Legal Priorities Project

Re-evaluating GPT-4's bar exam performance

Eric Martínez

LPP WORKING PAPER N° 2-2023

Re-Evaluating GPT-4's Bar Exam Performance

Eric Martínez[1]

Abstract

Perhaps the most widely touted of GPT-4's at-launch, zero-shot capabilities has been its reported 90th-percentile performance on the Uniform Bar Exam, with its reported 80-percentile-points boost over its predecessor, GPT-3.5, far exceeding that for any other exam. This paper investigates the methodological challenges in documenting and verifying the 90th-percentile claim, presenting four sets of findings that suggest that OpenAI's estimates of GPT-4's UBE percentile, though clearly an impressive leap over those of GPT-3.5, appear to be overinflated, particularly if taken as a "conservative" estimate representing "the lower range of percentiles," and moreso if meant to reflect the actual capabilities of a practicing lawyer.

First, although GPT-4's UBE score nears the 90th percentile when examining approximate conversions from February administrations of the Illinois Bar Exam, these estimates are heavily skewed towards repeat test-takers who failed the July administration and score significantly lower than the general test-taking population. Second, data from a recent July administration of the same exam suggests GPT-4's overall UBE percentile was below the 69th percentile, and ∼48th percentile on essays. Third, examining official NCBE data and using several conservative statistical assumptions, GPT-4's performance against first-time test takers is estimated to be ∼63rd percentile, including ∼42nd percentile on essays. Fourth, when examining only those who passed the exam (i.e. licensed or license-pending attorneys), GPT-4's performance is estimated to drop to ∼48th percentile overall, and ∼15th percentile on essays.

Taken together, these findings carry timely insights for the desirability and feasibility of outsourcing legally relevant tasks to AI models, as well as for the importance for AI developers to implement rigorous and transparent capabilities evaluations to help secure safe and trustworthy AI.

[1]PhD Student in Cognitive Science, Massachusetts Institute of Technology (MIT); JD, Harvard Law School. Email: ericmart@mit.edu.

Special thanks to Aixiu An, Christoph Winter, Matthijs Maas, Markus Anderljung, John Bliss, Daniel Katz, Jonas Schuett, and Rosemary Reshetar for comments and feedback. All remaining errors are my own.

Note that all code for this paper is available at the following repository link: code

1. Introduction

On March 14th, 2023, OpenAI launched GPT-4, said to be the latest milestone in the company's effort in scaling up deep learning [1]. As part of its launch, OpenAI revealed details regarding the model's "human-level performance on various professional and academic benchmarks."[1] Perhaps none of these capabilities was as widely publicized as GPT-4's performance on the Uniform Bar Examination, with OpenAI prominently displaying on various pages of its website and technical report that GPT-4 scored in or around the "90th percentile,"[1–3] or "the top 10% of test-takers,"[1, 2] and various prominent media outlets[4–8] and legal scholars[9] resharing and discussing the implications of these results for the legal profession and the future of AI.

Of course, assessing the capabilities of an AI system as compared to those of a human is no easy task,[10–15] and in the context of the legal profession specifically, there are various reasons to doubt the usefulness of the bar exam as a proxy for lawyerly competence (both for humans and AI systems), given that, for example: (a) the content on the UBE is very general and does not pertain to the legal doctrine of any jurisdiction in the United States,[16] and thus knowledge (or ignorance) of that content does not necessarily translate to knowledge (or ignorance) of relevant legal doctrine for a practicing lawyer of any jurisdiction; and (b) the tasks involved on the bar exam, particularly multiple-choice questions, do not reflect the tasks of practicing lawyers, and thus mastery (or lack of mastery) of those tasks does not necessarily reflect mastery (or lack of mastery) of the tasks of practicing lawyers.

Notwithstanding these concerns, the bar exam results appeared especially startling compared to GPT-4's other capabilities, for various reasons. Aside from the sheer complexity of the law in form[17–19] and content,[20–22] the first is that the boost in performance of GPT-4 over its predecessor GPT-3.5 (80 percentile points) far exceeded that of any other test, including seemingly related tests such as the LSAT (40 percentile points), GRE verbal (36 percentile points), and GRE Writing (0 percentile points).[2, 3]

The second is that half of the Uniform Bar Exam consists of writing essays,[16] and GPT-4 seems to have scored much lower on other exams involving writing, such as AP English Language and Composition (14th-44th percentile), AP English Literature and Composition (8th-22nd percentile) and GRE Writing ($\sim$54th percentile).[1, 2] In each of these three exams, GPT-4 failed to achieve a higher percentile performance over GPT-3.5, and failed to achieve a percentile score anywhere near the 90th percentile.

