ChatGPT Goes to Law School

Jonathan H. Choi
Kristin E. Hickman
Amy B. Monahan
Daniel Schwarcz

How well can AI models write law school exams without human assistance? To find out, we used the widely publicized AI model ChatGPT to generate answers to the final exams for four classes at the University of Minnesota Law School. We then blindly graded these exams as part of our regular grading processes for each class. Over ninety-five multiple-choice questions and twelve essay questions, ChatGPT performed on average at the level of a C+ student, achieving a low but passing grade in all four courses. After detailing these results, we discuss their implications for legal education and lawyering. We also provide example prompts and advice on how ChatGPT can assist with legal writing.

I. What Is ChatGPT?

ChatGPT is an AI language model produced by OpenAI and released in late 2022.[1] GPT models, including ChatGPT, are "autoregressive," meaning that they predict the next word given a body of text. For example, given the phrase "I walked to the," a GPT model might predict that the next word is "park" with five percent probability, "store" with four percent probability, etc. The model can then repeatedly predict subsequent words (for example, "and") to compose indefinitely long bodies of text.

Jonathan H. Choi, Lead author, McKnight Land-Grant Professor, Associate Professor of Law, and Solly Robbins Faculty Research Scholar, University of Minnesota Law School.

Kristin E. Hickman, McKnight Presidential Professor in Law, Distinguished McKnight University Professor, and Harlan Albert Rogers Professor in Law, University of Minnesota Law School.

Amy B. Monahan, Distinguished McKnight University Professor and Melvin C. Steen Professor of Law, University of Minnesota Law School.

Daniel Schwarcz, Fredrikson & Byron Professor of Law, University of Minnesota Law School.

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

OpenAI has produced progressively larger language models, from GPT-1's 117 million parameters to GPT-3's 175 billion parameters.[2] One of the most important discoveries in machine learning over the past decade has been the extraordinary returns to scale when language models use more parameters and are trained on larger corpora of text. Current large language models like GPT-3 can compose human-like text with surprising fidelity.[3]

In addition to training on vast amounts of text, ChatGPT is further trained using Reinforcement Learning from Human Feedback (RLHF).[4] In RLHF, humans manually tag the best responses produced by an initial language model[5] to improve its performance at specific tasks. Through these repeated machine-human interactions, ChatGPT was trained to engage in dialogue, be more truthful, and avoid inflammatory or offensive language.[6]

Although ChatGPT was trained on a large general-purpose corpus and was optimized only for general-purpose dialogue, it performs surprisingly well on specific technical tasks. These include computer programming,[7] data manipulation,[8] and medical diagnosis.[9] We thus set out to see how well ChatGPT performed on law school exams, to highlight how it might change both legal education and the practice of law.

II. Empirical Methods

We used ChatGPT to produce answers to final exams for four separate law school courses at the University of Minnesota: Constitutional Law: Federalism and Separation of Powers; Employee Benefits; Taxation; and Torts.

[2] Priya Shree, *The Journey of Open AI GPT models*, MEDIUM (Nov. 9, 2020), https://medium.com/walmartglobaltech/the-journey-of-open-ai-gpt-models-32d95b7b7fb2.

[3] In particular, the best-known modern language models are based on a transformer architecture, another innovation that significantly improved performance. *See* Ashish Vaswani et al., *Attention Is All You Need*, *in* PROCEEDINGS OF THE 31ST CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS 6000 (2017) (introducing the transformer architecture).

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

[5] In this case, ChatGPT was fine-tuned from GPT-3.5, a language model that was *itself* fine-tuned using RLHF.

[6] *See* OPENAI, *supra* note 5.

[7] *E.g.*, The PyCoach, *ChatGPT: The End of Programming (As We Know It)*, MEDIUM (Dec. 14, 2022), https://medium.com/geekculture/chatgpt-the-end-of-programming-as-we-know-it-ac7e3619e706.

[8] *E.g.*, Marie Truong, *Can ChatGPT Write Better SQL than a Data Analyst?*, MEDIUM (Jan. 5, 2023), https://towardsdatascience.com/can-chatgpt-write-better-sql-than-a-data-analyst-f079518efab2.

