2.93*	(1.58)	+15.1%	135
		o1-preview	1.83	(1.44)	+9.5%	135
Time Spent	50.30	Vincent	−7.15***	(2.54)	−14.2%	134
		o1-preview	−6.11***	(2.28)	−12.1%	134
Productivity	0.39	Vincent	0.22***	(0.05)	+55.0%	134
		o1-preview	0.13***	(0.05)	+34.3%	134

Notes: Effects shown relative to No AI control group. For quality criteria (Accuracy through Professionalism), the scoring scale is 1-7, with Total Score ranging from 5-35. Time Spent shows minutes (time limit: 60 minutes). Productivity measures points earned per minute. Percent changes calculated relative to control group mean. Robust standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 12: Treatment Effects for Draft Legal Memo

Outcome	Control Mean	Model	Effect	SE	% Change	N
Accuracy	2.95	Vincent	0.30	(0.27)	+10.2%	126
		o1-preview	0.10	(0.28)	+3.3%	126
Analysis	3.09	Vincent	0.37	(0.25)	+12.1%	126
		o1-preview	0.52**	(0.25)	+17.0%	126
Organization	4.05	Vincent	0.16	(0.34)	+4.1%	126
		o1-preview	0.98***	(0.31)	+24.2%	126
Clarity	3.27	Vincent	0.77***	(0.26)	+23.6%	126
		o1-preview	0.93***	(0.27)	+28.5%	126
Professionalism	3.55	Vincent	0.66**	(0.33)	+18.7%	126
		o1-preview	1.45***	(0.32)	+41.0%	126
Total Score	16.91	Vincent	2.28*	(1.21)	+13.5%	126
		o1-preview	3.99***	(1.19)	+23.6%	126
Time Spent	183.91	Vincent	−30.37***	(11.30)	−16.5%	125
		o1-preview	−25.98**	(12.10)	−14.1%	125
Productivity	0.10	Vincent	0.06***	(0.02)	+61.0%	125
		o1-preview	0.08***	(0.02)	+77.5%	125

Notes: Effects shown relative to No AI control group. For quality criteria (Accuracy through Professionalism), the scoring scale is 1-7, with Total Score ranging from 5-35. Time Spent shows minutes (time limit: 240 minutes). Productivity measures points earned per minute. Percent changes calculated relative to control group mean. Robust standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 13: Treatment Effects for Analysis of Complaint

Outcome	Control Mean	Model	Effect	SE	% Change	N
Accuracy	5.10	Vincent	0.26	(0.25)	+5.2%	127
		o1-preview	0.46*	(0.26)	+9.0%	127
Analysis	4.62	Vincent	0.33	(0.26)	+7.1%	127
		o1-preview	0.44	(0.30)	+9.6%	127
Organization	4.80	Vincent	0.47*	(0.25)	+9.8%	127
		o1-preview	0.53**	(0.26)	+10.9%	127
Clarity	5.05	Vincent	0.15	(0.23)	+3.1%	127
		o1-preview	0.18	(0.24)	+3.6%	127
Professionalism	4.83	Vincent	0.72***	(0.27)	+14.9%	127
		o1-preview	0.87***	(0.29)	+18.1%	127
Total Score	24.40	Vincent	1.94*	(1.10)	+8.0%	127
		o1-preview	2.48**	(1.21)	+10.2%	127
Time Spent	107.10	Vincent	−39.51***	(5.23)	−36.9%	126
		o1-preview	−29.92***	(5.72)	−27.9%	126
Productivity	0.24	Vincent	0.27***	(0.06)	+114.6%	126
		o1-preview	0.21***	(0.05)	+86.7%	126

Notes: Effects shown relative to No AI control group. For quality criteria (Accuracy through Professionalism), the scoring scale is 1-7, with Total Score ranging from 5-35. Time Spent shows minutes (time limit: 120 minutes). Productivity measures points earned per minute. Percent changes calculated relative to control group mean. Robust standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 14: Treatment Effects for Draft NDA

Outcome	Control Mean	Model	Effect	SE	% Change	N
Accuracy	5.70	Vincent	−0.08	(0.18)	−1.4%	127
		o1-preview	−0.19	(0.18)	−3.3%	127
Analysis	4.93	Vincent	−0.14	(0.19)	−2.7%	127
		o1-preview	−0.15	(0.20)	−3.1%	127
Organization	5.05	Vincent	0.03	(0.19)	+0.6%	127
		o1-preview	−0.25	(0.18)	−4.9%	127
Clarity	5.14	Vincent	0.07	(0.16)	+1.3%	127
		o1-preview	−0.21	(0.19)	−4.0%	127
Professionalism	5.58	Vincent	0.19	(0.22)	+3.4%	127
		o1-preview	−0.31	(0.25)	−5.6%	127
Total Score	26.40	Vincent	0.07	(0.82)	+0.3%	127
		o1-preview	−1.11	(0.86)	−4.2%	127
Time Spent	96.30	Vincent	−5.46	(10.59)	−5.7%	127
		o1-preview	−13.30	(9.30)	−13.8%	127
Productivity	0.37	Vincent	0.07	(0.07)	+18.7%	127
		o1-preview	0.04	(0.06)	+10.3%	127