Moreover, in its technical report, GPT-4 claims that its percentile estimates are "conservative" estimates meant to reflect "the lower bound of the percentile range,"[2, p. 6] implying that GPT-4's actual capabilities may be even greater than its estimates.

Methodologically, however, there appear to be various uncertainties related to the calculation of GPT's bar exam percentile. For example, unlike the administrators of other tests that GPT-4 took, the administrators of the Uniform Bar Exam (the NCBE as well as different state bars) do not release official

percentiles of the UBE,[23, 24] and different states in their own releases almost uniformly report only passage rates as opposed to percentiles,[25, 26] as only the former are considered relevant to licensing requirements and employment prospects.

Furthermore, unlike its documentation for the other exams it tested,[2, p. 25] OpenAI's technical report provides no direct citation for how the UBE percentile was computed, creating further uncertainty over both the original source and validity of the 90th percentile claim.

The reliability and transparency of this estimate has important implications on both the legal practice front and AI safety front. On the legal practice front, there is great debate regarding to what extent and when legal tasks can and should be automated.[27–30] To the extent that capabilities estimates for generative AI in the context law are overblown, this may lead both lawyers and non-lawyers to rely on generative AI tools when they otherwise wouldn't and arguably shouldn't, plausibly increasing the prevalence of bad legal outcomes as a result of (a) judges misapplying the law; (b) lawyers engaging in malpractice and/or poor representation of their clients; and (c) non-lawyers engaging in ineffective pro se representation.

Meanwhile, on the AI safety front, there appear to be growing concerns of transparency[2] among developers of the most powerful AI systems.[32, 33] To the extent that transparency is important to ensuring the safe deployment of AI, a lack of transparency could undermine our confidence in the prospect of safe deployment of AI.[34, 35] In particular, releasing models without an accurate and transparent assessment of their capabilities (including by third-party developers) might lead to unexpected misuse/misapplication of those models (within and beyond legal contexts), which might have detrimental (perhaps even catastrophic) consequences moving forward.[36, 37]

Given these considerations, this paper investigates some of the key methodological challenges in verifying the claim that GPT-4 achieved 90th percentile performance on the Uniform Bar Examination. The paper's findings are fourfold. First, although GPT-4's UBE score nears the 90th percentile when examining approximate conversions from February administrations of the Illinois Bar Exam, these estimates appear heavily skewed towards those who failed the July administration and whose scores are much lower compared to the general test-taking population. Second, using data from a recent July administration of the same exam reveals GPT-4's percentile to be below the 69th percentile on the UBE, and ∼48th percentile on essays. Third, examining official NCBE data and using several conservative statistical assumptions, GPT-4's performance against first-time test takers is estimated to be ∼63rd percentile, including 42nd percentile on essays. Fourth, when examining only those who passed the exam,

[2]Note that transparency here is not to be confused with the interpretability or explainability of AI systems themselves, as is often used in the AI safety literature. For a discussion of the term as used more along the lines of these senses, see [31, p. 2] (arguing that making an AI system "transparent to inspection" by the programmer is one of "many socially important properties").

GPT-4's performance is estimated to drop to ∼48th percentile overall, and ∼15th percentile on essays.

Taken together, these findings suggest that OpenAI's estimates of GPT-4's UBE percentile, though clearly an impressive leap over those of GPT-3.5, are likely overinflated, particularly if taken as a "conservative" estimate representing "the lower range of percentiles," and even moreso if meant to reflect the actual capabilities of a practicing lawyer. These findings carry timely insights for the desirability and feasibility of outsourcing legally relevant tasks to AI models, as well as for the importance for generative AI developers to implement rigorous and transparent capabilities evaluations to help secure safer and more trustworthy AI.

2. Evaluating the 90th Percentile Estimate

2.1. Evidence from OpenAI

Investigating the OpenAI website, as well as the GPT-4 technical report, reveals a multitude of claims regarding the estimated percentile of GPT-4's Uniform Bar Examination performance but a dearth of documentation regarding the backing of such claims. For example, the first paragraph of the official GPT-4 research page on the OpenAI website states that "it [GPT-4] passes a simulated bar exam with a score around the top 10% of test takers."[1] This claim is repeated several times later in this and other webpages, both visually and textually, each time without explicit backing.[3]

Similarly undocumented claims are reported in the official GPT-4 Technical Report.[4] Although OpenAI details the methodology for computing most of its percentiles in A.5 of the Appendix of the technical report, there does not appear to be any such documentation for the methodology behind computing the UBE percentile. For example, after providing relatively detailed breakdowns of its methodology for scoring the SAT, GRE, SAT, AP, and AMC, the report states that "[o]ther percentiles were based on official score distributions," followed by a string of references to relevant sources.[2, p. 25]

