

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.

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 *overperform* 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. Burkhardt, *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 typically⁸⁹ 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 inconsistencies 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

User Prompt:
Q: <Question>

A: Let's think step by step.

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

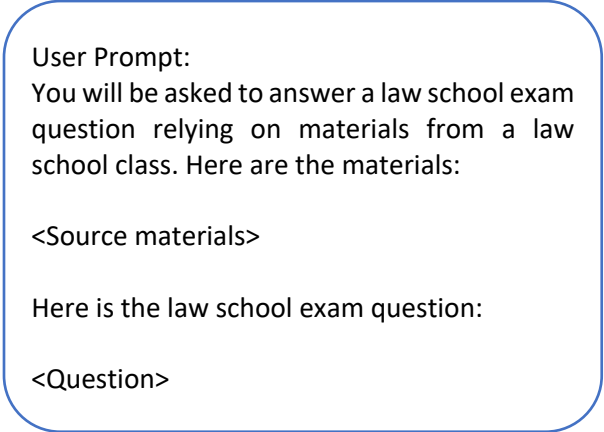
User Prompt:
<Question>

Assistant Prompt:
<Model answer>

User Prompt:
<Same as above, with a different question>

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

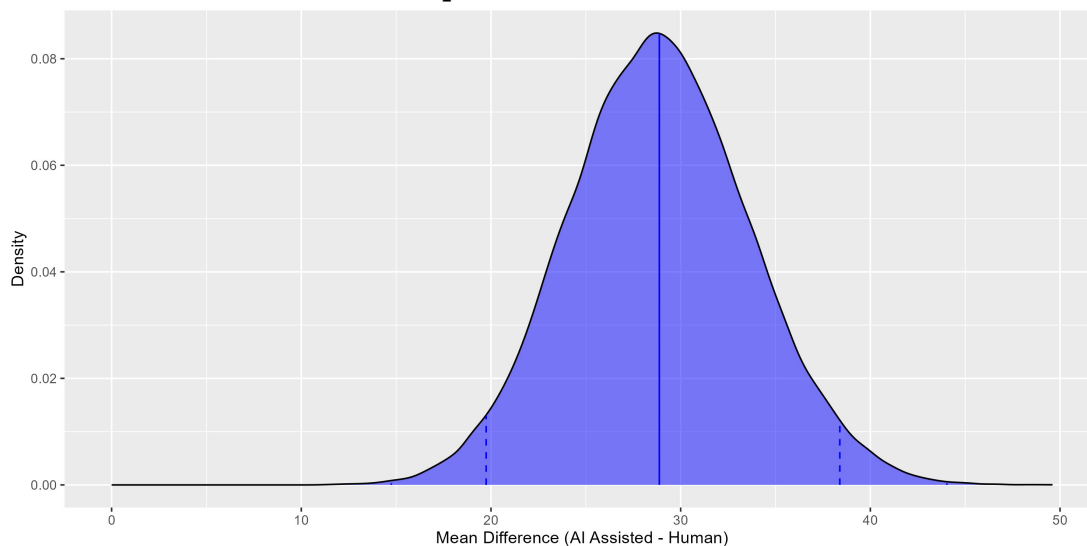


Figure 13: Introduction to American Law Essay – Mean Student Improvement Given Access to GPT-4

