INSTITUTE
FOR LAW & AI

Re-evaluating GPT-4's bar exam performance

LawAI Working Paper Series, No. 1-2023 | Eric Martínez | May 2023 law-ai.org

Artificial Intelligence and Law
https://doi.org/10.1007/s10506-024-09396-9

ORIGINAL RESEARCH

Re-evaluating GPT-4's bar exam performance

Eric Martínez[1]

Accepted: 30 January 2024

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. This paper begins by investigating the methodological challenges in documenting and verifying the 90th-percentile claim, presenting four sets of findings that indicate that OpenAI's estimates of GPT-4's UBE percentile are overinflated. 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 ~62nd 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. In addition to investigating the validity of the percentile claim, the paper also investigates the validity of GPT-4's reported scaled UBE score of 298. The paper successfully replicates the MBE score, but highlights several methodological issues in the grading of the MPT + MEE components of the exam, which call into question the validity of the reported essay score. Finally, the paper investigates the effect of different hyperparameter combinations on GPT-4's MBE performance, finding no significant effect of adjusting temperature settings, and a significant effect of few-shot chain-of-thought prompting over basic zero-shot prompting. 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.

Note that all code for this paper is available at the following repository link: https://osf.io/c8ygu/?view only=dcc617accc464491922b77414867a066.

Extended author information available on the last page of the article

Published online: 30 March 2024

Keywords NLP · Legal NLP · Legal analytics · Natural language processing · Machine learning · Artificial intelligence · Artificial intelligence and law · Law and technology · Legal profession

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 (OpenAI 2023a). As part of its launch, OpenAI revealed details regarding the model's "human-level performance on various professional and academic benchmarks" (OpenAI 2023a). 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," (OpenAI 2023a, b, n.d.) or "the top 10% of test-takers," (OpenAI 2023a, b) and various prominent media outlets (Koetsier 2023; Caron 2023; Weiss 2023; Wilkins 2023; Patrice 2023) and legal scholars (Schwarcz and Choi 2023) 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 (Hernandez-Orallo 2020; Burden and Hernández-Orallo 2020; Raji et al. 2021; Bowman 2022, 2023; Kojima et al. 2022), 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 (National Conference of Bar Examiners n.d.-h), 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.

Moreover, although the UBE is a closed-book exam for humans, GPT-4's huge training corpus largely distilled in its parameters means that it can effectively take the UBE "open-book", indicating that UBE may not only be an accurate proxy for lawyerly comptetence but is also likely to provide an overly favorable estimate of GPT-4's lawyerly capabilities relative to humans.

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 (Martinez et al. 2022a, b, in press) and content (Katz and Bommarito 2014; Ruhl et al. 2017; Bommarito and Katz 2017), 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) (OpenAI 2023b, n.d.).

The second is that half of the Uniform Bar Exam consists of writing essays (National Conference of Bar Examiners n.d.-h),[1] 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 (~54th percentile) (OpenAI 2023a, b). 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," (OpenAI 2023b, 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 (JD Advising n.d.-b; Examiner n.d.-b), and different states in their own releases almost uniformly report only passage rates as opposed to percentiles (National Conference of Bar Examiners n.d.-c; The New York State Board of Law Examiners n.d.), as only the former are considered relevant to licensing requirements and employment prospects.

Furthermore, unlike its documentation for the other exams it tested (OpenAI 2023b, 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 (Winter et al. 2023; Crootof et al. 2023; Markou and Deakin 2020; Winter 2022). 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.

[1] Note that Uniform Bar Exam (UBE) has multiple components, including: (a) the Multistate Bar Exam (MBE), a 6 h, 200-question multiple choice test (National Conference of Bar Examiners n.d.-c, d) the Multistate Essay Exam (MEE), a 3 h, six-part essay exam (National Conference of Bar Examiners n.d.-e); and (c) the Multistate Practice Exam (MPT), a 3 h, two-part "closed universe" essay exam (National Conference of Bar Examiners n.d.-f). The exam is graded on a scale of 400. The MBE and essays (MEE + MPT) are each graded on a scale of 200 (National Conference of Bar Examiners n.d.-g). Thus, essays and multiple choice are each worth half of an examinee's score.

Meanwhile, on the AI safety front, there appear to be growing concerns of transparency[2] among developers of the most powerful AI systems (Ray 2023; Stokel-Walker 2023). 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 (Brundage et al. 2020; Li et al. 2023). 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 (Ngo 2022; Carlsmith 2022).

Given these considerations, this paper begins by investigating 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 in this regard 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 ~62nd percentile, including 42 percentile on essays. Fourth, 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.

Next, whereas the above four findings take for granted the scaled score achieved by GPT-4 as reported by OpenAI, the paper then proceeds to investigate the validity of that score, given the importance (and often neglectedness) of replication and reproducibility within computer science and scientific fields more broadly (Cockburn et al. 2020; Echtler and Häußler 2018; Jensen et al. 2023; Schooler 2014; Shrout and Rodgers 2018). The paper successfully replicates the MBE score of 158, but highlights several methodological issues in the grading of the MPT + MEE components of the exam, which call into question the validity of the essay score (140).

Finally, the paper also investigates the effect of adjusting temperature settings and prompting techniques on GPT-4's MBE performance, finding no significant effect of adjusting temperature settings on performance, and some significant effect of prompt engineering on model performance when compared to a minimally tailored baseline condition.

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

[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 (Bostrom and Yudkowsky 2018, p. 2) (arguing that making an AI system "transparent to inspection" by the programmer is one of "many socially important properties").