Examining these references, however, none of the sources contains any information regarding the Uniform Bar Exam, let alone its "official score distributions."[2, p. 22-23] Moreover, aside from the Appendix, there are no other direct references to the methodology of computing UBE scores, nor any indirect references aside from a brief acknowledgement thanking "our collaborators at Casetext and Stanford CodeX for conducting the simulated bar exam."[2, p. 18]

[3] For example, near the top of the GPT-4 product page is displayed a reference to GPT-4's 90th percentile Uniform Bar Exam performance as an illustrative example of how "GPT-4 outperforms ChatGPT by scoring in higher approximate percentiles among test-takers."[3]

[4] As with the official website, the technical report (page 6) claims that GPT-4 "passes a simulated version of the Uniform Bar Examination with a score in the top 10% of test takers."[2] This attested result is presented visually in Table 1 and Figure 4. Furthermore, the caption of Figure 4 goes on to claim that its estimates aim to be "conservative" by "report[ing] the lower end of the range of percentiles," suggesting that GPT-4's capabilities may be much higher than those reported in the technical report.[2, p. 6]

2.2. Evidence from GPT-4 Passes the Bar

Another potential source of evidence for the 90th percentile claim comes from an early draft version of the paper, "GPT-4 passes the bar exam," written by the administrators of the simulated bar exam referenced in OpenAI's technical report.[38] The paper is very well-documented and transparent about its methodology in computing raw and scaled scores, both in the main text and in its comprehensive appendices. Unlike the GPT-4 technical report, however, the focus of the paper is not on percentiles but rather on the model's scaled score compared to that of the average test taker, based on publicly available NCBE data. In fact, one of the only mentions of percentiles is in a footnote, where the authors state, in passing: "Using a percentile chart from a recent exam administration (which is generally available online), ChatGPT would receive a score below the 10th percentile of test-takers while GPT-4 would receive a combined score approaching the 90th percentile of test-takers." [38, p. 10]

2.3. Evidence Online

As explained by [23], The National Conference of Bar Examiners (NCBE), the organization that writes the Uniform Bar Exam (UBE) does not release UBE percentiles.[5] Because there is no official percentile chart for UBE, all generally available online estimates are unofficial. Perhaps the most prominent of such estimates are the percentile charts from pre-July 2019 Illinois bar exam,[6] which provide an "approximate" conversion to the UBE given the similarity between the two exams.[23][7]

Examining these approximate conversion charts, however, yields conflicting results. For example, although the percentile chart from the February 2019 administration of the Illinois Bar Exam estimates a score of 300 (2-3 points higher thatn GPT-4's score) to be at the 90th percentile, this estimate is heavily skewed compared to the general population of July exam takers,[8] since the majority of those who take the February exam are repeat takers who failed the

[5] As the website JD Advising points out: "The National Conference of Bar Examiners (NCBE), the organization that writes the Uniform Bar Exam (UBE) does not release UBE percentiles."[23] Instead, the NCBE and state bar examiners tend to include in their press releases much more general and limited information, such as mean MBE scores and the percentage of test-takers who passed the exam in a given administration.[24–26]

[6] Note that Starting in July 2019, Illinois began administering the Uniform Bar Exam [39], and accordingly stopped releasing official percentile charts. Thus, the generally available Illinois percentile charts are based on pre-UBE Illinois bar exam data.

[7] In addition to the Illinois conversion chart, some sources often make claims about percentiles of certain scores without clarifying the source of those claims. See, for example, [40]. There are also several generally available unofficial online calculators, which either calculate an estimated percentile of an MBE score based on official NCBE data,[41] or make other non-percentile-related calculations, such as estimated scaled score.[42]

[8] For example, according to [25], the pass rate in Illinois for the February 2023 administration was 43%, compared to 68% for the July administration.

July exam,[43][9] and repeat takers score much lower[10] and are much more likely to fail than are first-timers.[11]

Indeed, examining the latest available percentile chart for the July exam estimates GPT-4's UBE score to be ∼68th percentile, well below the 90th percentile figure cited by OpenAI.[45].

3. Towards a More Accurate Estimate

Although using the July bar exam percentiles from the Illinois Bar would seem to yield a more accurate estimate than the February data, the July figure is also biased towards lower scorers, since approximately 23% of test takers in July nationally are estimated to be re-takers and score, for example, 16 points below first-timers on the MBE.[46] Limiting the comparison to first-timers would provide a more accurate comparison that avoids double-counting those who have are taking the exam again after failing once or more.

