AI ASSISTANCE IN LEGAL ANALYSIS: AN EMPIRICAL STUDY

Jonathan H. Choi* & Daniel Schwarcz**

Can artificial intelligence (AI) augment human legal reasoning? To find out, we designed a novel experiment administering law school exams to students with and without access to GPT-4, the best-performing AI model currently available. We found that assistance from GPT-4 significantly enhanced performance on simple multiple-choice questions but not on complex essay questions. We also found that GPT-4's impact depended heavily on the student's starting skill level; students at the bottom of the class saw huge performance gains with AI assistance, while students at the top of the class saw performance *declines*. This suggests that AI may have an equalizing effect on the legal profession, mitigating inequalities between elite and nonelite lawyers.

In addition, we graded exams written by GPT-4 alone to compare it with humans alone and AI-assisted humans. We found that GPT-4's performance varied substantially depending on prompting methodology. With basic prompts, GPT-4 was a mediocre student, but with optimal prompting it outperformed *both* the average student *and* the average student with access to AI. This finding has important implications for the future of work, hinting that it may become advantageous to entirely remove humans from the loop for certain tasks.

* Professor of Law, University of Southern California Gould School of Law.

** Fredrikson & Byron Professor of Law, University of Minnesota Law School.

Contents

INTRODUCTION

Within months of its public release in November 2022,[1] ChatGPT became the fastest-growing consumer application in history.[2] Legal scholars quickly embraced the trend; in the past nine months, academic studies have found that generative artificial intelligence (AI) performs well on law school exams,[3] the bar exam,[4] and other types of legal analysis.[5] Commentators have speculated that AI tools could soon replace lawyers entirely for many tasks, potentially both broadening access to justice and eliminating many legal jobs.[6]

---

[1] *Introducing ChatGPT*, OPENAI (Nov. 30, 2022), https://openai.com/blog/chatgpt.

[2] *See* Krystal Hu, *ChatGPT Sets Record for Fastest-Growing User Base*, REUTERS (Feb. 2, 2023), https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01.

[3] Jonathan H. Choi, Kristin E. Hickman, Amy B. Monahan, & Daniel Schwarcz, *ChatGPT Goes to Law School*, 72 J. LEG. ED. (forthcoming 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). For examples of work on the impact of ChatGPT in other fields, see Chung Kwan, *What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature*, 13 EDUC. SCI. 410 (2023); David A. Wood et al., *The ChatGPT Artificial Intelligence Chatbot: How Well Does It Answer Accounting Assessment Questions?*, 2023 ISSUES IN ACCOUNTING EDUC. 1; Gina Kolada, *When Doctors Use a Chatbot to Improve Their Bedside Manner*, N.Y. TIMES (June 13, 2023), https://www.nytimes.com/2023/06/12/health/doctors-chatgpt-artificial-intelligence.html; Steve Lohr, A.I. *May Someday Work Medical Miracles. For Now, It Helps Do Paperwork*, N.Y. TIMES (July 26, 2023), https://www.nytimes.com/2023/06/26/technology/ai-health-care-documentation.html; Harsha Nori et al., Capabilities of GPT-4 on Medical Challenge Problems (Apr. 12, 2023) (unpublished manuscript) (on file with authors); Alejandro Lopez-Lira & Yuehua Tang, Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models (May 12, 2023) (unpublished manuscript) (on file with authors).

[4] Daniel Martin Katz et al., GPT-4 Passes the Bar Exam (Apr. 5, 2023) (unpublished manuscript) (on file with authors). *But see* Eric Martínez, Re-Evaluating GPT-4's Bar Exam Performance (June 12, 2023) (unpublished manuscript) (on file with authors) (discussing potential methodological issues with the initial finding that GPT-4 surpassed the bar exam score of 90% of human test takers).

[5] John Ney et al., *Large Language Models as Tax Attorneys: A Case Study in Legal Capabilities Emergence*, 381 PHIL. TRANSACTIONS OF THE ROYAL SOC'Y A: MATHEMATICAL, PHYSICAL & ENG'G SCI. (forthcoming 2023).

[6] *See, e.g.*, JOSEPH BRIGGS ET AL., GOLDMAN SACHS, THE POTENTIALLY LARGE EFFECTS OF ARTIFICIAL INTELLIGENCE ON ECONOMIC GROWTH (2023) (suggesting that AI could automate nearly half of all legal tasks). David De

To date, these studies have narrowly focused on the performance of AI acting alone.[7] But at least in the near future, AI will likely *assist* humans rather than replacing them.[8] This is in part because AI is not yet sufficiently accurate[9] and in part because of legal and ethical qualms

---

Cremer, Nicola Morini Bianzino & Ben Falk, *How Generative AI Could Disrupt Creative Work*, HARV. BUS. REV., Apr. 13, 2023, https://hbr.org/2023/04/how-generative-ai-could-disrupt-creative-work; Daniel Schwarcz & Jonathan H. Choi, *AI Tools for Lawyers: A Practical Guide*, 108 MINN. L. REV. HEADNOTES (forthcoming 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. Scholars have also commented on other legal issues relating to ChatGPT, including intellectual property issues and issues surrounding contract interpretation. *See* Matthew Sag, *Copyright Safety for Generative AI*, 61 HOUSTON L. REV. (forthcoming 2024); David A. Hoffman & Yonathan Arbel, Generative Interpretation (July 31, 2023) (unpublished manuscript) (on file with authors). The potential for transformative AI tools in the legal industry is reflected in the hefty valuations of firms that roll out AI tools. *See, e.g.*, *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-tech-provider-casetext-650-mln-2023-06-27; Caroline Hill, *Investment: Harvey Raises $21m Series A and Dori Comes 'Out of Stealth Mode' with $2m Seed Raise*, LEGALTECHNOLOGY (Apr. 27 2023), https://legaltechnology.com/2023/04/27/investment-harvey-raises-21m-series-a-and-dori-comes-out-of-stealth-mode-with-2m-seed-raise.

[7] *See infra* Part I.

[8] *See, e.g.*, Gina Kolata, *When Doctors Use a Chatbot to Improve Their Bedside Manner*, N.Y. TIMES (June 12, 2023), https://www.nytimes.com/2023/06/12/health/doctors-chatgpt-artificial-intelligence.html; Lohr, *supra* note 3; Daniel Schwarcz & Jonathan H. Choi, *AI Tools for Lawyers: A Practical Guide*, 108 MINN. L. REV. HEADNOTES (forthcoming 2023).

[9] 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. *See also* Lance Eliot, *Lawyers Getting Tripped Up by Generative AI such as ChatGPT but Who Really Is to Blame, Asks AI Ethics and AI Law*, FORBES (May 29, 2023),

about removing humans from important legal judgments.[10] Existing scholarship about the performance of AI models in legal analysis ignores the interaction between humans and machines and therefore misses the most likely application of AI in the law.[11]

To fill this gap in the literature, we conducted an experiment to test the impact of AI assistance on legal reasoning. We recruited students in two law school classes at the University of Minnesota to take a second set of final exams with the assistance of GPT-4, the best-performing large language model (LLM) currently available.[12] Before doing so, participants received training on how to use GPT-4 effectively on an exam. By comparing their performance on this second set of exams (where they had access to GPT-4 and training on how to use it) with their performance on real exams (where access to GPT-4 was banned), we can explore how access to AI impacts human performance in professional work and educational settings generally and in legal contexts in particular.

We found that AI assistance had mixed results. For multiple-choice questions, it dramatically improved student results, producing a 29 percentile-point improvement relative to these students' performance

---

https://www.forbes.com/sites/lanceeliot/2023/05/29/lawyers-getting-tripped-up-by-generative-ai-such-as-chatgpt-but-who-really-is-to-blame-asks-ai-ethics-and-ai-law/?sh=6c8cc01c3212; Steve Lohr, *supra* note 3. *See also* Hussam Alkaissi & Samy I. McFarlane, *Artificial Hallucinations in ChatGPT: Implications in Scientific Writing*, 15 CUREUS J. MED. SCI. (forthcoming 2023).

[10] *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).

[11] The need for humans to work with AI rather than to allow AI to replace them has been understood long before ChatGPT was publicly released. *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). Since ChatGPT's release, the sentiment that humans must work with AI rather than simply allowing AI to replace them has become broadly accepted. *See e.g.*, 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*, JDSUPRA (Apr. 20, 2023), https://www.jdsupra.com/legalnews/ai-human-a-bright-future-for-legal-co-7452383.

[12] *See AlpacaEval Leaderboard*, GITHUB, https://tatsu-lab.github.io/alpaca_eval (last visited Aug. 5, 2023) (ranking GPT-4 as the best-performing model currently available).

on the real exam. However, AI assistance had no effect on student grades in the essay components of the two exams. Thus AI assistance seems to be most valuable in straightforward settings like multiple-choice questions and of little average value on difficult issue-spotter questions. Behind these average results, we also found significant variation in how useful AI assistance was to students depending on their baseline performance. The worst-performing students benefited enormously from AI, with gains of approximately 45 percentile points. By contrast, the best-performing students received *worse* grades when given access to AI, experiencing declines of approximately 20 percentile points.

We also tested how GPT-4 performed on its own using several different well-known prompting strategies, finding substantial variation in GPT-4's performance depending on the prompt. In a more basic law school class introducing undergraduates to legal analysis, AI alone achieved grades between B+ and A relative to a B+ median, representing performance at median or toward the top of the class. On an upper-level course covering advanced legal concepts, AI alone achieved grades between B and A- relative to a B+ median, representing solid performance but not at the top of the class. In both classes, "grounding" GPT-4 with relevant source material (in the form of instructor notes) caused it to obtain better grades than either human students alone or human students with access to GPT-4.

These findings have important implications for the future of lawyering. They suggest that AI assistance will most benefit those at the bottom of the skill distribution, potentially acting as an equalizing force in a notoriously unequal profession. But they also suggest that AI assistance might not be particularly useful on average in complex legal reasoning tasks (like essay-writing) that more closely resemble the difficult work of lawyering than multiple-choice questions. Finally, the fact that GPT-4 outperformed both humans and AI-assisted humans with optimal prompting suggests that AI might entirely remove humans from the loop for certain kinds of basic legal tasks. This may free up human lawyers for more enjoyable high-touch legal work, but it also may cause significant short-term disruption in the legal market.

I. EXISTING LITERATURE AND RESEARCH QUESTIONS

Since ChatGPT's release, dozens of research projects have evaluated the capacity of GPT models and other large language models (LLMs) to complete tasks traditionally performed by humans. Most have compared the performance of GPT-3.5 and humans on various exams.[13]

[13] We co-authored the most directly relevant study on GPT-3.5's performance in law school exams. Choi et al., *supra* note 3. We used GPT-3.5 to

And in the vast majority of cases, these studies have used few, if any, prompt-engineering strategies to evaluate GPT-3.5's capabilities, instead simply using the content of exam questions as the prompts.[14]

Because of heightened public interest in the performance of modern LLMs, a significant literature already exists on the usefulness of LLM models that are even more powerful than GPT-3.5. The most important such model is GPT-4, which OpenAI publicly released in March 2023.[15] Numerous recent studies have evaluated GPT-4's capabilities on medical exams,[16] generally finding that it outperforms

---

generate answers on final exams for classes on Torts, Employee Benefits, Tax Law, and Constitutional Law at the University of Minnesota Law School. Our study found that GPT-3.5's performance was highly uneven across different essay questions and types of multiple exam questions, but that on average it performed at the level of a C+ student. Additionally, the study found that GPT-3.5 achieved a low but passing grade in all four courses. *Id.* at *5-6. Common problems with ChatGPT-drafted exams included inaccurate, unreliable, and outdated information. *Id.* at *9-11.

See also Chung Kwan Lo, *What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature*, 13 EDUC. SCI. 410 (2023) (finding that GPT-3.5 had produced "outstanding" results in economics examinations but only middling results in computer programming and medical education, and "unsatisfactory" results in fields like mathematics and psychology); Lakshmi Varanasi, *ChatGPT Could Be a Stanford Medical Student, a Lawyer, or a Financial Analyst. Here's a List of Advanced Exams the AI Bot Has Passed So Far*, INSIDER (June 25, 2023), https://www.businessinsider.com/list-here-are-the-exams-chatgpt-has-passed-so-far-2023-1.

Finally, a related strand of literature has explored the possibility that LLM models like ChatGPT can operate as "smart readers" that allow individual consumers to better understand complex contracts. *See* Yonathan A. Arbel & Shmuel I. Becher, *Contracts in the Age of Smart Readers*, 90 GEO. WASH. L. REV. 83 (2022); Yonathan A. Arbel & Shmuel I. Becher, How Smart are Smart Readers? LLMs and the Future of the No-Reading Problem (June 25, 2023) (unpublished manuscript) (on file with authors).

[14] *See, e.g.*, Choi et al., *supra* note 3; Blair-Stanek et al., *supra* note 3; Katz et al., *supra* note 4.

[15] *See* Joshua J., *What Is the Difference Between the GPT-4 Models?*, OPENAI, https://help.openai.com/en/articles/7127966-what-is-the-difference-between-the-gpt-4-models (last visited Aug. 5, 2023) (describing GPT-4 and the different models available). GPT-4 consistently outperforms ChatGPT on most tasks. *See* GPT-4, OPENAI (Mar. 14, 2023), https://openai.com/research/gpt-4; OpenAI, GPT-4 Technical Report (Mar. 15, 2023) (unpublished manuscript) (on file authors). To be sure, there are exceptions to these results. For instance, OpenAI's own studies found that GPT-4 did not outperform GPT 3.5 in certain writing-related exams, such as the GRE writing exam and AP English and Composition Exam. *See id.*

[16] *See, e.g.*, Nori et al., *supra* note 3 (finding that GPT-4, without any specialized prompting passes a range of medical exams and out-performs both

average human test takers.[17] Other studies examining GPT-4's performance have produced similar results in fields ranging from logic, to engineering, to psychology.[18]

Not surprisingly, GPT-4 has also proven proficient at legal analysis. One prominent paper found that GPT-4 passed the Uniform Bar Examination (UBE).[19] Another recent study found that GPT-4

ChatGPT and LLM models specifically fine-tuned on medical knowledge); John C. Lin et al., *Comparison of GPT-3.5, GPT-4, and Human User Performance on a Practice Ophthalmology Written Examination*, EYE, May 8, 2023 ("GPT-4 but not GPT-3.5 achieved the passing threshold for a practice ophthalmology written examination"); Rohaid Ali et al., Performance of ChatGPT and GPT-4 on Neurosurgery Written Board Examinations (Mar. 25, 2023) (unpublished manuscript) (on file with authors) (finding that both GPT-4 and GPT-3.5 pass neurosurgery practice board exams at rates comparable to neurosurgery residents).