Notes: Effects shown relative to No AI control group. For quality criteria (Accuracy through Professionalism), the scoring scale is 1-7, with Total Score ranging from 5-35. Time Spent shows minutes (time limit: 180 minutes). Productivity measures points earned per minute. Percent changes calculated relative to control group mean. Robust standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 15: Treatment Effects for Draft Motion to Consolidate

Outcome	Control Mean	Model	Effect	SE	% Change	N
Accuracy	3.78	Vincent	0.08	(0.30)	+2.2%	126
		o1-preview	0.40	(0.27)	+10.6%	126
Analysis	3.42	Vincent	0.30	(0.27)	+8.7%	126
		o1-preview	0.83***	(0.26)	+24.4%	126
Organization	3.47	Vincent	0.77**	(0.31)	+22.1%	126
		o1-preview	1.48***	(0.25)	+42.8%	126
Clarity	3.42	Vincent	0.46**	(0.19)	+13.5%	126
		o1-preview	0.81***	(0.18)	+23.8%	126
Professionalism	3.40	Vincent	0.48	(0.35)	+14.2%	126
		o1-preview	1.50***	(0.30)	+44.0%	126
Total Score	17.49	Vincent	2.09*	(1.24)	+12.0%	126
		o1-preview	4.92***	(1.08)	+28.1%	126
Time Spent	95.80	Vincent	−17.40**	(8.14)	−18.2%	126
		o1-preview	−17.50**	(8.48)	−18.3%	126
Productivity	0.23	Vincent	0.09*	(0.05)	+38.3%	126
		o1-preview	0.17***	(0.06)	+73.3%	126

Notes: Effects shown relative to No AI control group. For quality criteria (Accuracy through Professionalism), the scoring scale is 1-7, with Total Score ranging from 5-35. Time Spent shows minutes (time limit: 150 minutes). Productivity measures points earned per minute. Percent changes calculated relative to control group mean. Robust standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Regression tables including control variables:

Table 16: Treatment Effects for Draft Persuasive Letter

Outcome	Control Mean	Model	Effect	SE	% Change	N
Accuracy	3.44	Vincent	−0.75*	(0.40)	−21.9%	126
		o1-preview	0.31	(0.38)	+9.0%	126
Analysis	3.46	Vincent	−0.35	(0.38)	−10.1%	126
		o1-preview	0.84**	(0.35)	+24.3%	126
Organization	4.00	Vincent	−0.52	(0.40)	−13.1%	126
		o1-preview	0.98**	(0.38)	+24.4%	126
Clarity	4.44	Vincent	0.18	(0.30)	+4.0%	126
		o1-preview	0.75***	(0.28)	+16.9%	126
Professionalism	4.23	Vincent	−0.30	(0.44)	−7.1%	126
		o1-preview	1.21***	(0.39)	+28.6%	126
Total Score	19.56	Vincent	−1.75	(1.70)	−8.9%	126
		o1-preview	4.09***	(1.56)	+20.9%	126
Time Spent	110.95	Vincent	−38.68***	(8.13)	−34.9%	126
		o1-preview	−29.14***	(7.71)	−26.3%	126
Productivity	0.19	Vincent	0.16***	(0.04)	+82.1%	126
		o1-preview	0.27***	(0.09)	+140.5%	126

Notes: Effects shown relative to No AI control group. For quality criteria (Accuracy through Professionalism), the scoring scale is 1-7, with Total Score ranging from 5-35. Time Spent shows minutes (time limit: 150 minutes). Productivity measures points earned per minute. Percent changes calculated relative to control group mean. Robust standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 17: Treatment Effects on Accuracy Across Tasks

Task	Control Mean	Model	Effect	SE	% Change	N
Draft Client Email	3.45	Vincent	0.10	(0.40)	+3.0%	119
		o1-preview	−0.47	(0.40)	−13.5%	119
Draft Legal Memo	3.05	Vincent	0.25	(0.27)	+8.1%	110
		o1-preview	−0.11	(0.30)	−3.7%	110
Analysis of Complaint	5.14	Vincent	0.24	(0.27)	+4.7%	111
		o1-preview	0.49*	(0.29)	+9.5%	111
Draft NDA	5.73	Vincent	−0.07	(0.19)	−1.2%	111
		o1-preview	−0.25	(0.21)	−4.4%	111
Draft Motion to Consolidate	3.77	Vincent	0.17	(0.35)	+4.5%	112
		o1-preview	0.39	(0.28)	+10.4%	112
Draft Persuasive Letter	3.49	Vincent	−0.73*	(0.42)	−21.0%	111
		o1-preview	0.43	(0.41)	+12.4%	111

Panel B: Control Variables

Variable	Coefficient	SE
GPA	0.68	(0.44)
3L Student (vs. 2L)	−0.04	(0.32)
LLM Student (vs. 2L)	0.00	(0.00)
1-5 times (vs. 0 Times)	0.71*	(0.43)
6-10 times (vs. 0 Times)	0.96*	(0.50)
11-20 times (vs. 0 Times)	0.60	(0.78)
More than 20 times (vs. 0 Times)	0.14	(0.50)