Relatedly, although (virtually) all licensed attorneys have passed the bar,[12] not all those who take the bar become attorneys. To the extent that GPT-4's UBE percentile is meant to serve as a proxy for its performance against other attorneys, a more valid comparison would not only limit the sample to first-timers but also to those who achieved a passing score.

Moreover, the data discussed above is based on purely Illinois Bar exam data, which (at the time of the chart) was similar but not identical to the UBE in its content and scoring,[23] whereas a more accurate estimate would be derived more directly from official NCBE sources.

3.1. Methods

To account for the issues with both OpenAI's estimate as well the July estimate, more accurate estimates (for GPT-3.5 and GPT-4) were sought to be computed here based on first-time test-takers, including both (a) first-time test-takers overall, and (b) those who passed.

To do so, the parameters for a normal distribution of scores were separately estimated for the MBE and essay components (MEE + MPT), as well as the UBE score overall.[13]

[9] According to [43], for the 2021 February administration in Illinois, 284 takers were first-time takers, as compared to 426 repeaters.

[10] For example, for the July administration, the 50th-percentile UBE-converted score was approximately 282[44], whereas for the February exam, the 50th-percentile UBE-converted score was approximately 264.[44]

[11] For example, according to [25], the pass rate among first-timers in the February 2023 administration in Illinois was 62%, compared to 35% for repeat takers.

[12] One notable exception was made in 2020 due to COVID, for example, as the Supreme Court of the state of Washington granted a "diploma privilege" which allowed recent law graduates "to be admitted to the Washington State Bar Association and practice law in the state without taking the bar exam.":[47]

[13] A normal distribution of scores was assumed, given that (a) standardized tests are normalized and aim for a normal distribution [48], (b) UBE is a standardized test, and

With regard to the MBE, although NCBE has publicly released the average MBE scores of first-time test takers for recent exam administrations,[46] it has not released other official information regarding the distribution of first-timer scores, such as the mean of MEE or MPT scores, nor percentile or standard-deviation data of any part of the exam.

Thus, to simulate the distribution of first-timers MBE scores, the official first-timer mean (143.8) was combined with the estimated standard deviation of official July MBE scores (this was computed using publicly available data on NCBE website).[49]

Given that the essay component is scaled to the MBE,[50] such that the mean and standard deviation of the essays are approximately equivalent to those of the MBE scores,[44, 45, 50] equivalent distribution was assumed in this study's computation, as well.[14]

With regard to the UBE, given that the MBE and essay components comprise the entirety of the UBE,[51] the overall mean of UBE scores was computed by summing the means of the MBE and essay components, leading to an overall mean of 287.6. Because the MBE and essay scores are unlikely to be statistically independent nor completely statistically dependent, the standard deviation of overall UBE scores was computed independently (as opposed to, for example, doubling or copying the standard-deviation of the MBE or essay score distributions), using the estimated standard deviation of Illinois Bar exam data (estimated by feeding the values and percentiles of the July Illinois Bar exam data into an optimization function in R).[15].

After simulating the distribution of each component of the UBE along with the distribution of the UBE overall, the percentiles of the performance of GPT-3.5 and GPT-4's on each component of the UBE, along with their scores on the UBE overall, could be directly computed to those of first-time test-takers.

Finally, to compute GPT's performance relative to qualified attorneys, a separate percentile was computed after removing all UBE scores below 270, which is the most common score cutoff for states using the UBE.[52] To compute models' performance on the individual components relative to qualified attorneys, a separate percentile was likewise computed after removing all subscores below

(c) official visual estimates of MBE scores, both for February and July, appear to follow an approximately normal distribution. [49]

[14]If anything, this assumption would lead to a conservative (that is, generous) estimate of GPT-4's percentile, since percentiles for a given essay score tend to be slightly lower than those for a given MBE score. For example, according to the conversion chart of the Illinois bar exam for the July administration, a score of 145 on the MBE was estimated to be at the 61st percentile, while the same score on the essay component was estimated to be at the 59th percentile.[45]

[15]This was computed using the optim() function using R's "stats" package.

135.[16]

3.2. Results

Table 1: Estimated Percentile of GPT-4's Uniform Bar Examination Performance

Test-Taking Population	Section of Exam		
	UBE	MBE	MEE + MPT
July Test-Takers	68th	86th	48th
All First-Timers	63rd	79th	42nd
Qualified Attorneys	48th	69th	15th

3.2.1. Performance against first-time test-takers

Results are visualized in Table 1. For each component of the UBE, as well as the UBE overall, GPT-4's estimated percentile among first-time July test takers is less than that of both the OpenAI estimate and the July estimate that include repeat takers.

