PREDICTING CONSUMER CONTRACTS

Noam Kolt*

37 BERKELEY TECH. L.J. __ (forthcoming 2022)

This Article empirically examines whether a computational language model can read and understand consumer contracts. Language models are able to perform a wide range of complex tasks by predicting the next word in a sequence. In the legal domain, language models can summarize laws, translate legalese into plain English, and, as this Article will explore, inform consumers of their contractual rights and obligations.

To showcase the opportunities and challenges of using language models to read consumer contracts, this Article studies the performance of GPT-3, the world's first commercial language model. The case study employs a novel dataset comprised of questions relating to the terms of service of popular U.S. websites. Although the results are not definitive, they offer several important insights. First, owing to its immense training data, the model can exploit subtle informational cues embedded in questions. Second, the model performed poorly on contractual provisions that favor the rights and interests of consumers, suggesting that it may contain an anti-consumer bias. Third, the model is brittle in unexpected ways. Performance was highly sensitive to the wording of questions, but surprisingly indifferent to variations in contractual language.

While language models could potentially empower consumers, they could also provide misleading legal advice and entrench harmful biases. Leveraging the benefits of language models in reading consumer contracts and confronting the challenges they pose requires a combination of engineering and governance. Policymakers, together with developers and users of language models, should explore technical and institutional safeguards to ensure that language models are used responsibly and align with broader social values.

[*] Doctoral Candidate and Vanier Scholar, University of Toronto Faculty of Law, Inaugural Graduate Fellow, Schwartz Reisman Institute for Technology and Society. I thank Anthony Niblett, David Hoffman, Gillian Hadfield, Albert Yoon, John Giorgi, and participants in seminars at the ETH Zurich Center for Law & Economics and University of Toronto Centre for Ethics for helpful comments and suggestions. OpenAI generously provided access to GPT-3 through the API Academic Access Program. Additional feedback on this draft can be sent to {my first name}{dot}{my last name}{@}mail.utoronto.ca. Any errors or oversights are mine alone.

INTRODUCTION 2
I. BACKGROUND 8
A. What Are Language Models? 8
B. How Is GPT-3 Different? 10
C. Opportunities for Law 13
II. EXPERIMENTAL DESIGN 17
A. Task Description 17
B. Evaluation Criteria 19
C. Challenges and Limitations 22
III. RESULTS AND DISCUSSION 25
A. Performance 25
B. Anti-Consumer Bias 27
C. Informational Cues 35
D. Brittleness 36
IV. BROADER IMPLICATIONS 39
A. Experimentation 39
B. Trust 40
C. Compounding Bias 42
D. Governance 44
CONCLUSION 48
APPENDIX 50
A. Test Conditions 50
B. Overall Performance 52
C. Calibration 53

INTRODUCTION

Consumer contracts govern important aspects of our lives.[1] The ability to communicate with people online, search for information, and make retail purchases are all mediated by consumer contracts. These contracts control access to services, dictate terms of payment, and determine the remedies available when consumers' rights are violated. Yet, we seldom read these agreements.[2] Ordinary people do not have the

[1] *See* MARGARET JANE RADIN, BOILERPLATE: THE FINE PRINT, VANISHING RIGHTS, AND THE RULE OF LAW 7–8 (2013). *See also* Robert A. Hillman & Jeffrey J. Rachlinski, *Standard-Form Contracting in the Electronic Age*, 77 N.Y.U. L. REV. 429, 463–69 (2002).

[2] *See* Yannis Bakos et al., *Does Anyone Read the Fine Print? Consumer Attention to Standard Form Contracts*, 43 J. LEGAL STUD. 1, 19–22 (2014); Florencia Marotta-Wurgler, *Will Increased Disclosure Help? Evaluating the Recommendations of the ALI's "Principles of the Law of Software Contracts*,*"* 78 U. CHI. L. REV. 165, 178–82 (2011). *See also* Ian Ayres & Alan Schwartz, *The No-Reading Problem in Consumer Contract Law*, 66 STAN. L. REV. 545, 546–48 (2014); OMRI BEN-SHAHAR & CARL E. SCHNEIDER, MORE THAN YOU WANTED TO KNOW: THE FAILURE OF MANDATED DISCLOSURE 79–93 (2014); RADIN, *supra* note 1, at 12–13, 252; Omri Ben-Shahar, *The Myth of the "Opportunity to Read" in Contract Law,* 5 EUR. REV. CONT. L. 1, 2–3 (2009).

time, expertise, or willingness to investigate how everyday consumer contracts affect their rights and interests.[3] Reading these contracts *ourselves* is simply unfeasible.

One rapidly developing technology—computational language models—could potentially offer a solution. These machine learning models are able to perform a wide range of complex tasks by predicting the next word in a sequence.[4] Language models are, essentially, a powerful autocomplete. A user provides the model with a portion of text, and the model uses machine learning to guess what words should follow. The results are surprisingly impressive. For example, GPT-3[5]—the world's first commercial language model[6]—can write fiction,[7] translate natural language into computer code,[8] and produce news articles that appear to be written by human authors.[9]

[3] *See* RESTATEMENT (THIRD) OF CONTRACTS 3 (AM. LAW INST., Tentative Draft, 2019). *See also* Meirav Furth-Matzkin & Roseanna Sommers, *Consumer Psychology and the Problem of Fine-Print Fraud*, 72 STAN. L. REV. 503, 506–7 (2020); Yehuda Adar & Shmuel I. Becher, *Ending the License to Exploit: Administrative Oversight of Consumer Contracts*, B.C. L. REV. (forthcoming 2022). *See also* Eric Goldman, *The Crisis of Online Contracts (As Told in 10 Memes)* (Santa Clara Univ. Legal Studies Research Paper) (Mar. 3, 2021), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3796519.

[4] *See infra* Part I.A (tracing the development of language model technology).

[5] *See* Tom B. Brown et al., *Language Models Are Few-Shot Learners*, PROC. 34TH CONF. NEURAL INFO. PROCESSING SYS. (2020) (introducing GPT-3). For discussions of the broader impact of GPT-3 and other language models, see Cade Metz, *Meet GPT-3. It Has Learned to Code (and Blog and Argue)*, N.Y. TIMES (Nov. 24, 2020), https://www.nytimes.com/2020/11/24/science/artificial-intelligence-ai-gpt3.html; Will Douglas Heaven, *Why GPT-3 is the Best and Worst of AI Right Now*, MIT TECH. REV. (Feb. 24, 2021), https://www.technologyreview.com/2021/02/24/1017797/gpt3-best-worst-ai-openai-natural-language/; Matthew Hutson, *Robo-Writers: The Rise and Risks of Language-Generating AI*, NATURE (Mar. 3, 2021).

[6] *See infra* Part I.B (discussing GPT-3's revenue model).

[7] *See, e.g.,* Gwern Branwen, *GPT-3 Creative Fiction,* GWERN (Sept. 28, 2020), https://www.gwern.net/GPT-3.

[8] *See, e.g.,* Sharif Shameem (@SharifShameem), TWITTER (Jul 13, 2020, 5:01PM), https://twitter.com/sharifshameem/status/1282676454690451457 (demonstrating that GPT-3 can generate JSX code). In mid-2021, OpenAI released a language model trained specifically to generate computer code. *See* Mark Chen et al., *Evaluating Large Language Models Trained on Code*, ARXIV (Jul. 14, 2021), https://arxiv.org/abs/2107.03374. Together with GitHub, OpenAI also developed a commercial code generation tool. *See* GITHUB COPILOT, https://copilot.github.com/.

[9] *See* Brown et al., *supra* note 5, at 25–26 (finding that study participants' ability to detect which articles which were produced by GPT-3 rather than by human beings was scarcely above random chance). *See also* Irene Solaiman et al., *Release Strategies and the Social Impacts of Language Models* 10–13 (OpenAI Report, Nov. 13, 2019), https://arxiv.org/abs/1908.09203; Ben Buchanan et al., *Truth, Lies, and Automation How Language Models Could Change Disinformation,* CENTER FOR SECURITY AND EMERGING TECHNOLOGY ch. 2 (May 2021), https://cset.georgetown.edu/publication/truth-lies-and-automation/.

In the legal domain, language models are highly versatile. GPT-3, for instance, can summarize laws,[10] draft legal documents,[11] and translate legalese into plain English.[12] These capabilities present significant opportunities for both lawyers and consumers of legal services.[13] In the future, lawyers could use language models to expedite routine tasks, such as document review and transactional drafting. Language models could also assist lawyers in conducting legal research, generating statements of claim, and even predicting case outcomes. If language models continue to improve, they have the potential to fundamentally alter the way in which legal services are performed.[14]

Above all, language models could improve access to justice. By automating onerous legal tasks, language models could directly assist consumers, especially those who cannot afford traditional legal services. For example, one start-up is experimenting with using GPT-3 to produce legal requests on behalf of tenants who might otherwise need to engage professional counsel.[15] This Article explores another possibility: using language models to read consumer contracts. Despite the ubiquity of these agreements, consumers often struggle to discover and exercise their contractual rights.[16] By analyzing the provisions of these contracts and explaining their legal ramifications, language models could inform and empower consumers.[17]

[10] *See* Daniel Gross (@DanielGross), TWITTER (Jun. 14, 2020, 9:42 PM), https://twitter.com/danielgross/status/1272238098710097920.

[11] *See* Francis Jervis (@f_j_j_), TWITTER (Jul. 17, 2020, 12:02 PM), https://twitter.com/f_j_j_/status/1284050844787200000.

[12] *See* Michael Tefula (@MichaelTefula), TWITTER (Jul. 21, 2020, 12:24 PM), https://twitter.com/michaeltefula/status/1285505897108832257.

[13] *See infra* Part I.C. For general accounts of the application of machine learning in law, see Harry Surden, *Machine Learning and Law*, 89 WASH. L. REV. 87, 101–14 (2014); John O. McGinnis & Russell G. Pearce, *The Great Disruption: How Machine Intelligence Will Transform the Role of Lawyers in the Delivery of Legal Services*, 82 FORDHAM L. REV. 3041 (2014); Dana Remus & Frank Levy, *Can Robots Be Lawyers: Computers, Lawyers, and the Practice of Law*, 30 GEO. J. LEGAL ETHICS 501 (2017); KEVIN D. ASHLEY, ARTIFICIAL INTELLIGENCE AND LEGAL ANALYTICS: NEW TOOLS FOR LAW PRACTICE IN THE DIGITAL AGE (2017); Benjamin Alarie et al., *How Artificial Intelligence Will Affect the Practice of Law*, 68 U. TORONTO L.J. 106 (2018); Harry Surden, *Artificial Intelligence and Law: An Overview*, 35 GA. ST. U. L. REV. 1305 (2019); LEGAL INFORMATICS (Daniel Martin Katz et al. eds., 2021); NOAH WAISBERG & ALEXANDER HUDEK, AI FOR LAWYERS (2021).

[14] *See, e.g.,* Rudy DeFelice, *What Does GPT-3 Mean for the Future of the Legal Profession?*, TECHCRUNCH (Aug. 28, 2020), https://techcrunch.com/2020/08/28/what-does-gpt-3-mean-for-the-future-of-the-legal-profession/; Caroline Hill, *GPT-3 and Another Chat About the End of Lawyers*, LEGAL IT INSIDER (Aug. 3, 2020), https://legaltechnology.com/gpt-3-and-another-chat-about-the-end-of-lawyers/.

[15] *See* Augrented: Rent Safer (@augrented), TWITTER (Jul. 20, 2020, 7:31 AM), https://twitter.com/augrented/status/1285069733818056704; Jervis, TWITTER (Oct. 28, 2020, 11:45 AM), https://twitter.com/f_j_j_/status/1321387632652283906.

[16] *See supra* notes 1–2.

[17] *See generally* Yonathan A. Arbel & Shmuel I. Becher, *Contracts in the Age of Smart*

These opportunities, however, are accompanied by a host of concerns. Like other machine learning tools, language models pose serious risks.[18] They can underperform on tasks, amplify harmful biases, and be used for malicious purposes. Since the release of GPT-3, researchers of language models have increasingly focused on issues traditionally sidelined by the computer science community.[19] Computational linguists have studied the extent to which language models can be prompted to generate racist, sexist, and other toxic content.[20] Social scientists have questioned whether language models can be deployed safely and responsibly.[21] If language models are to become part of our legal toolkit, we must confront these issues.

Readers, 90 GEO. WASH. L. REV. (forthcoming 2022); Abhilasha Ravichander et al., *Breaking Down Walls of Text: How Can NLP Benefit Consumer Privacy?*, PROC. 59TH ANN. MEETING ASS'N COMPUTATIONAL LINGUISTICS 4125 (2021).

[18] *See* Emily M. Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, ACM CONF. FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 610 (2021); Alex Tamkin et al., *Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models*, ARXIV (Feb. 4, 2021), https://arxiv.org/abs/2102.02503. For general discussions of the risks associated with machine learning, see FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015); Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671 (2016); CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY (2016); Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023 (2017); Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap,* 51 U.C. DAVIS L. REV. 399 (2017); SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM (2018); VIRGINIA EUBANKS, AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR (2018); MICHAEL KEARNS & AARON ROTH, THE ETHICAL ALGORITHM (2019); Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218 (2019); Jon Kleinberg et al., *Discrimination in the Age of Algorithms*, 10 J. LEGAL ANALYSIS 113 (2019); Ben Hutchinson & Margaret Mitchell, *50 Years of Test (Un)fairness: Lessons for Machine Learning,* CONF. FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 40 (2019); FRANK PASQUALE, NEW LAWS OF ROBOTICS: DEFENDING HUMAN EXPERTISE IN THE AGE OF AI (2020).

[19] However, the field of natural language processing ethics is not new. Seminal papers include Dirk Hovy & Shannon L. Spruit, *The Social Impact of Natural Language Processing*, PROC. 54TH ANN. MEETING ASS'N COMPUTATIONAL LINGUISTICS 591 (2016); Tolga Bolukbasi et al., *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*, PROC. 30TH INT'L CONF. NEURAL INFO. PROCESSING SYS. 4356 (2016).

[20] *See, e.g.,* Samuel Gehman et al., *RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models*, FINDINGS 2020 CONF. EMPIRICAL METHODS IN NLP 3356 (2020) (examining the toxicity of content generated by language models).

[21] *See, e.g.,* Zhijing Jin et al., *How* Good *Is NLP? A Sober Look at NLP Tasks through the Lens of Social Impact*, FINDINGS ASS'N COMPUTATIONAL LINGUISTICS 3099 (2021). *See also* Luciano Floridi & Massimo Chiriatti, *GPT-3: Its Nature, Scope, Limits, and Consequences*, 30 MINDS & MACHINES 681 (2020); Kevin LaGrandeur, *How Safe Is Our Reliance on AI, and Should We Regulate It?*, AI ETHICS 1, 4 (Oct. 6, 2020).

To properly unpack the opportunities and challenges of deploying language models in law, we need to understand their inner workings. The way language models operate and the data used to train them can have far-reaching consequences. Accordingly, this Article traces the technology's development, from the simplest models through to the most recent breakthroughs.[22] One feature, however, remains constant. All language models, including GPT-3, do primarily one thing: predict the probability of a word or sequence of words given the preceding text. They function as an autocomplete, guessing what words follow from a particular text. Seen in this light, the range of tasks that state-of-the-art models can perform is remarkable. At the same time, this feature of language models also explains some of their pitfalls.

A theoretical understanding of language model technology, however, does not guarantee reliable performance. To evaluate the ability of a language model to perform legal tasks, we need to test a model on legal tasks. This Article presents a preliminary case study in using GPT-3 to read consumer contracts.[23] The case study aims to examine the degree to which the model can understand certain consumer contracts. To conduct the case study, I created a novel dataset comprised of 200 yes/no legal questions relating to the terms of service of the 20 most-visited U.S. websites, and tested the model's ability to answer these questions. The results are illuminating.

First, owing to its immense training data, GPT-3 appears to be able to exploit subtle informational cues embedded in questions.[24] More specifically, the case study offers suggestive evidence that GPT-3 can recall information regarding specific companies from its training data, which in turn improves performance.

Second, GPT-3 performed considerably worse on contractual provisions that favor the rights and interests of consumers.[25] The model answered correctly nearly 84% of the questions concerning provisions that favor companies, but only 60% of the questions concerning provisions that favor consumers. This result is potentially disturbing. One possible explanation is that the model contains an anti-consumer bias that reflects the skewed data on which the model was trained—namely, website terms of service that disproportionately preference the rights and interests of companies over the rights and interests of consumers.

[22] *See infra* Part I.A.
[23] *See infra* Part II.
[24] *See infra* Part III.C.
[25] *See infra* Part III.B.