Notes: Effects are shown relative to the No AI control group. Robust standard errors (in parentheses) are reported. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table 18: Treatment Effects on Analysis Across Tasks

Task	Control Mean	Model	Effect	SE	% Change	N
Draft Client Email	3.76	Vincent	0.36	(0.36)	+9.4%	119
		o1-preview	−0.17	(0.37)	−4.5%	119
Draft Legal Memo	3.13	Vincent	0.33	(0.25)	+10.5%	110
		o1-preview	0.41*	(0.25)	+13.2%	110
Analysis of Complaint	4.71	Vincent	0.24	(0.27)	+5.0%	111
		o1-preview	0.43	(0.31)	+9.1%	111
Draft NDA	4.97	Vincent	−0.15	(0.19)	−3.0%	111
		o1-preview	−0.26	(0.21)	−5.3%	111
Draft Motion to Consolidate	3.40	Vincent	0.40	(0.29)	+11.8%	112
		o1-preview	0.82***	(0.28)	+24.1%	112
Draft Persuasive Letter	3.60	Vincent	−0.37	(0.41)	−10.4%	111
		o1-preview	0.77**	(0.39)	+21.5%	111

Panel B: Control Variables

Variable	Coefficient	SE
GPA	0.85**	(0.42)
3L Student (vs. 2L)	0.09	(0.29)
LLM Student (vs. 2L)	0.00	(0.00)
1-5 times (vs. 0 Times)	0.63	(0.39)
6-10 times (vs. 0 Times)	0.43	(0.48)
11-20 times (vs. 0 Times)	0.37	(0.68)
More than 20 times (vs. 0 Times)	0.31	(0.47)

Notes: Effects are shown relative to the No AI control group. Robust standard errors (in parentheses) are reported. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table 19: Treatment Effects on Clarity Across Tasks

Task	Control Mean	Model	Effect	SE	% Change	N
Draft Client Email	4.24	Vincent	0.85***	(0.30)	+20.2%	119
		o1-preview	1.11***	(0.29)	+26.2%	119
Draft Legal Memo	3.36	Vincent	0.78***	(0.27)	+23.1%	110
		o1-preview	0.82***	(0.29)	+24.5%	110
Analysis of Complaint	5.09	Vincent	0.04	(0.26)	+0.9%	111
		o1-preview	0.21	(0.28)	+4.0%	111
Draft NDA	5.24	Vincent	0.04	(0.16)	+0.9%	111
		o1-preview	−0.36*	(0.20)	−6.9%	111
Draft Motion to Consolidate	3.45	Vincent	0.43**	(0.21)	+12.6%	111
		o1-preview	0.75***	(0.20)	+21.8%	111
Draft Persuasive Letter	4.51	Vincent	0.22	(0.32)	+4.9%	111
		o1-preview	0.71**	(0.31)	+15.7%	111

Panel B: Control Variables

Variable	Coefficient	SE
GPA	0.36	(0.32)
3L Student (vs. 2L)	0.33	(0.23)
LLM Student (vs. 2L)	0.00***	(0.00)
1-5 times (vs. 0 Times)	0.59*	(0.32)
6-10 times (vs. 0 Times)	0.84***	(0.32)
11-20 times (vs. 0 Times)	0.61	(0.54)
More than 20 times (vs. 0 Times)	0.41	(0.40)

Notes: Effects are shown relative to the No AI control group. Robust standard errors (in parentheses) are reported. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table 20: Treatment Effects on Organization Across Tasks

Task	Control Mean	Model	Effect	SE	% Change	N
Draft Client Email	4.24	Vincent	0.37	(0.33)	+8.6%	119
		o1-preview	0.19	(0.33)	+4.5%	119
Draft Legal Memo	4.13	Vincent	0.12	(0.37)	+3.0%	110
		o1-preview	0.84**	(0.33)	+20.3%	110
Analysis of Complaint	4.80	Vincent	0.40	(0.28)	+8.3%	111
		o1-preview	0.52*	(0.29)	+10.7%	111
Draft NDA	5.14	Vincent	0.06	(0.19)	+1.2%	111
		o1-preview	−0.42**	(0.17)	−8.2%	111
Draft Motion to Consolidate	3.50	Vincent	0.86**	(0.35)	+24.5%	112
		o1-preview	1.39***	(0.29)	+39.8%	112
Draft Persuasive Letter	4.11	Vincent	−0.54	(0.44)	−13.2%	111
		o1-preview	0.94**	(0.42)	+22.8%	111

Panel B: Control Variables

Variable	Coefficient	SE
GPA	0.36	(0.38)
3L Student (vs. 2L)	0.08	(0.27)
LLM Student (vs. 2L)	0.00	(0.00)
1-5 times (vs. 0 Times)	0.36	(0.37)
6-10 times (vs. 0 Times)	0.62	(0.42)
11-20 times (vs. 0 Times)	0.12	(0.67)
More than 20 times (vs. 0 Times)	0.33	(0.43)