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" (OpenAI 2023a). 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 (OpenAI 2023b, p. 25).

Examining these references, however, none of the sources contains any information regarding the Uniform Bar Exam, let alone its "official score distributions" (OpenAI 2023b, pp. 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" (OpenAI 2023b, p. 18).

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

[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" (OpenAI n.d.).

[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" (OpenAI 2023b). This attested result is presented visually in Table 1 and Fig. 1.

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	62rd	79th	42nd
Qualified attorneys	45th	69th	15th

administrators of the simulated bar exam referenced in OpenAI's technical report (Katz et al. 2023). 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". (Katz et al. 2023, p. 10)

2.3 Evidence online

As explained by JD Advising (n.d.-b), 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. Pre-2019,[6] Illinois, unlike other states, provided percentile charts of their own exam that allowed UBE test-takers to estimate their approximate percentile given the similarity between the two exams (JD Advising n.d.-b).[7]

Examining these approximate conversion charts, however, yields conflicting results. For example, although the percentile chart from the February 2019

[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" (JD Advising n.d.-b). 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 (Examiner n.d.-c; National Conference of Bar Examiners n.d.-c; The New York State Board of Law Examiners n.d.)

[6] Note that Starting in July 2019, Illinois began administering the Uniform Bar Exam (University of Illinois Chicago n.d.), 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 (Lang 2023). There are also several generally available unofficial online calculators, which either calculate an estimated percentile of an MBE score based on official NCBE data (UBEEssays.com 2019), or make other non-percentile-related calculations, such as estimated scaled score (Rules.com n.d.)

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 July exam (Examiner n.d.-a),[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 (Illinois Board of Admissions to the Bar 2018).

3 Towards a more accurate percentile 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 (Reshetar 2022). Limiting the comparison to first-timers would provide a more accurate comparison that avoids double-counting those who have taken 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 reflect its performance against other attorneys, a more appropriate 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 (JD Advising n.d.-b), whereas a more accurate estimate would be derived more directly from official NCBE sources.

[8] For example, according to (National Conference of Bar Examiners n.d.-b), the pass rate in Illinois for the February 2023 administration was 43%, compared to 68% for the July administration.

[9] According to (Examiner n.d.-a), 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 (Illinois Board of Admissions to the Bar 2019), whereas for the February exam, the 50th-percentile UBE-converted score was approximately 264 (Illinois Board of Admissions to the Bar 2019)

[11] For example, according to (National Conference of Bar Examiners n.d.-b), 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.": (Washington State Bar Association 2020)

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]

Assuming that UBE scores (as well as MBE and essay subscores) are normally distributed, percentiles of GPT's score can be directly computed after computing the parameters of these distributions (i.e. the mean and standard deviation).

Thus, the methodology here was to first compute these parameters, then generate distributions with these parameters, and then compute (a) what percentage of values on these distributions are lower than GPT's scores (to estimate the percentile against first-timers); and (b) what percentage of values above the passing threshold are lower than GPT's scores (to estimate the percentile against qualified attorneys).

With regard to the mean, according to publicly available official NCBE data, the mean MBE score of first-time test-takers is 143.8 (Reshetar 2022).

As explained by official NCBE publications, the essay component is scaled to the MBE data (Albanese 2014), such that the two components have approximately the same mean and standard deviation (Albanese 2014; Illinois Board of Admissions to the Bar 2018, 2019). Thus, the methodology here assumed that the mean first-time essay score is 143.8.[14]

Given that the total UBE score is computed directly by adding MBE and essay scores (National Conference of Bar Examiners n.d.-h), an assumption was made that mean first-time UBE score is 287.6 (143.8 + 143.8).

With regard to standard deviations, information regarding the SD of first-timer scores is not publicly available. However, distributions of MBE scores for July scores (provided in 5 point-intervals) are publicly available on the NCBE website (The National Bar Examiner n.d.).

Under the assumption that first-timers have approximately the same SD as that of the general test-taking population in July, the standard deviation of first-time MBE scores was computed by (a) entering the publicly available distribution of MBE scores into R; and (b) taking the standard deviation of this distribution using

[13] A normal distribution of scores was assumed, given that (a) standardized tests are normalized and aim for a normal distribution (Kubiszyn and Borich 2016), (b) UBE is a standardized test, and (c) official visual estimates of MBE scores, both for February and July, appear to follow an approximately normal distribution. (The National Bar Examiner n.d.)

[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 (Illinois Board of Admissions to the Bar 2018)

the built-in sd() function (which calculates the standard deviation of a normal distribution).

Given that, as mentioned above, the distribution (mean and SD) of essay scores is the same as MBE scores, the SD for essay scores was computed similarly as above.

With regard to the UBE, Although UBE standard deviations are not publicly available for any official exam, they can be inferred from a combination of the mean UBE score for first-timers (287.6) and first-time pass rates.

For reference, standard deviations can be computed analytically as follows:

$$\sigma = \frac{x - \mu}{z}$$

where

- x is the quantile (the value associated with a given percentile, such as a cutoff score),
- μ is the mean,
- z is the z-score corresponding to a given percentile,
- σ is the standard deviation.

Thus, by (a) subtracting the cutoff score of a given administration (x) from the mean (μ); and (b) dividing that by the z-score (z) corresponding to the percentile of the cutoff score (i.e., the percentage of people who did not pass), one is left with the standard deviation (σ).