Third, the case study found that GPT-3 is brittle in unexpected ways. The model appears to be highly sensitive to how questions are worded, but surprisingly indifferent to variations in contractual language.[26] Performance decreased dramatically when the questions presented to the model were less readable (i.e., more difficult for a human to read). However, performance was not worse on longer or less readable contractual texts.

Notably, the case study offers only an initial exploratory analysis of the prospect of using language models to read consumer contracts. The analysis is subject to several limitations and, accordingly, the findings are not definitive. Nevertheless, the case study raises important questions regarding the potential advantages and pitfalls of using language models to inform consumers of their contractual rights, and proposes concrete directions for future research.

Subject to these qualifications, the case study paints a nuanced picture. On the one hand, it illustrates some of the mechanisms that could account for GPT-3's generally strong performance. On the other hand, the case study highlights some of the model's weaknesses, including the outsized impact of question wording on performance. In addition, poor performance on contractual provisions that favor consumers reinforces broader concerns regarding the effect of societal biases in machine learning.

These insights have implications for various stakeholders. Users of language models need to be aware of the technology's limitations. Developers of language models should explore technical methods for addressing these limitations and improving the reliability of language models. Finally, policymakers should design institutional safeguards to ensure that language models are used responsibly and align with broader social values.

This Article proceeds in four parts. Part I provides an introduction to language model technology and the opportunities it offers the legal domain. Part II describes the experimental design used in the case study. Part III presents and analyzes the results. Part IV discusses the case study's broader implications and proposes avenues for future work.

[26] *See infra* Part III.D.

I. BACKGROUND

A. What Are Language Models?

Language models assign probabilities to sequences of words and sentences.[27] By assigning these probabilities, language models can predict the probability of a word or sequence of words given the preceding text.[28] For example, given the sequence "she took the LSAT and applied to law . . ." an effective language model will assign probabilities to each word in the sequence and predict that the next word in the sequence is likely to be "school." Language models can also predict the content of longer sequences, and thereby generate lengthy synthetic texts. For example, GPT-3 can write Shakespearean sonnets.[29] The striking feature of advanced language models is that, merely by predicting the next word in a sequence, they can produce human-like texts that appear to exhibit genuine knowledge, understanding, and even emotion.[30]

How do language models assign probabilities and make predictions? The basic idea is that words can be understood by their context, that is, other nearby words.[31] Suppose, for example, we have the sequence "many legal questions concerning contracts are" and want to calculate the probability that the next word in the sequence is "difficult". One way to estimate this probability is to take a large corpus of text, such as millions of books or websites, and count the number of times that the sequence "many legal questions concerning contracts are" is followed by the word "difficult", and divide this by the total number of times that the initial sequence appears in the corpus.[32] If the corpus is sufficiently large, this method will produce a good estimate. However, this method will fail if the relevant sequence does not appear in the corpus. This issue is

[27] *See* DAN JURAFSKY & JAMES H. MARTIN, SPEECH AND LANGUAGE PROCESSING 29–52, 127–47, 173–201 (draft 3rd ed., revised Dec. 30, 2020) (providing an overview of *n*-gram language models, neural language models, and RNNs, respectively).

[28] Autoregressive models, such as GPT-3, process text from left to right and assign probabilities based only the preceding text. In contrast, bidirectional models learn from the surrounding text on both sides of the target word. *See, e.g.,* Jacob Devlin et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, PROC. 2019 ANN. CONF. N. AM. CH. ASS'N COMPUTATIONAL LINGUISTICS 4171 (2019) (introducing Google's BERT, which is a bidirectional language model).

[29] Gwern, *supra* note 7.

[30] *But see infra* note 119 (discussing whether language models can understand language).

[31] *See* J.R. Firth, *A Synopsis of Linguistic Theory, 1930–1955*, *in* STUDIES IN LINGUISTIC ANALYSIS 1, 11 (J.R. Firth et al. eds., 1957) (coining the canonical phrase "[y]ou shall know a word by the company it keeps"). *See also* LUDWIG WITTGENSTEIN, PHILOSOPHICAL INVESTIGATIONS § 43 (1953) (contending that "[t]he meaning of a word is its use in the language").

[32] This is known as a *relative frequency count*. *See* JURAFSKY & MARTIN, *supra* note 27, at 29–30.

commonplace. For instance, with the addition of just a few words to the above sequence—"many legal questions concerning *ancient Roman commercial* contracts are"—we may have a novel sentence that does not appear in any existing corpus. Accordingly, the above method could not be used to calculate the probability of the next word in the sequence.

A simple solution is to instead calculate the probability of a word based on only one or a few of the immediately preceding words, rather than on the entire preceding sequence. A *bigram* uses the one immediately preceding word. A *trigram* uses the two preceding words. This family of language models, known as *n*-grams, treats the probability of a word as depending only on the preceding *n* - 1 words.[33] Calculating the relative frequencies for *n*-grams is usually feasible. In a large corpus, there are likely to be sequences in which the word "Roman" follows the word "ancient", and the word "commercial" follows "Roman."

More sophisticated methods of language modeling have been developed in recent decades. The most prominent of these is neural language models.[34] Neural language models, which are based on neural networks, can use longer sequences of text to predict an upcoming word or sequence, and typically make these predictions with higher accuracy than *n*-gram models. Neural language models also generalize better to unseen contexts. One of the main ways in which neural language models differ from *n*-grams is that they represent text by semantic word embeddings,[35] i.e., mathematical representations that express the *meaning* of words.[36] For example, neural language models may make similar predictions regarding the sequence that follows the words "contract" and "agreement" because these two words have similar meanings.

[33] Formally, *n*-grams assume that the probability of a given word can be predicted based on only a limited number of preceding words (the Markov assumption). By multiplying together the probabilities of different words, *n*-grams can also be used to estimate the probabilities of entire sequences of text (as opposed to just single words).

[34] *See* Yoshua Bengio et al., *A Neural Probabilistic Language Model*, 3 J. MACH. LEARNING RES. 1137 (2003); Yoshua Bengio et al., *Neural Probabilistic Language Models*, *in* INNOVATIONS IN MACH. LEARNING 137 (D.E. Holmes & L.C. Jain eds., 2006) (introducing neural language models). For an overview of neural language models, see Yoav Goldberg, *A Primer on Neural Network Models for Natural Language Processing*, 57 J. AI RES. 345 (2017).

[35] *See, e.g.,* Tomás Mikolov et al., *Efficient Estimation of Word Representations in Vector Space*, 1ST INT'L CONF. LEARNING REPRESENTATIONS (2013); Tomás Mikolov et al., *Distributed Representations of Words and Phrases and Their Compositionality*, PROC. 26TH INT'L CONF. NEURAL INFO. PROCESSING SYS. 3111 (2013) (introducing the word2vec embeddings); Jeffrey Pennington et al., *GloVe: Global Vectors for Word Representation*, PROC. 2014 CONF. EMPIRICAL METHODS IN NLP 1532 (2014) (introducing the GloVe embeddings).

[36] By contrast, *n*-gram models use only the relative frequency count of words.

Neural language models, however, confront the problem of *long-range dependencies*.[37] Consider the following sequence: "the lawyers, who had been working at the firm for many years, are eager to . . ." A language model, when predicting the probability of the word following "years," may forget that the subject ("lawyers")—which appears much earlier in the sentence—is plural and should thus be followed by "are" (rather than "is"). By the end of a long sequence the model may fail to retain information contained in earlier parts of the sequence. The advent of *transformers*, the current state-of-the-art architecture for language modeling, has made significant progress in tackling this problem.[38]

Another notable development is *pretraining*. This involves training a general-purpose language model on a very large unlabeled dataset.[39] This process is compute-intensive and costly, and is typically carried out by a large organization. The resulting pretrained model, which is publicly released,[40] can then be fine-tuned on a smaller dataset that is relevant for a specific task. For example, Google's pretrained BERT model can be fine-tuned on case law and contracts in order to perform specialized legal tasks.[41] Fine-tuning is comparatively inexpensive. As a result, developers and researchers can now access and conveniently deploy powerful language models in their respective domains.

B. How Is GPT-3 Different?

GPT-3 is a powerful pretrained language model.[42] Although it is structurally similar to earlier models,[43] GPT-3 differs in important ways. First of all, GPT-3 can perform a diverse range of tasks without additional training or fine-tuning. For example, GPT-3 can out-of-the-box answer trivia questions, summarize text and translate between

[37] *See* JURAFSKY & MARTIN, *supra* note 27, at 186–87.

[38] *See* Ashish Vaswani et al., *Attention Is All You Need*, PROC. 30TH INT'L CONF. NEURAL INFO. PROCESSING SYS. 5998 (2017) (introducing transformers). *See also* Dzmitry Bahdanau et al., *Neural Machine Translation by Jointly Learning to Align and Translate*, 3RD INT'L CONF. LEARNING REPRESENTATIONS (2015) (introducing the attention mechanism, which is a key component of the transformer architecture).

[39] *See* Sebastian Ruder, *Recent Advances in Language Model Fine-tuning* (Feb. 24, 2021), https://ruder.io/recent-advances-lm-fine-tuning/.

[40] Most prominent pretrained models are available in the Hugging Face library. *See* HUGGING FACE, https://huggingface.co/models.

[41] *See* Ilias Chalkidis et al., *LEGAL-BERT: The Muppets Straight Out of Law School*, FINDINGS 2020 CONF. EMPIRICAL METHODS IN NLP 2898 (2020).

[42] *See* Brown et al., *supra* note 5. References to GPT-3 are to the largest model in the GPT-3 family of models, which has 175 billion parameters.

[43] It is especially similar to its predecessor, GPT-2. *See* Alec Radford et al., *Language Models are Unsupervised Multitask Learners* (OpenAI Working Paper, Feb. 2019) (introducing GPT-2).

languages.[44] In addition, users can teach the model to perform new tasks simply by providing instructions (in natural language) or presenting the model with several examples of the desired tasks. For instance, the following prompt could teach GPT-3 to correct the grammar of an English text:[45]

> Non-standard English: If I'm stressed out about something, I tend to have problem to fall asleep.
> Standard English: If I'm stressed out about something, I tend to have a problem falling asleep.
>
> Non-standard English: There is plenty of fun things to do in the summer when your able to go outside.
> Standard English: There are plenty of fun things to do in the summer when you are able to go outside.
>
> Non-standard English: She no went to the market.
> Standard English: She didn't go to the market.

Presented with another grammatically erroneous text, GPT-3 can learn to produce a grammatically correct version of that text.[46] This kind of learning—known as *few-shot learning*[47]—is highly intuitive and enables non-programmers to program the model.[48] For these and other reasons, some observers have suggested that GPT-3 is the closest attempt to achieving artificial general intelligence.[49]

[44] *See* Brown et al., *supra* note 5, at 10–29.

[45] This prompt is adapted from an example provided in the OpenAI API.

[46] Similar sets of examples can be used to teach GPT-3 to construct headlines for news articles, write professional emails, and convert English instructions into computer code. *See, e.g.,* Yaser Martinez Palenzuela et al., *Awesome GPT-3*, GitHub (Sep. 29, 2020), https://github.com/elyase/awesome-gpt3.

[47] The ability to learn from prompts is also known as *prompt-based learning*, *in-context learning* or *meta-learning*.

[48] *See* Vasili Shynkarenka, *How I Used GPT-3 to Hit Hacker News Front Page 5 Times in 3 Weeks*, Vasili Shynkarenka (Oct. 28, 2020) https://vasilishynkarenka.com/gpt-3/ ("If we teleport 50 years from now, it will seem barbaric that in 2020 we had an elite cast of hackers who knew how to write special symbols to control the computing power"). *See also* Chen et al, *supra* note 8, at 34 (discussing the impact of code generation on non-engineers).

[49] *See, e.g.,* Julien Lauret, *GPT-3: The First Artificial General Intelligence?*, Towards Data Science (Jul. 22, 2020), https://towardsdatascience.com/gpt-3-the-first-artificial-general-intelligence-b8d9b38557a1; Katherine Elkins & Jon Chun, *Can GPT-3 Pass a Writer's Turing Test?*, J. Cultural Analytics (Sept. 14, 2020). *Compare* Gary Marcus & Ernest Davis, *GPT-3, Bloviator: OpenAI's Language Generator Has No Idea What It's Talking About*, MIT Tech. Rev. (Aug. 22, 2020), https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion; Yann LeCun, Facebook (Oct. 27, 2020), https://www.facebook.com/yann.lecun/posts/10157253205637143. *See also infra* note 119 (discussing whether language models can understand language). For broader discussions

The second difference between GPT-3 and earlier language models—and the main factor accounting for GPT-3's strong performance—is scale.[50] GPT-3 contains 175 parameters (i.e., model weights or coefficients), which is an order of magnitude more than the previously largest language model.[51] More concretely, the cost of training GPT-3 is estimated to be between $4 million and $12 million.[52] The size of the dataset on which GPT-3 was trained is also immense. It includes over 570GB of raw web page data, online books corpora, and English-language Wikipedia[53]—which contain approximately 57 billion times the number of words perceived in an average human lifetime.[54]

The third feature that distinguishes GPT-3 from other language models is that it is proprietary. Most large language models, such as Google's BERT and Facebook's RoBERTA, are publicly available.[55] Researchers are free to inspect the code and weights of these models, and re-train or fine-tune them on new data. OpenAI, however, did not open-source the underlying GPT-3 model.[56] Instead, OpenAI released a closed-

regarding artificial general intelligence, see ARTIFICIAL GENERAL INTELLIGENCE (Ben Goertzel & Cassio Pennachin eds., 2007); NICK BOSTROM, SUPERINTELLIGENCE: PATHS, DANGERS, STRATEGIES (2014).

[50] *See* Jared Kaplan et al., *Scaling Laws for Neural Language Models*, ARXIV (Jan. 23, 2020), https://arxiv.org/abs/2001.08361; Tom Henighan et al., *Scaling Laws for Autoregressive Generative Modeling*, ARXIV (Oct. 28, 2020), https://arxiv.org/abs/2010.14701 (showing that language models improve with scale).

[51] Subsequently, however, even larger models have been developed. *See, e.g.,* William Fedus et al., *Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity*, ARXIV (Jan. 11, 2021), https://arxiv.org/abs/2101.03961 (introducing the first language model known to exceed one trillion parameters).

[52] These are estimates of the cost of compute only, not staff or other costs. *See* Kyle Wiggers, *OpenAI Launches an API to Commercialize Its Research*, VENTUREBEAT (Jun. 11, 2020), https://venturebeat.com/2020/06/11/openai-launches-an-api-to-commercialize-its-research/; Chuan Li, *OpenAI's GPT-3 Language Model: A Technical Overview*, LAMBDA (Jun. 3, 2020), https://lambdalabs.com/blog/demystifying-gpt-3/.

[53] This is roughly two orders of magnitude larger than the Corpus of Contemporary American English (COCA). *See* CORPUS OF CONTEMPORARY AMERICAN ENGLISH, https://www.english-corpora.org/coca/. However, words in COCA are annotated with additional linguistic information that facilitate using corpus linguistics techniques to analyze text. *See, e.g.,* ANNE O'KEEFE & MICHAEL MCCARTHY, THE ROUTLEDGE HANDBOOK OF CORPUS LINGUISTICS 433 (2010). In contrast, the training data for GPT-3 and other pretrained language models are not annotated or labeled.

[54] *See* Shana Lynch, *Is GPT-3 Intelligent? A Directors' Conversation with Oren Etzioni*, STANFORD UNIVERSITY HUMAN-CENTERED ARTIFICIAL INTELLIGENCE (Oct. 1, 2020), https://hai.stanford.edu/blog/gpt-3-intelligent-directors-conversation-oren-etzioni.

[55] *See* Devlin et al., *supra* note 28 (introducing BERT). *See* Yinhan Liu et al., *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, ARXIV (Jul. 26, 2019), https://arxiv.org/abs/1907.11692 (introducing RoBERTa). Notably, GPT-2 was subject to a staged release, in which increasingly large models in the GPT-2 family of models were made publicly available. *See GPT-2: 1.5B Release*, OPENAI (Nov. 5, 2019) https://openai.com/blog/gpt-2-1-5b-release/.

[56] The underlying model has been exclusively licensed to Microsoft. *See* Kevin Scott,

access application programming interface (API) that is powered by GPT-3.[57] Developers can request access and, if granted, pay for API credits.[58] This revenue model, which is the first of its kind, makes GPT-3 the world's first commercial language model.[59]

Finally, GPT-3 has distinctive shortcomings.[60] Because of its size, GPT-3 is slow to run compared with leaner task-specific language models. In addition, because users cannot access the underlying model, they cannot (themselves) fine-tune GPT-3.[61] Last of all, heightened capabilities present heightened risks of abuse. GPT-3 could be used to spread disinformation, generate spam, and achieve other nefarious purposes at unprecedented scale.[62]

C. Opportunities for Law

Language models have been deployed in the legal domain for decades. While language models were initially used to classify case law resources[63] and assist in legal search,[64] they are now being deployed in a broader range of legal applications.[65] Language models have been used to review

Microsoft Teams Up with OpenAI to Exclusively License GPT-3 Language Model, MICROSOFT (Sept. 22, 2020), https://blogs.microsoft.com/blog/2020/09/22/microsoft-teams-up-with-openai-to-exclusively-license-gpt-3-language-model/.