Notes: Effects are shown relative to the No AI control group. Robust standard errors (in parentheses) are reported. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table 21: Treatment Effects on Professionalism Across Tasks

Task	Control Mean	Model	Effect	SE	% Change	N
Draft Client Email	4.50	Vincent	0.60	(0.44)	+13.3%	119
		o1-preview	1.04**	(0.41)	+23.1%	119
Draft Legal Memo	3.56	Vincent	0.68*	(0.35)	+18.9%	110
		o1-preview	1.47***	(0.32)	+41.4%	110
Analysis of Complaint	4.77	Vincent	0.69**	(0.30)	+14.5%	111
		o1-preview	1.06***	(0.32)	+22.1%	111
Draft NDA	5.68	Vincent	0.20	(0.24)	+3.4%	111
		o1-preview	−0.48*	(0.28)	−8.5%	111
Draft Motion to Consolidate	3.45	Vincent	0.51	(0.38)	+14.7%	112
		o1-preview	1.45***	(0.33)	+42.0%	112
Draft Persuasive Letter	4.31	Vincent	−0.35	(0.47)	−8.1%	111
		o1-preview	1.22***	(0.43)	+28.2%	111

Panel B: Control Variables

Variable	Coefficient	SE
GPA	0.45	(0.47)
3L Student (vs. 2L)	−0.13	(0.34)
LLM Student (vs. 2L)	0.00	(0.00)
1-5 times (vs. 0 Times)	0.54	(0.50)
6-10 times (vs. 0 Times)	0.76	(0.62)
11-20 times (vs. 0 Times)	0.36	(0.86)
More than 20 times (vs. 0 Times)	0.51	(0.58)

Notes: Effects are shown relative to the No AI control group. Robust standard errors (in parentheses) are reported. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table 22: Treatment Effects on Time Spent Across Tasks

Task	Control Mean	Model	Effect	SE	% Change	N
Draft Client Email	51.00	Vincent	-8.52^{***}	(2.93)	-16.7%	118
		o1-preview	-8.22^{***}	(2.56)	-16.1%	118
Draft Legal Memo	186.79	Vincent	-30.09^{**}	(12.25)	-16.1%	111
		o1-preview	-29.89^{**}	(13.80)	-16.0%	111
Analysis of Complaint	107.41	Vincent	-41.85^{***}	(5.63)	-39.0%	110
		o1-preview	-27.85^{***}	(6.80)	-25.9%	110
Draft NDA	90.30	Vincent	-1.77	(11.28)	-2.0%	111
		o1-preview	-10.16	(9.51)	-11.3%	111
Draft Motion to Consolidate	95.75	Vincent	-17.72^{**}	(8.55)	-18.5%	112
		o1-preview	-18.43^{**}	(9.00)	-19.2%	112
Draft Persuasive Letter	110.94	Vincent	-40.69^{***}	(8.74)	-36.7%	111
		o1-preview	-26.57^{***}	(9.08)	-23.9%	111

Panel B: Control Variables

Variable	Coefficient	SE
GPA	2.62	(3.28)
3L Student (vs. 2L)	-3.47	(2.29)
LLM Student (vs. 2L)	-0.00	(0.00)
1-5 times (vs. 0 Times)	4.01	(2.99)
6-10 times (vs. 0 Times)	7.49^{**}	(3.78)
11-20 times (vs. 0 Times)	-5.03	(5.92)
More than 20 times (vs. 0 Times)	3.52	(4.00)

Notes: Effects are shown relative to the No AI control group. Robust standard errors (in parentheses) are reported. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table 23: Treatment Effects on Total Score Across Tasks

Task	Control Mean	Model	Effect	SE	% Change	N
Draft Client Email	20.18	Vincent	1.76	(1.70)	+8.7%	120
		o1-preview	1.71	(1.61)	+8.5%	120
Draft Legal Memo	17.23	Vincent	2.15*	(1.24)	+12.5%	110
		o1-preview	3.44***	(1.21)	+19.9%	110
Analysis of Complaint	24.51	Vincent	1.61	(1.21)	+6.6%	111
		o1-preview	2.70**	(1.35)	+11.0%	111
Draft NDA	26.76	Vincent	0.08	(0.79)	+0.3%	111
		o1-preview	−1.78**	(0.91)	−6.7%	111
Draft Motion to Consolidate	17.57	Vincent	2.34*	(1.42)	+13.3%	112
		o1-preview	4.71***	(1.20)	+26.8%	112
Draft Persuasive Letter	20.03	Vincent	−1.77	(1.84)	−8.9%	111
		o1-preview	4.07**	(1.74)	+20.3%	111

Panel B: Control Variables

Variable	Coefficient	SE
GPA	3.06*	(1.81)
3L Student (vs. 2L)	0.09	(1.29)
LLM Student (vs. 2L)	0.00	(0.00)
1-5 times (vs. 0 Times)	3.64*	(1.92)
6-10 times (vs. 0 Times)	4.50**	(2.22)
11-20 times (vs. 0 Times)	2.74	(3.30)
More than 20 times (vs. 0 Times)	2.50	(2.27)