Here, the standard deviation was calculated according to the above formula using the official first-timer mean, along with pass rate and cutoff score data from New York, which according to NCBE data has the highest number of examinees for any jurisdiction (National Conference of Bar Examiners 2023).[15]

After obtaining these parameters, distributions of first-timer scores for the MBE component, essay component, and UBE overall were computed using the built-in rnorm function in R (which generates a normal distribution with a given mean and standard deviation).

Finally, after generating these distributions, percentiles were computed by calculating (a) what percentage of values on these distributions were lower than GPT's scores (to estimate the percentile against first-timers); and (b) what percentage of values above the passing threshold were lower than GPT's scores (to estimate the percentile against qualified attorneys).

With regard to the latter comparison, percentiles were computed after removing all UBE scores below 270, which is the most common score cutoff for states using

[15] Note that in a previous version of the paper, the standard deviation of overall UBE scores was instead computed 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, using the optim() function using R's "stats" package). This analysis was supplanted by the current method due to the latter having fewer/more plausible statistical assumptions, though both versions of the analysis yield converging results. For robustness purposes, the results of the old version can be found and replicated using the code available in the OSF repository.

Table 2 Estimated percentiles of MBE, essay, and total UBE scores among first-time test takers of uniform bar exam

Scaled Score	MBE percentile	Essay percentile	Total scaled score	UBE percentile
185	99	99	370	99
180	98	98	360	98
175	96	96	350	96
170	93	93	340	93
165	88	88	330	89
160	82	82	320	82
155	74	74	310	74
150	64	64	300	64
145	53	53	290	53
140	42	42	280	41
135	31	31	270	31
130	22	22	260	22
125	14	14	250	14
120	9	9	240	9
115	5	5	230	5
110	3	3	220	3
105	1	1	210	1

the UBE (National Conference of Bar Examiners n.d.-a). To compute models' performance on the individual components relative to qualified attorneys, a separate percentile was likewise computed after removing all subscores below 135.[16]

3.2 Results

3.2.1 Performance against first-time test-takers

Results are visualized in Tables 1 and 2. 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 62nd percentile as compared to the ~90th percentile February estimate and the ~68th percentile July

[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 3 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	2nd–62rd	6th–79th	0th–42nd
Qualified attorneys	0th–45th	0th–69th	0th–15th

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 ~2nd 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 ~45th 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 (Table 3).

4 Re-evaluating the raw score

So far, this analysis has taken for granted the scaled score achieved by GPT-4 as reported by OpenAI—that is, assuming GPT-4 scored a 298 on the UBE, is the 90th-percentile figure reported by OpenAI warranted?

However, given calls for the replication and reproducibility within the practice of science more broadly (Cockburn et al. 2020; Echtler and Häußler 2018; Jensen et al. 2023; Schooler 2014; Shrout and Rodgers 2018), it is worth scrutinizing the validity of the score itself—that is, did GPT-4 in fact score a 298 on the UBE?

Moreover, given the various potential hyperparameter settings available when using GPT-4 and other LLMs, it is worth assessing whether and to what extent adjusting such settings might influence the capabilities of GPT-4 on exam performance.

To that end, this section first attempts to replicate the MBE score reported by OpenAI (2023a) and Katz et al. (2023) using methods as close to the original paper as reasonably feasible.

The section then attempts to get a sense of the floor and ceiling of GPT-4's out-of-the-box capabilities by comparing GPT-4's MBE performance using the best and worst hyperparameter settings.

Finally, the section re-examines GPT-4's performance on the essays, evaluating (a) the extent to which the methodology of grading GPT-4's essays deviated that from official protocol used by the National Conference of Bar Examiners during actual bar exam administrations; and (b) the extent to which such deviations might undermine one's confidence in the the scaled essay scores reported by OpenAI (2023a) and Katz et al. (2023).

4.1 Replicating the MBE score

4.1.1 Methodology

Materials

As in Katz et al. (2023), the materials used here were the official MBE questions released by the NCBE. The materials were purchased and downloaded in pdf format from an authorized NCBE reseller. Afterwards, the materials were converted into TXT format, and text analysis tools were used to format the questions in a way that was suitable for prompting, following Katz et al. (2023).

Procedure

To replicate the MBE score reported by OpenAI (2023a), this paper followed the protocol documented by Katz et al. (2023), with some minor additions for robustness purposes.

In Katz et al. (2023), the authors tested GPT-4's MBE performance using three different temperature settings: 0, .5 and 1. For each of these temperature settings, GPT-4's MBE performance was tested using two different prompts, including (1) a prompt where GPT was asked to provide a top-3 ranking of answer choices, along with a justification and authority/citation for its answer; and (2) a prompt where GPT-4 was asked to provide a top-3 ranking of answer choices, without providing a justification or authority/citation for its answer.

For each of these prompts, GPT-4 was also told that it should answer as if it were taking the bar exam.

For each of these prompts / temperature combinations, Katz et al. (2023) tested GPT-4 three different times ("experiments" or "trials") to control for variation.

The minor additions to this protocol were twofold. First, GPT-4 was tested under two additional temperature settings: .25 and .7. This brought the total temperature / prompt combinations to 10 as opposed to 6 in the original paper.

Second, GPT-4 was tested 5 times under each temperature / prompt combination as opposed to 3 times, bringing the total number of trials to 50 as opposed to 18.