With regard to the aggregate UBE score, GPT-4 scored in the 63rd percentile as compared to the ∼90th percentile February estimate and the ∼68th percentile July estimate. With regard to MBE, GPT-4 scored in the ∼79th percentile as compared to the ∼95th percentile February estimate and the 86th percentile July estimate. With regard to MEE + MPT, GPT-4 scored in the ∼42nd percentile as compared to the ∼69th percentile February estimate and the ∼48th percentile July estimate.

With regard to GPT-3.5, its aggregate UBE score among first-timers was in the ∼1st percentile, as compared to the ∼2nd percentile February estimate and ∼1st percentile July estimate. Its MBE subscore was in the ∼6th percentile, compared to the ∼10th percentile February estimate ∼7th percentile July estimate. Its essay subscore was in the ∼0th percentile, compared to the ∼1st percentile February estimate and ∼0th percentile July estimate.

3.2.2. Performance against qualified attorneys

Predictably, when limiting the sample to those who passed the bar, the models' percentile dropped further.

With regard to the aggregate UBE score, GPT-4 scored in the ∼48th percentile. With regard to MBE, GPT-4 scored in the ∼69th percentile, whereas for the MEE + MPT, GPT-4 scored in the ∼15th percentile.

With regard to GPT-3.5, its aggregate UBE score among qualified attorneys was 0th percentile, as were its percentiles for both subscores.

[16] Note that this assumes that all those who "failed" a subsection failed the bar overall. Since scores on the two portions of the exam are likely to be highly but not directly correlated, this assumption is implausible. However, its percentile predictions would still hold true, on average, for the two subsections–that is, to the extent that it leads to a slight underestimate of the percentile on one subsection it would lead to a commensurate overestimate on the other.

Table 2: Estimated Percentile Leap from GPT-3.5 to GPT-4 on Uniform Bar Examination

Test-Taking Population	Section of Exam		
	UBE	MBE	MEE + MPT
July Test-Takers	1st-68th	7th-86th	0th-48th
All First-Timers	1st-63rd	6th-79th	0th-42nd
Qualified Attorneys	0th-48th	0th-69th	0th-15th

4. Discussion

This paper has investigated the issue of OpenAI's claim of GPT-4's 90th percentile UBE performance, resulting in four main findings. The first finding is that although GPT-4's UBE score approaches the 90th percentile when examining approximate conversions from February administrations of the Illinois Bar Exam, these estimates are heavily skewed towards low scorers, as the majority of test-takers in February failed the July administration and tend to score much lower than the general test-taking population. The second finding is that using July data from the same source would result in an estimate of ∼68th percentile, including below average performance on the essay portion. The third finding is that comparing GPT-4's performance against first-time test takers would result in an estimate of ∼63rd percentile, including ∼42nd percentile on the essay portion. The fourth main finding is that when examining only those who passed the exam, GPT-4's performance is estimated to drop to ∼48th percentile overall, and ∼15th percentile on essays.

Taken together, these findings suggest that OpenAI's estimates of GPT-4's performance against human test-takers, though clearly an impressive leap over that of GPT-3.5, are a significant overestimate, particularly if taken as a "conservative" calculation representing "the lower range of percentiles,"[2] and particularly if intended to represent GPT-4's abilities compared to those of a practicing lawyer.

Of course, assessing the capabilities of an AI system as compared to those of a practicing lawyer is no easy task. Scholars have identified several theoretical and practical difficulties in creating accurate measurement scales to assess AI capabilities and have pointed out various issues with some of the current scales.[10–12] Relatedly, some have pointed out that simply observing that GPT-4 under- or over-performs at a task in some setting is not necessarily reliable evidence that it (or some other LLM) is capable or incapable of performing that task in general.[13–15]

In the context of legal profession specifically, there are various reasons to doubt the usefulness of UBE percentile as a proxy for lawyerly competence (both for humans and AI systems), given that, for example: (a) the content on the UBE is very general and does not pertain to the legal doctrine of any jurisdiction in the United States,[16] and thus knowledge (or ignorance) of that content does not necessarily translate to knowledge (or ignorance) of relevant legal doctrine for a practicing lawyer of any jurisdiction; (b) the tasks involved

on the bar exam, particularly multiple-choice questions, do not reflect the tasks of practicing lawyers, and thus mastery (or lack of mastery) of those tasks does not necessarily reflect mastery (or lack of mastery) of the tasks of practicing lawyers; and (c) given the lack of direct professional incentive to obtain higher than a passing score (typically no higher than 270),[52] obtaining a particularly high score or percentile past this threshold is less meaningful than for other exams (e.g. LSAT), where higher scores are taken into account for admission into select institutions.[53]

Setting these objections aside, however, to the extent that one believes the UBE to be a valid proxy for lawyerly competence, these results suggest GPT-4 to be substantially less lawyerly competent than previously assumed, as GPT-4's score against likely attorneys (i.e. those who actually passed the bar) is ∼48th percentile. Moreover, when just looking at the essays, which more closely resemble the tasks of practicing lawyers and thus more plausibly reflect lawyerly competence, GPT-4's performance falls in the bottom ∼15th percentile.