Notes: Effects are shown relative to the No AI control group. Robust standard errors (in parentheses) are reported. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table 24: Treatment Effects on Productivity Across Tasks

Task	Control Mean	Model	Effect	SE	% Change	N
Draft Client Email	0.41	Vincent	0.22***	(0.06)	+53.3%	118
		o1-preview	0.16***	(0.05)	+39.5%	118
Draft Legal Memo	0.10	Vincent	0.06***	(0.02)	+60.5%	109
		o1-preview	0.08***	(0.03)	+84.8%	109
Analysis of Complaint	0.24	Vincent	0.28***	(0.06)	+118.9%	110
		o1-preview	0.19***	(0.05)	+80.4%	110
Draft NDA	0.40	Vincent	0.06	(0.08)	+15.5%	111
		o1-preview	0.02	(0.07)	+6.0%	111
Draft Motion to Consolidate	0.24	Vincent	0.09	(0.05)	+36.7%	112
		o1-preview	0.18**	(0.07)	+75.2%	112
Draft Persuasive Letter	0.20	Vincent	0.17***	(0.05)	+85.3%	111
		o1-preview	0.27***	(0.09)	+139.2%	111

Panel B: Control Variables

Variable	Coefficient	SE
GPA	−0.02	(0.07)
3L Student (vs. 2L)	0.07	(0.05)
LLM Student (vs. 2L)	0.00	(0.00)
1-5 times (vs. 0 Times)	−0.01	(0.07)
6-10 times (vs. 0 Times)	−0.05	(0.07)
11-20 times (vs. 0 Times)	0.16	(0.18)
More than 20 times (vs. 0 Times)	−0.01	(0.08)

Notes: Effects are shown relative to the No AI control group. Robust standard errors (in parentheses) are reported. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

C. Grading Rubrics

General Rubric

1 Accuracy
How accurate and helpful was your research? 1<---------->7

- Factors
    - Cases
        - Did you include the relevant cases and properly state the holding?
        - Did you include any irrelevant cases?
        - Did you include any fictitious cases?
    - Facts
        - Did you include the relevant facts?
        - Did you include any irrelevant facts?
        - Did you include any fictitious "facts"?

2. Analysis
How sound and insightful was your analysis ?1<---------->7

- Factors
    - Logic
        - Were there any logical gaps in your reasoning?
        - Were your conclusions well-supported by the law?
        - Were your conclusions well-supported by the facts?
    - Nuance
        - Did you move beyond just a superficial analysis of the issues?
        - Did you address the key counterarguments?
        - Did you include any ideas that were both novel and helpful?

3. Organization
How organized and easy-to-follow was your work product? 1<---------->7

- Factors
    - Macro
        - Did you provide a coherent, user-friendly structure?
        - Did you prioritize the most relevant information?
        - Did you go off on any tangents?
    - Micro
        - Did you provide helpful headings?
        - Did each of your paragraphs build off the ones before it and set up the ones after it?
        - Did you end (or start) too abruptly?

4. Clarity

How clear and compelling was your writing? 1<---------->7

- Factors
    - Readability
        - Were your sentences clunky and convoluted?
        - Were your transitions awkward and choppy?
        - Did I have to read certain parts multiple times?
    - Polish
        - Did you have any typos?
        - Did you have any grammatical mistakes?
        - Did you have any punctuation mistakes?

5. Professionalism

How well did you follow the directions? 1<---------->7

- Factors
    - Word Count
        - Did you go over (or way under) the word count?
    - Document Type
        - Did you put together a persuasive memo when you were supposed to put together an objective memo (or vice versa)?
    - Time
        - Did you finish on time?

Specific Rubric for Assignment One

1 Accuracy

How accurate and helpful was your research? 1<---------->7

- Factors
    - Cases
        - Did you include the relevant cases and accurately state the holding?
        - Supreme Court
        - Tenth Circuit
        - Persuasive authority
        - Did you correctly identify the relative persuasiveness of the relevant cases?
        - Overwhelmingly against your client's preferred outcome
        - Did you include any irrelevant cases or fictitious cases?
    - Facts
        - Did you include the relevant facts?

- Defamation claim
- In litigation
- In actual filing
- Did you include any irrelevant facts or fictitious "facts"?

2. Analysis

How sound and insightful was your analysis ?1<---------->7

- Factors
  - Logic
    - Were there any logical gaps in your reasoning?
    - Were your conclusions well-supported by the law?
    - Did you avoid equivocation or offering false hope?
    - Were your conclusions well-supported by the facts?
  - Nuance
    - Did you move beyond just a superficial analysis of the issues?
    - Did you address the key counterarguments?
    - Did you distinguish cases not in your favor?
    - Did you include any ideas that were both novel and helpful?

3. Organization

How organized and easy-to-follow was your work product? 1<---------->7

- Factors
  - Macro
    - Did you provide a coherent, user-friendly structure?
      - Introduction, Analysis, Conclusion
      - Clear, firm conclusion
      - Unwavering analysis
    - Did you prioritize the most relevant information?
    - Did you go off on any tangents?
  - Micro
    - Did you provide helpful headings?
    - Did each of your paragraphs build off the ones before it and set up the ones after it?
    - Did you end (or start) too abruptly?