After prompting, raw scores were computed using the official answer key provided by the exam. Scaled scores were then computed following the method outlined

Table 4 Comparison of estimated percentiles of UBE scores for different groups

UBE score	February percentile	July percentile	First-timer percentile	Attorney percentile
370	99+	99+	99	99
360	99+	99+	98	97
350	99+	99+	96	94
340	99+	99	93	90
330	99	96	88	83
320	98	90	82	74
310	94	82	74	62
300	90	70	64	48
290	85	59	53	31
280	73	48	41	15
270	58	37	31	0
260	44	27	22	0
250	26	16	15	0
240	17	8	9	0
230	9	4	5	0
220	5	2	3	0

February and July scores are based on data from Illinois bar exam (Illinois Board of Admissions to the Bar 2018, 2019). First-timer and Attorney percentiles are based on original calculations here. Attorney percentiles are based on a UBE cutoff score of 270, which is the most common cutoff score in UBE jurisdictions

in JD Advising (n.d.-a), by (a) multiplying the number of correct answers by 190, and dividing by 200; and (b) converting the resulting number to a scaled score using a conversion chart based on official NCBE data.

After scoring, scores from the replication trials were analyzed in comparison to those from Katz et al. (2023) using the data from their publicly available github repository.

To assess whether there was a significant difference between GPT-4's accuracy in the replication trials as compared to the Katz et al. (2023) paper, as well as to assess any significant effect of prompt type or temperature, a mixed-effects binary logistic regression was conducted with: (a) paper (replication vs original), temperature and prompt as fixed effects[17]; and (b) question number and question category as random effects. These regressions were conducted using the lme4 (Bates et al. 2014) and lmertest (Kuznetsova et al. 2017) packages from R.

[17] All fixed effect predictors were coded as factors, with treatment coding.

Table 5 GPT-4's MBE performance across temperature and prompt settings

Prompt type	Temperature setting				
	0	.25	.5	.7	1
Answer without explanation	76.1	75.5	76.6	75.8	75.2
Answer with explanation	75.7	75.1	75.3	75.8	75.0

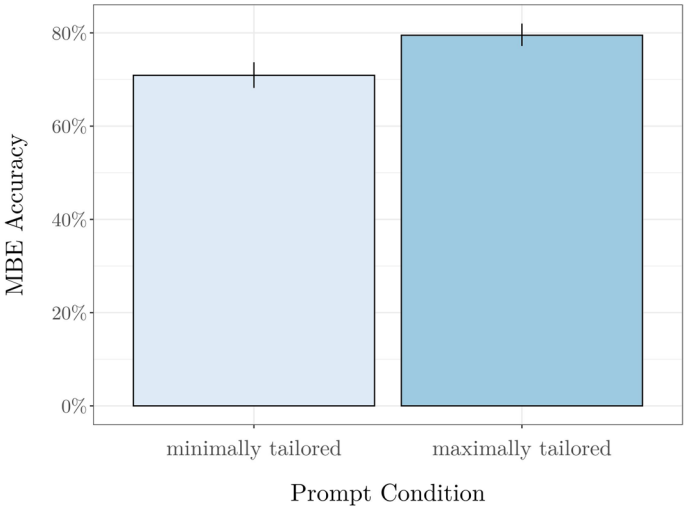

Fig. 1 GPT-4's MBE Accuracy in minimally tailored vs. maximally tailored prompting conditions. Bars reflect the mean accuracy. Lines correspond to 95% bootstrapped confidence intervals

4.1.2 Results

Results are visualized in Table 4. Mean MBE accuracy across all trials in the replication here was 75.6% (95% CI: 74.7 to 76.4), whereas the mean accuracy across all trials in Katz et al. (2023) was 75.7% (95% CI: 74.2 to 77.1).[18]

The regression model did not reveal a main effect of "paper" on accuracy ($p = .883$), indicating that there was no significant difference between GPT-4's raw accuracy as reported by Katz et al. (2023) and GPT-4's raw accuracy as performed in the replication here.

There was also no main effect of temperature ($p > .1$)[19] or prompt ($p = .741$). That is, GPT-4's raw accuracy was not significantly higher or lower at a given

[18] As a sanity check, note that the original mean accuracy originally reported by Katz et al. (2023) was also 75.7%, indicating that there were no errors here in reading the original data or computing the mean.

[19] Note that because temperature was coded as a factor (categorical variable) as opposed to numeric (continuous variable), there were multiple β coefficients and p values (one for each level, not including the reference level). The p values for all levels were higher than .1.

temperature setting or when fed a certain prompt as opposed to another (among the two prompts used in Katz et al. (2023) and the replication here) (Table 5).

4.2 Assessing the effect of hyperparameters

4.2.1 Methods

Although the above analysis found no effect of prompt on model performance, this could be due to a lack of variety of prompts used by Katz et al. (2023) in their original analysis.

To get a better sense of whether prompt engineering might have any effect on model performance, a follow-up experiment compared GPT-4's performance in two novel conditions not tested in the original (Katz et al. 2023) paper.

In Condition 1 ("minimally tailored" condition), GPT-4 was tested using minimal prompting compared to Katz et al. (2023), both in terms of formatting and substance.

In particular, the message prompt in Katz et al. (2023) and the above replication followed OpenAI's Best practices for prompt engineering with the API (Shieh 2023) through the use of (a) helpful markers (e.g. '```') to separate instruction and context; (b) details regarding the desired output (i.e. specifying that the response should include ranked choices, as well as [in some cases] proper authority and citation; (c) an explicit template for the desired output (providing an example of the format in which GPT-4 should provide their response); and (d) perhaps most crucially, context regarding the type of question GPT-4 was answering (e.g. "please respond as if you are taking the bar exam").