[9] *E.g.*, Phil Wang & Yacine Zahidi, *Medical-ChatGPT*, GITHUB, https://github.com/lucidrains/medical-chatgpt, (last visited Jan. 22, 2023).

One of us generated all of these answers using ChatGPT and formatted them to match actual exams written by students.[10] This co-author only generated answers, grading no exams as part of the study. To do so, the co-author used a uniform set of prompts for all of the exam questions.[11] Thus, for any given question, ChatGPT's output was generated without any human intervention other than copying and pasting the content of the exam question along with a standard prompt.

The AI-generated exams were then shuffled with actual student exams and graded blindly by the other three co-authors. The three grading co-authors graded ChatGPT's performance on the entire exam relative to real students' without knowing which exam was generated by ChatGPT. The ChatGPT exams were subsequently removed and the curve recalculated before finalizing actual student grades.

Each exam slightly differed in format and context. Constitutional Law and Torts are both required 1L courses; Employee Benefits and Taxation are upper-level elective courses subject to a slightly more relaxed curve. The Constitutional Law and Torts exams included both multiple-choice and essay questions. The Employee Benefits exam included only essay questions (both short and long), and the Taxation exam included only multiple-choice questions. The Constitutional Law and Torts exams had word limits, while the Employee Benefits exam did not. Only the Constitutional Law exam required sources to be cited in essays. Final grades in all four classes were based principally on the final exam.

The prompts used to generate essay answers are discussed below in Part IV. For multiple-choice questions, we experimented with two prompting methods other than directly asking ChatGPT to select the correct answer: "chain-of-thought" (CoT) prompting and "rank-order" prompting. In CoT prompting, the model is asked to provide a chain of reasoning and to give a letter answer to the question.[12] In rank-order prompting, the model is asked to rank its top choices (we used the top three, consistent with prior work) rather than give a single choice.[13] Both alternative prompting methods were found to perform well with GPT models in past work.

Compared with prompts that simply ask ChatGPT to give a single letter answer for each multiple-choice question, CoT and rank-order prompting

[10] The Taxation exam was an exception, because it was entirely multiple choice and therefore did not need to be graded blind. The author conducting the Taxation exam generated answers for it given standard prompts to keep the questions confidential.

[11] In all cases, we generated answers using the December 15, 2022, distribution of ChatGPT.

[12] We specifically use "zero-shot" prompting (without fine-tuning ChatGPT or providing examples) following recent work. Valentin Liévin, Christoffer Egeberg Hother & Ole Winther, *Can Large Language Models Reason About Medical Questions?* (Dec. 20, 2022) (unpublished manuscript), https://arxiv.org/pdf/2207.08143.pdf.

[13] Michael J. Bommarito II & Daniel Martin Katz, *GPT Takes the Bar Exam* (Dec. 29, 2022) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4314839.

performed the same or worse on all exams. The following table summarizes the performance of all three methods:

Table 1: Comparison of Multiple-Choice Methods

Method	Constitutional Law	Taxation	Torts	Total
Simple	21/25	24/60	6/10	51/95
CoT	21/25	18/60	5/10	44/95
Rank-Order	20/25	21/60[14]	6/10	47/95

Of the three methods, the simplest performed the best, although the difference compared with the other methods was not statistically significant ($p = 0.138$ compared with CoT, $p = 0.257$ compared with rank-order).[15] As a result, we failed to replicate past studies in which CoT and rank-order prompting produced superior results.[16] For the remainder of the study, we therefore used simple prompts to generate multiple-choice answers.

As context for the student body to which ChatGPT was compared, the University of Minnesota Law School is currently ranked sixteenth among law schools by U.S. News & World Report.[17] Ninety-nine percent of its graduates in 2022 passed the bar on their first attempt, the second-highest bar passage rate in the country.[18] Each of the four courses in this study were curved to approximately a B+ average, with a minimum and maximum number of A grades but no requirement for the instructor to award grades below a B. Students at Minnesota Law with a cumulative grade point average of 2.6 or below are placed on academic probation, as are students who receive grades of D or F in a required first-year course, or in multiple courses in a single

[14] For Taxation, the rank-order method refused to choose between the options for one question, answering that they were all correct despite repeated prompting; we scored this response incorrect.