4. Clarity

How clear and compelling was your writing? 1<---------->7

- Factors
  - Readability
    - Were your sentences clunky and convoluted?
    - Were your transitions awkward and choppy?
    - Did I have to read certain parts multiple times?
  - Polish
    - Did you have any typos?
    - Did you have any grammatical mistakes?
    - Did you have any punctuation mistakes?

5. Professionalism

How well did you follow the directions? 1<---------->7

- Factors
  - Word Count
    - Did you go over (or way under) 700 words?
  - Document Type
    - Did you put together a persuasive brief when you were supposed to put together an objective memo?
  - Time
    - Did you exceed 1.5 hours working on this problem?

Specific Rubric for Assignment Two

1 Accuracy

How accurate and helpful was your research? 1<---------->7

- Factors
  - Cases
    - Did you include the relevant cases and accurately state the holding?
      - *Massachusetts Law*
        - *Vermont Mut. Ins. Co v. Poirier*
      - *New Hampshire Law*
        - *Wallace v. Nautilus*

        - Did you correctly identify the relative persuasiveness of the relevant cases?
            - *Poirier (Mass SJC interpreting Mass Law) > Nautilus (unpublished, district court, pre-Poirer)*
        - Did you include any irrelevant cases or fictitious cases?
        - Did you include the relevant statute?
            - *Ontario Courts of Justice Act, R.S.O. 19990, c. C.43*
        - Did you include any irrelevant or fictitious statute?
        - Did you include the relevant language from the insurance contract?
            - *"All costs taxed against the 'insured' in any 'suit' against the 'insured' we defend."*
        - Did you include any irrelevant or fictitious language from the insurance contract?
    - Facts
        - Did you include the relevant facts?
            - *Accident occurred in Ontario*
            - *Several passengers severely injured*
            - *Insurer agrees to pay $2 million to settle*
            - *Lawyers for victim wants extra 200,000 for attorneys' fees*
        - Did you include any irrelevant facts or fictitious "facts"?

2. Analysis

How sound and insightful was your analysis ?1<---------->7

- Factors
    - Logic
        - Were there any logical gaps in your reasoning?
        - Were your conclusions well-supported by the law?
        - Were your conclusions well-supported by the facts?
    - Nuance
        - Did you move beyond just a superficial analysis of the issues?
            - *Did you note that even under the logic of Poirer, the attorneys' fee may be covered?*
        - Did you address the key counterarguments?
            - *Did you push back on the analogy to Poirer?*
        - Did you distinguish cases not in your favor?
        - Did you include any ideas that were both novel and helpful?

3. Organization
How organized and easy-to-follow was your work product? 1<---------->7

- Factors
    - Macro
        - Did you provide a coherent, user-friendly structure?
            - *Introduction, Argument, Conclusion*
        - Did you prioritize the most relevant information?
        - Did you go off on any tangents?
    - Micro
        - Did you provide helpful headings?
        - Did each of your paragraphs build off the ones before it and set up the ones after it?
        - Did you end (or start) too abruptly?

4. Clarity
How clear and compelling was your writing? 1<---------->7

- Factors
    - Readability
        - Were your sentences clunky and convoluted?
        - Were your transitions awkward and choppy?
        - Did I have to read certain parts multiple times?
    - Polish
        - Did you have any typos?
        - Did you have any grammatical mistakes?
        - Did you have any punctuation mistakes?

5. Professionalism
How well did you follow the directions? 1<---------->7

- Factors
    - Word Count
        - Did you go over (or way under) 1500 words?
    - Document Type
        - Did you put together a persuasive brief when you were supposed to put together an objective memo?
    - Time
        - Did you finish in under 4 hours?

Specific Rubric for Assignment Three

1. Accuracy: How accurate and relevant was your summary and assessment of the complaint? Rating: 1 ←----------→ 7
    a. Factors
        i. Claims:
            1. Did you correctly summarize the key allegations and claims under the Fair Labor Standards Act (FLSA) and Nevada state law?
            2. Did you accurately reflect the nature of the class and collective action?
            3. Did you identify all the relevant legal provisions, such as 29 U.S.C. § 207, Nevada Revised Statutes (NRS) §§ 608.016, 608.018?
        ii. Facts:
            1. Did you capture the key facts relevant to the allegations (e.g., unpaid work before and after shifts, failure to pay overtime)?
            2. Did you avoid including irrelevant or immaterial facts?
            3. Did you identify any factual inconsistencies or missing details?

2. Analysis: How insightful and sound was your analysis of the strength of the claims? Rating: 1 ←----------→ 7
    a. Factors:
        i. Strength of Claims:
            1. Did you evaluate the strength of the claims based on the facts provided?
            2. Did you discuss potential weaknesses in the plaintiff's case?
            3. Did you consider any defenses the defendant might raise, such as arguments about tracking hours worked or defenses under FLSA and state law?
        ii. Nuance:
            1. Did your analysis consider potential counterarguments or challenges, like the practical difficulties of tracking unpaid work?
            2. Did you consider potential risks for the defendant and strengths for the plaintiff?