[17] *See supra* note 16. Some of these studies also find that GPT-4 outperforms other LLM models that are specifically fine-tuned on a subject matter specific corpus of data, like medical journal articles. *See* Nori et al., *supra* note 3.

[18] *See, e.g.*, Hanmeng Liu et al., Evaluating the Logical Reasoning Ability of ChatGPT and GPT-4 (May 5, 2023) (unpublished manuscript) (on file with authors); Vinay Pursnani, Yusuf Sermet & Ibrahim Demir, Performance of ChatGPT on the US Fundamentals of Engineering Exam: Comprehensive Assessment of Proficiency and Potential Implications for Professional Environmental Engineering Practice (Apr. 20, 2023) (unpublished manuscript) (on file with authors).

[19] Katz et al., *supra* note 4. This result extended both to the multiple-choice portion of the exam as well as to the open-ended essay components of the exam. *Id.* at *2. Although the authors did not use any prompt-engineering strategies to generate multiple choice answers, they slightly modified essay questions by presenting each sub-question in an independent prompt, and by "lightly correcting the language" in the prompt so that it formed a complete sentence. *Id.* at *7. Additionally, the grading of the essay questions was preliminarily performed by two of the studies' authors, and then validated by several additional peer reviewers who were provided with blinded samples of the essay questions. *Id.* at *7. It is not entirely clear from preliminary versions of the paper how effective these techniques may have been to avoid confirmation bias. For instance, it is unclear if the additional reviewers were also asked to grade human-produced exams such that their grading efforts could be evaluated for consistency.

OpenAI interpreted these results to suggest that GPT-4 outperformed 90% of human test-takers, OpenAI, GPT-4 Technical Report (March 27, 2023) (unpublished manuscript) (on file with authors), though this estimate is likely inflated because it compares the AI's performance to a population that was heavily skewed toward repeat test-takers who failed one or more prior administrations of the exam, Eric Martínez, Re-Evaluating GPT-4's Bar Exam Performance (June 12, 2023) (unpublished manuscript) (on file with authors).

outperformed GPT-3.5 on several law school exams at the University of Maryland, achieving grades on six different exams in the B and C ranges.[20] But these studies, while specific to legal exams, did not experiment with different prompting methods and crucially did not evaluate the performance of students with AI assistance.

Only a handful of studies so far have evaluated how AI can improve human performance at professional writing tasks.[21] None of these studies, moreover, focus on legal writing or other legal tasks, and none of them study the most advanced LLM available, GPT-4.[22] One study found that giving college-educated professionals access to GPT-3.5 substantially improved their performance at a variety of writing tasks, with the greatest gains going to the least-skilled workers.[23] A second study produced similar results, finding that access to an AI model improved the productivity of call center employees, again with the greatest gains accruing to the least skilled.[24]

---

[20] *See* Blair-Stanek et al., *supra* note 3. *See also* Margaret Ryznar, *Exams in the Time of ChatGPT*, 80 WASH. & LEE L. REV. ONLINE 305 (2023) (reporting mixed results).

[21] There are some recent papers that evaluate how access to generative AI can improve professionals' ability to perform non-writing tasks, like computer coding. *See* Sida Peng et al., The Impact of AI on Developer Productivity: Evidence from GitHub Copilot (Feb. 13, 2023) (unpublished manuscript) (on file with authors).

[22] Additionally, none of these studies evaluate how more sophisticated prompting techniques can impact results.

[23] Shakked Noy & Whitney Zhang, *Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence*, 381 SCIENCE 187, 187 (2023). To reach this conclusion, the experimenters recruited over 400 participants in five professional categories: grant writers, consultants, data analysts, human resource professionals, and managers. Participants were then tasked with completing two short writing assignments comparable to those they would complete in their professional settings, such as drafting press releases, short reports or emails. After completing the first writing assignment, half of the participants were given access to ChatGPT for the second writing assignment. The study found that participants who were provided with access to ChatGPT completed their writing tasks faster and produced higher quality work than participants who were not provided access to this tool. Moreover, the participants who performed relatively poorly on the initial task (which took place prior to being instructed how to use ChatGPT) disproportionately benefited from access to AI, receiving both higher quality scores and taking decreased amounts of time to complete their writing task. By contrast, access to ChatGPT did not improve the quality of work for participants who scored well in the initial writing task, though it did increase the speed at which they could produce that work.

[24] Erik Brynjolfsson, Danielle Li & Lindsey R. Raymond, *Generative AI at Work* (Nat'l Bureau of Econ. Rsch., Working Paper No. 31161, 2023). This

Other empirical work has suggested that AI assistance can sometimes have negative consequences. One study found that giving professional recruiters access to high-quality AI-generated predictions actually resulted in less accurate assessments of job applications than did access to low-quality AI generated predictions. [25] This counter-intuitive result was driven by the tendency of the human recruiters who received lower quality AI-assistance to exert greater effort to individually evaluate resumes.[26] A second recent study found that giving radiologists access to AI did not improve their diagnostic accuracy, even though AI alone was a more accurate diagnostician than humans.[27] The principal explanation was that human radiologists disregarded the AI's analysis when its conflicted with their initial judgments.[28]

This empirical work on human use of LLMs builds on an older literature on human-machine interaction. Scholars have long hypothesized that while humans might outperform machines at some tasks and machines might outperform humans at others, we should generally expect the best performance of all from a combination of human and machine.[29] For example, human chess players generally outperformed computers until Deep Blue famously defeated world chess champion Gary Kasparov in 1997;[30] today, even the strongest human

---

study focused on a large software firm's introduction of a chatbot, which used similar technology to ChatGPT, to support the work of its customer service agents by providing suggested (but not required) responses to customer questions. It found that this technology increased the rate at which customer service agents were able to successfully resolve customer questions, with most productivity gains being experienced by the least skilled and experienced agents. By contrast, the productivity of relatively skilled and experienced agents was largely unaffected by the introduction of the generative AI technology, suggesting that its primary effect was to help lower performing employees to operate more like their high-performing coworkers.

[25] Fabrizio Dell'Acqua, Falling Asleep at the Wheel: Human/AI Collaboration in a Field Experiment on HR 1 (Dec. 2, 2021) (unpublished manuscript) (on file with authors).

[26] *Id.*

[27] Nikhil Agarwal et al., *Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology* 1 (Nat'l Bureau of Econ. Rsch., Working Paper No. 31422, 2023).

[28] *Id.*

[29] *See, e.g.*, Michael A. Peshkin et al., *Cobot Architecture*, 17 IEEE TRANSACTIONS ON ROBOTICS & AUTOMATION 377, 377-90 (2001) (proposing a framework for human-robot collaboration); Shirine El Zaatari et al., *Cobot Programming for Collaborative Industrial Tasks: An Overview*, 116 ROBOTS & AUTONOMOUS SYSTEMS 162, 162-80 (2019) (surveying attempts at human-robot collaboration).

[30] George Johnson, *Deep, Deeper, Deepest Blue*, N.Y. TIMES, May 18, 1997, at A6.

players cannot compete with machines.[31] But machine chess programs managed by humans still outperform both humans and machines.[32]

The literatures described above leave some substantial unanswered questions. Most notably, these studies fail to answer the most salient question, namely how useful AI will be at helping humans rather than replacing them. The studies that have focused on the benefits of AI assistance to humans have not studied legal analysis or an exam setting, and they have focused on GPT-3.5, a prior generation of LLM that is qualitatively different and substantially less useful than GPT-4. Additionally, scholars studying how well AI models tackle legal analysis have generally not employed different prompting techniques, potentially leading them to underestimate AI's real-world potential.

In this Article, we add to this existing literature by focusing on the following unanswered research questions. How well do humans, AI, and AI-assisted humans conduct legal analysis? Does their performance vary by type of analysis, and does the benefit of AI assistance vary depending on baseline human performance? Does the answer depend on whether we consider the quality of output or speed of output, and are these related? And how much does prompting matter in determining the performance of AI in conducting legal analysis?

II. DATA AND METHODS

To answer these questions, we recruited students from two different law school classes to participate in an experiment. This Part describes these study participants in more detail. It also describes our basic experimental design, as well as the training that we provided to prepare students to effectively use AI to complete the required tasks.[33] Finally, it describes the prompting methods we used to evaluate GPT-4's performance on its own.

A. Study Design

We conducted our study using real examinations from two classes at the University of Minnesota Law School taught by one of the co-authors. The University of Minnesota Law School is one of the top law

[31] H. Jaap van den Herik, *Computer Chess: From Idea to Deepmind*, 40 ICGA J. 160, 174 (2018).

[32] *See* Martin Fischer, *Better than an Engine: Leonardo Ljubicic*, CHESSNEWS (Feb. 21, 2016), https://en.chessbase.com/post/better-than-an-engine-leonardo-ljubicic-1-2 (interviewing one correspondence chess player who uses chess AI to outplay AI on its own).

[33] This experiment received approval from the University of Minnesota's institutional review board. IRB #STUDY00019012.

schools in the country, currently ranked 16th in the U.S. News ranking of law schools.[34] The two courses we studied were Introduction to American Law and Legal Reasoning, and Insurance Law. Introduction to American Law and Legal Reasoning is an introductory class offered to undergraduates at the University of Minnesota, covering selected topics in Contract Law, Tort Law, Criminal Law, Civil Procedure, Property Law and Constitutional Law. Although the class is geared toward undergraduates, it is taught in much the same ways as a typical law school class and is principally intended to develop students' basic legal reasoning skills. Insurance Law is a more typical law school class that is taught principally to upper-level 2L and 3L University of Minnesota Law students.[35]

To evaluate the effect of access to AI on student exam performance, we recruited students who were taking either of these classes in the Spring of 2023. 16 out of 52 students taking Insurance Law (roughly 31%) completed the study, while 32 out of 128 students taking Introduction to American law (roughly 23%) completed the study.[36]

To participate in the study, students first completed a one-hour online training course that we developed and taught on how to use GPT-4 effectively in legal analysis.[37] The course drew heavily on our previous

[34] *2023 Best Law Schools*, U.S. NEWS, https://www.usnews.com/best-graduate-schools/top-law-schools/law-rankings (last visited Aug. 5, 2023).

[35] Masters students and undergraduates who complete Law 3000 are also able to take Insurance Law with the Instructor's permission. One study participant for Insurance Law was an undergraduate, though the rest were upper-level law students. The main results in this study remain unchanged when this single undergraduate is excluded from the sample.

[36] Attrition between recruitment and experiment completion was relatively low for both classes. In Insurance Law, 18 students responded to recruiting emails, and 16 in fact completed the study. For Introduction to American Law and Legal Reasoning, 39 students indicated they would be part of the study, and 32 students ultimately completed the study.

[37] In addition, participants in the Introduction to American Law class completed a second 45-minute training module on how to use GPT-4 to answer multiple choice questions, which also drew on our prior work and tasked participants to practice their AI-skills using sample questions. With respect to the multiple-choice questions, we reminded students that GPT-4 can hallucinate, crowd-out independent thinking, and employ knowledge or sources that were outside of the scope of the class. We then encouraged participants to (1) Ask GPT-4 what the right answer to the question is assuming the question is not entirely dependent on the specific way that the class was taught; (2) While GPT-4 generates an answer, attempt to answer the question on your own without looking at the GPT-4 answer; (3) If your answer and GPT-4's answer match, move on; (4) If you gave a different answer than GPT-4's answer, evaluate whether GPT-4's answer is likely to be correct by asking GPT-4 why the answer you chose is incorrect and assessing your own confidence in your own

work, *AI Tools for Lawyers: A Practical Guide*.[38] It focused on how to apply active lawyering skills while using AI, rather than mechanically relying on the output of GPT-4. For example, we instructed participants to structure essays on their own and to identify the best answers to multiple choice questions while waiting for GPT-4 to generate answers.[39] We also provided participants with some training on basic prompting techniques, such as encouraging them to ask the AI questions that break down legal analysis into pieces (somewhat analogous to chain-of-thought prompting, discussed below) and encouraging them to supply GPT-4 with relevant legal rules (analogous to grounded prompting, discussed below).[40] Additionally, the training required participants to practice these skills by using GPT-4 to answer sample problems. After participants completed this training on how to use AI effectively, they also spent one hour reviewing the substantive material from the class.[41]

After their training and review, study participants used GPT-4 to help them take a portion of the real exams that were administered in the

---

answer; and (5) Keep track of time – don't let use of GPT-4 slow you down too much!

[38] Schwarcz & Choi, *supra* note 6.

[39] More specifically, the key points that we emphasized in this training video were that AIs can hallucinate, will not know the scope of material covered in class, and can crowd out independent thinking. For these reasons, we encouraged participants to consider the following tips: (1) First read the essay question and generate your own basic outline of issues; (2) Ask GPT-4 to identify all relevant issues raised by the fact pattern; if the essay identifies specific issues to address in its prompt, see if GPT-4 can break those issues down into sub-issues or identify any additional issues; (3) For each issue, ask GPT-4 to analyze that issue by applying the relevant legal rule to the facts. To the extent that you can easily access and identify the relevant rules provided in class, provide those rules to GPT-4; (4) Ask GPT-4 to provide the best arguments on both sides of the relevant issue, so you can use this material to generate counter-arguments; (5) Ask GPT-4 to analogize to relevant cases to the extent you can identify such cases and copy/paste brief explanations of those cases into the prompt; (6) Repeat steps 3-5 for each subsequent issue; (7) Assemble material produced above into a final essay question, remembering to keep track of time and to focus on the goal of producing a good essay question that uses an IRAC-style structure, precise rules, and thesis sentences, while highlighting key facts from the problem and making strong counter-arguments. Ultimately, we emphasized that the goal was for students to use GPT-4 as a tool to help them draft an excellent answer.