[57] *See* OPENAI BETA, https://beta.openai.com/.

[58] *See OpenAI API*, OPENAI (Jun. 11, 2020), https://openai.com/blog/openai-api/. *See* Cheng He, *Understand the Pricing of GPT-3*, MEDIUM (Sept. 14, 2020), https://chengh.medium.com/understand-the-pricing-of-gpt3-e646b2d63320. The open source community has attempted to reproduce a language model similar to GPT-3. *See* EleutherAI, *GPT-Neo*, GITHUB, https://github.com/EleutherAI/gpt-neo; *GPT-J-6B*, GITHUB, https://github.com/kingoflolz/mesh-transformer-jax/#gpt-j-6b.

[59] *See infra* Part IV.D (discussing access to language model technology).

[60] *See* Nostalgebraist, *Is GPT-3 Few-Shot Ready For Real Applications?*, LESSWRONG (Aug. 3, 2020), https://www.lesswrong.com/posts/LWdpWRyWu9aXsoZpA/is-gpt-3-few-shot-ready-for-real-applications-1; Max Woolf, *Tempering Expectations for GPT-3 and OpenAI's API*, MAX WOOLF'S BLOG (Jul. 18, 2020), https://minimaxir.com/2020/07/gpt3-expectations/.

[61] Although fine-tuning was not available when GPT-3 was released, OpenAI has subsequently offered fine-tuning through its API. Users, however, cannot themselves fine-tune the model. *See Fine-tuning,* OPENAI, https://beta.openai.com/docs/guides/fine-tuning.

[62] *See infra* Part IV.D (discussing the potential misuses of language models).

[63] *See, e.g.,* Stefanie Brüninghaus & Kevin D. Ashley, *Finding Factors: Learning to Classify Case Opinions Under Abstract Fact Categories*, PROC. 6TH INT'L CONF. AI & L. 123 (1997).

[64] *See, e.g.,* Jacques Savoy, *Searching Information in Legal Hypertext Systems*, 2 AI & L. 205 (1993).

[65] *See* Ilias Chalkidis & Dimitrios Kampas, *Deep Learning in Law: Early Adaptation and Legal Word Embeddings Trained on Large Corpora*, 27 AI & L. 171 (2019); Haoxi Zhong et al., *How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence*, PROC. 58TH ANN. MEETING ASS'N COMPUTATIONAL LINGUISTICS 5218 (2020).

documents in e-discovery,[66] predict case outcomes,[67] and generate patent claims.[68] In the realm of contracts, language models have been used to detect unfair or invalid clauses,[69] identify contractual provisions,[70] and draft investment agreements.[71] Language models also feature prominently in the emerging field of computational legal studies.[72]

Today, language models that perform legal tasks are typically trained or fine-tuned on legal data. For example, a language model designed to interpret tax legislation was trained on a corpus of tax cases and rulings.[73] Models like this have the advantage of being tailored to a particular task. However, assembling the necessary training data can be costly and time-consuming, especially if it involves recruiting legal experts.[74] Consequently, many legal organizations cannot effectively deploy language models.

[66] *See, e.g.,* Ngoc Phuoc An Vo et al., *Experimenting Word Embeddings in Assisting Legal Review*, PROC. 16TH INT'L CONF. AI & L. 189 (2017).

[67] *See, e.g.,* Haoxi Zhong et al., *Legal Judgment Prediction via Topological Learning*, PROC. 2018 CONF. EMPIRICAL METHODS IN NLP 3540 (2018).

[68] *See, e.g.,* Jieh-Sheng Lee & Jieh Hsiang, *Patent Claim Generation by Fine-Tuning OpenAI GPT-2*, 62 WORLD PATENT INFORMATION 101983 (2020).

[69] *See, e.g.,* Marco Lippi et al., *CLAUDETTE: An Automated Detector of Potentially Unfair Clauses in Online Terms of Service*, 27 AI & L. 117 (2019); Daniel Braun & Florian Matthes, *NLP for Consumer Protection: Battling Illegal Clauses in German Terms and Conditions in Online Shopping*, PROC. 1ST WORKSHOP ON NLP FOR POSITIVE IMPACT 93 (2021); Alfonso Guarino et al., *A Machine Learning-Based Approach to Identify Unlawful Practices in Online Terms of Service: Analysis, Implementation and Evaluation*, NEURAL COMPUTATION & APPLICATIONS (2021).

[70] *See, e.g.,* Ilias Chalkidis et al., *Neural Contract Element Extraction Revisited*, 33RD CONF. NEURAL INFO. PROCESSING SYS. (2019); Ilias Chalkidis et al., *Obligation and Prohibition Extraction Using Hierarchical RNNs*, PROC. 56TH ANN. MEETING ASS'N COMPUTATIONAL LINGUISTICS 254 (2018); Emad Elwany et al., *BERT Goes to Law School: Quantifying the Competitive Advantage of Access to Large Legal Corpora in Contract Understanding*, 33RD CONF. NEURAL INFO. PROCESSING SYS. (DOCUMENT INTELL. WORKSHOP) (2019); Dan Hendrycks et al., *CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review*, ARXIV (Mar. 10, 2021), https://arxiv.org/abs/2103.06268 [hereinafter Hendrycks et al., *CUAD*].

[71] *See, e.g.,* Wolfgang Alschner & Dmitriy Skougarevskiy, *Towards an Automated Production of Legal Texts Using Recurrent Neural Networks*, PROC. 16TH INT'L CONF. AI & L. 229 (2017).

[72] *See generally* LAW AS DATA: COMPUTATION, TEXT, AND THE FUTURE OF LEGAL ANALYSIS (Michael A. Livermore & Daniel N. Rockmore eds. 2019); Jens Frankenreiter & Michael A. Livermore, *Computational Methods in Legal Analysis*, 16 ANNU. REV. L. & SOC. SCI. 39 (2020); COMPUTATIONAL LEGAL STUDIES: THE PROMISE AND CHALLENGE OF DATA-DRIVEN RESEARCH (Ryan Whalen ed., 2020).

[73] *See* Nils Holzenberger et al., *A Dataset for Statutory Reasoning in Tax Law Entailment and Question Answering*, PROC. NATURAL LEGAL LANGUAGE PROCESSING WORKSHOP at 4–5 (2020).

[74] *See* Hendrycks et al., *CUAD*, *supra* note 70, at 1–2. *See also* Kevin D. Ashley, *Automatically Extracting Meaning from Legal Texts: Opportunities and Challenges*, 35 GA. ST. U. L. REV. 1117, 1138–44 (2019) (discussing the need for manual annotation in supervised learning).

GPT-3's powerful out-of-the-box performance could signal a change. Within weeks of its release, developers had used GPT-3 to summarize sections of the tax code,[75] translate legalese into plain English,[76] and prepare various legal documents.[77] These and other applications present significant commercial opportunities for lawyers and legal technology firms.[78] For example, language models could help lawyers conduct legal research more efficiently,[79] automate transactional drafting,[80] and generate synthetic legal data to train other machine learning models to perform legal tasks.

Above all, however, language models could improve access to justice.[81] For example, language models could democratize legal knowledge by providing simple explanations of complex legal texts.[82] Meanwhile, using language models to prepare legal documents could empower consumers who cannot afford traditional legal services.[83] Language models could

[75] *See* Gross, *supra* note 10.

[76] *See* Tefula, *supra* note 12. Others have used GPT-3 to translate plain English into legalese. *See, e.g.,* Ed Leon Klinger (@edleonklinger), TWITTER (Jul. 18, 2020, 1:19AM), https://twitter.com/edleonklinger/status/1284251420544372737.

[77] *See* Jervis, *supra* note 11 (using GPT-3 to generate requests for admission).

[78] *See GPT-3 – A Game Changer for Legal Tech?,* ARTIFICIAL LAWYER (Jul. 29, 2020), https://www.artificiallawyer.com/2020/07/29/gpt-3-a-game-changer-for-legal-tech/. But arguably OpenAI's control over the API precludes other companies from gaining a competitive advantage by using GPT-3. *See generally* Ben Dickson, *What It Takes to Create a GPT-3 Product,* VENTUREBEAT (Jan. 26, 2021), https://venturebeat.com/2021/01/26/what-it-takes-to-create-a-gpt-3-product/. For other applications of NLP in legal practice, see Brian S. Haney, *Applied Natural Language Processing for Law Practice*, 2020 B.C. INTELL. PROP. & TECH. F. 1 (2020); Robert Dale, *Law and Word Order: NLP in Legal Tech*, 25 NATURAL LANGUAGE ENG'G 211 (2019); Alarie et al., *supra* note 13, at 115–20; Ashley, *supra* note 74, at 1119.

[79] Language models can facilitate semantic search, i.e., search that uses the contextual meaning of search terms to identify a user's intent, rather than rely only on keywords. *See* R. Guha et al., *Semantic Search*, PROC. 12TH INT'L CONF. WORLD WIDE WEB 700 (2003).

[80] *See generally* William E. Forster & Andrew L. Lawson, *When to Praise the Machine: The Promise and Perils of Automated Transactional Drafting*, 69 S.C. L. REV. 597 (2018); Kathryn D. Betts & Kyle R. Jaep, The *Dawn of Fully Automated Contract Drafting: Machine Learning Breathes New Life into a Decades-Old Promise*, 15 DUKE L. & TECH. REV. 216 (2017).

[81] *See generally* Remus & Levy, *supra* note 13, at 551–52; Drew Simshaw, *Ethical Issues in Robo-Lawyering: The Need for Guidance on Developing and Using Artificial Intelligence in the Practice of Law*, 70 HASTINGS L.J. 173, 179–83 (2019).

[82] *See* Arbel & Becher, *supra* note 17, at 10–22 (showing that GPT-3 can simplify, personalize, interpret, and benchmark contracts).

[83] *See generally* DEBORAH L. RHODE, ACCESS TO JUSTICE 97, ch. 5 (2004) (discussing the legal needs of low-income communities). For explanations of the high cost of legal services, see Gillian K. Hadfield, *The Cost of Law: Promoting Access to Justice Through the (Un)Corporate Practice of Law*, 38 INT'L REV. L. & ECON. 43, 48–49 (2014); Albert H. Yoon, *The Post-Modern Lawyer: Technology and the Democratization of Legal Representation*, 66 U. TORONTO L.J. 456, 458–60 (2016).

even play a role in the justice system. For instance, a language model trained to evaluate the strengths and weaknesses of legal arguments could be used to assist *pro se* litigants in assessing the merits of their case before going to court.

Despite these opportunities, the performance of GPT-3 on legal tasks has not been rigorously tested. Many examples of its performance are demonstrations only. Although impressive, these may be subject to selection bias, i.e., cherry-picking instances of impressive performance.[84] At the same time, the findings in more systematic studies are equivocal. For example, one study that evaluated GPT-3 on a range of multiple-choice tests found that performance on bar exam questions was scarcely above random chance, while performance on international law and jurisprudence exams was exceptionally high.[85]

Turning to GPT-3's ability to understand legal texts, it is worth noting that the model performed poorly on general-purpose reading comprehension tasks.[86] However, there are notable differences between the language used in those tasks and legal language.[87] While we might assume that legal language is more technical or verbose—and that, therefore, GPT-3 will perform worse on legal texts—given GPT-3's unconventional method of learning,[88] it is problematic to make this assumption. To evaluate the degree to which the model can understand

[84] Arbel & Becher, *supra* note 17, at 6, 28 (noting that the examples they cite are cherry-picked). *See also* Raphaël Millière (@raphamilliere), TWITTER (Jul. 31, 2020, 6:50 PM), https://twitter.com/raphamilliere/status/1289226960279764992. Others may have cherry-picked *problematic* outputs. *See, e.g.,* Gary Marcus & Ernst Davis, *Experiments Testing GPT-3's Ability at Commonsense Reasoning: Results*, DEPT. COMP. SCI., N.Y.U. (Aug. 2020), https://cs.nyu.edu/faculty/davise/papers/GPT3CompleteTests.html (excluding prompts that resulted in "reasonable" outputs during preliminary tests).

[85] *See* Dan Hendrycks et al., *Measuring Massive Multitask Language Understanding*, 9TH INT'L CONF. LEARNING REPRESENTATIONS 1, 6 (2021) [hereinafter Hendrycks et al., *Measuring Understanding*].

[86] *See* Brown et al., *supra* note 5, at 18.

[87] *See, e.g.,* DAVID MELLINKOFF, THE LANGUAGE OF THE LAW (1963); Mary Jane Morrison, *Excursions into the Nature of Legal Language*, 37 CLEV. ST. L. REV. 271, 274 (1989); PETER M. TIERSMA, LEGAL LANGUAGE pt. 2 (1999); RUPERT HAIGH, LEGAL ENGLISH (2018). The distinctive features of legal language present challenges for machine learning. *See, e.g.,* Lucia Zheng et al., *When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset of 53,000+ Legal Holdings,* PROC. 18TH INT'L CONF. AI & L. 159, 161 (2021) [hereinafter Zheng et al., *CaseHOLD*]; Nguyen Ha Thanh & Nguyen Le Minh, *Sublanguage: A Serious Issue Affects Pretrained Models in Legal Domain,* ARXIV (Apr. 15, 2021), https://arxiv.org/abs/2104.07782.

[88] *See* Hendrycks et al., *Measuring Understanding, supra* note 85, at 7 ("GPT-3 acquires knowledge quite unlike humans. For example, GPT-3 learns about topics in a pedagogically unusual order. GPT-3 does better on College Medicine (47.4%) and College Mathematics (35.0%) than calculation-heavy Elementary Mathematics (29.9%). GPT-3 demonstrates unusual breadth, but it does not master a single subject. Meanwhile we suspect humans have mastery in several subjects but not as much breadth. . . . GPT-3 has many knowledge blindspots and has capabilities that are lopsided.")

legal texts, we need to test the model on legal texts.

II. EXPERIMENTAL DESIGN

The following Part outlines the methodology employed in the case study. I begin by describing the tasks on which GPT-3 was tested. Next, I explain the criteria used to evaluate performance. Finally, I discuss several notable challenges and limitations.

A. Task Description

In the field of natural language processing (NLP), evaluating a model's performance on real-world applications is instructive.[89] For example, testing whether GPT-3 can explain the meaning of contractual provisions sheds lights on the degree to which the model understands consumer contracts.[90] The problem, however, is that it is difficult to objectively evaluate responses to open-ended questions. What makes one explanation of a provision "better" than another (where, for the sake of argument, both are accurate)?[91] Unlike the fact-based trivia questions commonly used in NLP benchmark datasets, there is not necessarily a single "correct" answer to questions of legal analysis.[92] Even if there are specific criteria for measuring the quality of responses, different people (or AI systems) bring different perspectives and may reach different conclusions.[93]

[89] This is known as *extrinsic evaluation*. *See* JURAFSKY & MARTIN, *supra* note 27, at 35.

[90] *See* Arbel & Becher, *supra* note 17, at 10–19.

[91] *But see* Federico Ruggeri et al., *Detecting and Explaining Unfairness in Consumer Contracts through Memory Networks*, AI & L. (forthcoming 2021) (proposing a method for linking classifications of contractual clauses as unfair to corresponding legal rationale).

[92] *See generally* H.L.A. HART, THE CONCEPT OF LAW 124–25 (1961) (discussing the "open texture" of legal rules and language). *Compare* Ronald Dworkin, *No Right Answer?*, 53 N.Y.U. L. REV. 1 (1978); RONALD DWORKIN, A MATTER OF PRINCIPLE 119–45 (1985) (arguing in favor of legal determinacy). *See also* Brian Bix, *H.L.A. Hart and the "Open Texture" of Language*, 10 LAW & PHIL. 51 (1991); BRIAN BIX, LAW, LANGUAGE, AND LEGAL DETERMINACY ch. 4 (1995). For the impact of these ideas on machine learning, see Reuben Binns, *Analogies and Disanalogies between Machine-Driven and Human-Driven Legal Judgement*, 1 J. CROSS-DISCIPLINARY RES. COMPUTATIONAL L. 1, 7–8 (2021).

