GPT takes the Bar Exam

Michael J Bommarito II[a,b,c] and Daniel Martin Katz[a,b,c,*]

[a]Illinois Tech - Chicago Kent College of Law, Chicago IL, USA, [b]Bucerius Law School, Hamburg, DE and [c]CodeX - The Stanford Center for Legal Informatics, Palo Alto CA, USA
[*]To whom correspondence should be addressed: dkatz3@kentlaw.iit.edu

Abstract

We experimentally evaluate OpenAI's TEXT-DAVINCI-003 (GPT-3.5) and related models on the multistate multiple choice (MBE) section of the Bar Exam. Despite the fact that humans with seven years of post-secondary education and exam-specific training only answer 68% of questions correct, TEXT-DAVINCI-003 is able to achieve a correct rate of 50.3% for best prompt and parameters and achieved passing scores in the Evidence and Torts sections. In addition to exceeding the baseline guessing rate of 25%, the model's top two and top three choices are correct 71% and 88% of the time, respectively, indicating strong non-entailment performance. We find no benefit in fine-tuning over GPT-3.5's zero-shot performance at the scale of our training data, we document our observations related to prompt engineering, (hyper)parameters, and older GPT models as well. While our ability to interpret these results is limited by nascent scientific understanding of LLMs and the proprietary nature of GPT models, the trend in improvement for recent GPT models strongly suggests that an LLM will pass the MBE component of the Bar Exam in the near future.

Key words: natural language processing (NLP), artificial intelligence (AI)

Introduction

The legal system is becoming increasingly complex [1][2][3], leading to a need for technology to assist with the quantity, quality, and accessibility of legal services demanded by society. As in other domains, artificial intelligence and process engineering have promised help for decades to both non-professional and professional users of legal systems [4][5][6]. Significant research and development effort has been devoted to use cases such as automated argumentation or brief construction, pre- and post-execution contract process automation, AI-assisted due diligence and e-discovery, the statistical prediction of judicial decision making, and technology-aided search and legal aid for laypeople [7][8][9][10]. However, the complexity of legal language and vastness of legal knowledge has made it historically difficult to develop systems that understand the nuances of legal tasks, and many systems have failed to deliver desired results or reach adoption.

Naturally, law is heavily reliant on the use of language, producing massive volumes of textual data [11][12]. Documents such as briefs, memos, statutes, regulations, contracts, patents, and judicial decisions are continuously authored by lawyers, judges, and regulators [2]. To make matters even more difficult, legal language is notoriously complex; lawyers and other legal professionals undertake many years of education and professional training to understand and generate it.

Why is this language so "complex?" Why do so many proficient users of natural languages struggle with legal documents such as contracts, statutes and regulations, even in their native tongue, to the point that descriptors like "legalese" or "lawyer speak" have become common parlance? The answer is likely two-fold. First, for both technical and cultural reasons, the grammar of legal language is significantly different than the grammar of normal language, featuring both highly-stylized customs and pedantically-precise phrasing. The resulting sentence structures are typically much larger and more complex than normal language, as the number of clauses and "distance" over which clauses are connected exceeds the working memory of both human and non-human readers. Second, by the very nature of common law and precedent, legal language is full of semantic nuance and history. Words like "security" that have common meaning in normal language often have different, context-specific meanings in legal language. Many words that do not occur at all in normal language, like "estoppel" or "indemnitor," occur regularly in legal corpora. This semantic depth and breadth traditionally required systems that interact with legal text to embed a large amount of domain-specific knowledge. Viewed from this perspective, legal education and training is required to teach humans to understand and produce this very particular type of language, and it is no surprise that traditional models in NLP struggled in general legal task assessments.

In recent years, however, developments in natural language processing and computing have led to significant advances in state of the art performance. Leveraging advances in neural

network research [13][14], sophisticated efforts have been made to build quasi-semantic models. The age of neural NLP can be traced to [15], which has been followed by successive waves of embedding [16] [17] and transformer-based large language models (LLMs) [18] [19] [20] [21] [22]. In particular, transformer architectures, first introduced in [23], have revolutionized machine learning research, and have been most successfully applied to text and image modalities. The most famous and accessible of these LLMs is OpenAI's family of Generative Pre-trained Transformer models, commonly referred to as GPT.