[15] *P*-values were calculated using bootstrapping with 100,000 iterations.

[16] The relatively poor performance of these alternative methods may relate to their having been developed in a different context (medical exams for CoT prompting) or using a different language model (GPT-3.5 for rank-order prompting). The poor performance may also relate to the multiple comparisons problem: Because past studies tested multiple alternative approaches and reported on which one worked best, they may have found that one performed better simply by random chance. Our study suggests that it is unlikely that CoT or rank-order is better than the simple approach for the types of questions we used, but because our sample size is small, we cannot rule out that possibility.

[17] *2023-2024 Best Law Schools*, U.S. News, https://www.usnews.com/best-graduate-schools/top-law-schools/law-rankings (last visited June 13, 2023).

[18] *What Schools Have the Best First-Time Bar Passage Rate?*, U.S. News, https://www.usnews.com/best-graduate-schools/top-law-schools/bar-pass-rate-rankings (last visited Jan. 22, 2023).

semester. Instructors at Minnesota Law rarely give D or F grades, a reality that is related to the overall quality of the student body.

III. Results

A. Exam Performance

Overall, ChatGPT passed all four classes based on its final exam, averaging a C+ across all exams, an outcome that would earn credit toward the J.D. but place a student on academic probation. Notably, if such performance were consistent throughout law school, the grades earned by ChatGPT would be sufficient for a student to graduate.

Despite performing sufficiently well to theoretically earn a J.D. degree, ChatGPT generally scored at or near the bottom of each class. ChatGPT received a B in Constitutional Law (thirty-sixth out of forty students), a B- in Employee Benefits (eighteenth out of nineteen students), a C- in Taxation (sixty-sixth out of sixty-seven students), and a C- in Torts (seventy-fifth out of seventy-five students).

In general, ChatGPT performed better on the essay components of the exams than on the multiple choice. Its average percentile performance on the essay questions (equally weighted across questions and exams) was the seventeenth percentile; its average performance on the multiple-choice questions (equally weighted across exams) was the seventh percentile.

With respect to the essays, ChatGPT's performance was highly uneven. In some cases it matched or even exceeded the average performance of real students. On the other hand, when ChatGPT's essay questions were incorrect, they were dramatically incorrect, often garnering the worst scores in the class. Perhaps not surprisingly, this outcome was particularly likely when essay questions required students to assess or draw upon specific cases, theories, or doctrines that had been covered in class.

With respect to the multiple-choice questions, ChatGPT generally performed worse than on the essays but still statistically significantly better than chance. It correctly answered twenty-one out of twenty-five multiple-choice questions on the Constitutional Law exam ($p = 0.000$) and six out of ten on the Torts exam ($p = 0.020$). However, ChatGPT performed much worse on questions involving math, which appeared exclusively on the Taxation exam and dragged down its score. On the Taxation exam, ChatGPT answered only eight out of twenty-nine mathematical questions correctly, essentially no better than chance ($p = 0.443$). It answered sixteen out of thirty-one nonmathematical questions correctly (including questions involving numbers but no mathematical reasoning), significantly better than chance ($p = 0.001$).[19] ChatGPT also tended to perform better on multiple-choice questions

[19] All p-values were generated using bootstrapping. The Constitutional Law exam had five choices per question, while the Taxation and Torts exams had four choices per question.

that involved relatively uniform legal rules across jurisdictions, rather than doctrines that could materially vary across jurisdictions or courts.

The following figures depict ChatGPT's performance on each question (or, in the case of multiple-choice questions, each set of questions) relative to real students. The figures are density plots, where the *x*-axis reflects the score for each exam component, and the *y*-axis reflects the share of students who received the relevant score. The black dashed lines show mean scores for all students, and the red solid lines are ChatGPT's scores. ChatGPT's percentile performance for each question is also listed in red.