[40] *See* Schwarcz & Choi, *supra* note 6.

[41] To replicate the conditions in the real classes, we specifically instructed participants in the Introduction to American Law class to review the materials on Property Law and Civil Procedure, which would be tested in the essay question on their exam.

same classes in the prior year (Spring 2022).[42] These exams were not publicly available and had not been previously provided to students in the Spring 2023 versions of their classes. The exams tested different specific material than that covered on the students' real 2023 exams. However, the tested material was covered in the 2023 classes at a comparable level of depth as the material tested on the real 2023 exam. We required all study participants to complete the experiment within two weeks after they finished their actual final exam in their classes, to minimize memory decay of the course material.[43]

In addition to producing AI-assisted human exams, we also generated four exams that were produced by AI alone. Each of these four AI-only exams was produced using one of four distinct prompting methods described in the following Section.[44]

Exams produced by AI-assisted humans and by AIs alone were then mixed with exams written by real students in 2022 without AI assistance, and blindly graded.[45] Blind grading helped to ensure that our preconceptions would not affect the results of the study.[46] In addition to

[42] Most people can access GPT-4 by creating a paid ChatGPT Plus account on the OpenAI website. However, it was not administratively possible to create such an account for each study participant without requiring participants to outlay cash on the subscriptions themselves. We instead created a central ChatGPT "clone" website using the GPT-4 API and gave students access to that website. This clone website had a nearly identical user interface and used the same system prompt as the real ChatGPT Plus.

[43] Memory decay is a cognitive process where information stored in the brain is lost over time due to lack of use or retrieval. Although the precise speed of memory decay can vary greatly over time, research suggests that even a spaced review of the underlying material can significantly counteract memory decay. *See* Siddharth Reddy et al., *Unbounded Human Learning: Optimal Scheduling for Spaced Repetition*, *in* PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 1815 (2016).

[44] *See infra* Section II.B.

[45] For Introduction to American Law, as with the real class, upper-level law students did the grading. In particular, they blindly graded a mix of 32 real human exams from 2022, 32 exams written by humans with AI assistance as part of our study, and four AI-only exams. To do so, they used the same rubric that teaching assistants employed to grade the 2022 essay, which they first reviewed and applied to sample answers with the assistance of the instructor (as in the real class). For Insurance Law, the instructor blindly graded a mix of 16 exams produced by humans only in 2022, 16 exams written by humans with AI assistance as part of our study, and four exams written solely by AI.

[46] *See* Andrea A. Curcio, Gregory Todd Jones & Tanya M. Washington, *Does Practice Make Perfect? An Empirical Examination of the Impact of Practice Essays on Essay Exam Performance*, 35 FLA. ST. U. L. REV. 271, 291 (2008)

the grading data that we produced after participants' completion of the experiment, we also collected data on student-participants' performance on their actual final exams in 2023 as well as the exam performance of all students who took the 2022 exams.

By re-grading old essay exams from 2022, we were able to measure consistency of grading between years and correct for any discrepancies.[47] The correlation between grades assigned on these exams in 2022 (as part of the real classes) and in 2023 (as part of our experiment) for Introduction to American Law was 0.77, moderately high. For Insurance Law, this correlation was much higher, at 0.94.[48] (Introduction to American Law is graded by student research assistants while Insurance Law is graded by the professor teaching the course, which explains the difference in levels of consistency.) We then calculated the average difference between the re-graded essays year over year and applied this as a "correction factor" to account for systematically harsher or more lenient grading in 2023 as compared to 2022. The correction factor for Introduction to American Law essays was -1.42 points out of 53 points, and for Insurance Law it was +1.375 points out of 50 points.[49]

B. Prompting Techniques

As described in Part I, the existing literature evaluating AI performance on exams uses basic prompts, essentially taking an exam question and feeding it directly to a tool like ChatGPT.[50] However, the computer science literature has developed several prompt-engineering techniques that might substantially improve the performance of tools like GPT-4.[51] We applied several such techniques in our study, both to

(relying on blind grading of exams by instructors to empirically study the result of different pedagogical innovations in law school).

[47] The multiple-choice component of the Introduction to American Law exam was graded mechanically and therefore required no regrading. Prior empirical research on law school exams has similarly used blind regrading of prior exams to ensure consistency. *See* Carol Springer Sargent & Andrea A. Curcio, *Empirical Evidence That Formative Assessments Improve Final Exams*, 61 J. LEGAL EDUC. 379, 389 (2012).

[48] Both correlations are Spearman's correlation coefficient.

[49] Note that the magnitude of the correction factor is orthogonal to the magnitude of the correlation because Spearman's correlation measures only whether rank-order is consistent between re-gradings. For example, if upon regrading each exam received a score exactly 5 points higher, correlation would still be perfect.

[50] *See supra* Section I.A.

[51] *See, e.g.*, Dils, *How to Use ChatGPT: Advanced Prompt Engineering*, WGMI MEDIA (Jan. 31, 2023), https://wgmimedia.com/how-to-use-chatgpt-advanced-prompt-engineering; *Awesome ChatGPT Prompts*, GITHUB,

test the performance of AI acting alone and to better understand how human variation in prompting strategies can impact results. Specifically, we evaluated the performance of basic prompting, "chain-of-thought" prompting, "few-shot" prompting, and "grounded" prompting.

The basic prompt we used was simply copying and pasting the exam question directly from the exam.[52] This is the simplest prompting technique, consistent with those used in past studies.[53] A chain-of-thought prompt asks the AI model to "think step-by-step" prior to producing its result.[54] Few-shot prompting involves providing the AI model with examples of good responses that it can use to shape its response.[55] For few-shot prompting, we gave GPT-4 questions and model answers from prior exams that had been provided to students taking each class. Finally, grounded prompting involves providing the AI model with relevant sources.[56] In this case, we provided GPT-4 with excerpts from the lecture notes used to teach the underlying class. The lecture notes addressed the course material tested in the relevant exam questions, although they were not formatted for easy reading and therefore would potentially have been difficult for a human to apply.

Specific language for each of the prompts used, and additional information about GPT-4 specifications, is in Section A.4 of the Appendix.

III. RESULTS

---

https://github.com/f/awesome-chatgpt-prompts/#readme; Alan D. Thompson, *Microsoft Bing Chat (Sydney/GPT-4),* LIFE ARCHITECT (Feb. 22, 2023), https://lifearchitect.ai/bing-chat (quoting the prompts used to convert GPT's base model into Bing Chat); Tyler Cowen & Alexander T. Tabarrok, How to Learn and Teach Economics with Large Language Models, Including GPT (Mar. 17, 2023) (unpublished manuscript) (on file with authors). *See also AI and Machine Learning Experts, Experienced Attorneys, Thousands of Hours of Prompt Engineering—and That's Just to Launch*, CASETEXT (May 12, 2023), https://casetext.com/blog/building-an-ai-legal-assistant-lawyers-can-trust.

[52] For multiple choice questions, we also instructed the model to "Answer the following multiple-choice question from a law school exam."

[53] *See supra* Part I.

[54] There are many variations on chain-of-thought prompting, but the one we used was based on the most promising version from the prior literature. Jason Wei et al., *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*, *in* PROCEEDINGS OF THE 36TH INTERNATIONAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS 4356 (2022).

[55] Tom B. Brown et al., *Language Models Are Few-Shot Learners*, *in* ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 33 (2020).

[56] *See* Baolin Peng et al., Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback (Mar. 8, 2023) (unpublished manuscript) (on file with authors).

We found that access to AI substantially improved average student performance on multiple-choice questions but did not substantially improve average performance on essay questions. Beyond these averages, GPT-4 substantially improved the scores of students at the bottom of the class and negatively impacted the scores of students at the top of the class. The performance of GPT-4 alone varied substantially depending on exam and prompting methodology. In general, the AI performed best when grounded with source materials. It also performed better on relatively simple multiple-choice questions than on relatively complex essay questions.

A. Quantitative Results

1. Percentile Performance

Table 1 shows the relative performance of humans alone, AI alone, and AI-assisted humans on the various exam components. Results are reported separately for the Introduction to American Law multiple-choice questions, the Introduction to American Law essay question (which tested a single straightforward legal issue), and the Insurance Law exam (which featured two complex issue spotter questions and no multiple-choice questions). All results are stated in percentile points relative to a baseline of performance set by a sitting of the exam by human exam-takers who did not have access to AI. The percentiles in the "Human-Only Mean" column reflect the mean scores for the subset of students who volunteered for our study relative to the entire class in 2023. These data show that the average exam scores of the students who completed the study were similar to those of the students taking the class as a whole, though volunteers from Introduction to American Law performed slightly above average on the multiple-choice portions of their real 2023 exams as compared to the class as a whole.

For the "AI-Assisted Mean" column, the percentiles were calculated by evaluating where the exams produced by students given access to AI would have scored relative to all the students who completed this exam in 2022 without access to (and prior to the creation of) ChatGPT. These data reveal that, on average, access to AI had little impact on how well participants performed on either the Insurance Law exam or the essay portion of Introduction to American Law. By contrast, participants' access to GPT-4 substantially improved their ability to answer the multiple-choice questions in Introduction to American Law.

Each cell in the table includes both the value listed and a 95% confidence interval.[57]

Table 1: Performance of Humans and AI-Assisted Humans on Exams

	Intro to Am Law - Multiple Choice	Intro to Am Law - Essay	Insurance Law
Human-Only Mean	59.1 (49.9, 67.3)	49.5 (39.1, 59.2)	47.5 (33.4, 63.3)
AI-Assisted Mean	88.0 (81.0, 92.8)	53.0 (43.3, 62.1)	46.3 (32.2, 58.1)
Mean Change	+28.9 (+20.3, +38.4)	+3.5 (-6.7, +15.3)	-1.3 (-9.7, +7.7)
Human-Only Standard Deviation	25.8 (21.4, 31.7)	29.4 (25.1, 34.9)	31.6 (25.2, 39.6)
AI-Assisted Standard Deviation	16.9 (12.5, 20.9)	28.0 (23.8, 32.8)	27.2 (21.9, 33.8)
Standard Deviation Change	-8.9 (-16.1, -2.7)	-1.4 (-8.2, +5.2)	-4.4 (-12.4, +3.0)

Figure 1 through Figure 3 below show the mean performance of humans and AI-assisted humans on each exam component (in each case relative to students without AI assistance who took the same exam). Each curve includes a mean (solid lines) and a 95% confidence interval (dashed lines).

[57] All distributions, confidence intervals, and significance tests in this Article were conducted by bootstrapping. *See* B. Efron, *Bootstrap Methods: Another Look at the Jackknife*, 7 ANNALS OF STAT. 1 (1979) (proposing bootstrap methods). I used 50,000 bootstrap iterations to generate the graphs in this Article.