GPT models are proprietary reinforcement learning models in production for OpenAI's customers, so there is no guarantee that previously-published academic literature is still complete or accurate. However, as of July 2020, OpenAI reported that GPT-3 was "an autoregressive language model with 175 billion parameters" featuring 96 layers trained with a batch size of 3.2M. While these numbers may be difficult to contextualize, those who have trained their own models can easily appreciate the effort involved. Since then, OpenAI has also launched or published a number of derivative models, most notably InstructGPT-3 and Codex 12B. Colloquially, these recent models are referred to by many, including OpenAI, as GPT-3.5. More specifically, as described on OpenAI's website, "GPT-3.5 series is a series of models that was trained on a blend of text and code from before Q4 2021." Our results in this publication are based on TEXT-DAVINCI-003, which is "an improvement on TEXT-DAVINCI-002", which is "an InstructGPT model based on CODE-DAVINCI-002", which is "a base model [...] for pure code-completion tasks."

GPT-3 and derivative models are not, however, directly available for use in frameworks like PyTorch or Tensorflow; for both commercial and ethical reasons, access to OpenAI models has historically only been available through OpenAI's API, which is designed both to accomplish specific customer tasks and to provide a layer of legal and ethical moderation. As of this publication, OpenAI's APIs offers text completion, code completion, image generation, and embedding generation endpoints. In recent weeks, OpenAI has also released a public-facing chatbot version of GPT-3.5 known as ChatGPT, which reportedly resulted in over 1M user sign-ups within six days of release.

While GPT-3.5 and ChatGPT have demonstrated previously-unseen performance on zero-shot or few-shot tasks, they are not domain-specific models. As reported in [19], OpenAI's models are trained on a combination of curated CommonCrawl data and high-quality reference data that, if we consider The Pile V1 as reference [24], may have included some material from public legal sources. However, given the complex nature of legal language and GPT-3.5's training on general task performance, it is an open question as to whether state-of-the-art LLMs like GPT-3.5 can succeed in legal task assessments, let alone zero- or few-shot tasks. In order to evaluate this question, we decided to test GPT-3.5 on the multistate multiple choice section of the Bar Exam, known as the Multistate Bar Examination (MBE), using zero-shot prompts for the TEXT-DAVINCI-003 text completion API.

Data

Licensure exams like "the Bar Exam" are common across professional fields, including not just law, but also medicine, dentistry, pharmacy, accounting, and engineering. While each jurisdiction (e.g., state) in the United States is responsible for administering its own law licensure requirements, the National Conference of Bar Examiners (NCBE) is the organization responsible for designing most of the bar examination materials used across the United States. In this research, we follow colloquial convention and refer to the NCBE's standardized exam format as "the Bar Exam" or "the Bar," while abstract exams that may vary across countries or states are referred to as bar exams in the indefinite.

For the individual test-taker, such bar exams are the culmination of years of education as well as preparation specific to each exam component. Successful performance on these exams generally requires two things: (i) the acquisition of a large amount of accumulated theoretical knowledge (semantics) and (ii) the ability to understand and answer exam-specific questions that often feature unique syntax. Prior attempts to develop systems to take bar exams around the world have yielded mixed results, with significant exam-specific training required to even achieve such performance [25] [26] [27].

For most test-takers, the Bar Exam represents the most significant single challenge of their academic careers. In order to be eligible, the typical applicant is required to complete at least seven years of post-secondary education, including a four-year bachelors degree and successful completion of three years of study at an ABA-accredited law school. Following graduation from law school, most applicants also invest substantial amounts of time and money into post-graduation Bar preparation training [28]. This additional preparation is intended to not only solidify one's legal knowledge, but also critically to teach the applicant how to understand and answer the exam's questions. Despite the incredible effort of the average test-taker, approximately one out of every five still fails to pass their exam on their initial attempt.

As a historical matter, the specific components of the Bar exam once differed widely from state to state. Recently, however, most states have adopted the Uniform Bar Examination (UBE). The UBE features three components: (i) a multiple choice test, (ii) an essay test, and (iii) scenario-based performance test. The multiple choice component, referred to as the Multistate Bar Examination or MBE, is typically worth 50% of an overall bar exam score.