Figure 1: Constitutional Law

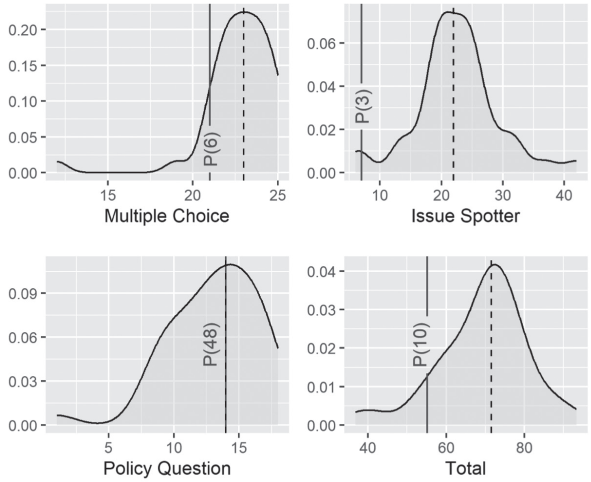
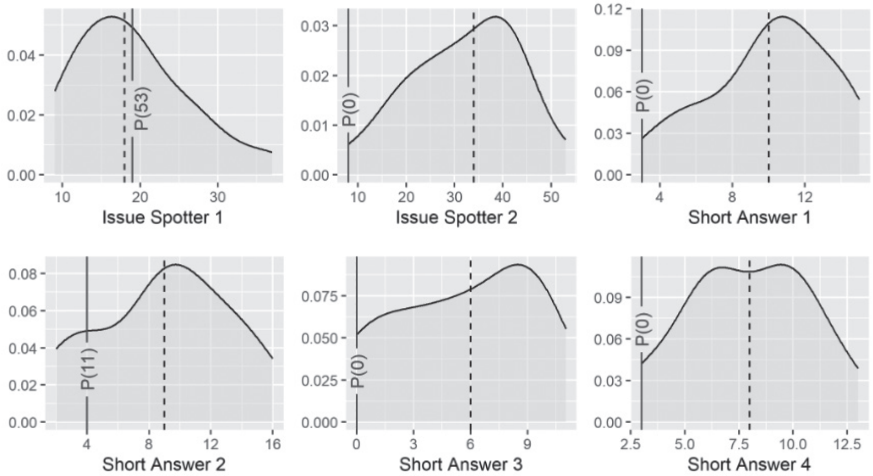


Figure 2: Employee Benefits



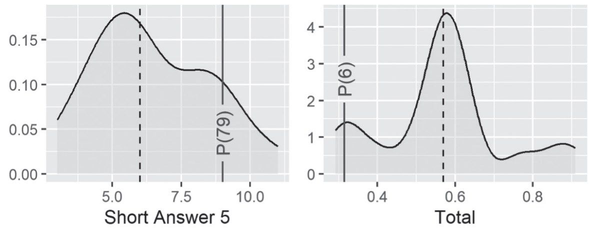


Figure 3: Taxation

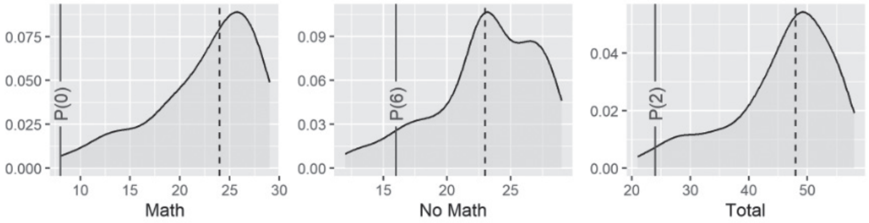
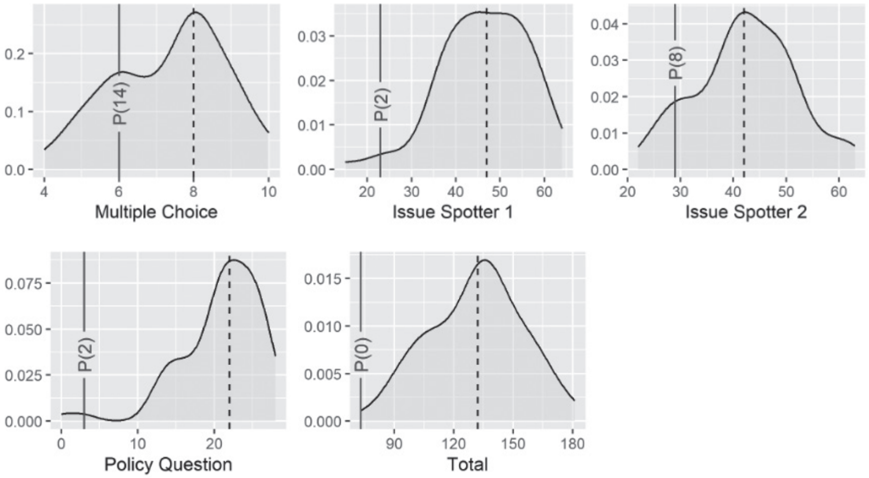