3. Did you include any novel or insightful ideas about how to defend against or settle the claims?

3. Organization: How well-organized and coherent was your memo? Rating: 1 ←----------→ 7
    a. Factors:
        i. Macro:
            1. Did you structure the memo logically, with clear sections (e.g., summary of allegations, analysis of claims, next steps)?
            2. Did you prioritize key information (e.g., strongest claims, biggest risks)?
            3. Did you avoid unnecessary digressions or irrelevant discussions?
        ii. Micro:
            1. Were your headings clear and helpful?
            2. Did your paragraphs flow logically from one to the next?
            3. Was your conclusion strong and appropriately placed?

4. Clarity: How clear and compelling was your writing? Rating: 1 ←----------→ 7
    a. Factors:
        i. Readability:
            1. Were your sentences concise and easy to follow?
            2. Were transitions between ideas smooth and logical?
            3. Did you avoid convoluted language that required rereading?
        ii. Polish:
            1. Did you avoid typos, grammatical errors, and punctuation mistakes?
            2. Was your writing professional and precise?

5. Professionalism: How well did you follow the assignment's guidelines? Rating: 1 ←----------→ 7
    a. Factors:
        i. Instructions:
            1. Did you summarize the key claims, assess the strength of the claims, and outline next steps as required by the assignment?

2. Did you avoid analyzing jurisdiction or venue, as instructed?
3. Did you keep the memo under the 1000-word limit?

ii. Time Management:

1. Did you complete the task within the two-hour time limit?

Specific Rubric for Assignment Four

1. Accuracy: How accurate and legally sound was your drafting of the nondisclosure agreement? Rating: 1 ←----------→ 7
    a. Factors:
        i. Legal Requirements:
            1. Did you correctly identify and include the necessary legal provisions relevant to Minnesota state law and other applicable laws governing NDAs?
            2. Did you ensure that the NDA complies with state-specific limitations on non-disclosure and non-compete clauses?
        ii. Protection of Trade Secrets:
            1. Did you accurately reflect Acme Co.'s interest in protecting its trade secrets and confidential information?
            2. Did you include relevant definitions, such as "Confidential Information," "Trade Secrets," and other key terms, that are both broad enough to protect Acme Co. and enforceable under the law?
        iii. Relevant Scope:
            1. Did you tailor the geographic and temporal scope of the restrictions to the company's business interests and the competitive landscape, ensuring enforceability?

2. Analysis: How sound and thoughtful was your approach to the drafting and enforceability of the NDA? Rating: 1 ←----------→ 7
    a. Factors:
        i. Risk Assessment:

1. Did you assess the potential risks to Acme Co. if Sue Scientist left for a competitor and draft the agreement accordingly?
2. Did you balance Acme Co.'s interests with the need to ensure that the NDA is not overly restrictive or vulnerable to being challenged in court?

ii. Strategic Choices:
1. Did you strategically choose provisions that favor Acme Co., such as the duration of confidentiality and non-compete clauses?
2. Did you demonstrate an understanding of how to protect the company without overreaching, ensuring enforceability in Minnesota and neighboring states?

iii. Customizations:
1. Did you move beyond simply copying the sample NDA and make thoughtful customizations specific to Acme Co. and Sue's role?

3. Organization: How well-organized and structured was your NDA draft? Rating: 1 ←----------→ 7
   a. Factors:
      i. Logical Flow:
         1. Did the agreement have a clear, logical structure, starting with definitions and progressing through confidentiality, non-compete, and assignment of inventions clauses?
         2. Were terms and obligations clearly organized so that each section built upon the one before it?
      ii. Prioritization:
         1. Did you focus on the most critical terms (e.g., confidentiality, non-compete, assignment of inventions) in a way that reflected Acme Co.'s priorities?
         2. Did you avoid unnecessary or irrelevant provisions that might make the NDA more complex than needed?
      iii. Formatting:
         1. Was the NDA formatted clearly, adhering to the required page length (three pages) and

formatting guidelines (single-spaced, 12-point font, 1" margins)?

4. Clarity: How clear and readable was your NDA? Rating: 1 ←----------→ 7
    a. Factors:
        i. Plain English Drafting:
            1. Did you follow the directive to use plain English, avoiding overly complex legal jargon?
            2. Was the language clear and understandable while maintaining legal enforceability?
        ii. Readability:
            1. Were sentences concise and straightforward?
            2. Did you avoid convoluted phrasing that could make the agreement difficult to understand for the average reader?

5. Professionalism: How well did you follow the assignment's instructions and professional drafting standards? Rating: 1 ←----------→ 7
    a. Factors:
        i. Instructions:
            1. Did you draft the NDA in accordance with the instructions, ensuring it was no more than three pages and properly formatted?
            2. Did you follow the directive to make the agreement favorable to Acme Co. while staying within enforceable legal limits?
        ii. Time Management:
            1. Did you complete the task within the three-hour time limit?