As the MBE is a single component of an exam, most jurisdictions do not require a minimum MBE score. The MBE is also scaled by jurisdictions and the NCBE after each exam window; for example, a raw score of roughly $\sim 60\%$ may yield an approximate scaled score of 133, which would be enough to pass in a significant number of jurisdictions, including New York, Illinois, and the District of Columbia.

Questions on the MBE are designed to test both legal knowledge and reading comprehension skills, requiring above-average semantic and syntactic command of the English language. Instead of posing direct legal questions as they might appear in a textbook or theory exam, most MBE questions present the test-taker with a fictional situation. Descriptions of the facts are typically embellished with details; some of these details are critically important, while others are added only to distract or confuse the reader. A public sample provided by the NCBE on their website is shown below:

Sample Question 2.1

```
Question: A man sued a railroad for personal injuries
suffered when his car was struck by a train at an
unguarded crossing.  A major issue is whether the train
sounded its whistle before arriving at the crossing.
```

```
The railroad has offered the testimony of a resident
who has lived near the crossing for 15 years.  Although
she was not present on the occasion in question, she will
testify that, whenever she is home, the train always
sounds its whistle before arriving at the crossing.

Is the resident's testimony admissible?

(A) No, due to the resident's lack of personal knowledge
regarding the incident in question.

(B) No, because habit evidence is limited to the conduct
of persons, not businesses.

(C) Yes, as evidence of a routine practice.

(D) Yes, as a summary of her present sense impressions.
```

The MBE portion of the Bar consists of approximately 200 questions like the sample above. As detailed in Table 1, real examinations present test-takers with 25 questions from each of eight categories, seven of which correspond to specific areas of law and one of which is used by the NCBE to experiment with test design. In some instances, a subset of these these questions are removed from final scoring of an exam by state bars or the NCBE; both individual state bars and the NCBE assess the performance of test-takers within and across states, dropping some questions and scaling the raw scores to maintain consistency across jurisdictions. As part of its role in exam design and preparation, the NCBE also maintains statistical information regarding exam performance. For comparison, we show their reported average accuracy of students by question category in Table 1. In absolute terms, this table makes clear the difficulty of the exam, as the average student answers more than one in four questions incorrectly.

Question Category	Questions	Correct Rate
Torts	25	71%
Contracts	25	70%
Evidence	25	65%
Real Property	25	65%
Civil Procedure	25	59%
Constitutional Law	25	72%
Criminal Law and Procedure	25	71%
Experimental Questions	25	N/A
	200	68%
	TOTAL	AVERAGE

Table 1. NCBE-Reported Average Student Performance by Question Category

For this research, we purchased the standard test preparation material offered by the NCBE, including practice questions and simulated exams for the MBE portion of the Bar Exam. While we cannot redistribute these materials, researchers interested in replicating the results contained in this paper can purchase these data for approximately 300 USD directly from the NCBE's online store.

All reported task assessments are based on the practice exam and answer key available in the downloadable MBE Study Aid purchased in December 2022, dated in the document as of 2019. The body of each question was automatically extracted with its four multiple choice options and stored separately from the answer key, which consisted solely of the correct letter answer for each question. The answer key is found at the back of the document in a simple table, provided without explanations of correct and incorrect answers. We also searched public search engines with excerpts of test questions to confirm that the exam was not publicly available. While we do not have detailed information about the provenance of GPT-3.5 and cannot rule out the possibility of training set information, we specifically chose this exam PDF for task assessment instead of other simulated exams because we believe the probability of training set inclusion is low and, even if it were to have occurred, it would likely not be learnable due to lack of answer explanations.

Methods

As discussed above, our experimental evaluation of GPT-3.5 involved using zero-shot prompts for the TEXT-DAVINCI-003 text completion API. In this section, we detail how we implemented this experiment, including the design and iteration of these prompts, related API hyperparameters, and an attempt at fine-tuning the mode. While replication of this research requires access NCBE's material and an OpenAI account, we have done our best to provide researchers with as much detail as we have ourselves.