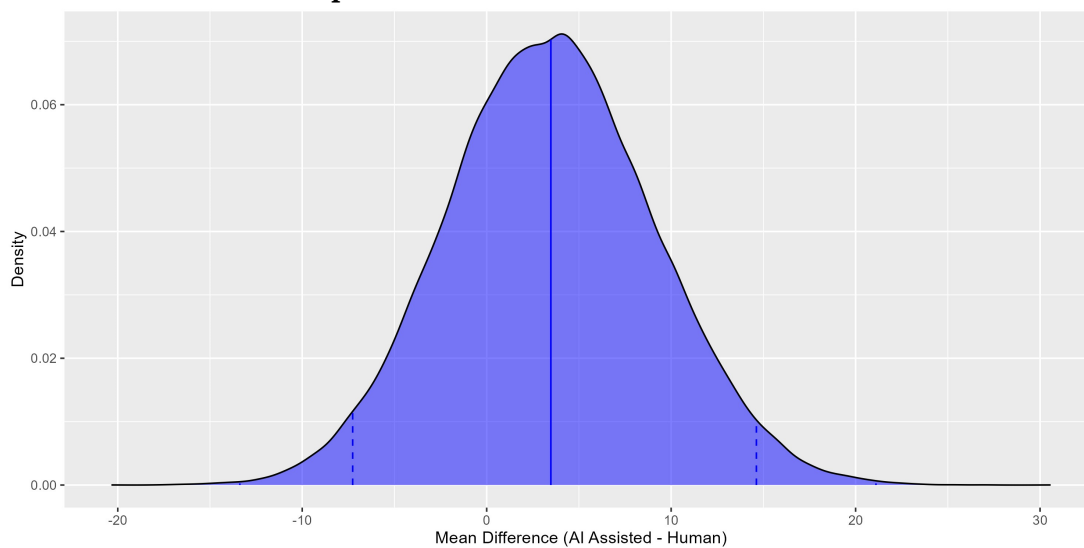
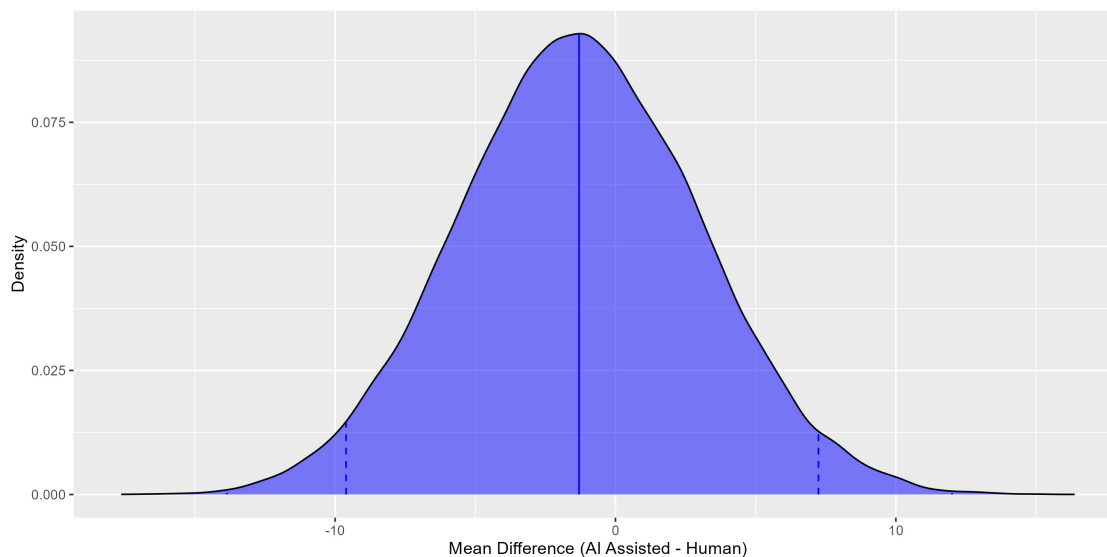