Specific Rubric for Assignment Five

1 Accuracy
How accurate and helpful was your research? 1<---------->7

- Factors
    - Cases
        - Did you include the relevant cases and accurately state the holding?
            - *Sinchuck v. Fullerton*

- *Shacter v. Richter*
- *Minnesota Person. Inj. Asbestos Cases v. Keene*
- *Anderson v. Connecticut*
- *Brooks Realty v. Aetna*

■ Did you include any irrelevant cases?
■ Did you include the relevant statute?
- *Minn. R. Civ. P. 42.01*

■ Did you include any irrelevant statute?
■ Did you include any fictitious case or statute?

○ Facts
■ Did you include the relevant facts?
- *Acme sued Beta in state court seeking declaratory judgment*
- *Owners of Beta sued Acme in different jurisdiction for fraud*
- *Same corse set of facts*
- *We present Beta in the first case and Beta's owners in the second*

■ Did you include any irrelevant facts?
■ Did you include any fictitious "facts"?

2. Analysis
How sound and insightful was your analysis ?1<---------->7

- Factors
  - Logic
    - Were there any logical gaps in your reasoning?
    - Were your conclusions well-supported by the law?
    - Were your conclusions well-supported by the facts?
  - Nuance
    - Did you move beyond just a superficial analysis of the issues?
    - Did you address the key counterarguments?
    - Did you distinguish cases not in your favor?
    - Did you include any ideas that were both novel and helpful?

3. Organization
How organized and easy-to-follow was your work product? 1<---------->7

- Factors
  - Macro
    - Did you provide a coherent, user-friendly structure?

- *Introduction, Argument, Conclusion*
    - Did you prioritize the most relevant information?
    - Did you go off on any tangents?
- Micro
    - Did you provide helpful headings?
    - Did each of your paragraphs build off the ones before it and set up the ones after it?
    - Did you end (or start) too abruptly?

4. Clarity
How clear and compelling was your writing? 1<---------->7

- Factors
    - Readability
        - Were your sentences clunky and convoluted?
        - Were your transitions awkward and choppy?
        - Did I have to read certain parts multiple times?
    - Polish
        - Did you have any typos?
        - Did you have any grammatical mistakes?
        - Did you have any punctuation mistakes?

5. Professionalism
How well did you follow the directions? 1<---------->7

- Factors
    - Word Count
        - Did you go over (or way under) 1000 words?
    - Document Type
        - Did you put together an objective memo when you were supposed to put together a persuasive brief?
    - Time
        - Did you finish in under 2.5 hours?

Specific Rubric for Assignment Six

1 Accuracy
How accurate and helpful was your research? 1<---------->7

- Factors
    - Cases
        - Did you include the relevant cases and accurately state the holding?
            - *Binding Indiana case law*

        - *Persuasive federal cases interpreting Indiana law*
    - Did you correctly identify the relative persuasiveness of the relevant cases?
    - Did you include any irrelevant cases or fictitious cases?
        - *All non-Indiana case law*
- Facts
    - Did you include the relevant facts?
        - *Restraint terms (three years; several specified counties, "private chef services")*
        - *Information and customer relations relevant to scope*
    - Did you include any irrelevant facts or fictitious "facts"?

2. Analysis
How sound and insightful was your analysis ?1<---------->7

- Factors
    - Logic
        - Were there any logical gaps in your reasoning?
            - *Included all three types of scope*
            - *Included direct solicitation prohibition*
        - Were your conclusions well-supported by the law?
        - Were your conclusions well-supported by the facts?
    - Nuance
        - Did you move beyond just a superficial analysis of the issues?
            - *Based analysis of scope types in measurable factors or protectable interests*
        - Did you address the key counterarguments?
        - Did you distinguish cases not in your favor?
        - Did you include any ideas that were both novel and helpful?
            - *Tolling, undue hardship, public harm*

3. Organization
How organized and easy-to-follow was your work product? 1<---------->7

- Factors
    - Macro
        - Did you provide a coherent, user-friendly structure?
            - *Introduction, Analysis, Conclusion*

        - *Separated Time, Geography, and Activity Restrained*
        - *Provided subparts as called for by the analysis*
    - Did you prioritize the most relevant information?
    - Did you go off on any tangents?
        - *Such as other CNTC elements, injunctive relief, or damages issues*
- Micro
    - Did you provide helpful headings?
    - Did each of your paragraphs build off the ones before it and set up the ones after it?
    - Did you end (or start) too abruptly?

4. Clarity

How clear and compelling was your writing? 1<---------->7

- Factors
    - Readability
        - Were your sentences clunky and convoluted?
        - Were your transitions awkward and choppy?
        - Did I have to read certain parts multiple times?
    - Polish
        - Did you have any typos?
        - Did you have any grammatical mistakes?
        - Did you have any punctuation mistakes?

5. Professionalism

How well did you follow the directions? 1<---------->7

- Factors
    - Word Count
        - Did you go over (or way under) 1500 words?
    - Document Type
        - Did you put together a persuasive brief when you were supposed to put together an objective memo?
    - Time
        - Did you exceed 2.5 hours of work on this problem?