In contrast, in the minimally tailored prompting condition, the message prompt for a given question simply stated "Please answer the following question," followed by the question and answer choices (a technique sometimes referred to as "basic prompting": Choi et al., 2023). No additional context or formatting cues were provided.

In Condition 2 ("maximally tailored" condition), GPT-4 was tested using the highest performing prompt settings as revealed in the replication section above, with one addition, namely that: the system prompt, similar to the approaches used in Choi (2023), Choi et al. (2023), was edited from its default ("you are a helpful assistant") to a more tailored message that included included multiple example MBE questions with sample answer and explanations structured in the desired format (a technique sometimes referred to as "few-shot prompting": Choi et al. (2023)).

As in the replication section, 5 trials were conducted for each of the two conditions. Based on the lack of effect of temperature in the replication study, temperature was not a manipulated variable. Instead, both conditions featured the same temperature setting (.5).

To assess whether there was a significant difference between GPT-4's accuracy in the maximally tailored vs minimally tailored conditions, a mixed-effects binary logistic regression was conducted with: (a) condition as a fixed effect; and (b) question number and question category as random effects. As above, these regressions

were conducted using the lme4 (Bates et al. 2014) and lmertest (Kuznetsova et al. 2017) packages from R.

4.2.2 Results

Mean MBE accuracy across all trials in the maximally tailored condition was descriptively higher at 79.5% (95% CI: 77.1–82.1), than in the minimally tailored condition at 70.9% (95% CI: 68.1–73.7).

The regression model revealed a main effect of condition on accuracy ($\beta = 1.395$, SE = .192, $p < .0001$), such that GPT-4's accuracy in the maximally tailored condition was significantly higher than its accuracy in the minimally tailored condition.

In terms of scaled score, GPT-4's MBE score in the minimally tailored condition would be approximately 150, which would place it: (a) in the 70th percentile among July test takers; (b) 64th percentile among first-timers; and (c) 48th percentile among those who passed.

GPT-4's score in the maximally tailored condition would be approximately 164—6 points higher than that reported by Katz et al. (2023) and OpenAI (2023a). This would place it: (a) in the 95th percentile among July test takers; (b) 87th percentile among first-timers; and (c) 82th percentile among those who passed.

4.3 Re-examining the essay scores

As confirmed in the above subsection, the scaled MBE score (not percentile) reported by OpenAI was accurately computed using the methods documented in Katz et al. (2023).

With regard to the essays (MPT + MEE), however, the method described by the authors significantly deviates in at least three aspects from the official method used by UBE states, to the point where one may not be confident that the essay scores reported by the authors reflect GPT models' "true" essay scores (i.e., the score that essay examiners would have assigned to GPT had they been blindly scored using official grading protocol).

The first aspect relates to the (lack of) use of a formal rubric. For example, unlike NCBE protocol, which provides graders with (a) (in the case of the MEE) detailed "grading guidelines" for how to assign grades to essays and distinguish answers for a given MEE; and (b) (for both MEE and MPT) a specific "drafters' point sheet" for each essay that includes detailed guidance from the drafting committee with a discussion of the issues raised and the intended analysis (Olson 2019), Katz et al. (2023) do not report using an official or unofficial rubric of any kind, and instead simply describe comparing GPT-4's answers to representative "good" answers from the state of Maryland.

Utilizing these answers as the basis for grading GPT-4's answers in lieu of a formal rubric would seem to be particularly problematic considering it is unclear even what score these representative "good" answers received. As clarified by the Maryland bar examiners: "The Representative Good Answers are not 'average' passing answers nor are they necessarily 'perfect' answers. Instead, they are

responses which, in the Board's view, illustrate successful answers written by applicants who passed the UBE in Maryland for this session" (Maryland State Board of Law Examiners 2022).

Given that (a) it is unclear what score these representative good answers received; and (b) these answers appear to be the basis for determining the score that GPT-4's essays received, it would seem to follow that (c) it is likewise unclear what score GPT-4's answers should receive. Consequently, it would likewise follow that any reported scaled score or percentile would seem to be insufficiently justified so as to serve as a basis for a conclusive statement regarding GPT-4's relative performance on essays as compared to humans (e.g. a reported percentile).

The second aspect relates to the lack of NCBE training of the graders of the essays. Official NCBE essay grading protocol mandates the use of trained bar exam graders, who in addition to using a specific rubric for each question undergo a standardized training process prior to grading (Gunderson 2015; Case 2010). In contrast, the graders in Katz et al. (2023) (a subset of the authors who were trained lawyers) do not report expertise or training in bar exam grading. Thus, although the graders of the essays were no doubt experts in legal reasoning more broadly, it seems unlikely that they would have been sufficiently ingrained in the specific grading protocols of the MEE + MPT to have been able to reliably infer or apply the specific grading rubric when assigning the raw scores to GPT-4.

The third aspect relates to both blinding and what bar examiners refer to as "calibration," as UBE jurisdictions use an extensive procedure to ensure that graders are grading essays in a consistent manner (both with regard to other essays and in comparison to other graders) (Case 2010; Gunderson 2015). In particular, all graders of a particular jurisdiction first blindly grade a set of 30 "calibration" essays of variable quality (first rank order, then absolute scores) and make sure that consistent scores are being assigned by different graders, and that the same score (e.g. 5 of 6) is being assigned to exams of similar quality (Case 2010).