[93] This can partly be explained by the inherent vagueness and ambiguity of contractual language. *See, e.g.,* Lawrence M. Solan, *Pernicious Ambiguity in Contracts and Statutes*, 79 CHI.-KENT L. REV. 859, 861–63 (2004); E. Allan Farnsworth, *"Meaning" in the Law of Contracts*, 76 YALE L.J. 939, 952–57 (1967); George C. Christie, *Vagueness and Legal Language*, 48 MINN. L. REV. 885, 885–86 (1964).

In light of these challenges, legal AI systems are often evaluated using yes/no questions with relatively uncontroversial answers. For example, the annual Competition on Legal Information Extraction and Entailment (COLIEE) includes yes/no bar exam questions.[94] Similarly, a contract analytics firm tested an AI system's ability to understand commercial agreements by asking it yes/no questions about those agreements.[95]

The case study presented in this Article adopts a similar method. It tests GPT-3 on yes/no questions relating to consumer contracts. Although the questions aim to be as objective as possible, contract interpretation always involves a degree of subjective judgment.[96] Some might argue that the better answer to a certain question is "sometimes," "possibly," or "it depends" (rather than "yes" or "no"). But including these or other more nuanced responses in the test would face the same problems posed by using open-ended questions.

For the case study, I created a novel question set comprised of 200 yes/no questions relating to the terms of service of the 20 most-visited U.S. websites (10 questions per document).[97] The questions relate to a wide range of legal issues arising in the terms of service, including eligibility to access services, payment for services, limitations of liability, intellectual property rights, and dispute resolution procedures. Answers to all questions can be obtained from the applicable website terms of service. Table 1 displays a sample of the questions.[98]

[94] *See* Juliano Rabelo et al., *COLIEE 2020: Methods for Legal Document Retrieval and Entailment*, 14TH INT'L WORKSHOP ON JURIS-INFORMATICS (2020).

[95] *See* Radha Chitta & Alexander K. Hudek, *A Reliable and Accurate Multiple Choice Question Answering System for Due Diligence*, PROC. 17TH INT'L CONF. AI & L. 184 (2019).

[96] *See supra* note 92 (discussing the open texture and indeterminacy of law).

[97] According to the Alexa rankings, the 20 most-visited U.S. websites as of November 17, 2020 are, in descending order: Google.com, Youtube.com, Amazon.com, Facebook.com, Yahoo.com, Zoom.us, Reddit.com, Wikipedia.org, Myshopify.com, eBay.com, Office.com, Instructure.com, Netflix.com, CNN.com, Bing.com, Live.com, Microsoft.com, NYtimes.com, Twitch.tv, and Apple.com. *See Top Sites in United States*, ALEXA, https://www.alexa.com/topsites/countries/US. Because the terms of service for Live.com and Microsoft.com are the same as the terms of service for Office.com, I instead used the terms of service of Instagram.com and ESPN.com, which are the 21st and 23rd most-visited websites, respectively. (The 22nd most-visited website is Microsoftonline.com, the terms of service of which are the same as for Microsoft.com.) The companies referred to in the relevant terms of service are, in some instances, parent companies. For example, the terms of service for Yahoo.com and ESPN.com are Verizon and Disney, respectively. All terms of service were accessed during November 10–17, 2020, copies of which are on file with the author.

[98] The full list of questions used in the case study can be found in the Online Appendix.

Table 1: Sample of Questions

Question	Correct Answer
Will Google always allow me to transfer my content out of my Google account?	No
Does Amazon sometimes give a refund even if a customer hasn't returned the item they purchased?	Yes
Can I sue Zoom in a small claims court?	Yes
Is the length of the billing cycle period the same for all Netflix subscribers?	No
Do I need to use my real name to open an Instagram account?	No

B. Evaluation Criteria

1. Accuracy

The study reports the percentage of yes/no questions that GPT-3 answered correctly and compares this against three baselines. The first baseline is *random chance*. Random guessing yields, on average, 50% accuracy. The second baseline is the *majority class*. The correct answer to 55% of the questions in the case study is "no"; the correct answer to 45% of the questions is "yes." Responding with the majority class ("no") to every question yields the majority class baseline, i.e., 55% accuracy. The third baseline—which I call *contract withheld*—involves querying GPT-3 on the questions without displaying the contract excerpts, i.e., testing the model on all 200 questions while withholding the corresponding website terms of service. If accuracy is not higher when GPT-3 is shown both the contract and the question (compared with when it is shown only the question), then the model would fail to demonstrate that it understands the contracts. Instead, GPT-3 could simply be responding to cues in the questions and/or relying on its training data.[99]

2. Calibration

While high accuracy is necessary for strong performance, it is not sufficient. For a model to be reliable it must also be *well-calibrated*, i.e., it should assign high probabilities to its correct predictions and low probabilities to its incorrect predictions.[100] In other words, there should

[99] *See infra* Part III.C (discussing the memorization of training data).

[100] Put differently, a model is well-calibrated if its confidence in a prediction

be a strong correlation between its confidence and its competence. Well-calibrated models can also achieve higher accuracy if predictions below a certain confidence threshold are discarded, and only predictions whose confidence exceeds that threshold are used. Filtering the predictions of a well-calibrated model in this way separates the wheat from the chaff; the remaining predictions are, on average, more accurate.

As explained, GPT-3 operates by predicting the next word in a sequence.[101] It assigns a probability to (what it calculates to be) the most likely word. For example, following a certain yes/no question, GPT-3 might assign a 43% probability that the next word is "yes," a 29% probability that the next word is "no," and assign the remaining probability (summing to a total of 100%) to various other words. Then, if for example the highest probability is assigned to "yes," GPT-3 will output "yes."

In order to assess GPT-3's calibration, it is not enough to measure only the probability assigned to the model's output (i.e., the word assigned the highest probability). It is also important to measure the probability assigned to the alternative answer (i.e., the word assigned the second highest probability). Compare the following cases:

Case 1: GPT-3 assigns "yes" a 43% probability and "no" a 29% probability.

Case 2: GPT-3 assigns "yes" a 43% probability and "no" a 42% probability.

Despite the same probability being assigned to the output in both cases (43%), GPT-3 is surely less confident in Case 2—as the difference between the two probabilities is only 1 percentage point (43% minus 42%), as opposed to 14 percentage points in the Case 1 (43% minus 29%). Consequently, in addition to reporting the probability assigned to the output, the study also reports the *difference* between (i) the probability assigned to the output, and (ii) the probability assigned to alternative answer. Accordingly, in Case 1 the confidence score would equal 14 and in Case 2 the confidence score would equal 1. This measure better captures the considerable difference in confidence between the two cases.

Yet, measuring the difference between probabilities also has some shortcomings. To illustrate this, consider another pair of cases:

Case 3: GPT-3 assigns "yes" a 35% probability and "no" a 30% probability.

Case 4: GPT-3 assigns "yes" a 15% probability and "no" a 10% probability.

(expressed as a probability) is a good estimate of the actual probability that the prediction is correct. *See generally* Chuan Guo et al., *On Calibration of Modern Neural Networks*, PROC. 34TH INT'L CONF. MACH. LEARNING 1321 (2017).

[101] Technically, predictions are of tokens (not words) and probabilities are log probabilities (not raw probabilities).

Measuring the difference between probabilities within each case would treat Case 3 and Case 4 as equivalent (the difference between probabilities in each case is 5 percentage points). But the two cases are not fully equivalent. The *relative* difference between the respective probabilities is different. To express this difference, the study also reports the *ratio* between (i) the probability assigned to the output, and (ii) the probability assigned to the alternative answer. Accordingly, in Case 3 the confidence score would be 1.17 (35/30) and in Case 4 the confidence score would be 1.50 (15/10). Whether or not this measure better describes the model's confidence is open to debate.

To accommodate different perspectives, and improve robustness, the study reports all three measures of confidence: (i) the probability assigned to the output (*Measure 1*); (ii) the difference between the probability assigned to the output and the probability assigned to the alternative answer (*Measure 2*); and (iii) the ratio between the probability assigned to the output and the probability assigned to the alternative answer (*Measure 3).* Then, by measuring the correlation between accuracy and confidence, we are able to evaluate GPT-3's calibration. If the correlation between accuracy and confidence is positive, it would suggest that GPT-3 is generally more confident in its correct responses than in its incorrect responses, i.e., that GPT-3 is well-calibrated and thus more reliable. The converse is also true.

3. Overall Performance

To assess overall performance, we need a score that accounts for both accuracy and calibration. This can be calculated by multiplying the sign of accuracy (+1 for correct and -1 for incorrect) by the confidence score. For example, if GPT-3 answers a question correctly and exhibits a confidence score of 28, the overall performance score for that question would be +28. Alternatively, if GPT-3 answers a question incorrectly and exhibits a confidence score of 28, the overall performance score would be -28. Because there are three different measures of confidence, there will also be three measures of overall performance, corresponding to each of the measures of confidence. The overall performance scores are instructive. They reward high confidence correct answers (large positive scores) and penalize high confidence mistakes (large negative scores). As with accuracy, surpassing the *contract withheld* baseline would offer the best indication that the model can, at least to some degree, understand the contracts presented to it.

C. Challenges and Limitations

This Part discusses the main methodological challenges facing the case study, as well as the steps taken to confront these challenges. In addition, it highlights several limitations and opportunities for future work.

1. Challenges

One common concern with using pre-existing tests to evaluate language models trained on vast internet corpora is *question answer contamination*, i.e., the risk that a model has already seen the answers to the test questions.[102] For example, if the answers to certain bar exam questions are available on a website, and that website is included in a language model's training data, then the model may "memorize" the answers to those questions.[103] Testing the model's performance on those questions could misrepresent the model's actual abilities. To address this concern, all questions in the case study were newly prepared and do not appear in GPT-3's training data.

Another challenge in evaluating AI systems is that their performance can change, and hopefully improve, as people interact with them. For example, if multiple questions were presented to GPT-3 in a continuous dialogue, then the earlier questions (and corresponding responses) would comprise part of the prompt for later questions and thereby affect the model's responses to those questions. To tackle this concern, all questions in the case study were presented as standalone prompts (and not a continuous dialogue), such that performance on each question was independent of performance on other questions.

A further challenge is that there is some randomness in the outputs of neural language models.[104] For instance, it is possible that if presented with a particular yes/no question on two occasions, GPT-3 will answer "yes" on one occasion and "no" on another, which would undermine the replicability of any test. Fortunately, there is a straightforward solution. The degree of randomness in a model's predictions can be controlled using a hyperparameter called *temperature*.[105] In simple terms, the lower the temperature, the more confident a model will be in its predictions, resulting in more conservative predictions; the higher the temperature,

[102] *See* Brown et al., *supra* note 5, at 29–33, 43–44 (investigating whether GPT-3's performance on certain benchmarks was contaminated by its training data).

[103] *See infra* Part III.C (discussing the memorization of training data).

[104] The technical term is *stochasticity*.

[105] *See* Geoffrey Hinton et al., *Distilling the Knowledge in a Neural Network*, ARXIV (Mar. 9, 2015), https://arxiv.org/abs/1503.02531 (introducing model distillation, the machine learning technique in which temperature was first used).

the more "excited" a model will be, resulting in more diverse and "adventurous" predictions. In the case study, GPT-3's temperature was set to zero, which reduces randomness in the model's predictions and, thereby, improves replicability.[106]

Finally, some demonstrations of GPT-3's capabilities have not been especially transparent. For example, it is not always clear how many different prompts a user tested before achieving the desired output, or which hyperparameters they used. The case study takes several steps to improve transparency. First, all questions presented to GPT-3 are listed in the Online Appendix.[107] Second, the entire priming prompt is disclosed in the Appendix. Third, the hyperparameters were held constant across all questions, as detailed in Table 6 in the Appendix. Fourth, each question was asked only once. No re-sampling took place.

2. Limitations

Despite addressing the above challenges, the case study has several notable limitations. To begin with, the case study evaluates only the *behavior* of GPT-3, that is, the model's observable outputs.[108] There is a lively debate in the computer science and linguistics communities regarding whether GPT-3, or indeed any language model, can understand language in manner analogous to how humans understand language.[109] This important debate is beyond the scope of this Article.

[106] *See* Benn Mann, *How to Sample from Language Models*, TOWARDS DATA SCIENCE (May 25, 2019), https://towardsdatascience.com/how-to-sample-from-language-models-682bceb97277 (explaining that setting temperature to zero is equivalent to argmax sampling, i.e., maximum likelihood sampling).

[107] To be made available.

[108] Anthropomorphic references to language models in this Article, such as "understand" and "memorize," are used only by way of analogy, and should not be interpreted as suggesting that language models are agentive. *See* Melanie Mitchell, *Why AI is Harder Than We Think*, ARXIV at 5 (Apr. 26, 2021), https://arxiv.org/abs/2104.12871.

[109] *See* Emily M. Bender & Alexander Koller, *Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data*, PROC. 58TH ANN. MEETING ASS'N COMPUTATIONAL LINGUISTICS 5185 (2020) (contending that language models trained only on "form", such as text or pixels, cannot learn or understand meaning); Yonatan Bisk et al., *Experience Grounds Language*, PROC. 2020 CONF. EMPIRICAL METHODS IN NLP 8718 (2020) (suggesting that broader physical and social context is necessary for language models to genuinely understand language); Marcus & Davis, *GPT-3, Bloviator*, *supra* note 49; Gary Marcus & Ernest Davis, *Insights for AI from the Human Mind*, 64 COMM. ACM 38, 39 (2021); Shannon Vallor, *GPT-3 and the Missing Labor of Understanding*, DAILY NOUS (Jul. 30, 2020), https://dailynous.com/2020/07/30/philosophers-gpt-3/#vallor (arguing that GPT-3 does not exhibit understanding); Lynch, *supra* note 54 (Oren Etzioni positing that language models do not achieve understanding); Gary Marcus, *GPT-2 and the Nature of Intelligence*, THE GRADIENT (Jan. 25, 2020), https://thegradient.pub/gpt2-and-the-nature-of-intelligence/ (contenting that prediction should not be equated with understanding). *Compare* William Merrill et al., *Provable Limitations of Acquiring Meaning from Ungrounded Form: What Will Future Language Models Understand?*,

Next, the sample size in the case study (200 questions) is small compared with NLP benchmark datasets for general-purpose question answering, which typically contain thousands of questions.[110] That being said, NLP datasets in the legal domain are often smaller, containing fewer than 200 questions.[111] Seen in this light, the sample size in the case study is not especially problematic. Nevertheless, we should aspire to create larger legal datasets in the future.[112]

A further limitation is that the case study does not include unanswerable questions, i.e., questions for which there is no answer or the answer to which cannot be found in the corresponding document. Responding inappropriately to such questions would cast doubt on a model's reliability. While some general-purpose NLP benchmark datasets include unanswerable questions,[113] many benchmarks in the legal domain do not.[114] Of course, expanding legal datasets to include unanswerable questions would be a worthwhile project.

Finally, the case study has a narrow objective: to examine whether GPT-3 can answer a certain type of question relating to a certain type of contract. The study does not aim to test the model's broader legal knowledge or its performance on other legal tasks. Likewise, the case study does not attempt to compare the performance of GPT-3 to the performance of human lawyers or examine how providers and consumers of legal services are likely to interact with these models in practice.[115]

ARXIV (Jun. 22, 2021), https://arxiv.org/abs/2104.10809; Christopher Potts, *Is It Possible for Language Models to Achieve Language Understanding?*, MEDIUM (Oct. 5, 2020), https://chrisgpotts.medium.com/is-it-possible-for-language-models-to-achieve-language-understanding-81df45082ee2.

[110] For example, the WebQuestions dataset consists of 6,642 questions and the TriviaQA dataset consists of over 650,000 questions. *See* Jonathan Berant et al., *Semantic Parsing on Freebase from Question-Answer Pairs*, PROC. 2013 CONF. EMPIRICAL METHODS IN NLP (2013); Mandar Joshi et al., *TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension*, PROC. 55TH ANN. MEETING ASS'N COMPUTATIONAL LINGUISTICS 1601 (2017).

[111] For example, the Jurisprudence and International Law tests used by Hendrycks et al. consist of 108 and 121 multiple-choice questions, respectively. *See* Dan Hendrycks, *Test,* GITHUB, https://github.com/hendrycks/test. Similarly, the 2020 COLIEE competition's legal textual entailment task contains 112 yes/no questions. *See* Rabelo et al., *supra* note 94, at 9–10.

[112] Recent efforts, which post-date the case study, include Hendrycks et al., *CUAD*, *supra* note 70; Zheng et al., *CaseHOLD*, *supra* note 87.

[113] *See, e.g.,* Pranav Rajpurkar et al., *Know What You Don't Know: Unanswerable Questions for SQuAD*, PROC. 56TH ANN. MEETING ASS'N COMPUTATIONAL LINGUISTICS 784 (2018) (introducing 50,000 unanswerable questions to an existing dataset).