Prompt Engineering and Responses
Our scientific understanding of large language models is nascent, and we often do not understand how or why they produce the outputs they do. However, despite this scientific gap, we do know that LLMs are often highly sensitive to the prompts they are provided. The "art" of crafting such prompts is typically referred to as "prompt engineering," and details of prompt engineering are critical to replication of studies involving LLMs. In this research, we experimented substantially with prompt engineering. The following prompt types were tested:

1. Single choice only: Ask the model for a single multiple choice answer only.
2. Single choice and explanation: Ask the model for a single multiple choice answer with an explanation of its reasoning.
3. Top two choices only: Ask the model for its best answer and a backup answer.
4. Top two choices and explanation: Ask the model for its best and backup answer with an explanation of its reasoning.
5. Top two choices and re-prompt: Ask the model for its best and backup answer, then re-prompt the model between these two choices, similar to the iterative "strike-out" heuristic that many human test-takers are taught.
6. Rank order all choices: Ask the model to rank order all four multiple choice answers.
7. Rank order top three choices: Ask the model to rank order its top three multiple choice answers.

Results ranged between 42% and 46% for average correct rate for most of these prompts. However, the prompt #7, rank-ordering of the top three choices, improved model correctness substantially. Unfortunately, because we have no direct insight into the head layers of GPT-3.5, we have no ability to comment further on why this prompt variation impacted the model's behavior in ways that other prompts did not. We speculate that this prompt best combined non-entailment performance, i.e., rejection of most incorrect answer, with probabilistic

entailment and recall for remaining choices. Below is an example of how this prompt manifested with a partially-redacted version of a real NCBE question:

```
Please answer the following Bar Exam question in the
following rank order format:
First Choice: <LETTER>
Second Choice: <LETTER>
Third Choice: <LETTER>

Question: A plaintiff domiciled in State A has brought a
federal diversity action in State A against a defendant
domiciled in State B, [...]
(A) Move for discovery [...]
(B) Move for judgment on the pleadings, [...]
(C) Move for sanctions against the plaintiff [...]
(D) Move to dismiss the action for lack of [...]

Answer:
```

Upon querying the text completion API endpoint, we then received back responses like those below. When explanations were requested in the prompt, the model's choices were unchanged from prompts without explanations.

```
First Choice: D
Second Choice: B
Third Choice: A
```

As an example of explanation's produced by the model, we return to the public NCBE Sample Question 2.1 reproduced in the Data section above. When we ask the model to explain its choices in that question, we receive the following API response:

```
First Choice: C
Second Choice: D
Third Choice: A

Explanation: The resident's testimony is admissible as
evidence of a routine practice (C) and as a summary of
her present sense impressions (D). However, it is not
admissible due to the resident's lack of personal
knowledge regarding the incident in question (A).
Habit evidence is limited to the conduct of persons,
not businesses (B).
```

Pedantically, TEXT-DAVINCI-003's confusing explanation reveals the inconsistency between the multiple choice options and question. The question asks whether the evidence is admissible, and as such, should only have two possible options - yes or no. Instead, the question provides four statements to evaluate for entailment. The sample question should actually read something like "Which of the following statements regarding admissibility is true?" There are many examples of "incorrectly"-specified questions throughout the exam, and while useful systems must be robust to such examples as they occur frequently in the real world, GPT-3.5 might perform even better with properly-specified questions.

The prompt and complete JSON response, including the OpenAI API request ID, were logged for all simulated exams. Each line of the text completion response was parsed and stored for scoring or qualitative analysis. In a small number of cases ($< 1\%$), responses included natural language or format variations such as "My first choice is (D)" and these variations were handled through exception cases in our parser. No responses were manually altered or evaluated by humans.

From a technical perspective, all of these prompts are related to traditional textual entailment tasks where a model must evaluate whether a statement is truthful or non-truthful. In most extant research on the topic, this problem is formulated relative to another statement or body of knowledge, and tasks are assessed by independently evaluating single claims in a binary setting. In our zero-shot exam simulation, unlike most extant research on entailment problems, we have little control over the framing of the hypothesis, claim, or body of knowledge. We have no insight into any knowledge graphs or state models, explicit or implicit, that exist in GPT. Furthermore, in some cases, multiple choices may be correct from an entailment perspective, and test-takers must rank order their choice based on knowledge of exam design. As such, there are elements of this test that are more similar to search and relevancy scoring than simple binary entailment/non-entailment problems.