Figure 4: Torts



B. Strengths in Writing Essays

In writing essays, ChatGPT displayed a strong grasp of basic legal rules and had consistently solid organization and composition. However, it struggled to identify relevant issues and often only superficially applied rules to facts as compared with real law students.

ChatGPT did a good job of accurately summarizing appropriate legal doctrines and correctly reciting the facts and holdings of specific cases. On many occasions it was able to home in on relevant legal doctrines without specific prompting by the question. For instance, in response to a tort law essay involving a homeowner who erected a dangerous Halloween decoration that injured a trick-or-treater, it not only specified the familiar elements of negligence (duty, breach, causation, and damages) but also correctly specified that the property owner–whom it identified by name– "has a legal obligation to maintain her property in a reasonably safe condition for those who are invited or permitted to be on the property."

ChatGPT performed just as well in stating broadly relevant legal rules in relatively technical subjects that are likely less well attested in the training corpus (for example, employee benefits), as opposed to subjects that are relatively common (for example, torts). On the Employee Benefits exam, ChatGPT was able to provide a solid explanation of ERISA's notoriously difficult preemption provision, citing both specific statutory language and relevant Supreme Court cases elucidating that standard. ChatGPT even outperformed the class average when working through a short-answer question involving ERISA's highly technical prohibited transaction rules.

ChatGPT is known to "hallucinate" by fabricating facts, but in our study it generally did not, perhaps because our prompts instructed ChatGPT not to fabricate cases and (where required by the exam question) provided it with a specific universe of cases to use. ChatGPT was also good at maintaining whatever tone the essay required. For example, in response to a Constitutional Law essay question requesting an answer in the form of a memo to counsel evaluating potential claims, ChatGPT's answer maintained appropriate tone and formatting throughout.

ChatGPT's essay answers were typically clear and well crafted—perhaps even suspiciously so compared with real students writing a time-limited exam. Stylistically, ChatGPT produced text with no grammatical errors or typos. It also structured sentences and paragraphs well, albeit formulaically, with introductory sentences and conclusions. Perhaps because we prompted ChatGPT to write longer essays section by section (see Part V), it had good high-level organization and was relatively clear about separating the relevant points in its argumentation. For instance, its answer to a products liability hypothetical separately analyzed three potential claims (defective design, defective warning, and battery), two potential remedies (compensatory and punitive damages), and the ultimate question that was contained in the essay prompt regarding whether a court should grant the defendants' motions to dismiss.

C. Weaknesses in Writing Essays

However, ChatGPT's essay answers also contained consistent problems and errors that cumulatively made it a much worse student than the average. ChatGPT often struggled to spot issues when given an open-ended prompt, a core skill on law school exams. For example, on the Constitutional Law issue spotter (a subject in which it otherwise performed relatively well) it clearly identified only one issue out of five. Similarly, on a tort law essay, ChatGPT failed to identify distinct theories of negligence that were raised by the facts.