Figure 1: Introduction to American Law Multiple Choice – Mean Performance

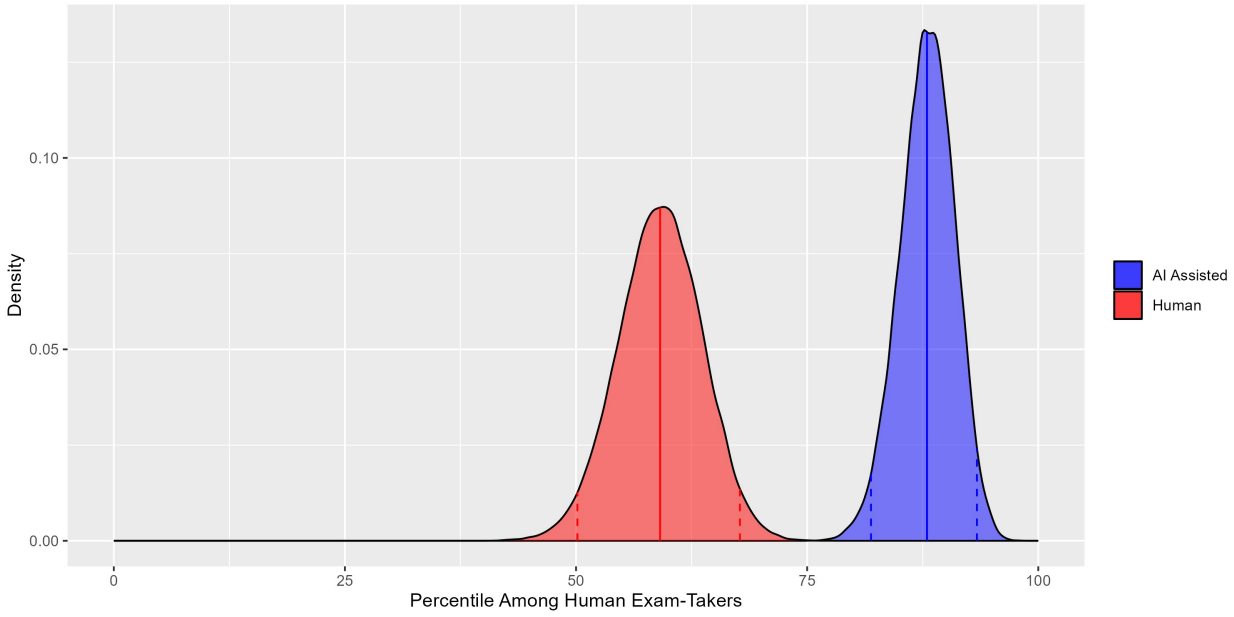


Figure 2: Introduction to American Law Essay – Mean Performance

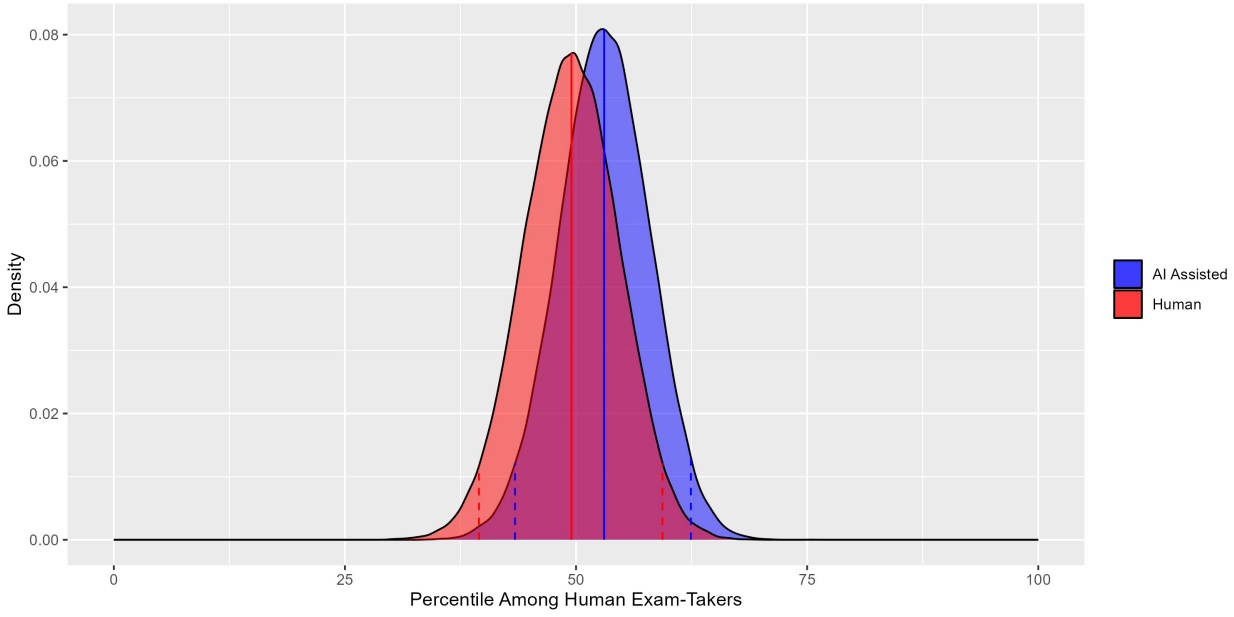
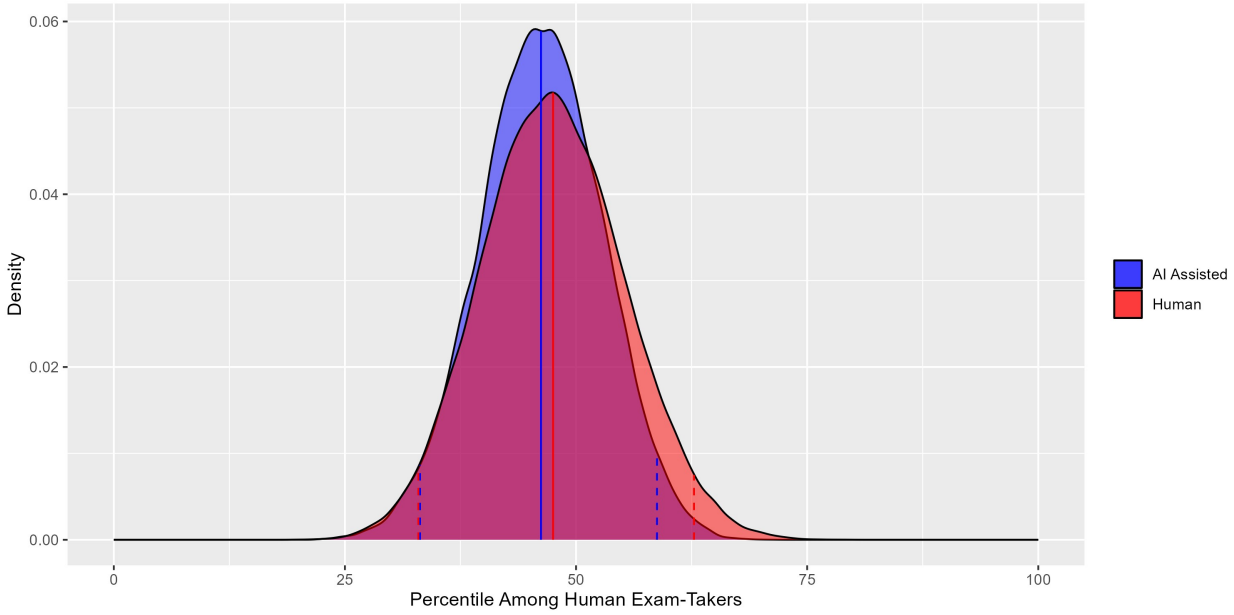


Figure 3: Insurance Law – Mean Performance



These Figures show that mean performance significantly increased for the multiple-choice questions but did not significantly increase for the essay questions.

In addition to changes in *mean* performance, we should also consider whether access to AI helps some students more than others. Consistent with the nascent literature on AI-enhanced professional writing outside of the legal setting,[58] we might hypothesize that access to AI would substantially help low-performing students by establishing a floor (the AI's own performance) below which their performance should not fall. In contrast, we might imagine that high-performing students might not benefit from AI nearly as much as low-performing students because their innate skills already surpass the AI's capabilities. Even further, at least one study suggests that high performers might be harmed by access to AI by becoming less creative or using the AI as a crutch.[59]

Figure 4 addresses this issue by showing the per-student performance improvement when given access to AI (on the *y*-axis) relative to unassisted student performance. Each dot represents one student, and the curve represents the line-of-best-fit with 95% confidence intervals estimated by bootstrapping.

[58] *See supra* Section I.B.
[59] *See id.*

Figure 4: All Exams – Performance Improvement with Access to AI Relative to Baseline Performance

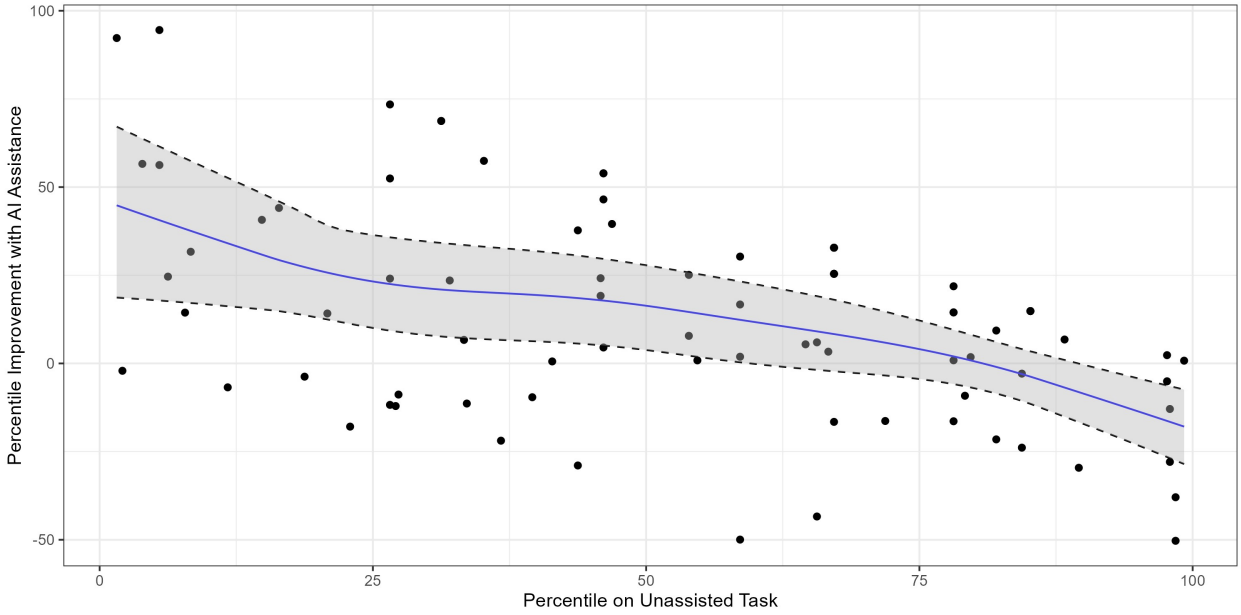


Figure 4 corroborates the hypothesis mentioned above. It suggests that the lowest-performing students see large and statistically significant improvements when given access to AI, whereas the best-performing students see moderate and statistically significant declines when given access to AI. Note that the results in Figure 4 may largely be driven by mean reversion—that is, if student exam results are randomly drawn from some distribution, we would expect that students who had underperformed in one sitting of the exam would outperform in the next sitting relative to their prior sitting. These results are therefore most meaningful to the extent that student performance on exams is consistent between sittings.[60] Figure 15 through Figure 17 in the

[60] In theory, we could either parametrically or non-parametrically (based on the empirical distribution of exam scores in our sample) estimate both a baseline of the expected amount of mean reversion between exam sittings and the degree to which our results differed from this baseline. However, constructing the baseline would require prior probabilities of the parallel forms reliability (also known as equivalent forms reliability) of law school exams, meaning the correlation between a student's scores in different sittings of the same exam type with different specific questions. To our knowledge, no empirical research on the parallel forms reliability of law school exams currently exists. We therefore present Figure 4 as is, noting that if one interprets Figure 4 to say that we can reject the null hypothesis for certain levels of human-only performance that access to AI does not affect performance, this interpretation crucially assumes perfect parallel forms reliability.

Appendix provide additional detail on the effect of AI assistance on the variation in performance between students.

Table 2 shows the performance of GPT-4 alone on each of the exam components, using four different prompting techniques.[61] Again, the results are percentile points relative to the performance of humans in a prior exam sitting. Note first that GPT-4's improvements based on prompting were consistent across settings: the basic prompt consistently performed the worst, with some improvement for the chain-of-thought prompt, more improvement for the few-shot prompt, and optimal performance for the grounded prompt.

Focusing on exam type rather than prompting strategy, Table 2 shows that the performance of GPT-4 alone varied widely. For the multiple-choice questions in the Introduction to American Law exam, GPT-4 alone consistently performed well above the median human test taker, consistent with its performance in other multiple choice legal settings, like the bar exam.[62] This was true with all four prompting strategies, and with few-shot and grounded prompting it received a perfect score. GPT-4's capacity to correctly answer the types of multiple-choice questions contained in the Introduction to American Law class explains why participants' access to GPT-4 substantially improved their performance on this component of the exam: It makes sense that human performance would improve more when given access to a high-performing tool.[63]

For the essay exams, GPT-4's performance varied significantly both by prompting strategy and by exam type. Overall, GPT-4 performed significantly better on the relatively straightforward Introduction to American Law essay than it did on the much more complex and specialized Insurance Law essays. With respect to the former, all four prompting strategies resulted in exams that scored above the class mean, and the best AI model – which, recall, grounded its answer in the relevant instructor lecture notes for the class – would have performed near the top of the Introduction to American Law class. For Insurance Law, by contrast, basic prompting produced an exam toward the bottom of the grade distribution.[64] Notably, however, improved prompting strategies generated progressively better results in Insurance Law, with grounded prompting allowing GPT-4 to produce an exam that would have performed moderately above-average in the class.

---

[61] *See supra* Section II.B.

[62] *See* Katz et al., *supra* note 4.

[63] *But see* Agarwal et al., *supra* note 27 (finding that the diagnostic quality of radiologists does not improve even when they are given access to an AI tool that outperforms humans on its own).

[64] This is consistent with earlier results showing that GPT-3.5 averaged a C+ on several different law school exams. *See* Choi et al., *supra* note 3.

Table 2: Performance of GPT-4 on Exams

Prompting Method	Intro to Am Law - MC	Intro to Am Law - Essay	Insurance Law
Basic	79	60	5
Chain of Thought	79	56	15
Few Shot	100	56	30
Grounded	100	93	65

Figure 5 through Figure 7 elaborate by showing the relative performance of humans alone (the red density plots with mean performance represented by the red vertical line), humans with access to AI (the blue density plots with mean performance represented by the blue vertical line), and AI alone (the dashed vertical lines) on each component of the exams. The graphs are density plots showing performance relative to all human exam-takers. Note that as with the tables above, the density plots include only students who participated in our study, not all students who took the relevant exams without the assistance of AI, who we use as the control group to calculate percentiles.