[114] *See, e.g.,* Zheng et al., *CaseHOLD*, *supra* note 87; Lippi et al., *supra* note 69. *But see* Hendrycks et al., *CUAD*, *supra* note 70, at 5 (allowing certain questions to have no answers).

[115] *See generally* Kawin Ethayarajh & Dan Jurafsky, *Utility is in the Eye of the User: A Critique of NLP Leaderboards,* PROC. 2020 CONF. EMPIRICAL METHODS IN NLP 4846 (2020) (emphasizing the importance of real-world evaluation). But new benchmarks and

Future work will need to grapple with these important issues.

III. Results and Discussion

This Part presents the results of the case study and attempts to explain some of the key findings. First, I discuss the performance of GPT-3 on the test questions. Next, I analyze whether certain characteristics of the contracts and questions presented to GPT-3 are associated with an increase or decrease in performance. Finally, I evaluate how variations in question wording can impact performance.

A. Performance

1. Accuracy

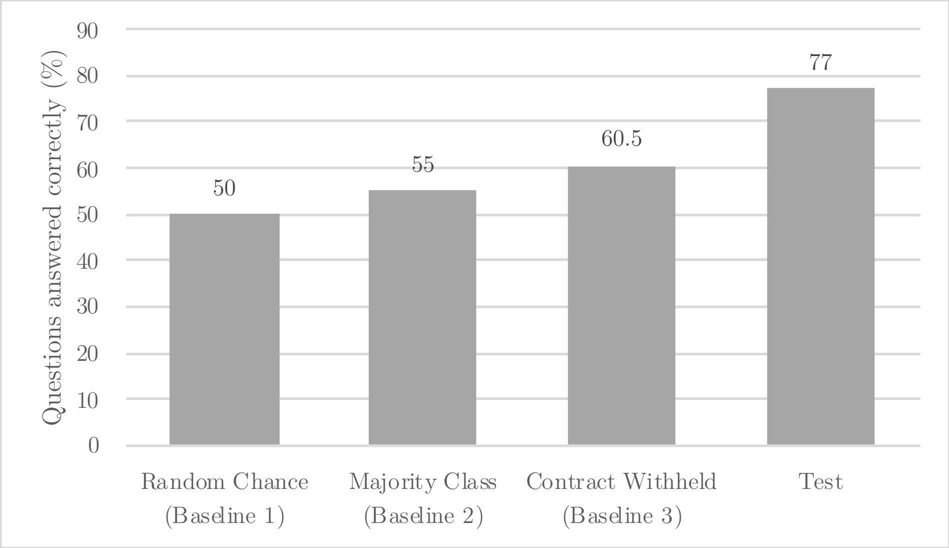
GPT-3 answered correctly 77% of the questions in the case study.[116] In terms of accuracy, performance exceeded all three baselines, as illustrated in Figure 1. That is, performance in the test was better than (i) random chance (randomly guessing answers); (ii) the majority class (answering "no" to all questions); and (iii) the contract withheld baseline (responding to questions without being shown the contract excerpts). Beating this final baseline by 16.5 percentage points indicates that performance was considerably better when GPT-3 was shown the contract excerpt, compared with when GPT-3 was not shown the contract excerpt. This result suggests that GPT-3 uses the contract to answer the questions and does not simply rely on cues in the questions or on data memorized during pretraining.[117]

evaluation platforms are being developed to better simulate real-world conditions. *See, e.g.,* Douwe Kiela et al., *Dynabench: Rethinking Benchmarking in NLP*, PROC. 2021 ANN. CONF. N. AM. CH. ASS'N COMPUTATIONAL LINGUISTICS 4110 (2021). [*See also* Marco Tulio Ribeiro et al., *Beyond Accuracy: Behavioral Testing of NLP Models with CheckList*, PROC. 58TH ANN. MEETING ASS'N COMPUTATIONAL LINGUISTICS 4902 (2020) (proposing a comprehensive framework for testing the real-world performance of language models).]

[116] GPT-3 did not provide a yes/no response to four questions and, instead, outputted the name of the relevant company. Given these responses fail to answer the question, the study omits these responses and reports the word assigned the second highest probability—"yes" or "no", which may be either correct or incorrect, as the case may be—and the corresponding probability. Notably, a similar filter would be applied if GPT-3 were deployed in practice: non-yes/no answers would be discarded and the response assigned the next-highest probability that actually answers the question (i.e., "yes" or "no") would be retained.

[117] *See infra* Part III.C (discussing the memorization of training data).

Figure 1: Comparison of Accuracy with Baselines



2. Calibration

In terms of calibration, there was a positive correlation between the model's accuracy and the model's confidence in its predictions.[118] That is, on average GPT-3 was more confident in its correct responses than in its incorrect responses. This result suggests that GPT-3's performance in the test was well-calibrated and, accordingly, encourages us to trust its predictions. The complete calibration results are shown in Figures 4A, 4B and 4C in the Appendix.

3. Overall Performance

Combining accuracy and calibration, average overall performance in the test exceeded average overall performance in the contract withheld baseline, across all three measures of overall performance.[119] Surpassing the contract withheld baseline in terms of overall performance provides further suggestive evidence that GPT-3 uses the contracts to answer the questions.

* * *

[118] The correlation coefficients (Pearson's r) between accuracy and the measures of confidence (described in Part II.B) are, respectively: $r = 0.226$** (Measure 1); $r = 0.258$*** (Measure 2); and $r = 0.205$** (Measure 3), where * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

[119] *See* Table 8A in the Appendix. *See also supra* Part II.B (explaining how overall performance is calculated).

The performance of GPT-3 in the case study appears encouraging. Despite the unintuitive and unhuman-like way in which language models operate—predicting the next word in a sequence—GPT-3 was able to answer correctly nearly four of five questions, and was generally well-calibrated. The results suggest that, contrary to conventional wisdom,[120] an NLP system can answer questions about contracts without directly extracting information from contracts. In addition, GPT-3's strong performance in the case study challenges the assumption that pretrained language models must be fine-tuned on legal data in order to effectively carry out legal tasks.[121]

B. Anti-Consumer Bias

But do these indications of strong performance apply equally to *all* questions in the case study? Or did GPT-3 perform better on some questions than on others?

The provisions in the terms of service in the case study can be categorized as follows. First, some provisions are *pro-company*, i.e., they favor the rights and interests of the relevant companies. Examples include provisions that exempt a company from liability, grant a company the right to refrain from assisting consumers, or enable a company to take certain actions without consumer consent. Second, some provisions are *pro-consumer*, i.e., they favor the rights and interests of consumers. Examples include provisions that grant consumers rights or protections, obligate a company to seek consumer consent in order to take certain actions, or require that a company provide notice to consumers. Third, some provisions are *neutral*, i.e., they do not favor either companies or consumers. Examples include provisions that stipulate eligibility requirements for accessing a service (e.g., age of user) or describe payment process (e.g., length of billing cycle), or provisions that do not explicitly favor either side (e.g., severability clauses).[122]

[120] *See, e.g.,* Ryan Catterwell, *Automation in Contract Interpretation*, 12 L. INNOVATION & TECH. 81, 100 (2020); Ashley, *supra* note 74, at 1135–37. *See also* JURAFSKY & MARTIN, *supra* note 27, at 464–65 (describing traditional NLP question answering systems, including information retrieval systems), 484–85 (describing the use of language models for question answering).

[121] *See, e.g.,* Hendrycks et al., *Measuring Understanding*, *supra* note 85, at 8; Zheng et al., *CaseHOLD*, *supra* note 87, at 167. GPT-3's training data, however, are likely to include many website terms of service, which are precisely the kind of legal document used in the case study.

[122] This categorization draws on principles from existing consumer contract classification schemes. *See, e.g.,* Florencia Marotta-Wurgler, *What's in a Standard Form Contract? An Empirical Analysis of Software License Agreements*, 4 J. EMPIRICAL LEGAL STUD. 677, 689–702 (2007) (proposing a "bias index" for end-user software license agreements); Lippi et al., *supra* note 69, at 121–27 (proposing a classification scheme based on EU consumer law). *See also* Guarino et al., *supra* note 69, at 6–9 (expanding the

Correspondingly, questions relating to pro-company provisions, pro-consumer provisions, and neutral provisions can be considered *pro-company questions*, *pro-consumer questions*, and *neutral questions*, respectively. In the case study, there are 110 pro-company questions (55%), 45 pro-consumer questions (22.5%), and 45 neutral questions (22.5%). Table 2 provides an example from each category.[123]

Table 2: Sample of Question Categories

Category	Contract Provision and Question	Correct Answer
Pro-Company	"The Service may contain links to third-party websites and online services that are not owned or controlled by YouTube. YouTube has no control over, and assumes no responsibility for, such websites and online services." *Does Youtube take responsibility for links to third party websites on Youtube?*	No
Pro-Consumer	"In no event, however, will you be charged for access to the Services unless we obtain your prior agreement to pay such charges." *Can NYT [New York Times] ever charge me without my consent?*	No
Neutral	"Unless you are the holder of an existing account in the United States that is a Yahoo Family Account, you must be at least the Minimum Age to use the Services." *If I'm below the minimum age but have a US Yahoo Family Account, can I use the services?*	Yes

The first provision in Table 2 is classified as *pro-company* because it shields the company from liability. The second provision is classified as *pro-consumer* because it protects consumers' interests by requiring their consent to payments. The third provision is classified as *neutral* because

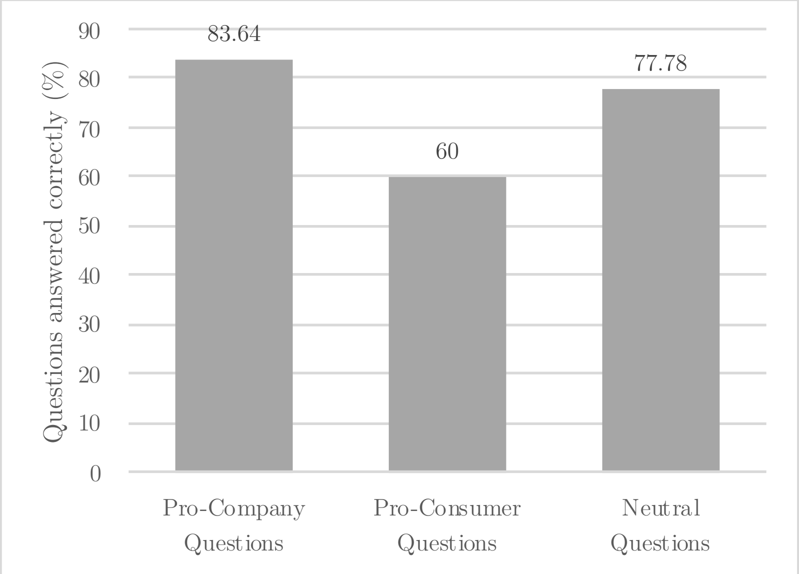
latter classification scheme). Importantly, classification invariably involves a degree of subjective judgment. For example, some might argue that a provision that grants a company the right to perform an action that could benefit a consumer (such as backing up personal data) without that consumer's consent should be classified as pro-consumer, not pro-company. Others might argue that the appropriate classification depends on the facts of the particular situation and individual consumer's personal preferences. However, for a classification scheme to be effective, it must avoid such open-ended categories.

[123] The contract excerpts in Table 2 are for demonstrative purposes only, and were extracted from the longer excerpts that were presented to the model in the case study. *See infra* Appendix pt. A.2.

it does not favor the interests of either the company or the consumer; it simply stipulates who may access the services.

How did GPT-3 perform across the different question categories? As illustrated in Figure 2, the model's accuracy on pro-company questions exceeded its accuracy on pro-consumer questions by approximately 24 percentage points. Meanwhile, accuracy on pro-consumer questions was approximately 18 percentage points lower than on neutral questions.[124]

Figure 2: Comparison of Accuracy across Question Categories



There was also a considerable disparity in calibration across the question categories. While GPT-3 was generally well-calibrated (i.e., on average, it expressed higher confidence in questions that it answered correctly),[125] this was not the case for pro-consumer questions.[126] Moreover, a disproportionately large number of questions with the poorest overall performance scores—that is, where GPT-3 answered incorrectly and with high confidence—were pro-consumer questions.[127] A sample of questions yielding high confidence anti-consumer mistakes is shown in Table 3.

[124] These results are consistent with the overall performance scores (that combine both accuracy and calibration), which are listed in Table 8B in the Appendix.

[125] *See supra* Part III.A.

[126] None of the correlation coefficients (Pearson's *r*) between accuracy and the measures of confidence (described in Part II.B) was found to be statistically significant.

[127] Six of the ten (60%) questions with the poorest overall performance scores (across all three measures of overall performance) were pro-consumer questions, despite the fact that pro-consumer questions comprise only 22.5% of the question set.

Table 3: Sample of High Confidence Anti-Consumer Mistakes

Question	Model's Output	Correct Answer
Are there any potential exceptions which would allow me to copy a Disney product?	No	Yes
Can Instructure back up my data without asking me?	Yes	No
Will Google help me if I think someone has taken and used content I've created without my permission?	No	Yes

To make these findings more concrete, consider the first question in Table 3: "Are there any potential exceptions which would allow me to copy a Disney product?" The correct (pro-consumer) answer according to the terms of service is "yes." The terms explicitly state that if Disney provides consent, then a consumer is permitted to copy a Disney product.[128] But GPT-3 answered "no," suggesting that a consumer *never* has the right to copy a Disney product. In other words, GPT-3 provided an anti-consumer (or pro-company) response, misrepresenting a contractual provision that has the potential to favor consumers.

Despite these notable findings, simply comparing performance across the question categories might not tell the whole story. It is possible that pro-consumer questions are systematically different to other questions. For example, perhaps pro-consumer provisions are longer or more complex than pro-company provisions, making them more difficult to answer, and thus leading to poorer performance. Or perhaps pro-company questions borrow more substantially from the language of the corresponding contract, making it easier to locate the answer, and thus leading to better performance.

To test whether there is in fact a relationship between performance and question category (the variable of interest), we need to control for other factors that could potentially influence performance. Accordingly, I employ an ordinary least squares (OLS) regression model, regressing performance (the dependent variable) on several characteristics of the questions and contracts, including the variable of interest (the independent variables). If the variable of interest has an independent effect on performance, then we would expect to see a significant relationship between the variable of interest and performance, even after controlling for other variables.

[128] Of course, in reality, such consent might not be especially forthcoming.

The regression analysis controls for the following variables:

(i) *Company Name in Question*: This variable describes whether the name of the relevant company[129] appears in the question. The rationale for including this variable is that the appearance of the company's name in a question may provide GPT-3 with a cue to recall information relating to that company that is contained in the model's training data, thereby improving performance.[130]

(ii) *Length of Contract*: This variable describes the length of the contract excerpt shown to GPT-3.[131] The rationale for including this variable is that, due to the problem of long-range dependencies, where the text presented to the model is longer, the model may be more likely to "forget" content contained earlier in the text, resulting in poorer performance.[132]

(iii) *Readability of Contract*: This variable describes the ease with which a human reader can understand the contract excerpt.[133] The rationale for including this variable is that GPT-3 may be expected to perform worse on texts that are more difficult for humans to read and understand.

(iv) *Similarity between Contract and Question*: This variable describes the degree to which the language in a question is similar to the language in the corresponding contract excerpt.[134] The rationale for

[129] That is, the company whose terms of service the question relates to. However, company name also includes products and services that are clearly identified with a particular company, such as Wikipedia (Wikimedia) and Xbox (Microsoft).

[130] *See infra* Part III.C (discussing the memorization of training data).

[131] That is, the total length of the contract excerpt presented to GPT-3, which is measured in characters (including spaces). In the regression, length was divided by 100 in order to avoid producing very small coefficients. Similar results are observed if we measure the distance between the end of the question (at the end of the prompt) and the part of the contract excerpt containing the information needed to answer the question.

[132] *See supra* Part I.A (discussing long-range dependencies).

[133] That is, the Flesch Reading Ease score for the contract excerpt, which is calculated as follows: 206.835 − (1.015 × ASL) − (84.6 × ASW), where ASL is the average sentence length and ASW is the average word length in syllables. Similar results are observed if we use the FORCAST Grade Level score, which is especially appropriate for non-prose texts (such as terms of service), and is calculated as follows: 20 − (N / 10), where N is the number of single-syllable words in a 150-word sample. The papers introducing the Flesch Reading Ease and FORCAST scores, respectively, are Rudolph Flesch, *A New Readability Yardstick*, 32 J. APPL. PSYCHOL. 221, 229 (1948); John S. Caylor et al., *Methodologies for Determining Reading Requirements of Military Occupational Specialties* 15 (Human Resources Research Organization, Technical Report No. 73-5, 1973).