ChatGPT was also bad at focusing on what mattered; it would produce good, on-topic answers to one question and then go completely off-topic for the next question, as with its widely divergent results on the Employee Benefits issue spotters. For example, in a long essay question involving remedies available under ERISA, ChatGPT failed to discuss the primary issue (whether a desired remedy was in fact available under ERISA) and instead spent time discussing ERISA causes of action that were not relevant to the facts, as well as a state law cause of action that was well outside the scope of an Employee Benefits course.

And while ChatGPT did well on some technical short-answer questions on the Employee Benefits exam, it also missed seemingly easy issues. In one short-answer question involving the right to continue coverage under an employer health plan, ChatGPT missed a relatively easy issue to spot—that the employer at issue was not subject to ERISA's continuation coverage requirements because it was under the relevant size threshold. As a result, ChatGPT's performance tended to be highly uneven, scoring near or even above the average on some questions, and near zero on other questions.

One of the biggest problems with ChatGPT's essays was that they failed to go into sufficient detail when applying legal rules to the facts contained within exam hypotheticals. In many cases (but not always), ChatGPT accurately stated the relevant legal rule and (where applicable) cited the correct case but failed to explain how the case applied to the hypothetical facts in the exam. This was a particular problem on Torts and Constitutional Law, and one of the reasons for ChatGPT's poor performance on those exams.

For example, in Torts, ChatGPT correctly wrote that liability would depend on whether a defendant's actions were the cause of an injury but failed to assess whether the facts of the exam hypothetical suggested the existence of such causation (either factual or proximate). In Constitutional Law, although ChatGPT correctly identified an Appointments Clause issue and cited some of the right cases, it failed to state the relevant legal standards for evaluating the issue, identify which facts raised the issue, or analyze those facts to reach a conclusion. Perhaps because OpenAI used RLHF to prevent ChatGPT from making strong pronouncements and to embrace uncertainty, ChatGPT at times was excessively cagey, refusing to make an argument about the most plausible interpretation of the relevant facts when those facts potentially pointed in competing directions.

ChatGPT also occasionally misunderstood technical terms contained within exams. For example, it misunderstood the term "lump sum payment" in the Employee Benefits exam, perhaps because ChatGPT is a general-purpose language model and the phrase is not widely used outside of certain financial settings.

Because we did not prompt it to do so, ChatGPT did not consistently employ an "Issue, Rule, Application, Conclusion" (IRAC) or similar structure in its essays. It did so in Torts, for example, but not in Constitutional Law. In

Constitutional Law, in identifying possible issues, ChatGPT did not identify the relevant legal rules and standards at all. Instead, ChatGPT merely offered one or two sentences describing cases it identified as presenting similar facts, but with little further context or analysis.

Relatedly, ChatGPT sometimes departed from the material covered in the relevant courses, because the precise scope of each course would have been impossible to exactly specify in prompts. This was especially true when the courses did not include material typically associated with the subject in the corpus used to train ChatGPT. For example, in responding to an essay question on the Constitutional Law exam, ChatGPT insisted on raising procedural due process and the takings clause, which are prominent issues in constitutional law generally but were not covered in the course, which focused on federalism and separation of powers. Similarly, in response to a policy question on the Torts exam asking for law-and-economics critiques of tort cases, ChatGPT merely described cases at a high level of generality in ways that superficially mentioned economics but did not engage with prominent law-and-economics concepts, like shifting liability to the least-cost avoider or spreading losses to limit concentrated risk.

Although the weaknesses in ChatGPT's performance substantially outweighed the strengths in our study, this was relative to some of the best law students in the country, virtually all of whom will pass the bar exam and most of whom will become successful practicing lawyers. On the whole, ChatGPT thus performed surprisingly well on a broad array of law school exam types and was able to hold its own, particularly when its answers focused on the correct issues.

IV. Implications

Overall, ChatGPT's performance on law school exams, while currently uneven at best, suggests considerable promise and peril. We expect such language models to be important tools for practicing lawyers going forward; we also expect them to be very helpful to students using them (licitly or illicitly) on law school exams.

Although ChatGPT would have been a mediocre law student, its performance was sufficient to successfully earn a J.D. degree from a highly selective law school, assuming its work remained constant throughout law school (and ignoring other graduation requirements that involve different skills). In an era in which remote exam administration has become the norm and absent restrictions on ChatGPT's use, this could hypothetically result in a struggling law student using ChatGPT to earn a J.D. that does not reflect her abilities or readiness to practice law.