Figure 5: Introduction to American Law Multiple Choice – Raw Performance

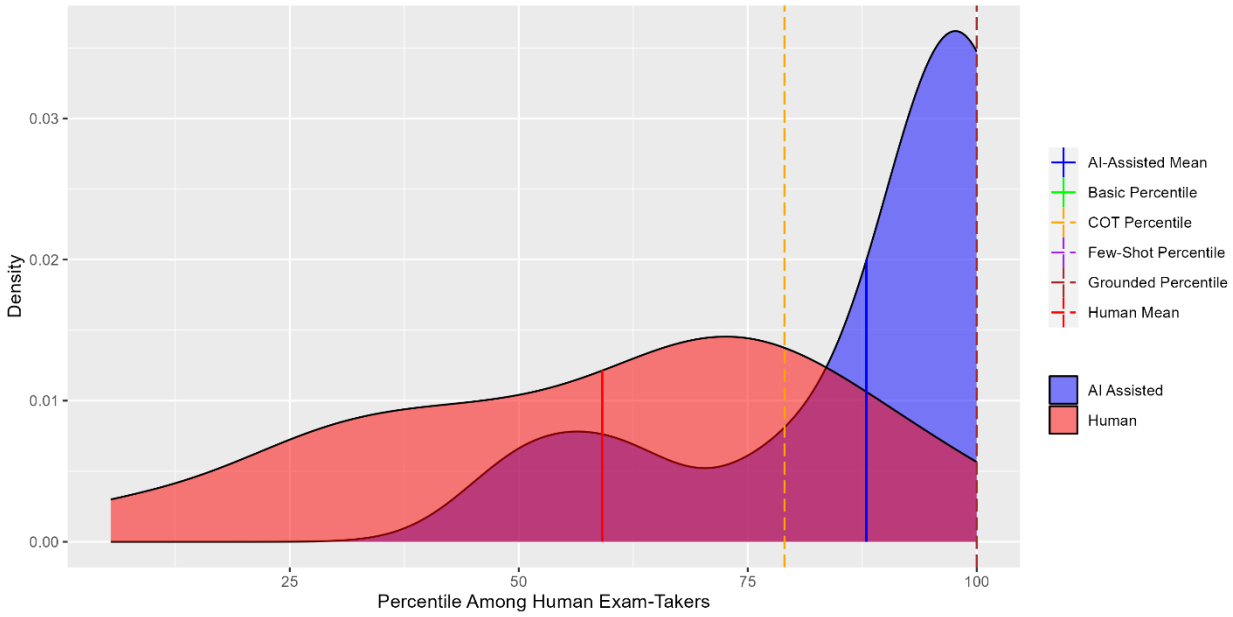


Figure 6: Introduction to American Law Essay – Raw Performance

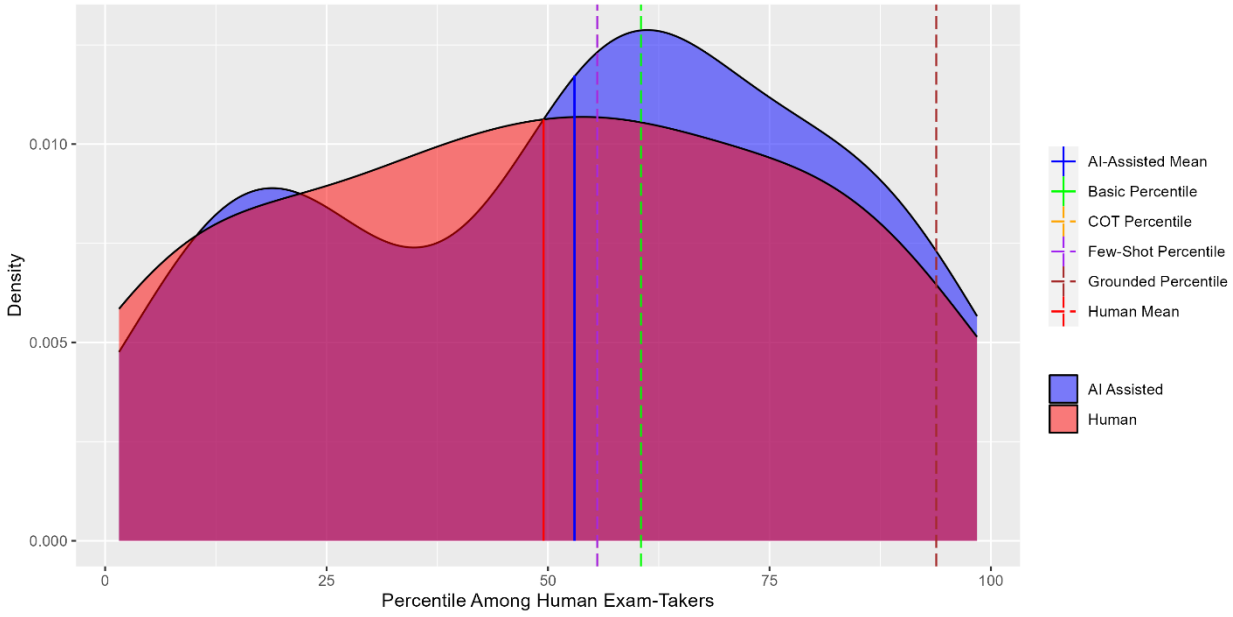
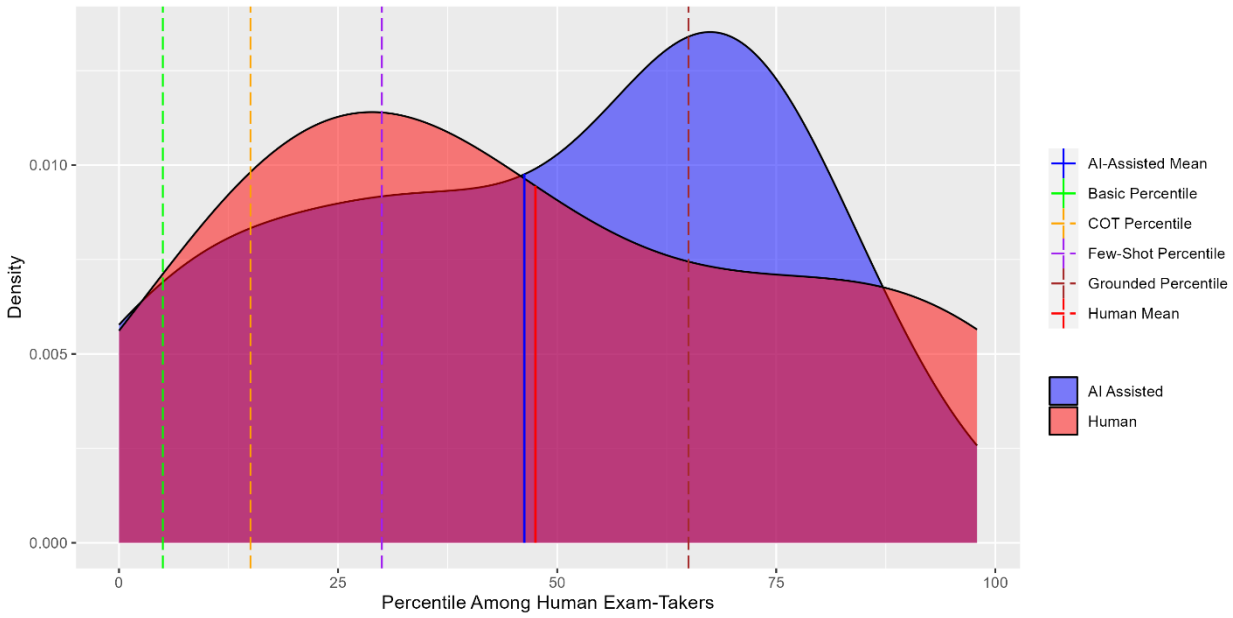


Figure 7: Insurance Law – Raw Performance



The solid and dashed lines in Figure 5 through Figure 7 report the same data as is contained in Tables 1-2. But the curves themselves tell a slightly more nuanced story than do the averages alone: consistent with prior research on the impact of AI on professional writing, human

performance improves somewhat with access to AI at the bottom end of the grade distribution, but does not significantly improve at the top end.[65] And for the more complex exam given to more sophisticated students, access to AI may even cause the performance of the top students to decline, as seems to be the case in Insurance Law. Section B of the Appendix contains additional discussion of the relationship between baseline performance and the benefit of AI assistance depending on the specific exam questions being analyzed.

2. Letter-Grade Performance

In addition to comparing humans, AI, and AI-assisted humans in percentile terms, we can construct letter grades for each of the exams based on our findings. To do so, we simply looked at the actual letter grades that real students received who performed at the class percentile of the relevant exam.

In Introduction to American Law, the median grade received by students in our study without AI assistance was a B. With AI assistance, the median grade improved to an A-, a result that was driven by the improvement in students' multiple-choice answers, which counted for 50% of the overall exam score. In Insurance Law, the median grade received by students in our study without AI assistance was a B/B+. Applying this grading curve to the exam taken with AI-assistance, that median grade was unchanged.[66]

On its own in Introduction to American Law, GPT-4 attained an A- grade with basic prompting, a B+ grade with chain-of-thought prompting, and an A grade with both few-shot and grounded prompting. This performance was relative to a whole-class median grade of B+ in Introduction to American Law, suggesting that GPT-4 systematically outperformed real students in that class. In Insurance Law, GPT-4 achieved a B grade with basic prompting, a B/B+ grade with chain-of-thought prompting, a B+ grade with few-shot prompting, and an A- grade with grounded prompting. This was also relative to a whole-class median grade of B+, suggesting that GPT-4's performance fluctuated around median depending on prompting techniques.

---

[65] *See supra* Part I.

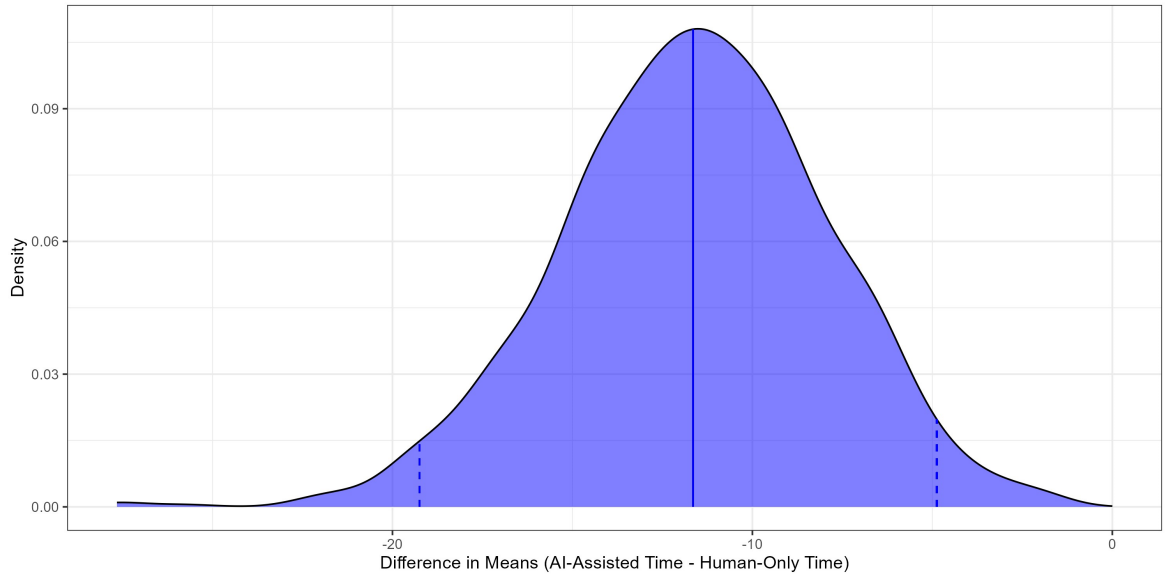
[66] In fact, the 2022 and 2023 Insurance Law classes were not subject to the same grading curve under University of Minnesota Law School rules. That was because the 2023 version of the class was significantly larger than the 2022 version, which triggered the application of a mandatory curve. That curve resulted in lower grades on average for students taking the 2023 exam than the 2022 exam.

3. Speed

Human access to AIs like GPT-4 can impact not only quality of writing, but also the speed that it takes humans to write.[67] For that reason, we examined how access to GPT-4 impacted the speed with which our participants completed their exams. On the real exam without AI assistance, our participants took an average of 74.5 minutes out of 75 allotted minutes to finish the final exam for Introduction to American Law. With access to GPT-4, our participants took an average of 62.9 minutes to finish the exam. We were not able to track how long study participants took to complete their actual final exam in Insurance Law, but with access to GPT-4, they took an average of 115.4 minutes out of the time limit of 120 minutes.

Figure 8 below shows the difference between the amount of time spent on the real exam and the AI-assisted exam for Introduction to American Law, with the curve representing the mean difference generated through bootstrapping.[68] The solid line denotes the bootstrapped mean of means, and the dashed line denotes the 95% confidence interval. The Figure shows a large and statistically significant drop in the amount of time students took to complete the exam when provided with AI assistance.

Figure 8: Change in Speed of Exam Completion with Access to GPT-4, Introduction to American Law



[67] *See supra* Section I.B.

[68] We generated this graph by bootstrapping with 10 million iterations.

Three participants completed the AI-assisted Introduction to American Law exam very quickly, in less than 25 minutes (out of the 75 minutes allotted). When we exclude these three participants, the average time taken on the AI-assisted exam increases to 66.9 minutes, and the average time taken on the real human-only exam decreases to 73.1 minutes, still a statistically significant difference. In contrast, the participants in the Insurance Law study seemed to exert consistently high effort. The fastest finisher (at 88 minutes) received the second-lowest score on the AI-assisted exam, but the second-fastest finisher (at 97 minutes) received a reasonably high A- grade.

B. Qualitative Results

Relative to human-only exams, exams written with AI assistance more consistently provided clear and reasonable answers to issues identified in exam prompts. At the same time, AI-assisted exams were more likely than human-only exams to have organizational problems, to overlook hidden legal issues, and to ignore specific cases and rule-variations.[69] AI-only exams had many similar strengths and weaknesses to AI-assisted exams. However, they were generally more clearly organized than AI-assisted exams, but also more likely to include conclusory analysis.[70] Notably, GPT-4 did not have these weaknesses when grounded with sources as part of its prompting.

Because these trends differed somewhat over the two exams, it is helpful to consider them in more detail based on the underlying class.

1. Insurance Law

Insurance Law exams written with AI assistance were particularly likely to provide clear and reasonable answers to the principal legal questions tested by the questions. These answers tended to do a better job, on average, than human-only exams at threading through their analysis many of the specific relevant facts from the exam hypothetical. Moreover, they tended to accurately characterize these exam facts and to accurately link them to the relevant legal issues. A related strength of AI-assisted exams was their tendency to quote

[69] Notably, several of these problems with AI-assisted exams mirror the problems that Dell'Acqua found to occur when human recruiters were provided with high-quality AI guidance. *See* Dell'Acqua, *supra* note 25. In particular, access to GPT-4 seems to have at times crowded out some of the higher-order legal reasoning that requires significant effort, such as spotting hidden issues or considering the implications of rule variations.

[70] This is consistent with some of the shortcomings of GPT-3.5 in drafting law school exams. *See* Choi et al., *supra* note 3.

relevant portions of key legal texts, such as excerpts from the applicable insurance policy text. Additionally, AI-assisted exams were generally better written at the sentence level than human-only exams, containing easily understandable sentence constructions and few spelling or grammatical errors.