Unlike this approach, as well as efforts to assess GPT models' law school performance (Choi et al. 2021), the method reported by Katz et al. (2023) did not initially involve blinding. The method in Katz et al. (2023) did involve a form of inter-grader calibration, as the authors gave "blinded samples" to independent lawyers to grade the exams, with the assigned scores "match[ing] or exceed[ing]" those assigned by the authors. Given the lack of reporting to the contrary, however, the method used by the graders would presumably be plagued by issue issues as highlighted above (no rubric, no formal training with bar exam grading, no formal intra-grader calibration).

Given the above issues, as well as the fact that, as alluded in the introduction, GPT-4's performance boost over GPT-3 on other essay-based exams was far lower than that on the bar exam, it seems warranted not only to infer that GPT-4's relative performance (in terms of percentile among human test-takers) was lower than that reported by OpenAI, but also that GPT-4's reported scaled score on the essay may have deviated to some degree from GPT-4's "true" essay (which, if true, would imply that GPT-4's "true" percentile on the bar exam may be even lower than that estimated in previous sections).

Indeed, Katz et al. (2023) to some degree acknowledge all of these limitations in their paper, writing: "While we recognize there is inherent variability in any qualitative assessment, our reliance on the state bars' representative "good" answers and the multiple reviewers reduces the likelihood that our assessment is incorrect enough to alter the ultimate conclusion of passage in this paper".

Given that GPT-4's reported score of 298 is 28 points higher than the passing threshold (270) in the majority of UBE jurisdictions, it is true that the essay scores would have to have been wildly inaccurate in order to undermine the general conclusion of Katz et al. (2023) (i.e., that GPT-4 "passed the [uniform] bar exam"). However, even supposing that GPT-4's "true" percentile on the essay portion was just a few points lower than that reported by OpenAI, this would further call into question OpenAI's claims regarding the relative performance of GPT-4 on the UBE relative to human test-takers. For example, supposing that GPT-4 scored 9 points lower on the essays, this would drop its estimated relative performance to (a) 31st percentile compared to July test-takers; (b) 24th percentile relative to first-time test takers; and (c) less than 5th percentile compared to licensed attorneys.

5 Discussion

This paper first 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 ~62nd 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.

In addition to these four main findings, the paper also investigated the validity of GPT-4's reported UBE score of 298. Although the paper successfully replicated the MBE score of 158, the paper also highlighted several methodological issues in the grading of the MPT + MEE components of the exam, which call into question the validity of the essay score (140).

Finally, the paper also investigated the effect of adjusting temperature settings and prompting techniques on GPT-4's MBE performance, finding no significant effect of adjusting temperature settings on performance, and some effect of prompt engineering when compared to a basic prompting baseline condition.

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

(Hernandez-Orallo 2020; Burden and Hernández-Orallo 2020; Raji et al. 2021). 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 (Bowman 2022, 2023; Kojima et al. 2022).

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 (National Conference of Bar Examiners n.d.-g), 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) (National Conference of Bar Examiners n.d.-a), 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 (US News and World Report 2022).

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. These findings align with recent research work finding that GPT-4 performed below-average on law school exams (Blair-Stanek et al. 2023).

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[20] 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 (Brundage et al. 2020), 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 a false sense of alarm or security (Zoe et al. 2021). 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" (OpenAI 2018).

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" (Katz et al. 2023, 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. The results also indicate that law schools and the legal profession should prioritize instruction in areas such as law and technology and law and AI, which, despite their importance, are currently not viewed as descriptively or normatively central to the legal academy (Martínez and Tobia 2023).

Acknowledgements Acknowledgements omitted for anonymous review.

Funding 'Open Access funding provided by the MIT Libraries'.

[20] 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.

Declarations

Conflict of interest The author declares no financial nor non-financial interests that are directly or indirectly related to the work submitted for publication.

References

Albanese MA (2014) The testing column: scaling: it's not just for fish or mountains. Bar Exam 83(4):50–56

Bates D, Mächler M, Bolker B, Walker S (2014) Fitting linear mixed-effects models using LME4. arXiv preprint arXiv:1406.5823

Blair-Stanek A, Carstens A-M, Goldberg DS, Graber M, Gray DC, Stearns ML (2023) Gpt-4's law school grades, Partnership tax b, property b-, tax b. Crim C-, Law & Econ C, Partnership Tax B, Property B-, Tax B

Bommarito MJ II, Katz DM (2017) Measuring and modeling the us regulatory ecosystem. J Stat Phys 168:1125–1135

Bostrom N, Yudkowsky E (2018) The ethics of artificial intelligence. Artificial intelligence safety and security. Chapman and Hall/CRC, New York, pp 57–69

Bowman S (2022) 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 (vol 1: Long papers) pp 7484–7499

Bowman SR (2023) Eight things to know about large language models. arXiv preprint arXiv:2304.00612

Brundage M, Avin S, Wang J, Belfield H, Krueger G, Hadfield G, et al (2020) Toward trustworthy AI development: mechanisms for supporting verifiable claims. arXiv preprint arXiv:2004.07213

Burden J, Hernández-Orallo J (2020) 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). pp 36–40

Carlsmith J (2022) Is power-seeking AI an existential risk? arXiv preprint arXiv:2206.13353

Caron P (2023) GPT-4 Beats 90% of aspiring lawyers on the bar exam. TaxProf Blog. https://taxprof.typepad.com/taxprof_blog/2023/03/gpt-4-beats-90-of-aspiring-lawyers-on-the-bar-exam.html. Accessed on 24 Apr 2023