[134] Measuring similarity involved three steps: (i) The question text and the part of the contract containing the information needed to answer the question were preprocessed by converting all characters to lowercase, removing punctuation, splitting the text into individual words, removing morphological affixes, and removing stop words. (ii) The resulting texts were then converted into vectors using term frequency-inverse document frequency (TF-IDF). (iii) Similarity was calculated by measuring the cosine of the angle between the vector representing the question and the vector representing the contract.

including this variable is that where there is considerable overlap in language between the question and the relevant part of the contract, GPT-3 might be expected to more successfully utilize information contained in the contract, thereby improving performance.[135]

Importantly, the regression only controls for these four variables. It is possible that regressing performance on additional variables could produce different results. This problem, known as *omitted variable bias*, affects all regression analyses and cannot be altogether avoided or dismissed.[136] Testing *every* plausible additional variable is beyond the scope of this Article and would introduce statistical problems.[137]

Table 4 displays the results of three specifications of the OLS model, regressing the three measures of overall performance on the above variables.[138] All three specifications indicate that there is a statistically significant negative correlation between performance and the classification of a question as pro-consumer. In other words, the regression analysis shows that, on average, GPT-3 performed worse on pro-consumer questions than on other questions.

Similar results are observed if we omit steps (ii) and (iii) and, instead, calculate the Jaccard similarity between the question and the contract, which measures the size of the intersection of the words in the two texts, divided by the size of the union of the words in the two texts.

[135] *See supra* note 120 (discussing traditional question answering systems, including information retrieval systems).

[136] *See* JAMES H. STOCK & MARK W. WATSON, INTRODUCTION TO ECONOMETRICS 211–16, 334–35 (4th ed. 2019).

[137] *See id.* at 516–18.

[138] *See supra* Part II.B (explaining how overall performance is calculated).

Table 4: Regression Analysis of Overall Performance

	Dependent Variable		
	Overall Performance (Measure 1)	Overall Performance (Measure 2)	Overall Performance (Measure 3)
Pro-Company Question	1.767	-0.775	-1.731*
	(5.018)	(3.393)	(0.769)
Pro-Consumer Question	-17.024**	-12.911**	-3.512***
	(5.938)	(4.015)	(0.910)
Company Name in Question	14.572*	9.564*	1.890*
	(5.974)	(4.040)	(0.916)
Length of Contract	-0.165	-0.099	-0.016
	(0.125)	(0.084)	(0.019)
Readability of Contract	0.080	-0.032	-0.013
	(0.170)	(0.115)	(0.026)
Similarity between Question and Contract	4.507	-1.830	-0.288
	(17.370)	(11.746)	(2.663)
Number of Observations	200	200	200
R^2	0.108	0.106	0.095

Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Pro-company question is a dummy variable equal to 1 if the question is pro-company. Pro-consumer question is a dummy variable equal to 1 if the question is pro-consumer. Neutral question is the omitted category (baseline).

There are several possible explanations for this result. One possibility is *bias in the model's training data.* GPT-3 might have performed worse on the pro-consumer questions because it replicates a systematic anti-consumer bias in its training data.[139] 82% of GPT-3's training data is

[139] Training data often contain biases that affect a language model's parameters. *See* Brown et al., *supra* note 5, at 36–39 (analyzing GPT-3's biases related to race, gender, and religion); Abubakar Abid et al., *Persistent Anti-Muslim Bias in Large Language Models*, arXiv (Jan. 18, 2021), https://arxiv.org/abs/2101.05783; Abubakar Abid et al., *Large Language Models Associate Muslims with Violence*, Nature Mach. Intell. (Jun. 17, 2021). For an overview, see Su Lin Blodgett et al., *Language (Technology) is Power: A Critical Survey of "Bias" in NLP*, Proc. 58th Ann. Meeting Ass'n Computational Linguistics 5454 (2020); Emily Sheng et al., *Societal Biases in Language Generation: Progress and Challenges*, Proc. 59th Ann. Meeting Ass'n Computational Linguistics

comprised of webpages extracted from the Common Crawl and Webtext2 datasets, which are likely to include many website terms of service and other consumer contracts.[140] These documents are typically drafted by company counsel and designed to favor the rights and interests of the relevant company, not consumers.[141] By performing worse on pro-consumer questions and producing a disproportionate number of anti-consumer responses, the model arguably overfits its training data.[142]

Another possibility is *bias in prompt-based learning*. GPT-3 might have performed worse on pro-consumer questions because of biases learned from the corresponding contract excerpt (i.e., the excerpt presented alongside a given question). Although the specific part of the contract excerpt needed to answer a pro-consumer question is (by definition) pro-consumer, the full contract excerpt presented to the model is, on the whole, likely to be pro-company.[143] This broader pro-company context—although not directly relevant to the question being asked—could inadvertently teach a model to provide incorrect, anti-consumer responses.

A further possibility is *bias in the questions*. GPT-3 might have performed worse on the pro-consumer questions because those questions and corresponding contract excerpts are systematically different to other questions and provisions in the case study. For example, perhaps the pro-consumer questions are legally or linguistically more complex than other questions. Alternatively, perhaps the pro-company provisions have been tested in litigation more often than pro-consumer provisions, resulting in

4275 (2021).

[140] For an analysis of certain aspects of the Common Crawl dataset, see Alexandra (Sasha) Luccioni & Joseph D. Viviano, *What's in the Box? A Preliminary Analysis of Undesirable Content in the Common Crawl Corpus*, ARXIV (May 31, 2021), https://arxiv.org/abs/2105.02732.

[141] *See* RADIN, *supra* note 1, at pt. 1; Omri Ben-Shahar, *Foreword to Boilerplate: Foundations of Market Contracts Symposium*, 104 MICH. L. REV. 821, 822 (2006) [hereinafter Ben-Shahar, *Boilerplate*] (discussing the one-sidedness of boilerplate contracts). *See also* Marotta-Wurgler, *What's in a Standard Form Contract?*, *supra* note 122, at 702–12 (finding that end-user software license agreements generally favor the interests of software companies).

[142] There is, however, some disagreement regarding whether replication of biases is *always* problematic. *See* Yoav Goldberg, *A Criticism of "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big"*, GITHUB (Jan. 23, 2021), https://gist.github.com/yoavg/9fc9be2f98b47c189a513573d902fb27 ("there are many good reasons to argue that a model of language use should reflect how the language is actually being used"). *Compare* Abeba Birhane & Vinay Uday Prabhu, *Large Image Datasets: A Pyrrhic Win for Computer Vision?*, PROC. IEEE/CVF WINTER CONF. APPLICATIONS OF COMPUT. VISION. 1537, 1541 (2021) ("[f]eeding AI systems on the world's beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy"). *See also* Anna Rogers, *Changing the World by Changing the Data*, PROC. 59TH ANN. MEETING ASS'N COMPUTATIONAL LINGUISTICS 2182, 2184–85 (2021).

[143] *See* RADIN, *supra* note 1, at pt. 1; Ben-Shahar, *Boilerplate*, *supra* note 141, at 822.

pro-company provisions using clearer, more accessible language than pro-consumer provisions.[144] It is difficult to measure, or control for, such differences.[145]

The case study cannot determine which (if any) of these explanations is correct. Nevertheless, the observed disparity in performance across the different question categories raises noteworthy issues and maps out avenues for future work. Identifying the source of anti-consumer biases—in training data, evaluation, and elsewhere—will be critical to improving the safety and reliability of language models in the legal domain.

C. Informational Cues

Another notable finding in the regression is that, on average, GPT-3 performed better on questions that explicitly contain the name of the relevant company.[146] A likely explanation is that the appearance of the company's name in a question provides GPT-3 with a cue to recall information regarding that company that was learned during pretraining. For example, the question "Does *Microsoft* undertake to inform users of all changes to the terms?" might prompt GPT-3 to recall information regarding Microsoft that is contained in the model's training data and stored in its parameters. GPT-3 is then able to use this information when answering a question that relates to Microsoft.

Although further testing is required to validate this explanation,[147] there is a body of research illustrating that language models "memorize" highly specific information during pretraining.[148] In the case of GPT-3, its training data is replete with information that may be relevant to

[144] I thank David Hoffman for suggesting this possibility. *Compare* Michelle E. Boardman, *Contra Proferentem: The Allure of Ambiguous Boilerplate*, 104 MICH. L. REV. 1105, 1111 (2006) (suggesting that judicial interpretation of contracts may in fact entrench *ambiguous* pro-company language). Note, however, that the regression did not find a statistically significant relationship between performance the classification of a question as pro-company.

[145] One reason is that readability scores are not reliable for short texts. *See* Thomas Oakland & Holly B. Lane, *Language, Reading, and Readability Formulas: Implications for Developing and Adapting Tests*, 4 INTL. J. TESTING 239, 245 (2004). Consequently, measurements of the readability of an individual question, or the specific part of a contract excerpt containing the answer to a question, are not reliable.

[146] *See supra* note 129 (describing the variable used in the regression analysis).

[147] *See generally* Zhengbao Jiang et al., *How Can We Know What Language Models Know?*, 8 TRANSACTIONS ASS'N COMPUTATIONAL LINGUISTICS 423, 423–25 (2020) (outlining the challenges involved in examining the knowledge contained in language models).

[148] *See* Nicholas Carlini et al., *Extracting Training Data from Large Language Models*, ARXIV (Dec. 14, 2020), https://arxiv.org/abs/2012.07805 (demonstrating that language models can memorize specific examples found in their training data); Vered Shwartz et al., *"You are Grounded!": Latent Name Artifacts in Pre-trained Language Models*, PROC. 2020 CONF. EMPIRICAL METHODS IN NLP 6850 (2020) (showing that memorization of training data can dramatically affect a model's predictions).

answering questions about consumer contracts. Specifically, the model's training data are likely to include many companies' terms of service, as well as other company-specific legal and business information.[149] It is therefore plausible that informational cues embedded in certain questions enable GPT-3 to "access" this information and achieve better performance on those questions.

But informational cues also offer a cautionary tale. If certain informational cues can improve performance, it is possible that other informational cues could cause performance to deteriorate or, worse still, subtly manipulate a model's outputs.[150] For example, perhaps companies could draft consumer contracts that language models cannot understand or systematically interpret in a manner that favors companies' interests.[151] Seen in this light, informational cues cut both ways. While they can potentially improve performance, informational cues also reinforce concerns that language models are disturbingly brittle.[152]

D. Brittleness

To further investigate the issue of brittleness, the case study tested GPT-3 on an alternative wording of all 200 test questions. The alternatively worded questions are, by design, less readable,[153] that is, more difficult for a human reader to understand.[154] The content of the questions is substantially the same, but their wording differs markedly. Table 5 depicts the original wording of a question (which is more readable) alongside the alternative wording of that question (which is less readable).

[149] Because GPT-3's training data are not publicly available we cannot ascertain precisely which documents they contain, let alone pinpoint the particular documents that assist the model in answering certain questions.

[150] *See generally* Moustafa Alzantot et al., *Generating Natural Language Adversarial Examples*, PROC. 2018 CONF. EMPIRICAL METHODS IN NLP 1890 (2018) (showing that language models are susceptible to adversarial attacks, i.e., imperceptible changes to model inputs designed to elicit incorrect or harmful responses); Keita Kurita et al., *Weight Poisoning Attacks on Pre-trained Models*, PROC. 58TH ANN. MEETING ASS'N COMPUTATIONAL LINGUISTICS 2793 (2020) (demonstrating that the outputs of language models can be manipulated by strategically injecting trigger words during pretraining).

[151] *See* Arbel & Becher, *supra* note 17, at 28–32, 47–48 (discussing potential adversarial attacks on language models in legal applications).

[152] *See, e.g.,* Robin Jia et al., *Certified Robustness to Adversarial Word Substitutions*, PROC. 2019 CONF. EMPIRICAL METHODS IN NLP 4129, 4129 (2020) (surveying studies that demonstrate the brittleness of language models).

[153] *See* Table 7 in the Appendix (listing the applicable readability scores).

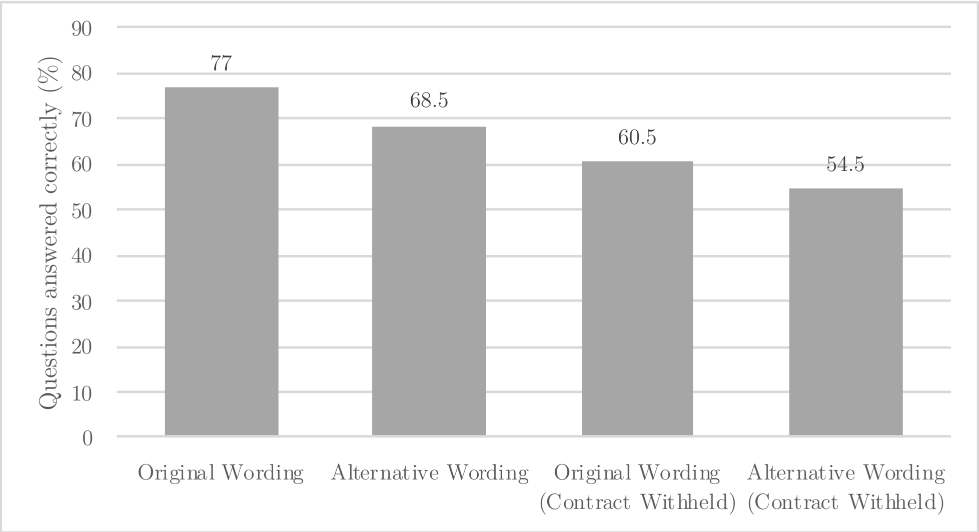
[154] *See, e.g., supra* note 133 (explaining how certain readability scores are calculated).

Table 5: Sample of Question Wordings

Original Wording (More Readable)	Alternative Wording (Less Readable)
Am I allowed to be paid for writing a Wikipedia article, assuming I disclose who's paying me?	Are Wikipedia contributors permitted to receive payment in respect of their contributions, provided they disclose the identity of the person or institution providing such payment?

In terms of accuracy, performance was nearly ten percentage points lower on the alternatively worded questions,[155] as illustrated in Figure 3.[156] (A smaller decrease in accuracy is observed in the corresponding contract withheld baselines.) These results suggest that GPT-3 is highly sensitive to the wording of questions, even if the substance of the questions is unchanged.

Figure 3: Comparison of Accuracy across Question Wordings



[155] These results are consistent with the overall performance scores (that combine both accuracy and calibration), which are listed in Table 8B in the Appendix. The observed disparity in performance is also consistent with previous studies demonstrating the sensitivity of language models to perturbations. *See, e.g.,* Jia et al., *supra* note 152.

[156] GPT-3 did not provide a yes/no response to seven of the alternatively worded questions. Given these responses fail to answer the question, the study omits these responses and reports the word assigned the second highest probability —"yes" or "no", which may be either correct or incorrect, as the case may be—and the corresponding probability. Notably, a similar filter would be applied if GPT-3 were deployed in practice: non-yes/no answers would be discarded and the response assigned the next-highest probability that actually answers the question (i.e., "yes" or "no") would be retained.

A related issue concerns whether GPT-3 is similarly sensitive to variations in the language of the contracts presented to it in the case study. To investigate this, the regression analysis controlled for the length and readability of the contract excerpts, as well as the similarity in language between the question and the corresponding contract excerpt. The regression did not find a statistically significant relationship between performance and any of these variables. Put simply, the analysis did not find that the contracts' length, readability, or similarity to the questions is associated with an increase or decrease in performance.

This result is surprising. Given the problem of long-range dependencies, one might assume that the longer the contract excerpt presented to GPT-3, the more likely the model is to "forget" content contained in earlier parts of the excerpt, which would result in poorer performance on longer excerpts. Similarly, one might expect GPT-3 to perform worse on contracts that humans find more difficult to understand. Finally, one might assume that greater overlap in language between the question and the contract would assist GPT-3 in understanding the contract. But none of these assumptions was borne out in the case study.[157]

On the one hand, this result is encouraging. It suggests that GPT-3 can cope well with longer and less readable texts (contracts are often long and less readable),[158] and does not require that the language of the question mirror the language of the contract in order to perform well.[159] On the other hand, given that performance is so sensitive to the wording of the questions, it is somewhat puzzling that performance does not appear to be at all sensitive to the language of the contracts.[160] One possible explanation is that GPT-3, like other language models, operates by predicting the next word in a sequence. The question is the final part of the prompt and, therefore, has an outsized impact on performance.[161]

[157] *See supra* Part III.B (discussing each of these assumptions).