AI-assisted Insurance Law exams also had several notable weaknesses relative to human-only exams. They often contained conclusory analysis, or failed to clearly articulate the relevant legal doctrines, particularly when those doctrines varied across jurisdictions. Another common issue in AI-assisted exams involved the macro-organization of answers. AI-assisted answers were particularly likely, for instance, to introduce relevant legal rules midway through the analysis, to repeat analysis, or to supply arguments whose relationships to one another were unclear. Apart from these organizational problems, AI-assisted exams were more likely to miss hidden issues, particularly when other issues were called out by the relevant exam prompt. Yet another common downside of AI-assisted exams was their sometimes-excessive length, which often reflected either repetition or attention to irrelevant issues. Finally, these exams tended to engage less with specific cases that were covered in the course relative to human-only exams.

The AI-only exams in Insurance Law had many of the same strengths and weaknesses as the AI-assisted exams, with several of the weaknesses being particularly notable depending on the prompting strategy that was used. AI-only exams were especially likely to provide conclusory analysis that did not fully explore the various ways in which the exam facts might be leveraged in support of arguments. This tendency was most visible in the basic AI exam, less visible in the COT exam, less visible still in the few-shot exam, and least visible in the grounded exam. Another notable feature of the AI-only exams was their tendency not to explicitly state relevant rules covered in class, though once again this tendency differed by prompting strategy, with the grounded AI performing best. A common feature of all of the AI-only exams was their tendency to miss somewhat hidden legal issues that were not explicitly alluded to in the facts.

Relative to both human and AI-assisted exams, all of the AI-only exams exhibited especially strong writing and organization of analysis. Many of the organizational problems contained in the AI-assisted exams, including repetition and the introduction of relevant rules in the middle of an analysis, did not appear in the AI-only exams.

2. Introduction to American Law

Relative to human-only exams, the AI-assisted exams for Introduction to American Law did a good job of clearly and reasonably

analyzing the principal legal issue identified in the exam prompt. On average, these exams were better written than human-only exams, both at the individual sentence level and the paragraph level. They often adhered to the basic "IRAC" structure that students were taught to employ.[71] Additionally, AI-assisted exams were generally more likely than human-only exams to precisely specify the central legal issue raised by the question. Another strength of the AI-assisted exams relative to the human-only exams was their direct application of the relevant legal rules to the facts of the problem; as with the insurance exam, AI-assisted exams for Introduction to American Law generally did a good job of accurately highlighting key facts in connection with the appropriate elements of the applicable rule. AI-assisted exams were particularly good at articulating strong counterarguments to the positions that essays ultimately endorsed.

Perhaps not surprisingly, the principal weakness of AI-assisted exams relative to human-only exams was their ability to draw from relevant caselaw covered in class. This weakness was particularly consequential for the essay exam in Introduction to American Law, which instructed students to analogize or distinguish two specific cases studied in class.

Many of these same strengths and weaknesses were evident in the AI-only exams for Introduction to American Law exams. For instance, these exams were well written and did a good job of directly answering the question asked using key facts from the prompt and the relevant legal rules. And, with the notable exception of the grounded AI-only exam, these AI only exams scored relatively poorly when it came to analogizing and distinguishing the specific cases identified in the exam prompt. The grounded AI-only exam, however, did an excellent job both at identifying some of the nuances in the relevant rules that were discussed in class as well as analogizing and distinguishing the specific cases that were identified in the exam prompt. For that reason, the grounded exam received a nearly perfect score, and outperformed nearly all of the AI-assisted exams.

C. Study Limitations

Our work represents an early attempt to gauge how student performance on exams improves with access to AI. However, the questions considered in this paper would benefit from subsequent research and replication attempts, especially in extrapolating these

[71] Jeffrey Metzler, *The Importance of IRAC and Legal Writing*, 80 U. DET. MERCY L. REV. 501 (2003).

results to other settings. There are several reasons to be cautious about the external validity of this study.

First, the training we provided to students may have been inadequate.[72] As students increasingly use ChatGPT and similar technologies in their everyday lives, and as universities begin explicitly to instruct students to use AI models more effectively,[73] students' performance improvement from access to AI may increase. This is an inherent limitation of the sort of study we conducted, which can only gauge the impact of AI in a single setting.

Second, students may have been less motivated to exert maximum effort in our study than they were in their real graded exam, causing them to underperform and causing us to underestimate the benefit of access to AI. Conversely, students might have been better prepared, more familiar with the format of the exam, or more at ease during our study than during their real exam, which would cause them to *over*perform and therefore cause us to overestimate the benefit of access to AI. One piece of evidence suggesting that effort was not the main driver of our results was the amount of time taken in the actual exam versus the exam in our study. In Insurance Law, our study participants took almost all of the time allotted to them, but their performance still did not improve on average.[74] In Introduction to American Law, the participants did take

---

[72] Although we relied on our prior work to train students, numerous tools increasingly purport to help individuals generally, and lawyers and law students in particular, use AI effectively. *See, e.g.*, Tammy Pettinato Oltz, *ChatGPT, Professor of Law*, 2023 U. ILL. J.L. TECH. & POL'Y 207 (2023); Andrew Perlman, *The Implications of ChatGPT for Legal Services and Society*, HARV. L. SCH. CTR. FOR THE LEGAL PRO. (March/April 2023), https://clp.law.harvard.edu/knowledge-hub/magazine/issues/generative-ai-in-the-legal-profession/the-implications-of-chatgpt-for-legal-services-and-society; Ashley B. Armstrong, Who's Afraid of ChatGPT? An Examination of ChatGPT's Implications for Legal Writing (Jan. 23, 2023) (unpublished manuscript) (on file with authors). Perhaps more importantly, tools for using AI more effectively are currently under development by numerous different firms, including Casetext and Harvey. *See supra* note 6. All this suggests both that there is no "right" way to train humans to use AI effectively for legal writing, and that the capacity of humans to effectively use AI to assist with legal writing may well increase significantly over time.

[73] *See* Jason Pohl, *From Tort Law to Cheating, What Is ChatGPT's Future in Higher Education?*, BERKELEY NEWS (Mar. 13, 2023), https://news.berkeley.edu/2023/03/21/from-tort-law-to-cheating-what-is-chatgpts-future-in-higher-education (noting that several Berkeley Law professors are encouraging their students to use AI tools because that is where the future of lawyering will be); Ethan R. Mollick & Lilach Mollick, Assigning AI: Seven Approaches for Students, with Prompts (June 12, 2023) (unpublished manuscript) (on file with authors).

[74] Part III.A, *supra*.

systematically less time on our study exam versus the real exam, but their performance *did* systematically improve on the multiple-choice component, while remaining roughly unchanged on average in the essay component.[75] This suggests that the type of exam question was the primary determinant of the benefit of AI assistance, rather than effort as proxied by time taken.

Third, we should be cautious about extrapolating the results of our study to other settings, particularly non-legal settings. Law school exams may differ systematically from exams in other subjects, and law school exams may not accurately reflect the kind of work a lawyer would do in real life.[76] We are currently working on a follow-on study that tests performance improvements on more realistic lawyering tasks when given access to AI.

Fourth and finally, the methodologies we have labelled "AI-only" vary in the extent to which they actually require some human input. Although we did not edit or review the responses produced by GPT-4, generating the few-shot prompts required access to model exams and the selection of appropriate source materials. If the AI model were asked to select its own model answers and source materials, its performance would likely have been worse.[77] Thus the performance of these models might be regarded as a ceiling on what is possible with current technology.

IV. IMPLICATIONS

Overall, we found that GPT-4 wrote reasonably good law school exams on its own and even better ones with appropriate prompting. We also found that access to GPT-4 improved average student performance only on straightforward multiple-choice questions, with essentially no change to performance on essay questions. However, the effect of AI assistance varied significantly depending on baseline student performance; low-performing students received a substantial boost, while top-performing students may have been harmed by access to AI. Finally, access to GPT-4 significantly decreased the time required for students in the Introduction to American Law course to complete exams

---

[75] *Id.*

[76] *See, e.g.*, Paul Brest, *The Responsibility of Law Schools: Educating Lawyers as Counselors and Problem Solvers*, 58 L. & CONTEMP. PROBS. 5, 6-7 (1995) (exploring how law school exams tend to ignore many important elements of legal practice, such as counseling, problem solving, and designing legal structures); Nancy L. Schultz, *How Do Lawyers Really Think?*, 42 J. LEGAL EDUC. 57, 71-72 (1992); Joan W. Howarth, *What Law Must Lawyers Know?*, 19 CONN. PUB. INT. L.J. 1, 2-3 (2019-2020).

[77] Ney et al., *supra* note 5.

while also significantly improving their grades, suggesting that access to AI can improve both the quality and speed of output. These findings have a variety of implications for the future of law and legal education.

First, the fact that AI helps with simple legal analysis but stumbles over complex legal reasoning complicates the conventional story that AI will be generally useful to practicing lawyers. Based on our results, we believe that current versions of AI like GPT-4 are best suited to the sort of simple tasks that are already frequently outsourced to assistants and paralegals. Our findings are consistent with our own prior work on the performance of AI on law school exams, in which we found that AI performed best at organization, composition, and simple analysis of legal rules, struggling with more complex legal judgments and issue-spotting.[78]

One reason why AI might not be particularly useful at complex legal problems is that GPT-4 itself is worse at these problems (as demonstrated by the relative performance of AI alone on multiple-choice versus essay questions). By analogy, a crib sheet with mostly correct answers will be less helpful than a half-correct and half-incorrect crib sheet. Alternatively, the process of integrating AI insights with human insights might be more difficult with complex essay questions, and in this setting AI responses might be more likely to crowd out human ingenuity.[79] For multiple-choice questions, we suggested that students make an initial guess as to the correct answer and use GPT-4 as a gut check, with additional introspection when the student and GPT-4 disagreed.[80] For essay questions, GPT-4's output was not so simple to integrate, requiring synthesis with the students' own writing and creating the potential for conflicting styles and organization. An interesting question for future research would be the extent to which variation in the quality of GPT-4's responses drove the results in this Article, versus variation in the difficulty of synthesizing human and AI answers.

Second, the fact that GPT-4 helped the worst-performing students significantly more than the best-performing students has important leveling implications for society in general. The legal profession has a well-known bimodal separation between "elite" and "nonelite" lawyers in pay and career opportunities.[81] By helping to bring up the bottom (and

[78] Choi et al., *supra* note 3, at *8-11.

[79] *See* Part III.B, *supra* (finding that, relative to human-only exams, AI-assisted exams often were poorly organized and did a poor job at spotting hidden issues or considering the potential relevance of rule variations).

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

[81] *See, e.g.*, *Salary Distribution Curves*, NALP, https://www.nalp.org/salarydistrib (last visited Aug. 5, 2023) (discussing the bimodal distribution of starting salaries for new law school graduates).

even potentially bring down the top), AI tools could be a significant force for equality in the practice of law.[82]

Of course, a major question that we cannot answer in this paper is precisely *why* the best performers did worse when given access to AI. Although this finding is preliminary (especially in light of the issue regarding mean reversion discussed above), access to AI might discourage effort when used as a crutch. In particular, access to AI might stifle creativity or lead users to settle for easy answers rather than exerting themselves and spotting more difficult issues. This finding is consistent with an emerging literature on the possible limitations of human-AI interaction in complex professional settings. For example, one study mentioned in Part I found that radiologists struggled to appropriately incorporate AI assistance in their decisionmaking and that unless they learn to do so, "the optimal solution involves assigning cases either to humans or to AI, but rarely to a human assisted by AI."[83] However, more research is needed to confirm this effect and to see whether it generalizes to the broader practice of law.[84]

Third, we found that with good prompting but no human supervision other than selecting relevant sources, GPT-4 alone

---

[82] *See* ORLY LOBEL, THE EQUALITY MACHINE: HARNESSING DIGITAL TECHNOLOGY FOR A BRIGHTER, MORE INCLUSIVE FUTURE (2022).

[83] Agarwal et al., *supra* note 27. Another study found that access to high-quality AI assistance induced workers to exert less effort and that, paradoxically, "maximizing human/AI performance may require lower quality AI." Dell'Acqua, *supra* note 25, at 1.

[84] The fact that AI assistance seemed to harm top performers might appear to conflict with the neoclassical non-satiation assumption in economics, under which it is always better to have additional options. See Lorenzo Garbo, *Early Evolution of the Assumption of Non-Satiation*, 24 REV. OF POL. ECON. 15 (2012) (discussing the history of the non-satiation assumption). One explanation might be that while we allowed the students to complete the extra exam in any way they saw fit, the framing of the study likely pushed them toward actually using the AI tools. In the first place, we provided them with training on how best to use these tools. In the second place, we induced students to participate with the promise that they would learn how to use GPT-4 and test them out, and it is hard to imagine students would have volunteered if they had not anticipated actually using AI. Given these factors, even if a student knew that using GPT-4 might worsen her exam performance, she might nevertheless have used GPT-4 in order to test its capabilities or to comply with the study guidelines. In contrast, a real-life lawyer would not use AI tools unless she believed they would improve her performance. Thus even though we found that top performers did worse with AI assistance, real-world lawyers would always have the option simply not to use AI tools. If so, this could cause an additional schism in the legal profession between nonelite lawyers making extensive use of AI assistance and elite lawyers forgoing AI assistance to engage in bespoke, complicated legal tasks.

outperformed both humans *and* AI-assisted humans on average. This was despite the fact that we provided study participants explicit instructions on how to provide relevant sources to GPT-4. Our results raise the possibility that humans soon may be entirely removed from the loop in certain legal tasks, especially given work by legal tech companies to automate the prompt engineering techniques described in this Article.[85] The fact that GPT-4 can outperform humans with access to GPT-4 has ominous implications for the paraprofessionals (like paralegals and law firm assistants) who often conduct simple legal analysis under the status quo. It is possible that these paraprofessionals will soon be entirely replaced.

Of course, an hour of training may have been insufficient for students to master prompt engineering. Thus one could also take our findings as evidence that lawyers and law students should invest in learning how to use AI tools effectively. And if new technologies effectively automate prompt engineering for specific legal tasks, lawyers may enjoy significant benefits from AI assistance even on complex legal tasks.