The lack of precision and transparency in OpenAI's reporting of GPT-4's UBE performance has implications for both the current state of the legal profession and the future of AI safety. On the legal side, there appear to be at least two sets of implications. On the one hand, to the extent that lawyers put stock in the bar exam as a proxy for general legal competence, the results might give practicing lawyers at least a mild temporary sense of relief regarding the security of the profession, given that the majority of lawyers perform better than GPT on the component of the exam (essay-writing) that seems to best reflect their day-to-day activities (and by extension, the tasks that would likely need to be automated in order to supplant lawyers in their day-to-day professional capacity).

On the other hand, the fact that GPT-4's reported "90th percentile" capabilities were so widely publicized might pose some concerns that lawyers and non-lawyers may use GPT-4 for complex legal tasks for which it is incapable of adequately performing, plausibly increasing the rate of (a) misapplication of the law by judges; (b) professional malpractice by lawyers; and (c) ineffective pro se representation and/or unauthorized practice of law by non-lawyers. From a legal education standpoint, law students who overestimate GPT-4's UBE capabilities might also develop an unwarranted sense of apathy towards developing critical legal-analytical skills, particularly if under the impression that GPT-4's level of mastery of those skills already surpasses that to which a typical law student could be expected to reach.

On the AI front, these findings raise concerns both for the transparency[17] of capabilities research and the safety of AI development more generally. In particular, to the extent that one considers transparency to be an important prerequisite for safety,[34] these findings underscore the importance of implementing rigorous transparency measures so as to reliably identify potential warning signs of transformative progress in artificial intelligence as opposed to creating

[17]As noted above, "transparency" here is not to be confused with the interpretability or explainability of the AI system, as is often used in the AI safety literature.

a false sense of alarm or security.[54] Implementing such measures could help ensure that AI development, as stated in OpenAI's charter, is a "value-aligned, safety-conscious project" as opposed to becoming "a competitive race without time for adequate safety precautions."[55]

Of course, the present study does not discount the progress that AI has made in the context of legally relevant tasks; after all, the improvement in UBE performance from GPT-3.5 to GPT-4 as estimated in this study remains impressive (arguably equally or even more so given that GPT-3.5's performance is also estimated to be significantly lower than previously assumed), even if not as flashy as the 10th-90th percentile boost of OpenAI's official estimation. Nor does the present study discount the seemingly inevitable future improvement of AI systems to levels far beyond their present capabilities, or, as phrased in *GPT-4 Passes the Bar Exam*, that the present capabilities "highlight the floor, not the ceiling, of future application."[38, 11]

To the contrary, given the inevitable rapid growth of AI systems, the results of the present study underscore the importance of implementing rigorous and transparent evaluation measures to ensure that both the general public and relevant decision-makers are made appropriately aware of the system's capabilities, and to prevent these systems from being used in an unintentionally harmful or catastrophic manner.

References

[1] OpenAI.: GPT 4. Accessed: 2023-04-24. https://openai.com/research/gpt-4.

[2] OpenAI.: GPT-4 Technical Report. Preprint submitted to arXiv. Available from: https://arxiv.org/abs/2303.08774.

[3] OpenAI.: GPT-4 is OpenAI's most advanced system, producing safer and more useful responses. Accessed: 2023-04-24. Available from: https://openai.com/product/gpt-4.

[4] Koetsier J.: GPT-4 Beats 90% of Lawyers Trying to Pass the Bar. Forbes. Available from: https://www.forbes.com/sites/johnkoetsier/2023/03/14/gpt-4-beats-90-of-lawyers-trying-to-pass-the-bar/?sh=b40c88d30279.

[5] Caron P.: GPT-4 Beats 90% Of Aspiring Lawyers On The Bar Exam. TaxProf Blog. Accessed: 2023-04-24. Available from: https://taxprof.typepad.com/taxprof_blog/2023/03/gpt-4-beats-90-of-aspiring-lawyers-on-the-bar-exam.html.

[6] Weiss DC.: Latest version of ChatGPT aces bar exam with score nearing 90th percentile. ABA Journal. Accessed: 2023-04-24. Available from: https://www.abajournal.com/web/article/latest-version-of-chatgpt-aces-the-bar-exam-with-score-in-90th-percentile.