(Hyper)parameters for GPT-3

The results of machine learning and computational research generally are often highly sensitive to model parameters or hyperparameters. Due to the zero-shot nature of the task and GPT-3.5's proprietary nature, it is not possible to identify all (hyper)parameters, or even to identify which parameters are truly hyperparameters. Despite this limitation, in addition to varying prompts as detailed above, we also evaluated how (hyper)parameters like model "temperature" impacted the performance of the model. We evaluated the following parameters for at least one prompt:

1. TEMPERATURE: Sampling temperature; 0.0 is deterministic, higher is more "random." We tested values in $\{0.0, 0.25, 0.5, 0.75, 1.0\}$.
2. TOP_P: Nucleus sampling probability. We tested values in $\{0.75, 1.0\}$.
3. BEST_OF: "Generates [N] completions server-side and returns the "best" (the one with the highest log probability per token)." We tested values in $\{1, 2, 4\}$.
4. MAX_TOKENS: Maximum number of tokens to generate. For prompts without an explanation, we tested values in $\{16, 32\}$. For prompts with an explanation, we tested values in $\{128, 256, 1024\}$.

In general, we do find some statistically significant differences between parameters, but most differences are trivial in relative performance terms. Two qualitative observations are notable. First, the model performs better when *some* stochasticity is introduced, but not too much. For both TEMPERATURE and TOP_P, correct rates increased when moving from deterministic to non-deterministic sampling. Second, the model performs better when at least two candidates are sampled via the BEST_OF parameter. Our ability to speculate or test further is limited in the context of GPT-3.5, but these findings may have applications in other assessments or for other models.

Fine-tuning

LLMs like GPT-3.5 have received so much interest in part because their zero-shot or few-shot performance is so good. Despite this, in some circumstances, subsequent supervised or unsupervised re-training of some or all layers of an LLM may improve performance [29][30]. OpenAI does make some retraining or "fine-tuning" capabilities available through its API, and these API endpoints do allow for some control of

the training process like learning rates or batch sizes. We did attempt to fine tune TEXT-DAVINCI-003 by providing it with 200 unseen, simulated MBE bar exam questions with correct and incorrect explanations. We provided the training samples both with and without explanatory text from the answer guide. In total, we trained six fine-tuned models, altering training prompts, training responses, batch size, learning rate, and prompt weighting. However, in all cases, the fine-tuned model significantly underperformed TEXT-DAVINCI-003 itself. Due to the scarcity of high-quality data for training and assessment, we did not pursue fine-tuning of GPT-3.5 models further, and these results possibly confirm LLM fine-tuning risks observed by others [31]. It is also possible that the decline in performance is related to the fact that fine-tuning through OpenAI's API results in a 50% reduction in the size of the maximum token count, which may be related to real contraction of the head layer(s) of the model. We are unable to speculate further without direct access to fine-tuning process or the architecture and weights of resulting models.

Other Models

OpenAI's API makes available multiple models, including smaller and older iterations of the GPT family. In addition to testing with TEXT-DAVINCI-003, we also repeated our best prompt and parameters with the TEXT-DAVINCI-001, TEXT-CURIE-001, TEXT-BABBAGE-001, and TEXT-ADA-001 models through the OpenAI API. We also attempted to test the GPT-2 model open-sourced by OpenAI through GitHub and Huggingface [32] and the CODE-DAVINCI-003 model available through OpenAI's API, but neither model followed the prompt instructions. A comparison of these models is included in the Results section below.

Results

In total, we executed 107 sample exams across the prompts and parameter values described above. Out of these prompts, prompt style #7 - rank-ordering of the top three choices - performed best, and we collected 41 sample runs across parameter combinations for this prompt. The performance of these runs is summarized in Figure 1 and Table 2, including a comparison with baseline student and passing rates.[1]

The table and figure clearly show that GPT-3.5 is not yet passing the overall multiple choice exam. However, GPT-3.5 is significantly exceeding the baseline random chance rate of 25%. Furthermore, GPT-3.5 has reached the average passing rate for at least two categories, Evidence and Torts.