Figure 14: Insurance Law – Mean Student Improvement Given Access to GPT-4



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

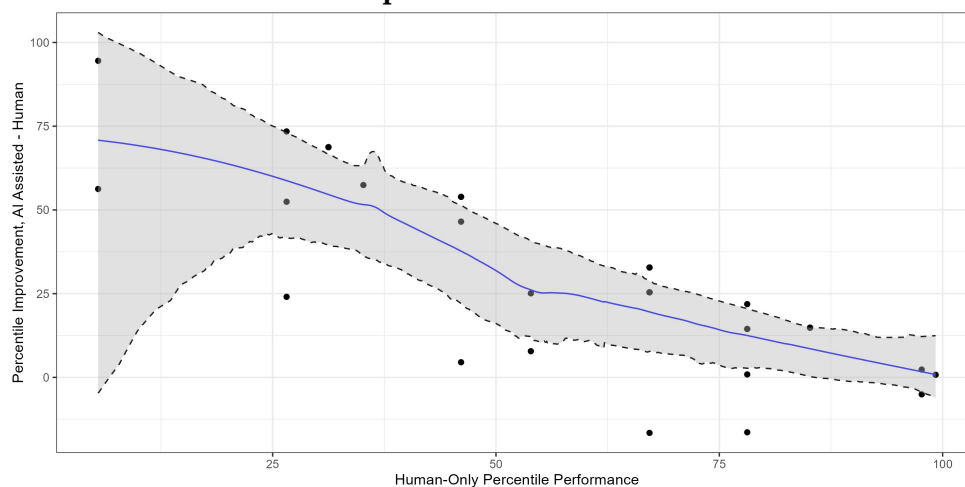


Figure 16: Introduction to American Law Essay – Mean Student Improvement Given Access to GPT-4

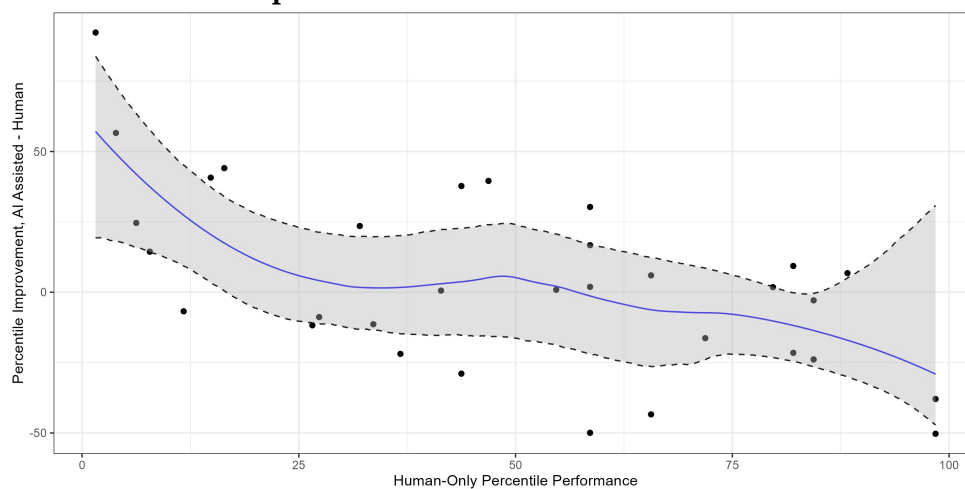
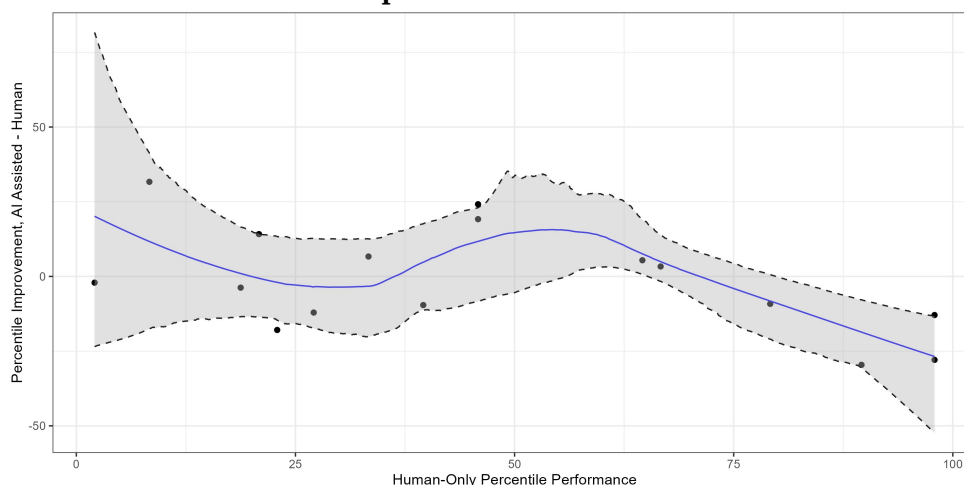