[7] Wilkins S.: How GPT-4 Mastered the Entire Bar Exam, and Why That Matters. Law.com. Accessed: 2023-04-24. Available from: https://www.law.com/legaltechnews/2023/03/17/how-gpt-4-mastered-the-entire-bar-exam-and-why-that-matters/?slreturn=20230324023302.

[8] Patrice J.: New GPT-4 Passes All Sections Of The Uniform Bar Exam. Maybe This Will Finally Kill The Bar Exam. Above the Law. Available from: https://abovethelaw.com/2023/03/new-gpt-4-passes-all-sections-of-the-uniform-bar-exam-maybe-this-will-finally-kill-the

[9] Schwarcz D, Choi JH. AI Tools for Lawyers: A Practical Guide. Available at SSRN. 2023;.

[10] Hernandez-Orallo J. AI evaluation: On broken yardsticks and measurement scales. In: Workshop on Evaluating Evaluation of Ai Systems at AAAI; 2020. .

[11] Burden J, Hernández-Orallo J. Exploring ai safety in degrees: Generality, capability and control. In: Proceedings of the Workshop on Artificial Intelligence Safety (SafeAI 2020) co-located with 34th AAAI Conference on Artificial Intelligence (AAAI 2020). ceur-ws. org; 2020. p. 36–40.

[12] Raji ID, Bender EM, Paullada A, Denton E, Hanna A. AI and the everything in the whole wide world benchmark. arXiv preprint arXiv:211115366. 2021;.

[13] Bowman S. The dangers of underclaiming: Reasons for caution when reporting how NLP systems fail. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2022. p. 7484–7499.

[14] Bowman SR. Eight things to know about large language models. arXiv preprint arXiv:230400612. 2023;.

[15] Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. arXiv preprint arXiv:220511916. 2022;.

[16] : Uniform Bar Examination. National Conference of Bar Examiners. Accessed: 2023-04-24. Available from: https://www.ncbex.org/exams/ube/.

[17] Martinez E, Mollica F, Gibson E. Poor writing, not specialized concepts, drives processing difficulty in legal language. Cognition. 2022;224:105070.

[18] Martinez E, Mollica F, Gibson E. So much for plain language: An analysis of the accessibility of United States federal laws (1951-2009). In: Proceedings of the Annual Meeting of the Cognitive Science Society. vol. 44; 2022. .

[19] Martinez E, Mollica F, Gibson E. Even lawyers don't like legalese. Proceedings of the National Academy of Sciences. forthcoming;.

[20] Katz DM, Bommarito MJ. Measuring the complexity of the law: the United States Code. Artificial intelligence and law. 2014;22:337–374.

[21] Ruhl J, Katz DM, Bommarito MJ. Harnessing legal complexity. Science. 2017;355(6332):1377–1378.

[22] Bommarito II MJ, Katz DM. Measuring and modeling the US regulatory ecosystem. Journal of Statistical Physics. 2017;168:1125–1135.

[23] Advising J.: July 2018 UBE Percentiles Chart. JD Advising. Accessed: 2023-04-24. Available from: https://jdadvising.com/july-2018-ube-percentiles-chart/.

[24] Examiner TB.: Statistics. The Bar Examiner. Accessed: 2023-04-24. Available from: https://thebarexaminer.ncbex.org/statistics/.

[25] of Bar Examiners NC.: Bar Exam Results by Jurisdiction. National Conference of Bar Examiners. Accessed: 2023-04-24. Available from: https://www.ncbex.org/statistics-and-research/bar-exam-results/.

[26] of Law Examiners TNYSB.: NYS Bar Exam Statistics. The New York State Board of Law Examiners. Available from: https://www.nybarexam.org/examstats/estats.htm.

[27] Winter C, Hollman N, Manheim D. Value alignment for advanced artificial judicial intelligence. American Philosophical Quarterly. 2023;60(2):187–203.

[28] Crootof R, Kaminski ME, Price II WN. Humans in the Loop. Vanderbilt Law Review, Forthcoming. 2023;.

[29] Markou C, Deakin S. Is Law Computable? From Rule of Law to Legal Singularity. From Rule of Law to Legal Singularity (April 30, 2020) University of Cambridge Faculty of Law Research Paper. 2020;.

[30] Winter CK. The Challenges of Artificial Judicial Decision-Making for Liberal Democracy. In: Judicial Decision-Making: Integrating Empirical and Theoretical Perspectives. Springer; 2022. p. 179–204.

[31] Bostrom N, Yudkowsky E. The ethics of artificial intelligence. In: Artificial intelligence safety and security. Chapman and Hall/CRC; 2018. p. 57–69.