[158] *See, e.g.,* Uri Benoliel & Shmuel I. Becher, *The Duty to Read the Unreadable*, 60 B.C.L. REV. 2255, 2270–2284 (2019); Shmuel I. Becher & Uri Benoliel, *Law in Books and Law in Action: The Readability of Privacy Policies and the GDPR*, *in* CONSUMER LAW & ECONOMICS 179, 191–200 (Klaus Mathis & Avishalom Tor eds., 2021). *See also* Alan M. White & Cathy Lesser Mansfield, *Literacy and Contract*, 13 STAN. L. & POL'Y REV. 233, 260 (2002).

[159] The finding that performance does not deteriorate on longer input texts is encouraging with respect to the prospect of using few-shot learning where the relevant examples of tasks are long, such as contracts and corresponding question-answer pairs. *But see infra* note 216 (showing that the model's context window constrains few-shot learning).

[160] However, as explained, performance does deteriorate when the contract is not shown to the model. *See supra* Part III.A (describing the contract withheld baseline).

[161] *See* Tony Z. Zhao et al., *Calibrate Before Use: Improving Few-Shot Performance of Language Models*, 38TH INT'L CONF. MACH. LEARNING at 4 (2021), https://arxiv.org/abs/2102.09690 (illustrating that content near the end of a prompt can

IV. BROADER IMPLICATIONS

Taken together, the results of the case study illustrate that language models present strengths and weaknesses in reading consumer contracts. Owing to its immense training data, GPT-3 can potentially draw on informational cues in questions to achieve relatively strong performance. At the same time, GPT-3 is very sensitive to how questions are worded and might contain an anti-consumer bias. These insights offer valuable guidance to various stakeholders and highlight some of the challenges facing the deployment of language models. Users of language models, developers of language models, and policymakers will need to work together to address these challenges and ensure that language models are used responsibly and align with broader social values.

A. Experimentation

The successful deployment of language models requires experimentation. At the very least, users should attempt to phrase questions in different ways. The case study suggests that simpler, more readable language elicits better performance. But we do not know if this finding generalizes to other contexts. In addition, the case study offers suggestive evidence that informational cues could improve performance. To test these and other hypotheses, users will need to present language models with different lexical and logical variations of questions and other prompts.

Experimentation, however, is onerous. Many users are unlikely to have the time or expertise required to rigorously test language models. For example, how many sample questions does a model need to see in order to learn the principles of contractual interpretation? Can prompts be rephrased to dampen the impact of a particular legal or societal bias? Clearly, users need guidance. The emerging field of *prompt design* aims to provide such guidance.[162] By systematically exploring methods to develop prompts that optimize performance, prompt design could help users leverage the benefits of language models and mitigate the associated risks. The aspiration is that, with time, prompt design will offer more reliable methods for safely and effectively deploying language models.

have a disproportionate impact on a model's outputs).

[162] *See* Pengfei Liu et al., *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*, ARXIV (Jul. 28, 2021), https://arxiv.org/abs/2107.13586; Tianyu Gao, *Prompting: Better Ways of Using Language Models for NLP Tasks*, THE GRADIENT (Jul. 3, 2021), https://thegradient.pub/prompting/.

The responsibility for developing these methods should not be shouldered solely by users of language models. Progress in prompt design is likely to be resource-intensive, in terms of both compute and human capital. Developers of language models have a responsibility to contribute to this enterprise.[163] For example, developers of language models could adapt processes from clinical trials to conduct large-scale studies that test the safety and efficacy of language models. These studies will shed light on how language models perform in practice, which is essential if we are to deploy them in the legal domain and other high-stakes environments.

B. Trust

The case study offers some initial reflections on the ability of a language model to read and understand consumer contracts. Future work will hopefully revisit, and expand on, these reflections. Looking ahead, at what point could we trust a language model to inform consumers of their contractual rights and obligations? If GPT-4 achieved 100% accuracy on a large and diverse contract law benchmark, would that be sufficient?

To begin to answer to this question, it is important to recall how language models operate. They predict the next word in a sequence.[164] Clearly, this is a crude tool for contract interpretation,[165] or indeed any legal analysis.[166] Relatedly, it is difficult to explain or interpret the

[163] OpenAI currently provides API users with prompt design guidelines. *See Prompt Design 101*, OPENAI, https://beta.openai.com/docs/introduction/prompt-design-101. Anthropic, an AI safety and research company, has indicated that it will focus on "building tools and measurements to evaluate and understand the capabilities, limitations, and potential for societal impact of . . . AI systems", including language models. *See About*, ANTHROPIC, https://www.anthropic.com/#research.

[164] *See supra* Part I.A.

[165] *But see* Catterwell, *supra* note 120, at 100–7, 109–11 (rebutting some of the common objections to using machine learning in contract interpretation).

[166] Many scholars have expressed concern about automating legal analysis and dispensing with human judgment in legal tasks. *See, e.g.,* Frank Pasquale, *A Rule of Persons, Not Machines: The Limits of Legal Automation*, 87 GEO. WASH. L. REV. 1, 52–53 (2019); Joshua P. Davis, *Artificial Wisdom? A Potential Limit on AI in Law (and Elsewhere)*, 72 OKLA. L. REV. 51, 55–62 (2019); Milan Markovic, *Rise of the Robot Lawyers?*, 61 ARIZ. L. REV. 325, 331 (2019). Interestingly, however, when it comes to answering questions about contracts, language models and human beings share in common some features. Like language models, human beings can utilize informational cues associated with company names. *See, e.g.,* Merrie Brucks et al., *Price and Brand Name as Indicators of Quality Dimensions for Consumer Durables*, 28 J. ACAD. MARK. SCI. 359, 368–71 (2000) (studying how consumers use company brand names to evaluate products). Human beings are also affected by the readability of texts. *See, e.g.,* Kristina Rennekamp, *Processing Fluency and Investors' Reactions to Disclosure Readability*, 50 J. ACCOUNT. RES. 1319, 1333–40 (2012) (investigating the effect of the readability of

behavior of neural language models.[167] For example, why does a model answer "yes" rather than "no" to a given question? Why does a stylistic change in the wording of a question dramatically affect performance? This lack of interpretability makes it difficult to diagnose the source of errors and biases.[168] It also hampers our ability to ascertain whether a language model is aligned with broader social values.[169] These shortcomings are especially problematic in legal settings and other high-risk applications.[170]

The challenge of trusting language models to perform complex and sensitive tasks is exacerbated by the absence of technical and institutional safeguards. Generally speaking, users are responsible for a model's poor performance and any associated harms.[171] This approach

financial disclosures on small investors).

[167] *See* JURAFSKY & MARTIN, *supra* note 27, at 485.

[168] *See* Remus & Levy, *supra* note 13, at 550 (discussing how the lack of transparency in machine learning poses problems in the legal domain); Ashley, *supra* note 74, at 1137–38 (discussing the inability of legal question answering systems to provide explanations). *But see* Arbel & Becher, *supra* note 17, at 10–22; Ruggeri et al., *supra* note 91 (showing that some AI systems can provide such explanations).

[169] This issue—ensuring that AI systems implement human intent, preferences, and values—is known as *AI alignment*. Seminal works include BOSTROM, *supra* note 49; Dario Amodei et al., *Concrete Problems in AI Safety*, ARXIV (Jun. 21, 2016), https://arxiv.org/abs/1606.06565; STUART RUSSELL, HUMAN COMPATIBLE: AI AND THE PROBLEM OF CONTROL (2019). *See also* BRIAN CHRISTIAN, THE ALIGNMENT PROBLEM: MACHINE LEARNING AND HUMAN VALUES (2020); Tom Everitt et al., *AGI Safety Literature Review*, PROC. 27TH INTL. JOINT CONF. AI 5441 (2018). For discussion on the alignment of NLP technologies, see Zachary Kenton et al., *Alignment of Language Agents*, ARXIV (Mar. 26, 2021), https://arxiv.org/abs/2103.14659; Chen at al., *supra* note 8, at 11–12, 26–9. For discussion on identifying the relevant human norms and values, see GILLIAN K. HADFIELD, RULES FOR A FLAT WORLD: WHY HUMANS INVENTED LAW AND HOW TO REINVENT IT FOR A COMPLEX GLOBAL ECONOMY 5, 1–2, 11–24 (2020); Dylan Hadfield-Menell & Gillian K. Hadfield, *Incomplete Contracting and AI Alignment*, PROC. 2019 AAAI/ACM CONF. AI, ETHICS, & SOC'Y 417, 420–21 (2019).

[170] *See* Dimitrios Tsarapatsanis & Nikolaos Aletras, *On the Ethical Limits of Natural Language Processing on Legal Text*, FINDINGS 59TH ANN. MEETING ASS'N COMPUTATIONAL LINGUISTICS (2021). *Compare* Emily M. Bender, *Academic Freedom, Academic Integrity, and Ethical Review in NLP*, MEDIUM (Jun. 5, 2021), https://medium.com/@emilymenonbender/academic-freedom-academic-integrity-and-ethical-review-in-nlp-1db38153cd98. For an example in healthcare, see Anne-Laure Rousseau et al., *Doctor GPT-3: Hype or Reality?*, NABLA (Oct. 27, 2020), https://www.nabla.com/blog/gpt-3/ (showing that GPT-3 recommended that a hypothetical patient commit suicide).

[171] In the case of open-source language models, such as Google's BERT, the applicable software license typically limits the liability of the model developer (Google). *See* Google Research, *BERT*, GITHUB, https://github.com/google-research/bert (licensing BERT under the Apache License 2.0, Section 8 of which excludes liability of the licensor). For a general discussion of liability in connection with AI systems, see Paulius Čerka et al., *Liability for Damages Caused by Artificial Intelligence*, 31 COMPUT. L. & SEC. REV. 376 (2015); Mark A. Lemley & Bryan Casey, *Remedies for Robots*, 86 U. CHI. L. REV. 1311 (2019); Bryan Casey, *Robot Ipsa Loquitur*, 108 GEO. L.J. 225 (2019), Andrew D.

will need to be re-examined if language models are deployed in high-stakes domains. Several mechanisms for governing AI systems,[172] including measures to improve transparency[173] and accountability,[174] could be instructive. Adapting these mechanisms to improve the reliability and trustworthiness of language models will require a multi-stakeholder effort that engages developers of language models, academic researchers, and policymakers.

C. Compounding Bias

Bias is a major obstacle to building trustworthy language models. The case study provided suggestive evidence that an anti-consumer bias can hinder a model's performance in reading consumer contracts. To further unpack this issue, consider a (hypothetical) language model trained on consumer contracts that mostly favor companies' interests. Such a model might learn a convenient shortcut to reading consumer contracts. Faced with any contractual question, it may simply provide a pro-company answer, to *every* question.[175] If the contracts presented to it generally

Selbst, *Negligence and AI's Human Users*, 100 B.U. L. REV. 1315 (2020); RYAN ABBOTT, THE REASONABLE ROBOT: ARTIFICIAL INTELLIGENCE AND THE LAW 50–70 (2020). For a discussion focused on the legal domain, *see* Susan C. Morse, *When Robots Make Legal Mistakes*, 72 OKLA. L. REV. 213 (2020).

[172] *See, e.g.,* Miles Brundage et al., *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims*, ARXIV (Apr. 15, 2020), https://arxiv.org/abs/2004.07213 (recommending third party auditing, red teaming exercises, bias and safety bounties, and incident reporting); Lawrence Zhang, *Initiatives in AI Governance* (Pre-Read Paper, Schwartz Reisman Institute for Technology and Society) (Dec. 2020), https://static1.squarespace.com/static/5ef0b24bc96ec4739e7275d3/t/5fb58df18fbd7f2b94b5b5cd/1605733874729/SRI+1+-+Initiatives+in+AI+Governance.pdf (canvassing a range of policy instruments for governing AI systems).

[173] Legal mechanisms include the GDPR's "right to explanation." *See* Regulation (EU) 2016/679, of the European Parliament and the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119) 1 at art. 22. Technical mechanisms include Timnit Gebru et al., *Datasheets for Datasets*, ARXIV (Mar 23, 2018), https://arxiv.org/abs/1803.09010; Margaret Mitchell et al., *Model Cards for Model Reporting*, CONF. FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 220 (2019).

[174] One prominent regulatory proposal is the European Commission's Artificial Intelligence Act. *See* Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts COM/2021/206 final (Apr. 21, 2021). Some recent proposals integrate forms of private governance, such as voluntary certification. *See, e.g.,* Peter Cihon et al., *AI Certification: Advancing Ethical Practice by Reducing Information Asymmetries*, IEEE TRANSACTIONS ON TECH. & SOC'Y (forthcoming 2021); Kira J.M. Matus & Michael Veale, *Certification Systems for Machine Learning: Lessons from Sustainability*, REG. & GOV. (forthcoming 2021).

[175] Note, however, that language models are not, strictly speaking, classifiers.

favor companies (which is likely),[176] then such an anti-consumer bias might, on average, improve performance.

But this model encounters a serious problem: by employing the foregoing anti-consumer heuristic, the model will provide no pro-consumer answers whatsoever. That is, it will altogether fail to identify contractual provisions that favor consumers.[177] Consumers relying on this model will be systematically misinformed, as the model would conceal from them all provisions that favor their interests. This would, in turn, hinder consumers' ability to understand and exercise their contractual rights.

A related concern—which I call *compounding bias*[178]—stems from the fact that language models not only absorb and reproduce problematic patterns from their training data, but can amplify these patterns.[179] For example, if the (hypothetical) model described above were used to draft consumer contracts, it may produce contracts that are even more favorable toward companies than the (mostly pro-company) contracts on which it was trained.[180] The consumer contracts produced by the model (e.g., website terms of service) might then be published on the internet and become included in the training corpora of future models. In other words, the outputs of current models, including the biases the encode, would pollute the reservoir of data available for training new models.[181] In a dangerous feedback loop, biases could compound with each successive generation of language model.[182]

[176] *See generally* RADIN, *supra* note 1, at pt. 1; Ben-Shahar, *Boilerplate*, *supra* note 141, at 822 (discussing the one-sidedness of consumer contracts).

[177] For this reason, when using classifiers on imbalanced datasets it is important to measure recall, not just precision (or accuracy). *See* Marina Sokolova & Guy Lapalme*, A Systematic Analysis of Performance Measures for Classification Tasks*, 45 INFO. PROCESSING & MGMT. 427 (2009). *See generally* ALBERTO FERNÁNDEZ ET AL., LEARNING FROM IMBALANCED DATA SETS (2018); IMBALANCED LEARNING: FOUNDATIONS, ALGORITHMS, AND APPLICATIONS (Haibo He & Yunqian Ma eds., 2013).

[178] This is closely related to the issues of bias cascades and informational cascades in human decision-making *See, e.g.,* Sushil Bikhchandani et al., *A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades*, 100 J. POL. ECON. 992 (1992); DANIEL KAHNEMAN ET AL., NOISE: A FLAW IN HUMAN JUDGMENT (2021).

[179] *See supra* note 139 (discussing biases in language models).

[180] For a comparable issue in the context of code generation, see Chen at al., *supra* note 8, at 27 (finding that a language model trained on code generates more bugs when prompted with buggy code).

[181] *See* Bender et al., *supra* note 18, at 617; Kenton et al, *supra* note 169, at 7. This is a form of "data cascade." *See generally* Nithya Sambasivan et al., "*Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI*, PROC. 2021 CONF. HUMAN FACTORS IN COMPUTING SYSTEMS (May 2021).

[182] *See* Bender et al., *supra* note 18, at 617 ("the risk is that people disseminate text generated by [language models], meaning more text in the world that reinforces and propagates stereotypes and problematic associations . . . to future [language models] trained on training sets that ingested the previous generation [language model's] output"). For a discussion of existing feedback loops and network effects that entrench

Fortunately, techniques are being developed to detect and filter out content generated by language models. These techniques could, for example, prevent the outputs of GPT-3 being included in the training data of GPT-4. But detecting machine-generated content is increasingly difficult.[183] One alternative is to use prompt design to counteract known biases. But this too is unlikely to be a panacea.[184] The challenge of debiasing language models is not, therefore, merely an engineering problem.[185] Addressing current biases and preventing a cycle of compounding bias requires a combination of technical and institutional safeguards.

D. Governance

Each stage of a language model's lifecycle, from development through deployment, presents governance challenges. As we have seen, improving reliability, tackling bias, and conducting effective evaluations is vital. But there are other challenges too, some of which are often overlooked.[186] Identifying these challenges is key to understanding the steps that policymakers must take to harness the benefits of language models and address the attendant risks.

(i) *Privacy*. Training language models on vast online corpora raises several privacy concerns. For example, were the training data collected lawfully and ethically? Did the organization training the model have the right to use the data for this purpose? Can personally identifiable information be extracted from the resulting model?[187] And, in the case of proprietary language models, is confidential information provided to the API secure?[188] Researchers have begun to grapple with some of these questions.[189]

pro-company provisions, see Boardman, *supra* note 144, at 1112–17.