Figure 17: Introduction to American Law Multiple Choice – Mean Student Improvement Given Access to GPT-4



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

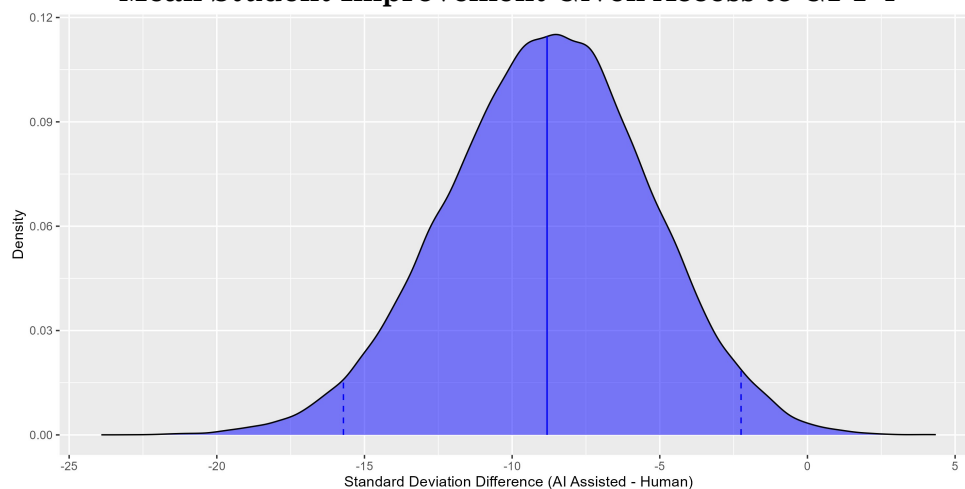
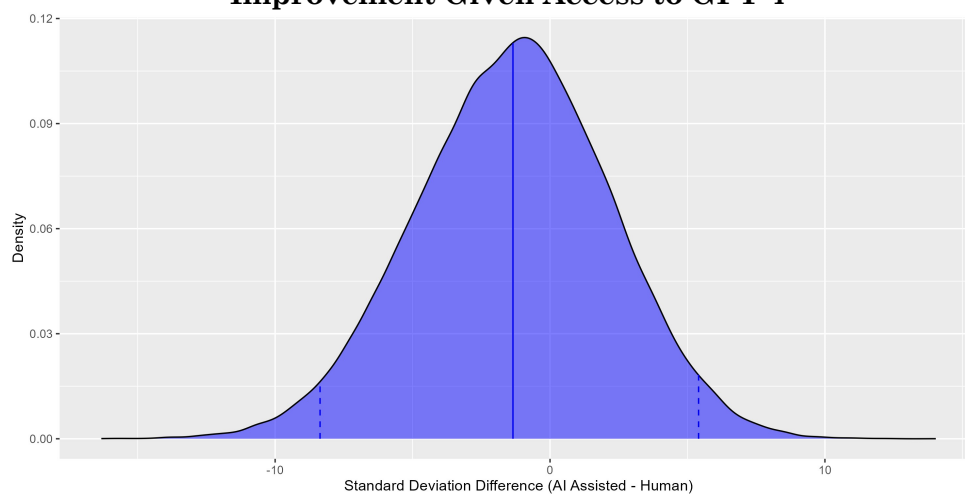


Figure 19: Introduction to American Law Essay – Mean Student Improvement Given Access to GPT-4



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