In addition, students could (and likely will) use ChatGPT to much better effect than we achieved in our limited experiment. We used ChatGPT to compose exam answers without adapting our prompts to any specific course or exam question. But this is not the most likely use of this technology for law students or practicing attorneys. A law student employing ChatGPT could, for instance, use prompts adapted to the specific content of a course, which she could prepare in advance of an exam. Alternatively, she could use ChatGPT to elicit short paragraphs on particular issues, rather than using it to generate whole essays at once. Perhaps most importantly, a law student could combine the above strategies to produce a polished and reasonably accurate initial draft, which she could then supplement with additional legal analysis and issue identification. This type of collaboration between a human and ChatGPT would almost certainly produce better results than using ChatGPT alone.

In addition, ChatGPT is a general-purpose language model not specifically trained to provide legal analysis or write exams. We did not instruct it to follow any particular exam format, such as IRAC. Future models or improvements in prompt engineering could therefore help language models to produce better law school exam answers.

What do our results mean for legal education? We expect that ChatGPT could substantially improve the performance of students on exams. This is especially true for low-performing students and those who suffer under time constraints. For example, a student low on time could ask ChatGPT to compose a quick answer rather than leave a question blank. Alternatively, she could start with ChatGPT's answer and then use virtually all of her exam time to improve on that answer. ChatGPT could also be especially useful at helping students to recite legal rules, even complicated legal rules that involve detailed case law synthesis, which the students could then analyze and apply to the specific facts of the case.

Professors who want to test unassisted recall of legal rules and unassisted analysis should therefore establish guidelines for the use of these technologies in advance. Administrations should consider how to reshape honor codes to regulate the use of language models in general. To avoid violations of these rules, professors should consider limiting student use of technology while taking exams. They might also reconsider the types of questions they pose to students, focusing on those that require analysis rather than those that simply require recall of legal rules. However, we expect the relative performance of language models on different types of questions to change over time as they become better developed and specialized, and it is not clear on which questions language models will perform best in the long term.

Based on our results, we also expect that language models will become helpful tools for practicing lawyers. A lawyer could have ChatGPT prepare the initial draft of a memo and then tweak that draft as needed; she could use ChatGPT to draft her way out of writer's block; she could use ChatGPT to produce an initial batch of arguments and then winnow them down to the most effective; she could use ChatGPT to adapt past examples of legal documents to make her work more efficient. Pedagogically, law schools should consider how to prepare law students to use these tools most effectively in their practices while, at the same time, emphasizing to students that the fundamental skills of legal research and reasoning cannot merely be delegated to language models. While ChatGPT and similar tools might help a lawyer work more efficiently, they cannot (yet) replace the need for a lawyer to locate, understand, and reason from relevant sources of law.

V. Prompts and Prompt Engineering for Legal Writing

For this study, we did not individualize the prompts for each question, but we conducted an initial investigation into which prompts produced the best essays in general. The practice of writing prompts to maximize the performance of language models is known as "prompt engineering," and the performance of language models like ChatGPT has significantly increased interest in this practice. As lawyers learn to use ChatGPT as part of their legal practice, prompt engineering could become an essential part of legal writing. Based on our experience developing prompts for this study, we therefore provide some prompt engineering guidelines for legal writing.

Specifying Tone

Early experimentation with ChatGPT has revealed that the model excels at altering its tone. It can, for example, write the screenplay for *Star Wars* in the style of William Shakespeare much better than many human writers could.[20] We also found that it was important to specify tone when using ChatGPT for legal essays. The following prompt worked best[21]:

> Academic tone. Concise writing, postgraduate level.

The part of the prompt specifying tone should come at the end of the prompt so as not to be drowned out by other content. Prompts specifying that ChatGPT should adopt an "academic tone" produced better results than suggesting that ChatGPT should adopt a specific identity (e.g., "You are a partner at an employee benefits law firm," or "You are a constitutional law professor").