Case SM (2010) Procedure for grading essays and performance tests. The Bar Examiner. https://thebarexaminer.ncbex.org/wp-content/uploads/PDFs/790410_TestingColumn.pdf

Choi JH (2023) How to use large language models for empirical legal research. J Instit Theor Econ (Forthcoming)

Choi JH, Monahan A, Schwarcz D (2023) Lawyering in the age of artificial intelligence. Available at SSRN 4626276

Choi JH, Hickman KE, Monahan AB, Schwarcz D (2021) Chatgpt goes to law school. J Legal Educ 71:387

Cockburn A, Dragicevic P, Besançon L, Gutwin C (2020) Threats of a replication crisis in empirical computer science. Commun ACM 63(8):70–79

Crootof R, Kaminski ME, Price II WN (2023) Humans in the loop. Vanderbilt Law Review, (Forthcoming)

Echtler F, Häußler M (2018) Open source, open science, and the replication crisis in HCI. Extended abstracts of the 2018 chi conference on human factors in computing systems. pp 1–8
Examiner TB (n.d.-a) First-time exam takers and repeaters in 2021. The Bar Examiner. https://thebarexaminer.ncbex.org/2021-statistics/first-time-exam-takers-and-repeaters-in-2021/. Accessed on 24 Apr 2023
Examiner TB (n.d.-b) Statistics. The Bar Examiner. https://thebarexaminer.ncbex.org/statistics/. Accessed on 24 Apr 2023
Gunderson JA (2015) The testing column: essay grading fundamentals. Bar Exam 84(1):54–56
Hernandez-Orallo J (2020) AI evaluation: on broken yardsticks and measurement scales. In: Workshop on evaluating evaluation of AI systems at AAAI
Illinois Board of Admissions to the Bar. (2018) https://www.ilbaradmissions.org/percentile-equivalent-charts-july-2018. Accessed on 24 Apr 2023
Illinois Board of Admissions to the Bar. (2019) https://www.ilbaradmissions.org/percentile-equivalent-charts-february-2019. Accessed on 24 Apr 2023
JD Advising (n.d.) MBE raw score conversion chart. https://jdadvising.com/mbe-raw-score-conversion-chart/. Accessed on 01 Jan 2024
JD Advising. (n.d.) https://jdadvising.com/july-2018-ube-percentiles-chart/. Accessed on 24 Apr 2023
Jensen TI, Kelly B, Pedersen LH (2023) Is there a replication crisis in finance? J Finance 78(5):2465–2518
Katz DM, Bommarito MJ, Gao S, Arredondo P (2023) GPT-4 passes the bar exam. Available at SSRN 4389233
Katz DM, Bommarito MJ (2014) Measuring the complexity of the law: the United States code. Artif Intell Law 22:337–374
Koetsier J (2023) GPT-4 Beats 90% of Lawyers Trying to Pass the Bar. Forbes. https://www.forbes.com/sites/johnkoetsier/2023/03/14/gpt-4-beats-90-of-lawyers-trying-to-pass-the-bar/?sh=b40c88d30279
Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y (2022) Large language models are zero-shot reasoners. arXiv preprint arXiv:2205.11916
Kubiszyn T, Borich GD (2016) Educational testing and measurement. John Wiley & Sons, Hoboken
Kuznetsova A, Brockhoff PB, Christensen RHB (2017) lmertest package: tests in linear mixed effects models. J Stat Software 82:13
Lang C (2023) What is a good bar exam score? Test Prep Insight. https://www.testprepinsight.com/what-is-a-good-bar-exam-score
Li B, Qi P, Liu B, Di S, Liu J, Pei J, Zhou B (2023) Trustworthy AI: From principles to practices. ACM Comput Surv 55(9):1–46
Markou C, Deakin S (2020) Is law computable? From rule of law to legal singularity. From Rule of Law to Legal Singularity. University of Cambridge Faculty of Law Research Paper
Martínez E, Tobia K (2023) What do law professors believe about law and the legal academy? Geo LJ 112:111
Martinez E, Mollica F, Gibson E (2022) Poor writing, not specialized concepts, drives processing difficulty in legal language. Cognition 224:105070
Martinez E, Mollica F, Gibson E (2022b) 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
Martinez E, Mollica F, Gibson E (in press) Even lawyers don't like legalese. In: Proceedings of the national academy of sciences
Maryland State Board of Law Examiners (2022) July 2022 uniform bar examination (UBE) in maryland—representative good answers. https://mdcourts.gov/sites/default/files/import/ble/examanswers/2022/202207uberepgoodanswers.pdf
National Conference of Bar Examiners (2023) Bar exam results by jurisdiction. https://www.ncbex.org/statistics-research/bar-exam-results-jurisdiction. Accessed on 01 Jan 2024
National Conference of Bar Examiners (n.d.-a) https://www.ncbex.org/exams/ube/scores/. Accessed on 03 May 2023
National Conference of Bar Examiners (n.d.-b) https://www.ncbex.org/exams/ube/score-portability/minimum-scores/. Accessed on 24 Apr 2023
National Conference of Bar Examiners (n.d.-c) Bar Exam Results by Jurisdiction. National Conference of Bar Examiners. https://www.ncbex.org/statistics-and-research/bar-exam-results/. Accessed on 24 Apr 2023
National Conference of Bar Examiners (n.d.-d) Multistate bar exam. https://www.ncbex.org/exams/mbe. Accessed on 01 Jan 2024