[32] Ray T.: With GPT-4, OpenAI opts for secrecy versus disclosure. ZDNet. Available from: https://www.zdnet.com/article/with-gpt-4-openai-opts-for-secrecy-versus-disclosure/.

[33] Stokel-Walker C.: Critics denounce a lack of transparency around GPT-4's tech. Fast Company. Available from: https://www.fastcompany.com/90866190/critics-denounce-a-lack-of-transparency-around-gpt-4s-tech.

[34] Brundage M, Avin S, Wang J, Belfield H, Krueger G, Hadfield G, et al. Toward trustworthy AI development: mechanisms for supporting verifiable claims. arXiv preprint arXiv:200407213. 2020;.

[35] Li B, Qi P, Liu B, Di S, Liu J, Pei J, et al. Trustworthy ai: From principles to practices. ACM Computing Surveys. 2023;55(9):1–46.

[36] Ngo R. The alignment problem from a deep learning perspective. arXiv preprint arXiv:220900626. 2022;.

[37] Carlsmith J. Is Power-Seeking AI an Existential Risk? arXiv preprint arXiv:220613353. 2022;.

[38] Katz DM, Bommarito MJ, Gao S, Arredondo P. Gpt-4 passes the bar exam. Available at SSRN 4389233. 2023;.

[39] : Overview of Illinois Bar Exam. University of Illinois Chicago. Accessed: 2023-04-24. Available from: https://law.uic.edu/student-support/academic-achievement/bar-exam-information/illinois-bar-exam/.

[40] Lang C.: What is a good bar exam score? Test Prep Insight. Available from: https://www.testprepinsight.com/what-is-a-good-bar-exam-score.

[41] : February MBE Scaling and Percentiles. UBEEssays.com. Available from: https://ubeessays.com/feb-mbe-percentiles/.

[42] com MR.: Bar Exam Calculators. Accessed: 2023-05-02. Available from: https://mberules.com/bar-exam-calculators/?__cf_chl_tk=lTwxFyYWOZqBwTAenLs0TzDfAuvawkHeH2GaXU1PQo0-1683060961-0-gaNycGzNDBA.

[43] Examiner TB.: First-Time Exam Takers and Repeaters in 2021. The Bar Examiner. Accessed: 2023-04-24. Available from: https://thebarexaminer.ncbex.org/2021-statistics/first-time-exam-takers-and-repeaters-in-2021/.

[44] : Illinois February 2019 Bar Examination Percentile Equivalents of MBE, Essay, and Total Scale Scores. Illinois Board of Admissions to the Bar. Accessed: 2023-04-24. Available from: https://www.ilbaradmissions.org/percentile-equivalent-charts-february-2019.

[45] : Illinois July 2018 Bar Examination Percentile Equivalents of MBE, Essay, and Total Scale Scores. Illinois Board of Admissions to the Bar. Accessed: 2023-04-24. Available from: https://www.ilbaradmissions.org/percentile-equivalent-charts-july-2018.

[46] Reshetar R. The Testing Column: Why Are February Bar Exam Pass Rates Lower than July Pass Rates? The Bar Examiner. 2022;91(1):51–53.

[47] : Supreme Court Grants Diploma Privilege. Washington State Bar Association. Accessed: 2023-04-24. Available from: https://wsba.org/news-events/latest-news/news-detail/2020/06/15/state-supreme-court-grants-diploma-privilege.

[48] Kubiszyn T, Borich GD. Educational testing and measurement. John Wiley & Sons; 2016.

[49] : 2022 MBE National Score Distributions. The National Bar Examiner. Accessed: 2023-04-24. Available from: https://thebarexaminer.ncbex.org/2022-statistics/the-multistate-bar-examination-mbe/#step3.

[50] Albanese MA. The Testing Column: Scaling: It's Not Just for Fish or Mountains. The Bar Examiner. 2014 12;83(4):50–56.

[51] : UBE scores. National Conference of Bar Examiners. Accessed: 2023-05-03. Available from: https://www.ncbex.org/exams/ube/scores/.

[52] : Minimum Passing UBE Score by Jurisdiction. National Conference of Bar Examiners. Accessed: 2023-04-24. Available from: https://www.ncbex.org/exams/ube/score-portability/minimum-scores/.

[53] : Best Law Schools. US News and World Report. Available from: https://www.usnews.com/best-graduate-schools/top-law-schools/law-rankings.

[54] Zoe Cremer C, Whittlestone J. Artificial Canaries: Early Warning Signs for Anticipatory and Democratic Governance of AI. 2021;.

[55] OpenAI.: OpenAI Charter. Available from: https://openai.com/charter.