On average across all categories, GPT-3.5 is trailing human test-takers by approximately 17%. In the case of Evidence, Torts, and Civil Procedure, this gap is negligible or in the single digits; at 1.5 times the standard error of the mean across our test runs, GPT-3.5 is already at parity with humans for Evidence questions. However, for the remaining categories of

[1] As the MBE is just one component of an overall bar exam, students cannot "pass" an exam solely by achieving 58-62% on the multiple choice; however, across a plurality of states, this score, in combination with adequate performance on other components, produces a satisfactory result.

Additional tables summarizing variation within and across hyperparameters is provided in the Supplementary Information section.
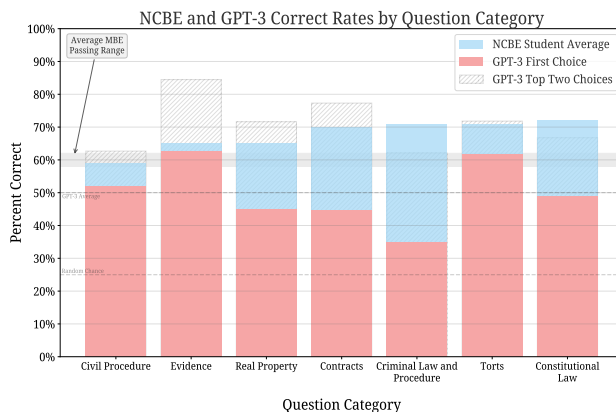

Fig. 1. Summary of performance by question category for GPT-3.5 and NCBE-Reported Students

	GPT-3.5	**Top 2**	**Top 3**	**NCBE**
Evidence	63%	84%	98%	65%
Torts	62%	72%	93%	71%
Civil Procedure	52%	63%	79%	59%
Constitutional Law	49%	67%	87%	72%
Real Property	45%	72%	85%	65%
Contracts	45%	77%	86%	70%
Criminal Law & Procedure	35%	62%	86%	71%
AVERAGE	50%	71%	88%	68%

Table 2. Summary of performance by question category for GPT-3.5, including second- and third-best choices, with NCBE-Reported Students

Constitutional Law, Real Property, Contracts, and Criminal Law, the gap is much more material, rising as high as 36% in the case of Criminal Law.

This performance gap may be attributable to at least two issues. First, it is possible that GPT-3.5's poor performance corresponds to bodies of knowledge that were absent from its training data or removed during subsequent model compression or fine-tuning. Our ability to speculate further is limited by lack of information about GPT-3.5's original provenance or subsequent architecture or re-training changes. Second, it is possible that GPT-3.5's poor performance on these categories is a result of the complex or purposefully-confusing language used by the exam's designers.

In order to explore these two possibilities, we next examine how "close" GPT-3.5 is to correct. If GPT-3.5 truly lacks knowledge about an area of law, then we should expect it to have low correlation between the rank of its answers and correctness. If, on the other hand, its second or third best choices are very often correct, then we can infer that the design of the questions may be responsible for poor performance. As shown by Table 1, some sections of the Bar are "trickier" than others, and so this finding may itself confirm what is held to be common knowledge by many human test-takers.

To understand this rank order performance, Figure 1 and Table 2 also include information about the performance of the model including its second-best and third-best answers. As shown by the gray dashed region in the figure and the "Top 2" column in the table, GPT-3.5's second best answer

is highly correlated with correctness. In all categories, the top two answers exceed the baseline random chance rate of 50%, and in five out of seven categories, exceed the NCBE-reported averages. The table also includes a summary of the top three GPT-3.5 answer performance in the "Top 3" column, which similarly shows strong overall correlation. Except for Civil Procedure, which is notably also the worst category for human test-takers, GPT-3.5's answers significantly exceed the baseline random chance rate.