National Conference of Bar Examiners (n.d.-e) Multistate essay exam. https://www.ncbex.org/exams/mee. Accessed on 01 Jan 2024
National Conference of Bar Examiners (n.d.-f) Multistate performance test. https://www.ncbex.org/exams/mpt. Accessed on 01 Jan 2024
National Conference of Bar Examiners (n.d.-g) Uniform bar exam. Accessed on 01 Jan 2024
National Conference of Bar Examiners (n.d.-h) Uniform Bar Examination. National Conference of Bar Examiners. https://www.ncbex.org/exams/ube/. Accessed on 24 Apr 2023
Ngo R (2022) The alignment problem from a deep learning perspective. arXiv preprint arXiv:2209.00626
Olson S (2019) 13 best practices for grading essays and performance tests. Bar Exam 88(4):8–14
OpenAI (2018) OpenAI Charter. https://openai.com/charter
OpenAI (2023) GPT 4. https://openai.com/research/gpt-4. Accessed on 24 Apr 2023
OpenAI (2023) GPT-4 Technical Report. arXiv:2303.08774. (Preprint submitted to arXiv)
OpenAI (n.d.) GPT-4 is OpenAI's most advanced system, producing safer and more useful responses. https://openai.com/product/gpt-4. Accessed on 24 Apr 2023
Patrice J (2023) New GPT-4 Passes All Sections Of The Uniform Bar Exam. Maybe This Will Finally Kill The Bar Exam. Above the Law. https://abovethelaw.com/2023/03/new-gpt-4-passes-all-sections-of-the-uniform-bar-exam-maybe-this-will-finally-kill-the-bar-exam/
Raji ID, Bender EM, Paullada A, Denton E, Hanna A (2021) Ai and the everything in the whole wide world benchmark. arXiv preprint arXiv:2111.15366
Ray T (2023) With GPT-4, OpenAI opts for secrecy versus disclosure. ZDNet. https://www.zdnet.com/article/with-gpt-4-openai-opts-for-secrecy-versus-disclosure/
Reshetar R (2022) The testing column: Why are February bar exam pass rates lower than July pass rates? Bar Exam 91(1):51–53
Ruhl J, Katz DM, Bommarito MJ (2017) Harnessing legal complexity. Science 355(6332):1377–1378
Rules.com M (n.d.) Bar Exam Calculators. https://mberules.com/bar-exam-calculators/?__cf_chl_tk=lTwxFyYWOZqBwTAenLs0TzDfAuvawkHeH2GaXU1PQo0-1683060961-0-gaNycGzNDBA. Accessed on 02 May 2023
Schooler JW (2014) Metascience could rescue the replication crisis. Nature 515(7525):9
Schwarcz D, Choi JH (2023) Ai tools for lawyers: a practical guide. Available at SSRN
Shieh J (2023) Best practices for prompt engineering with openai api. https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-openai-api. OpenAI. Accessed on 01 Jan 2024
Shrout PE, Rodgers JL (2018) Psychology, science, and knowledge construction: broadening perspectives from the replication crisis. Ann Rev Psychol 69:487–510
Stokel-Walker C (2023) Critics denounce a lack of transparency around GPT-4's tech. Fast Company. https://www.fastcompany.com/90866190/critics-denounce-a-lack-of-transparency-around-gpt-4s-tech
The National Bar Examiner (n.d.) https://thebarexaminer.ncbex.org/2022-statistics/the-multistate-bar-examination-mbe/#step3. Accessed on 24 Apr 2023
The New York State Board of Law Examiners (n.d.) NYS Bar Exam Statistics. The New York State Board of Law Examiners. https://www.nybarexam.org/examstats/estats.htm
UBEEssays.com. (2019) https://ubeessays.com/feb-mbe-percentiles/
University of Illinois Chicago (n.d.) https://law.uic.edu/student-support/academic-achievement/bar-exam-information/illinois-bar-exam/. Accessed on 24 Apr 2023
US News and World Report (2022) https://www.usnews.com/best-graduate-schools/top-law-schools/law-rankings
Washington State Bar Association (2020) https://wsba.org/news-events/latest-news/news-detail/2020/06/15/state-supreme-court-grants-diploma-privilege. Accessed on 24 Apr 2023
Weiss DC (2023) Latest version of ChatGPT aces bar exam with score nearing 90th percentile. ABA Journal. https://www.abajournal.com/web/article/latest-version-of-chatgpt-aces-the-bar-exam-with-score-in-90th-percentile. Accessed on 24 Apr 2023
Wilkins S (2023) How GPT-4 mastered the entire bar exam, and why that matters. Law.com. https://www.law.com/legaltechnews/2023/03/17/how-gpt-4-mastered-the-entire-bar-exam-and-why-that-matters/?slreturn=20230324023302. Accessed on 24 Apr 2023
Winter CK (2022) The challenges of artificial judicial decision-making for liberal democracy. Judicial decision-making: Integrating empirical and theoretical perspectives. Springer, Berlin, pp 179–204
Winter C, Hollman N, Manheim D (2023) Value alignment for advanced artificial judicial intelligence. Am Philos Quart 60(2):187–203

Zoe Cremer C, Whittlestone J (2021) Artificial canaries: early warning signs for anticipatory and democratic governance of AI

Authors and Affiliations

Eric Martínez[1]

✉ Eric Martínez
ericmart@mit.edu

[1] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology (MIT), Cambridge, MA 02138, USA