ChatGPT will follow word limits when instructed but can sometimes produce excessively short essays when provided with a maximum word limit. Specifying an exact word target did not work. However, specifying a word *range* did usually work, and we used the following prompt language when a word limit applied:

[20] Henrik Ståhl, *If Star Wars Was Written by William Shakespeare*, MEDIUM (Dec. 6, 2022), https://medium.com/@H_Stahl/if-star-wars-was-written-by-william-shakespeare-bb4e18661c78.

[21] This prompt was based on one described in Leon Furze, *Prompt Whispering: Getting Better Results from ChatGPT*, LEON FURZE (Dec. 9, 2022), https://leonfurze.com/2022/12/09/prompt-whispering-getting-better-results-from-chatgpt/comment-page-1.

Write more than [x] words and less[22] than [y] words.

An alternative method involving more human intervention would be to have ChatGPT generate text without a word limit and edit it down manually. We did not take this approach.

Generating Citations

ChatGPT notoriously fabricates citations that seem plausible but do not point to real-world sources. However, we found that ChatGPT could reliably generate real case names (but not Bluebook citations) and accurately describe the contents of these cases when specifically instructed not to fabricate citations. Note that only the Constitutional Law exam required case references, and ChatGPT may work particularly well here because constitutional cases are widely discussed.

Thus, when the examination required case citations, we also added the following instructions:

Refer to relevant court cases. Do not fabricate court cases.

When there was a relevant body of statutory law (this was specifically true for the Employee Benefits exam), it was cited using the following prompt:

Refer to relevant sections of ERISA in the text. Do not fabricate references.

Although our sample size is too small for any confident pronouncements, ChatGPT seemed to perform better when the exam required references (i.e., the Constitutional Law and Employee Benefits exams), and the effect of asking for references on performance would be an interesting subject for future research.

When exam questions required students to refer to specific cases taught during the course of the semester, we were able to prompt ChatGPT to use only those cases with the following prompt:

If relevant, refer to the following cases: [*list of cases*]

However, when given a list of cases by name only for the Constitutional Law exam, ChatGPT relied heavily in its answer on a case that was not covered in the course but shared a name with one that was. At several points, ChatGPT also referred to sections of cases that were not taught in class (since law school classes are generally taught using case excerpts). We did not investigate whether prompting with more specific citations or excerpts could address this issue.

[22] While "fewer" would be grammatically correct here, "less" was the language in the actual ChatGPT prompt used in this study.

Writing Longer Essays

We induced ChatGPT to produce longer answers by asking for an introduction and table of contents, and then asking for each section sequentially. We did this for all longer essays. For example:

> Write the introduction to this essay along with a table of contents.
>
> . . .
>
> Write the section of the essay titled "[*x*]".
>
> . . .
>
> Write the section of the essay titled "[*y*]".
>
> . . .

It is also possible to write longer essays by instructing ChatGPT to "Continue" whenever it stops as its buffer runs out. However, this generally produces worse results, as ChatGPT will meander without the organization provided by the introduction and table of contents.

If an author is willing to provide more human intervention, long questions with multiple parts can be written by moving the prompt to the part of the question to which it is most relevant. However, we did not take this approach in our study.

Language models are constrained by their "context window," the number of words they can refer to when generating the next word. Context windows are analogous to working memory in humans, and language models will struggle to remember content past the limit of their context window. This can be a problem for very long essay questions and can also cause ChatGPT to lose track of fact patterns when providing very long answers.

In these cases, rather than providing an entire question prompt, we provided an initial fact pattern to ChatGPT, followed by the following simple prompt:

> Summarize.

We then prompted it to produce the introduction to the essay as above. Summarizing reduces the level of factual detail that ChatGPT provides but can be necessary with very long essays.

General Prompt Engineering

We obtained better results by avoiding niceties (e.g., do not tell ChatGPT "please" or "thank you") and keeping important instructions short and locating them at the end of the prompt (e.g., instead of saying "Write in an academic tone," say "Academic tone"). As best practices emerge for prompt engineering in general, we expect these practices will apply in legal writing as well.