Trends in model performance

As noted in the Methods section above, we also repeated these best prompt and parameters for other OpenAI models and attempted to test GPT-2. The results of this comparison are summarized in Figure 2, which charts the headline correct rate across GPT-2 and TEXT-ADA-001, TEXT-BABBAGE-001, TEXT-CURIE-001, TEXT-DAVINCI-001, and TEXT-DAVINCI-003. GPT-2 failed to respond to the prompt and also would not have been able to answer prompts over 1,024 tokens.
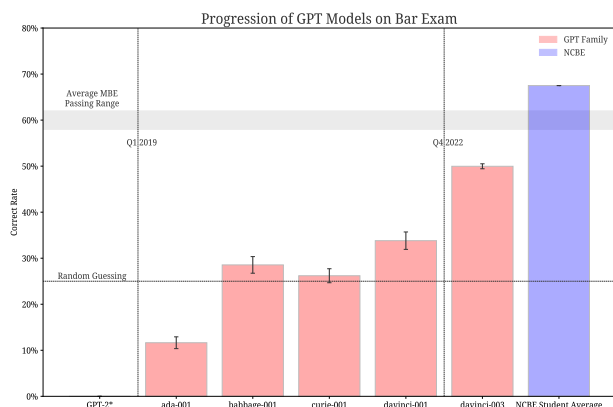
Fig. 2. Summary of performance across models and NCBE-Reported Students. (GPT-2 was unable to execute the prompt or handle longer input questions.)

This figure provide fascinating insight not only into GPT's performance on the Bar Exam specifically, but also into LLM capabilities generally. In under 36 months, state of the art LLMs have gone from 0% to achieving near-parity with humans in two categories. Both ADA and CURIE, trained sometime between 2019 and 2021, are worse than or indistinguishable from chance; BABBAGE and DAVINCI-001, while significantly above chance, did not perform much better. Meanwhile, the worst prompts and parameters for DAVINCI-003 are at least 10% better than DAVINCI-001. For best prompt and parameter, we have gone from 33.8% for DAVINCI-001 (+8.8% vs. chance) to 50.3% for DAVINCI-003 (+25.8% vs. chance). While it is clearly impossible for this absolute or relative rate to continue for more than one model generation at most, it stands to reason that a "future GPT" or similar LLM is on track to meet or exceed human performance within one to two model generations.

Conclusion and Future Work

In this research, we document our experimental evaluation of GPT-3.5 on the MBE portion of NCBE's model Bar Exam. Across all prompts and hyperparameter values, GPT-3.5

significantly outperformed the baseline rate of random guessing. Without any fine-tuning, it currently achieves a passing rate on two categories of the Bar and achieves parity with human test-takers on one. Its rank-ordering of possible choices is strongly correlated with correctness in excess of random chance, confirming its general understanding of the legal domain.

Overall, we find that GPT-3.5 significantly exceeds our expectations for performance on this task. Despite thousands of hours on related tasks over the last two decades between the authors, we did not expect GPT-3.5 to demonstrate such proficiency in a zero-shot settings with minimal modeling and optimization effort. While our ability to interpret how or why GPT-3.5 chooses between candidate answers is limited by understanding of LLMs and the proprietary nature of GPT, the history of similar problems strongly suggests that an LLM may soon pass the Bar. Based on anecdotal evidence related to GPT-4 and Bloom family of models, it is quite possible that this will occur within the next 0-18 months.

Many of the outstanding questions or improvements on this problem require either collaboration with OpenAI or the use of an alternative model that can be directly inspected, such as those maintained by EleutherAI, BigScience, or LAION communities. We intend to replicate our experimental design and continue fine-tuning with models from the GPT-J, GPT-Neo, and Bloom families. Separately, as noted above, the MBE is just one component of the overall Bar exam; we intend to assess both GPT-3.5 and other models mentioned above on both the essay (MEE) and situational performance (MPT) sections of the Exam in future work.

Acknowledgments

Supplementary Material

Supplementary material is available online at `https://github.com/mjbommar/gpt-takes-the-bar-exam`.

Author contributions statement

M.B. and D.K. contributed equally in this research.

Preprints

A preprint of this article is published at
`https://arxiv.org/abs/2212.14402`
`https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4314839`

Data availability

While exam data required to reproduce this publication cannot be redistributed by the authors, it can be purchased directly from the NCBE for under 300 USD. Results and analysis that do not reproduce copyrighted material are also available in the online SI.