Fourth, the discussion above about the *quality* of legal output fails to account for an equally important finding in our study about the *speed* of legal output. Significant speed gains in Introduction to American Law suggest that AI could substantially improve lawyer efficiency, at least where straightforward lawyering tasks are concerned. In this way, AI could be a double boost to productivity, both increasing the quality of output and decreasing the time that it takes to produce that output. Limited lawyer time is an important current impediment in access to justice, and efficiency improvements could dramatically expand the scope of legal services to the public.[86]

Finally, these results have important pedagogical implications for universities. If AI benefits the worst-performing students the most, assignments on which students use AI (whether allowed to or not) may have compressed grading curves, potentially making it harder for instructors to draw granular distinctions between the performance of different students.[87] In addition, the fact that AI tends to be more useful

[85] *See supra* note 6.

[86] *See* Geoffrey T. Burkhart, *How to Leverage Public Defense Workload Studies*, 14 OHIO ST. J. CRIM. L. 403, 403 (2017); Peter A. Joy, *Ensuring the Ethical Representation of Clients in the Face of Excessive Caseloads*, 75 MO. L. REV. 771, 791 (2010).

[87] *See infra* Appendix Section B (noting that the standard deviation of scores significantly decreased in the multiple-choice component of the Introduction to American Law Exam). In our limited study, grade compression seems most apparent when AI assistance actually improves student performance.

on multiple-choice questions rather than issue-spotters might induce professors to move toward essay questions, both to reduce the benefits of cheating and to focus pedagogical attention on the sorts of tasks at which humans have comparative advantage. More generally, our research can help to inform law schools about what skills remain uniquely human and therefore most likely to benefit law school students as they enter the labor market.[88]

V. CONCLUSION

We conducted an experiment to test how AI assistance affects legal reasoning, by comparing student performance on law school exams with and without access to AI. We found that students consistently benefited from AI assistance only on simple multiple-choice questions and that GPT-4 alone outperformed both students alone and students with AI assistance with effective prompt engineering. We also found large variation in the effect of AI assistance, with the worst-performing students seeing the largest gains, and the best-performing students seeing declines in performance. These findings have significant implications for the future of lawyering, legal education, and professional work more generally.

[88] One speculative possibility that is worth exploring in future research is whether law school instructors can use tools like GPT-4 to provide more frequent and consistent feedback on student work-product. For instance, it may be possible to use grounded and few-shot prompting that incorporates grading rubrics and model instructor comments to facilitate effective feedback. If so, the implications could be significant for legal education, as research suggests that individualized formative feedback on law school exams can produce significant and generalizable benefits for law students. *See* Daniel Schwarcz & Dion Farganis, *The Impact of Individualized Feedback on Law Student Performance*, 67 J. LEGAL EDUC. 139, 140 (2017).

APPENDIX

A. Additional Information on Data and Methods

1. Background on Courses

The principal goal of Introduction to American Law is to introduce undergraduates to legal analysis. To that end, the class is taught in much the same way as a typical law school class; readings consist principally of edited judicial opinions, and exams test students' capacity to apply legal principles to new situations. Four such exams are given throughout the semester. These exams typically[89] include both multiple choice questions and essay questions that require students to analyze a novel fact pattern by clearly stating the relevant legal issue, specifying the applicable legal rules, applying those rules to the facts, analogizing and/or distinguishing to relevant case law, and considering key counterarguments. The essay portion of exams are graded by 2L and 3L law school research assistants, who use a grading rubric produced by the instructor and who are trained by the instructor to apply that rubric consistently.

Insurance Law grades are based principally on a single final exam, which consists of between 2 and 4 essay questions. These questions generally involve elaborate and novel hypothetical scenarios that require students to analyze multiple legal issues while drawing on caselaw, statutes, and regulations studied in class. Final exams also occasionally require students to perform a policy analysis of legal rules or proposals.

The substantive material covered in both Introduction to American Law and Insurance Law was virtually identical in the Spring of 2023 and the Spring of 2022, when the co-author instructor taught these same classes.

2. Exams and Grading

Several practical considerations affected the exams that we administered to study participants differing in some respects from the real exams that we administered in 2022. For Insurance Law, study participants answered two of the three questions from the Spring 2022 Insurance Law final exam, each of which counted toward 25% of students' final grades in 2022. Ultimately, then, students from the

[89] The fourth and final exam includes 30 multiple choice questions that cover all the material taught in the class. By contrast, the first three exams are non-cumulative and consist of 15 multiple choice questions and one essay question.

Insurance Law class took half of the real 2022 final exam. We used this approach both because we believed it would be easier to recruit students to spend two, rather than four, hours on the AI-assisted exam, and because students in the 2023 Insurance Law class had previously been given the third question from the 2022 exam for practice. For students in Introduction to American Law, we used 15 multiple choice questions from the 2022 final exam question, and an essay question from the third 2022 exam, which covered property and civil procedure. We used this approach because we wanted the exam students completed to include both an essay and a multiple-choice component, as three of the four exams in the class do. It was not possible to accomplish this solely by relying on the final exam in the class from 2022, because (as noted earlier) the final exam in the class is entirely multiple choice.

For Introduction to American Law, we used the same process for training RAs to accurately and consistently use the grading rubrics in this study as is used in the actual class. First, the instructor developed a detailed grading rubric that required graders to evaluate how well essays frame the relevant issues, describe the relevant legal rules, apply those rules to the specific facts of the exam, analogize or distinguish to relevant caselaw, and consider counterarguments. Specific descriptions, such as "The answer notes both the question of whether a motion to dismiss should be granted and whether Charlie adversely possessed the property but reflects a somewhat confused understanding of the interaction of these two questions or inconsistent usage," are associated with specific numeric scores. After reviewing the rubric together, the RAs and the instructor independently applied it to four test essays, and then reviewed the consistency of the results in a meeting. Differences in scoring were discussed and resolved, and the rubric itself was adjusted accordingly. After these test exams were scored, RAs were instructed to use them as anchors during their grading if they were unsure how to score a particular component of an exam. Subsequent to this process, a "lead" RA reviewed all of the grades produced by the individual RAs to ensure that they were consistent and flagged any potential inconsistences in scoring for further review by the instructor.

For each exam that was graded for this study, the grader filled in a detailed grading rubric that assigned specific points for different elements of the exam answer. Once the initial grading was complete and the exams were unblinded, the co-author who taught the classes examined these rubrics along with the underlying exams to identify trends in the relative strengths and weaknesses of AI-assisted and AI-only exams.

3. Recruitment of Study Participants

To recruit study participants, we sent emails to all students enrolled in Introduction to American Law and Insurance Law. We paid each participant a flat fee for their participation, unrelated to their exam performance and conditional only on successful completion of the trainings and the exam. Participants agreed to take part in the study prior to receiving their final grades for the relevant courses.

As Figure 5 through Figure 7 above show, the students who volunteered for our study were not a representative sample across performance levels. If they were, the curves for human-only performance (representing how well our participants did in percentile terms compared to the class as a whole) would have been roughly flat. Instead, they were inverse-U shaped, suggesting that our sample under-represented both low-performing and high-performing students. Thus, average treatment effects in our study are most representative of the effect on roughly average students.

4. GPT-4 Parameters and Prompts

To produce AI-only exam responses, we used the 8K GPT-4 model through OpenAI's API,[90] using the March 14, 2023 version (gpt-4-0314). For optimal reproducibility, we set temperature to 0, as recommended by OpenAI.[91] We used the following system prompt for essays: "You are an experienced legal academic like Cass Sunstein or William Eskridge writing a model answer to a law school exam." And we used the following system prompt for multiple choice questions: "You are an experienced legal academic like Cass Sunstein or William Eskridge answering a multiple choice question on a law school exam." These system prompts (and all of the following prompts) were validated using a small training set of old exam questions and then used out-of-sample to produce the results described in the "Results" section above.[92]

---

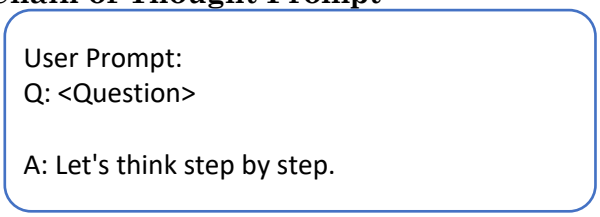
[90] An API, or Application Programming Interface, is a set of rules and protocols for building and interacting with software applications. The GPT-4 API is provided by OpenAI to allow developers to access and interact with the GPT-4 model in their applications or services. *See generally* Greg Brockman et al., *OpenAI API*, OPENAI (June 11, 2020), https://openai.com/blog/openai-api (describing the GPT-4 API).

[91] Boris Power, *OpenAI Cookbook*, GITHUB, https://github.com/openai/openai-cookbook/blob/main/examples/Fine-tuned_classification.ipynb (last visited Aug. 5, 2023).

[92] This is consistent with the use of training sets and validation/test sets more broadly in the literature on natural language processing. *See, e.g.*, *Sklearn.model_selection.train_test_split*, SCIKIT-LEARN, https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html (last visited Aug. 5, 2023) (discussing the training/testing split procedure in Scikit-Learn, a popular natural language processing software package).

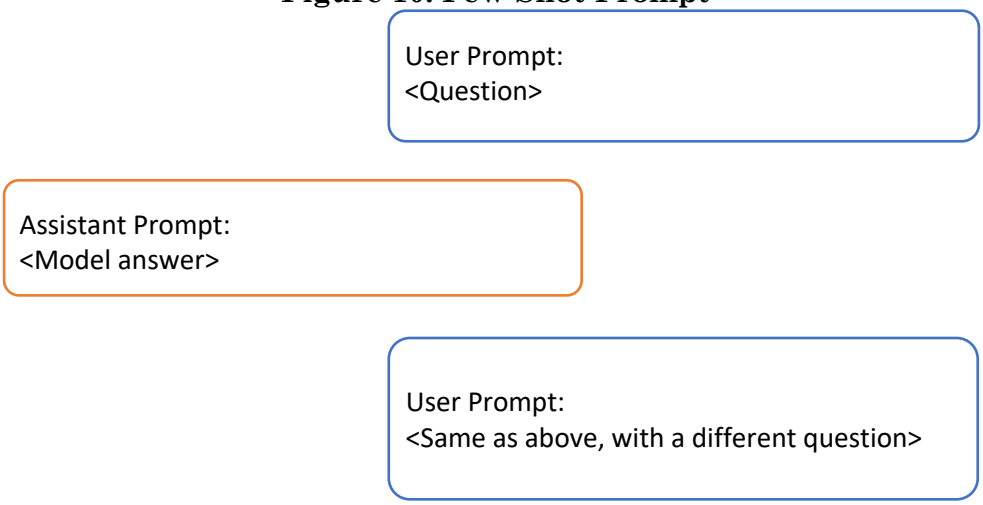
Each of the following prompts was the version used to generate responses for the essay questions. We used the same prompt with minor appropriate modifications for the multiple choice questions.

Figure 9: Chain-of-Thought Prompt



The few-shot prompting method we used requires specifying both user and assistant prompts within the GPT-4 API; thus it is currently not possible to perfectly replicate this method without API access.[93] In our study, we provided the AI model with a single model exam and answer; this might properly be described as "one-shot" prompting, as opposed to the "zero-shot" prompting used in the basic and chain-of-thought prompts.

Figure 10: Few-Shot Prompt



[93] Users can attempt to imperfectly mimic this approach within the conventional ChatGPT interface in various ways, such as by simply describing the question and model answer in the user prompt.

Figure 11: Grounded Prompt

User Prompt:
You will be asked to answer a law school exam question relying on materials from a law school class. Here are the materials:

<Source materials>

Here is the law school exam question:

<Question>

B. Additional Figures

The following Figures show the mean improvement of humans when given access to GPT-4 on exams in our study. As above, curves are generated through bootstrapping and 95% confidence intervals are shown by dashed lines.

Figure 12: Introduction to American Law Multiple Choice – Mean Student Improvement Given Access to GPT-4

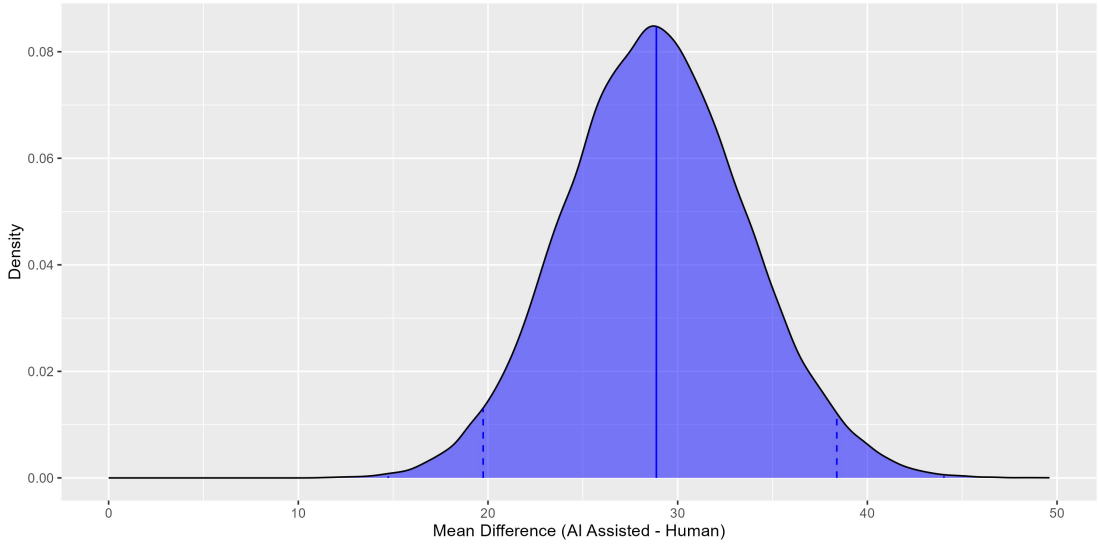


Figure 13: Introduction to American Law Essay – Mean Student Improvement Given Access to GPT-4

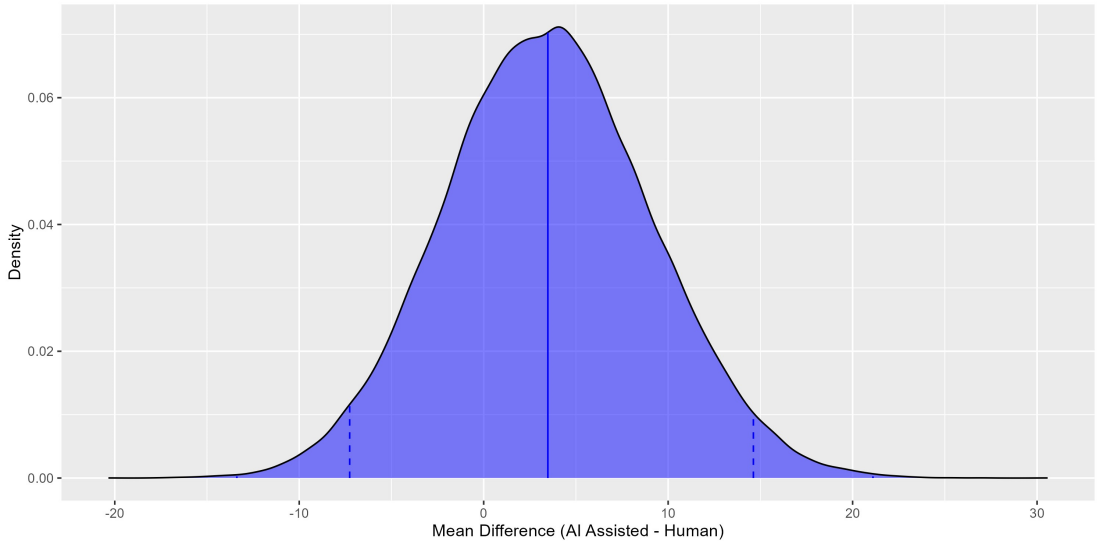
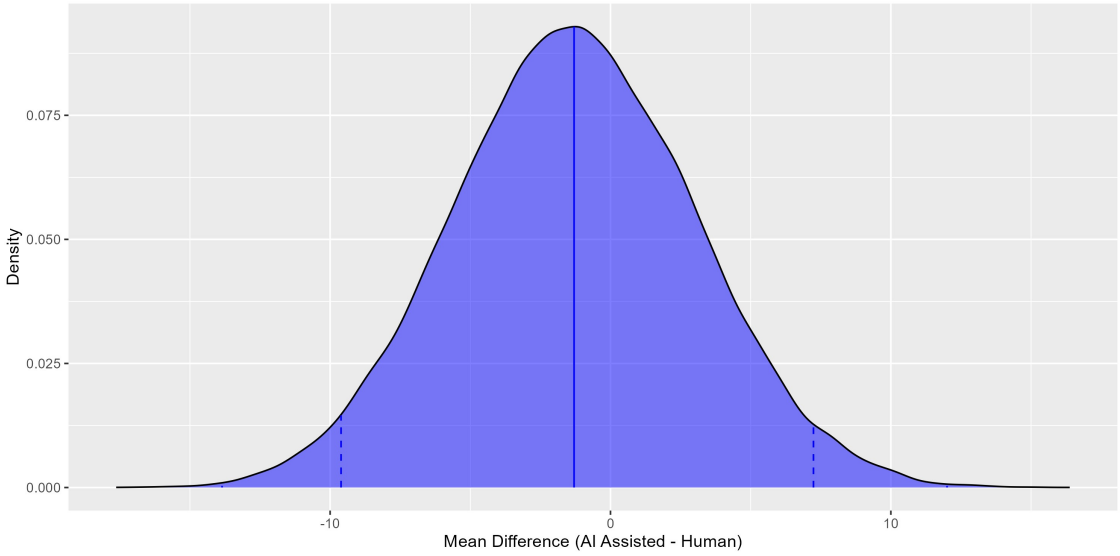


Figure 14: Insurance Law – Mean Student Improvement Given Access to GPT-4



The following Figures show performance improvements from AI assistance in relation to the initial performance of students without AI assistance (similar to Figure 4), broken down by exam and exam component. The Figures show a consistent inverse correlation between performance without AI and the boost a student receives from AI assistance. However, the specifics differ by exam component. In the multiple-choice section of Introduction to American Law (where GPT-4 performed best on its own), students toward the bottom of the class saw enormous performance gains in excess of 50 percentile points, while students toward the top of the class saw no performance losses (just smaller gains). The essay section of Introduction to American Law saw both gains for the bottom students and declines for the top students. And the Insurance Law exam generally saw modest gains for the students toward the bottom and noticeable declines for students toward the top. Overall, these findings add nuance to the general story described in Section III.A.1 that assistance from GPT-4 helped struggling students and harmed top students; specifically, it seemed to help struggling students the most with easier (especially multiple-choice) questions and harm top students the most with more complex (especially essay) questions.

Figure 15: Introduction to American Law Multiple Choice – Mean Student Improvement Given Access to GPT-4

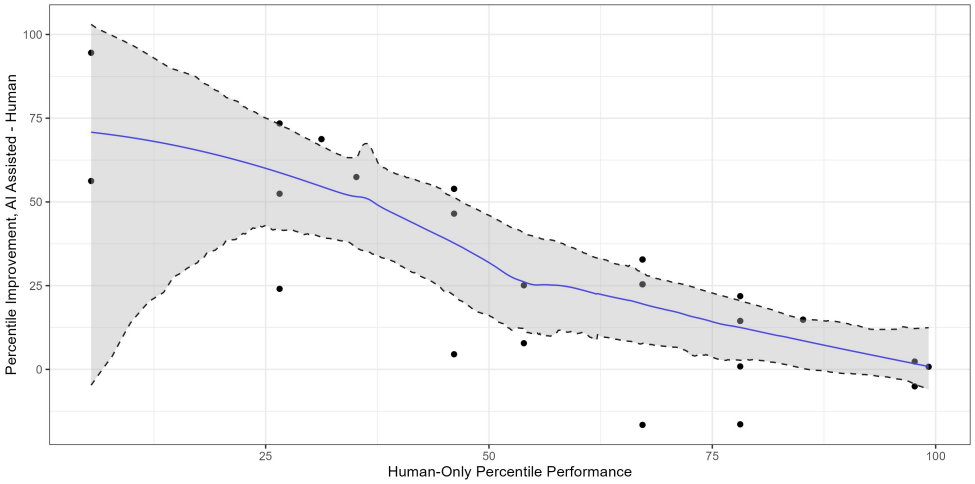


Figure 16: Introduction to American Law Essay – Mean Student Improvement Given Access to GPT-4

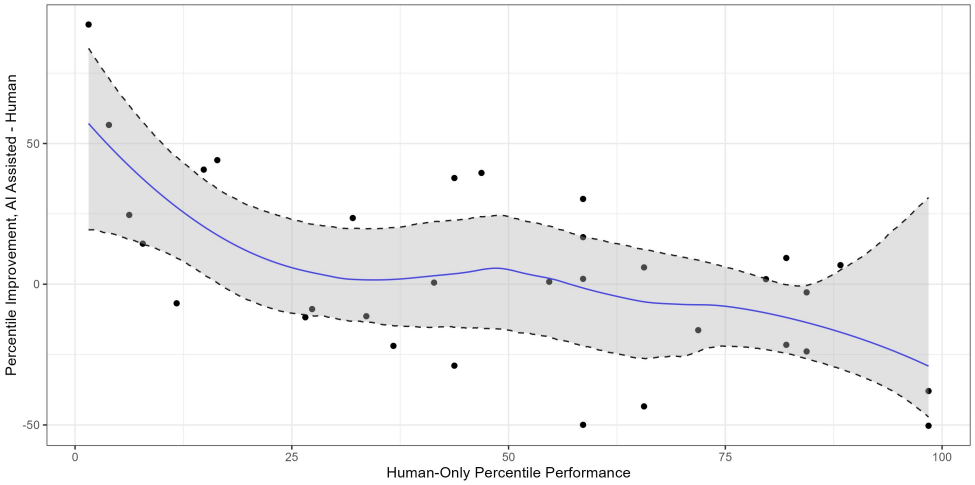
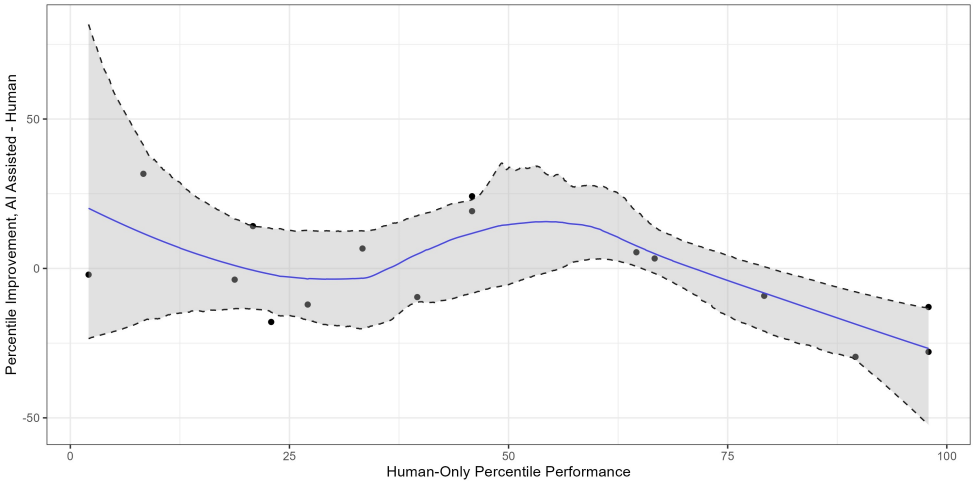


Figure 17: Introduction to American Law Multiple Choice – Mean Student Improvement Given Access to GPT-4



Finally, another way to consider how AI assistance affects the distribution of student exam performance is to look at changes in the standard deviation of scores when AI assistance is provided. The following Figures plot this difference, with 95% confidence intervals denoted by dashed lines. While standard deviation is a measure of variation in student performance, asking whether AI assistance decreases the standard deviation of exam percentiles is subtly different than asking whether AI assistance benefits bottom students more than top students. For example, if AI assistance simply cause bottom students and top students to switch places—turning the 1st-percentile student into the 99th-percentile student and vice versa, 2nd-percentile into 98th-percentile and vice versa, etc.—we would see strong variation in improvement from AI assistance, as above, without seeing *any* change in standard deviation.

The following Figures suggest that something like this may be happening—that the standard deviation of performance decreased only in the multiple-choice component of Introduction to American Law, with essentially no change in the essay component of either exam. Although further research is needed, this might be because skill in traditional legal analysis is orthogonal to skill at properly employing AI assistance, so that the best students at legal analysis may "switch places" with the best students at AI collaboration. In general, though, AI assistance may only compress the raw amount of variation between students when it actually improves performance.

Figure 18: Introduction to American Law Multiple Choice – Mean Student Improvement Given Access to GPT-4

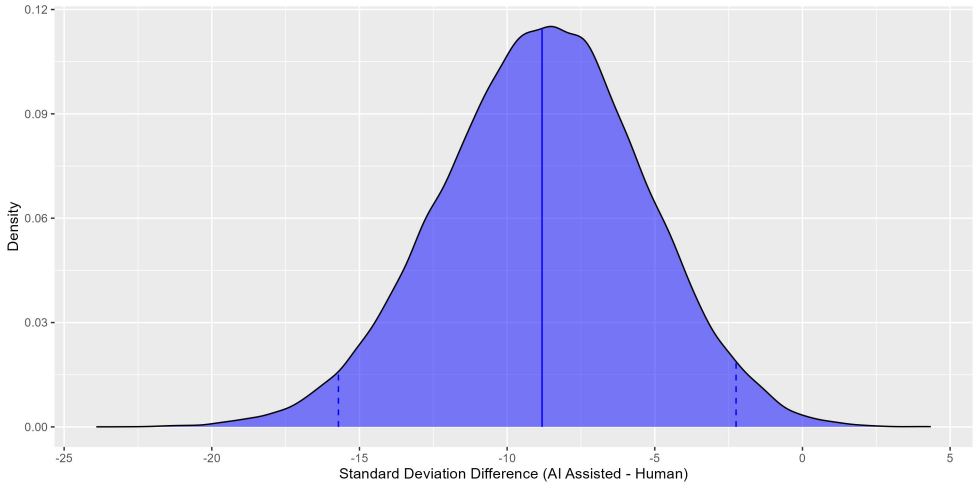


Figure 19: Introduction to American Law Essay – Mean Student Improvement Given Access to GPT-4

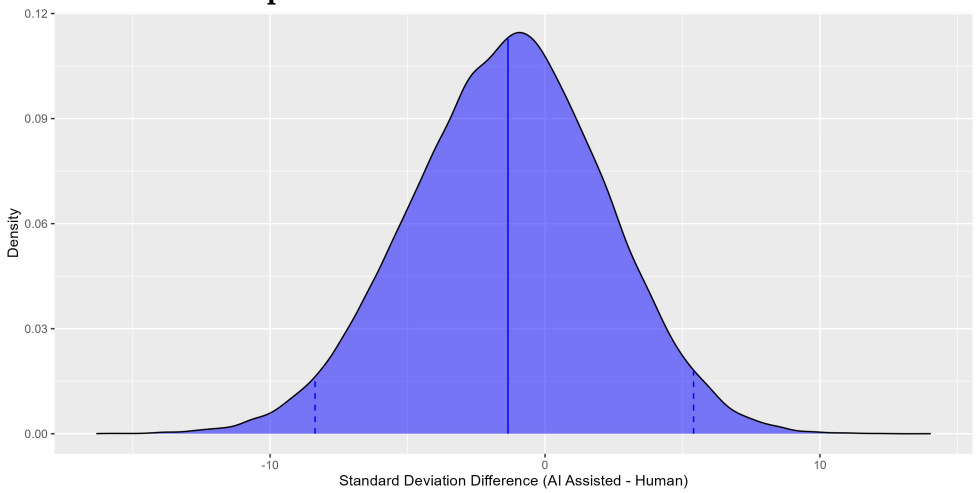


Figure 20: Introduction to American Law Multiple Choice – Mean Student Improvement Given Access to GPT-4