References

1. MJ Bommarito II and DM Katz. Measuring and modeling the us regulatory ecosystem. *Journal of Statistical Physics*, 168(5):1125–1135, 2017.

2. C Coupette, J Beckedorf, D Hartung, M Bommarito, and DM Katz. Measuring law over time: A network analytical framework with an application to statutes and regulations in the united states and germany. *Frontiers in Physics*, 9, 2021.

3. JB Ruhl, DM Katz, and MJ Bommarito. Harnessing legal complexity. *Science*, 355(6332):1377–1378, 2017.

4. Richard Susskind. *The end of lawyers*. Oxford: Oxford University Press, 2008.

5. Richard E. Susskind. *Tomorrow's Lawyers: An Introduction to Your Future*. Oxford University Press, 2017.

6. Daniel M. Katz, Ron Dolin, and Michael J. Bommarito, editors. *Legal Informatics*. Cambridge University Press, 2021.

7. John Armour, Rachel Parnham, and Mariko Sako. Augmented lawyering. *U. Ill. L. Rev.*, page 71, 2022.

8. David F. Engstrom and Jonah B. Gelbach. Legal tech, civil procedure, and the future of adversarialism. *U. Pa. L. Rev.*, 169:1001, 2020.

9. Daniel M. Katz. Quantitative legal prediction-or-how i learned to stop worrying and start preparing for the data-driven future of the legal services industry. *Emory LJ*, 62:909, 2012.

10. Ronald W Staudt. All the wild possibilities: Technology that attacks barriers to access to justice. *Loy. LAL Rev.*, 42:1117, 2008.

11. I. Chalkidis, A. Jana, D. Hartung, M. Bommarito, I. Androutsopoulos, D. M. Katz, and N. Aletras. Lexglue: A benchmark dataset for legal language understanding in english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, May 2022.

12. D. M. Katz, C. Coupette, J. Beckedorf, and D. Hartung. Complex societies and the growth of the law. *Scientific Reports*, 10(1):1–4, Oct 2020.

13. D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, Oct 1986.

14. Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

15. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

16. J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, Oct 2014.

17. M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proceedings of the NAACL HLT 2018-2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies-Proceedings of the Conference*, 2018.

18. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186, 2018.

19. T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, and S. Agarwal. Language models are few-shot learners. In *Advances in neural information processing systems*, volume 33, pages 1877–1901. 2020.

20. M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, and A. Ahmed. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297, 2020.

21. T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, and J. Tow. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

22. R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H. T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, and Y. Li. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

23. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, volume 30, 2017.

24. Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

25. A. Z. Wyner, B. J. Fawei, and J. Z. Pan. Passing a usa national bar exam: a first corpus for experimentation. In *LREC 2016, Tenth International Conference on Language Resources and Evaluation*, pages 23–30, 2016.

26. Yoshinobu Kano, Mi-Young Kim, Masaharu Yoshioka, Yao Lu, Juliano Rabelo, Naoki Kiyota, Randy Goebel, and Ken Satoh. Coliee-2018: Evaluation of the competition on legal information extraction and entailment. In *New Frontiers in Artificial Intelligence*, pages 177–192. Springer International Publishing, 2019.

27. H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and M. Sun. Jec-qa: a legal-domain question answering dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9701–9708, 2020.

28. J. W. Howarth. *Shaping the Bar: The Future of Attorney Licensing*. Stanford University Press, 2022.

29. A. Dunn, J. Dagdelen, N. Walker, S. Lee, A. S. Rosen, G. Ceder, K. Persson, and A. Jain. Structured information extraction from complex scientific text with fine-tuned large language models. *arXiv preprint arXiv:2212.05238*, 2022.

30. J. Huang, S. S. Gu, L. Hou, Y. Wu, X. Wang, H. Yu, and J. Han. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*, 2022.

31. Y. Wang, S. Si, D. Li, M. Lukasik, F. Yu, C. J. Hsieh, I. S. Dhillon, and S. Kumar. Preserving in-context learning ability in large language model fine-tuning. *arXiv preprint arXiv:2211.00635*, 2022.

32. Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.