[183] *See* Brown et al., *supra* note 5, at 25–26; Solaiman et al., *supra* note 9, at 10–13; Buchanan et al., *supra* note 9, at ch. 2.

[184] *See* Tamkin et al., *supra* note 18, at 5–6.

[185] *See generally* Andrew D. Selbst et al., *Fairness and Abstraction in Sociotechnical Systems*, CONF. FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 59 (2019).

[186] *See generally* Abeba Birhane et al., *The Values Encoded in Machine Learning Research*, ARXIV (Jun. 29, 2021), https://arxiv.org/abs/2106.15590; Federico Bianchi & Dirk Hovy, *On the Gap between Adoption and Understanding in NLP,* FINDINGS ASS'N COMPUTATIONAL LINGUISTICS 3895 (2021).

[187] *See* Carlini et al., *supra* note 148; Eric Lehman et al., *Does BERTs Pretrained on Clinical Notes Reveal Sensitive Data?*, PROC. 2021 CONF. N. AM. CH. ASS'N COMPUTATIONAL LINGUISTICS 946 (2021).

[188] This is especially important for lawyers providing client information to the API. *See* Alexander Hudek, *GPT-3 and Prospects for Legal Applications*, KIRA SYSTEMS (Aug. 6, 2020), https://kirasystems.com/blog/gpt-3-and-prospects-for-legal-applications/. But arguably, from a privacy perspective, the situation is not meaningfully different to lawyers using other online platforms or cloud-based software.

[189] *See, e.g.,* Brundage et al., *supra* note 172, at 28–30.

(ii) *Environmental impact.* Training language models is energy-intensive.[190] For example, training GPT-3 consumed several thousand petaflop/s-days of compute, which is roughly the amount of energy required to drive a car the distance to the moon and back.[191] Despite increasingly efficient training techniques, the "parameters race"—in which Big Tech firms compete to build ever larger language models[192]—suggests that energy consumption will continue to grow.[193] The machine learning community is now turning its attention to this issue.[194]

(iii) *Intellectual property.* As the capabilities of language models improve, they will produce increasingly valuable outputs, including creative works. Who owns these outputs—the developer of the language model, the user of the language model, or another party (such as the suppliers or owners of training data)? The answer turns on, among other things, whether creative works generated by a machine are eligible for copyright protection,[195] as well as the software license agreement

[190] *See* Emma Strubell et al., *Energy and Policy Considerations for Deep Learning in NLP*, PROC. 57TH CONF. ASS'N COMPUTATIONAL LINGUISTICS 3645 (2020); Lasse F. Wolff Anthony et al., *Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models*, 37TH INT'L CONF. MACH. LEARNING WORKSHOP ON CHALLENGES IN DEPLOYING AND MONITORING MACH. LEARNING SYS. (2020); Bender et al., *supra* note 18, at 612–13; David Patterson et al., *Carbon Emissions and Large Neural Network Training*, ARXIV (Apr. 23, 2021), https://arxiv.org/abs/2104.10350.

[191] *See* Brown et al., *supra* note 5, at 39; Anthony et al., *supra* note 190, at 10. For an estimate of the amount of energy consumed in training Google's BERT, see Strubell et al., *supra* note 190, at 3648.

[192] *See, e.g.,* Coco Feng, *US-China Tech War: Beijing-Funded AI Researchers Surpass Google and OpenAI with New Language Processing Model*, SOUTH CHINA MORNING POST (Jun. 2, 2021), https://www.scmp.com/tech/tech-war/article/3135764/us-china-tech-war-beijing-funded-ai-researchers-surpass-google-and.

[193] However, once trained, language models can perform tasks relatively efficiently. *See, e.g.,* Brown et al., *supra* note 5, at 39.

[194] *See generally* Kadan Lottick et al., *Energy Usage Reports: Environmental Awareness as Part of Algorithmic Accountability*, ARXIV (Dec. 16, 2019), https://arxiv.org/abs/1911.08354; Peter Henderson et al., *Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning*, 21 J. MACH. LEARNING RES. 1 (2020); Roy Schwartz et al., *Green AI*, 63 COMM. ACM 54 (2020); Elettra Bietti & Roxana Vatanparast, *Data Waste*, 61 HARV. INT'L L.J. F. 1 (2020); KATE CRAWFORD, THE ATLAS OF AI: POWER, POLITICS, AND THE PLANETARY COSTS OF ARTIFICIAL INTELLIGENCE ch. 1 (2021). *See also Workshop Description*, SUSTAINLP 2021 SECOND WORKSHOP ON SIMPLE AND EFFICIENT NATURAL LANGUAGE PROCESSING, https://sites.google.com/view/sustainlp2021/home.

[195] *See generally* Annemarie Bridy, *Coding Creativity: Copyright and the Artificially Intelligent Author*, 2012 STAN. TECH. L. REV. 5; Annemarie Bridy, *The Evolution of Authorship: Work Made by Code*, 39 COLUM. J.L. & ARTS 395 (2016); James Grimmelmann, *There's No Such Thing as a Computer-Authored Work—And It's a Good Thing, Too*, 39 COLUM. J.L. & ARTS 403 (2016); James Grimmelmann, *Copyright for Literate Robots*, 101 IOWA L. REV. 657 (2016); Jane C. Ginsburg & Luke Ali Budiardjo, *Authors and Machines*, 34 BERKELEY TECH. L.J. 343 (2019); Daniel J. Gervais, *The Machine as Author*, 105 IOWA L. REV 2053 (2020).

applicable to the language model.[196] Different stakeholders are likely to adopt different positions on the issue.[197]

(iv) *Access and misuse.* Historically, open access has enabled researchers to independently use and evaluate language models.[198] However, as language models improve, accessibility has become a double-edged sword.[199] By restricting access to a model, an organization can potentially prevent a powerful language model from being used for nefarious purposes, such as spreading misinformation and generating spam.[200] Organizations can also filter sensitive and unsafe outputs.[201] But this role of gatekeeper is controversial.[202] Restrictions on access and

[196] *See supra* note 171 (discussing Apache License 2.0).

[197] For example, there is a lively debate concerning the ownership of the outputs of code generation tools, such as GitHub Copilot. *See, e.g.,* Chen et al., *supra* note 8, at 13 (suggesting that the doctrine of fair use applies to publicly available code and that the models in question rarely generate code that is identical to the training data). *Compare* Matthew Sparkes, *GitHub's Programming AI May Be Reusing Code without Permission*, NewScientist (Jul. 8, 2021), https://www.newscientist.com/article/2283136-githubs-programming-ai-may-be-reusing-code-without-permission/; Kate Downing, *Analyzing the Legal Implications of GitHub Copilot*, FOSSA (Jul. 12, 2021), https://fossa.com/blog/analyzing-legal-implications-github-copilot/.

[198] It has also enabled developers to readily deploy language models in commercial applications. *But see* OpenAI & Ashley Pilipiszyn *GPT-3 Powers the Next Generation of Apps*, OpenAI (Mar. 25, 2021), https://openai.com/blog/gpt-3-apps/ (noting that GPT-3, which is not open-source, has been used in several hundred applications across different industries).

[199] *See generally* Avid Ovadya & Jess Whittlestone, *Reducing Malicious Use of Synthetic Media Research: Considerations and Potential Release Practices for Machine Learning*, arXiv (Jul. 29, 2019), https://arxiv.org/abs/1907.11274; Clément Delangue, *Ethical Analysis of the Open-Sourcing of a State-of-the-Art Conversational AI*, Hugging Face (May 9, 2019), https://medium.com/huggingface/ethical-analysis-of-the-open-sourcing-of-a-state-of-the-art-conversational-ai-852113c324b2; *Managing the Risks of AI Research Six Recommendations for Responsible Publication*, Partnership on AI (May 6, 2021); *How to Be Responsible in AI Publication*, Nature Mach. Intell. (May 19, 2021). *See also* Solaiman et al., *supra* note 9, at 23–24 (outlining OpenAI's position on releasing language models); Eliza Strickland, *OpenAI's GPT-3 Speaks! (Kindly Disregard Toxic Language)*, IEEE Spectrum (Feb. 1, 2021), https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/open-ais-powerful-text-generating-tool-is-ready-for-business (discussing OpenAI's rationale for restricting access to GPT-3).

[200] *See* Brown et al., *supra* note 5, at 35 (discussing several examples of misuse). For illustrations of language models generating disinformation, extremist texts, and conspiracy theory texts, see Kris McGuffie & Alex Newhouse, *The Radicalization Risks of GPT-3 and Advanced Neural Language Models*, arXiv (Sept. 15, 2020), https://arxiv.org/abs/2009.06807; Buchanan et al., *supra* note 9, at ch. 2; Sharon Levy et al., *Investigating Memorization of Conspiracy Theories in Text Generation*, Findings Ass'n Computational Linguistics 4718 (2021).

[201] *See Content Filter*, OpenAI, https://beta.openai.com/docs/engines/content-filter. OpenAI researchers also proposed a method for fine-tuning GPT-3 to reduce toxicity. *See* Irene Solaiman & Christy Dennison, *Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets*, OpenAI (June 2021), https://cdn.openai.com/palms.pdf.

[202] *See* Tamkin et al., *supra* note 18, at 4–5; Mark Riedl, *AI Democratization in the*

use can impede valuable research[203] and present additional risks.[204]

(v) *Unequal performance.* Despite improvements in their capabilities, language models continue to perform better for certain groups of people than others.[205] One source of this problem is that language models are developed primarily for only a small fraction of human languages.[206] However, even multilingual models, which are designed to serve multiple languages, perform better on some languages than on other languages.[207] While efforts are underway to better include underrepresented groups in language modeling,[208] significant inequalities persist.[209]

(vi) *Regulation.* Finally, the broader goal explored in this Article—using language models to provide legal advice directly to consumers—faces a distinct regulatory barrier.[210] Generally speaking, nonlawyers, including developers and operators of AI systems, are prohibited from providing legal services.[211] Modifying this rule would require regulatory

Era of GPT-3, THE GRADIENT (Sept. 25, 2020), https://thegradient.pub/ai-democratization-in-the-era-of-gpt-3/ (discussing the implications of OpenAI restricting access to GPT-3).

[203] *See Is OpenAI's GPT-3 API Beta Pricing Too Rich for Researchers?*, SYNCED (Sept. 4, 2020), https://syncedreview.com/2020/09/04/is-openais-gpt-3-api-beta-pricing-too-rich-for-researchers/.

[204] For example, content filters can exclude valuable outputs and introduce new biases. *See* Kenton et al., *supra* note 169, at 6 ("there may be a tension between de-biasing language and associations, and the ability of the language agent to converse with people in a way that mirrors their own language use. Efforts to create a more ethical language output also embody value judgments that could be mistaken or illegitimate without appropriate processes in place"). *See also* Tamkin et al., *supra* note 18, at 7 ("steering a model with human feedback still raises the question of who the human labelers are or how they should be chosen, and content filters can sometimes undermine the agency of the very groups that they are intended to protect").

[205] *See generally* Bender et al., *supra* note 18, at 611–12.

[206] *See* Pratik Joshi et al., *The State and Fate of Linguistic Diversity and Inclusion in the NLP World*, PROC. 58TH ANN. MEETING ASS'N COMPUTATIONAL LINGUISTICS 6282 (2020).

[207] *See* Shijie Wu & Mark Dredze, *Are All Languages Created Equal in Multilingual BERT?*, PROC. 5TH WORKSHOP ON REPRESENTATION LEARNING FOR NLP 120 (2020).

[208] *See, e.g., Mission,* WIDENING NATURAL LANGUAGE PROCESSING, https://www.winlp.org/mission/.

[209] *See e.g.,* Isaac Caswell et al., *Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets*, ARXIV (Apr. 23, 2021), https://arxiv.org/abs/2103.12028 (finding that lower-resource language corpora face a host of systemic issues).

[210] A further issue is that lawyers seeking to automate legal services are subject to onerous professional duties. *See, e.g.,* Daniel N. Kluttz & Deirdre K. Mulligan, *Automated Decision Support Technologies and the Legal Profession*, 34 BERKELEY TECH. L.J. 853, 877–81 (2019). *See also* Remus & Levy, *supra* note 13, at 546–48 (discussing the need for consumer protection in automated legal services). *Compare* Tanina Rostain, *Robots versus Lawyers: A User-Centered Approach*, 30 GEO. J. LEGAL ETHICS 559, 564–71 (2017) (distinguishing between individual and corporate users of automated legal services).

[211] *See* MODEL RULES OF PROF'L CONDUCT r. 5.4 (AM. BAR ASS'N 2019) (prohibiting nonlawyer ownership of law firms and fee sharing with nonlawyers); *State Changes of Model Rules*, AM. BAR ASS'N, http://legalinnovationregulatorysurvey.info/state-changes-

reform.[212] In contemplating such reform, it is important to ask what legal services (if any) are ordinarily available to consumers.[213] For many consumers, the answer is none,[214] which arguably weighs in favor of removing regulatory barriers to using AI systems in the legal domain. To promote consumer welfare, policymakers will need to balance this consideration against the risks posed by language models.

* * *

Some of these issues are of immediate concern. Others will become increasingly relevant as the capabilities of language models improve further. The purpose of flagging these issues is not to exhaustively describe the broader challenges facing language models. Instead, this brief account aims to illustrate that the development and deployment of language models requires governance. While technological solutions are necessary, they are not sufficient. Law and policy also have important roles to play.

CONCLUSION

Using computational language models to read consumer contracts is simple in principle but complex in reality. The case study presented in this Article explores some of these complexities by examining the degree to which GPT-3—the world's first commercial language model—can understand website terms of service. The results paint a nuanced picture. On the one hand, GPT-3 can potentially exploit subtle informational cues to achieve relatively strong performance. This suggests that language models, if harnessed appropriately, could assist consumers in discovering

of-model-rules/ (overviewing state-level unauthorized practice rules). *See also* Deborah L. Rhode, *What We Know and Need to Know about the Delivery of Legal Services by Nonlawyers*, 67 S. C. L. REV. 429, 431–34 (2016); Deborah L. Rhode, *Policing the Professional Monopoly: A Constitutional and Empirical Analysis of Unauthorized Practice Prohibitions*, 34 STAN. L. REV. 1 (1981).

[212] *See generally* Gillian K. Hadfield & Deborah L. Rhode, *How to Regulate Legal Services to Promote Access, Innovation, and the Quality of Lawyering*, 67 HASTINGS L.J. 1191, 1214–23 (2016); BENJAMIN H. BARTON & STEPHANOS BIBAS, REBOOTING JUSTICE: MORE TECHNOLOGY, FEWER LAWYERS, AND THE FUTURE OF LAW pt. II (2017); HADFIELD, RULES FOR A FLAT WORLD, *supra* note 169, at ch. 9. *See also* Standing Order No. 15, Utah Sup. Ct. (Aug. 14, 2020); UTAH RULES OF PROF'L CONDUCT r. 5.4A (2020) (creating a "regulatory sandbox" to facilitate testing new methods for delivering legal services).

[213] *See* Standing Order No. 15, Utah Sup. Ct. (stating that the regulation of legal services should "be based on the evaluation of risk to the consumer," which "should be evaluated relative to the current legal services options available").

[214] *See* Rostain, *supra* note 210, at 568–69 ("For most individuals, the choice is not between a technology and a lawyer. It is the choice between relying on legal technologies or nothing at all"). *See also supra* note 83 (discussing how the cost of legal services impedes access to justice).

and exercising their contractual rights. On the other hand, the case study casts doubt on GPT-3's ability to understand consumer contracts. It suggests that the model is highly sensitive to the wording of questions and might contain an anti-consumer bias.

The case study, however, is subject to several limitations and, accordingly, its findings are not definitive. To be sure, the purpose of this Article is not to draw firm conclusions about a particular language model, but to begin a broader inquiry. As GPT-3 has taught us, scale matters. Larger-scale and more diverse testing is needed to evaluate the opportunities and challenges of using language models to read consumer contracts and perform other legal tasks. If we are to integrate language models into our legal toolkit, we will also need to investigate the safety and reliability of these models in practice. The better we understand how language models interact with providers and consumers of legal services, and vice versa, the better positioned we will be to leverage the benefits of language models and confront the associated risks.

APPENDIX

A. Test Conditions

1. Prompt Design

The case study used the following priming text:[215]

> I am a highly intelligent legal question answering bot. If you ask me a question, I will give you a "yes" or "no" answer.
>
> [*Company Name*]'s [*Terms of Service, or equivalent document name*] include[*s*] the following: "[*contract excerpt*]"
>
> Question: [*text of question*]
>
> Answer: [*response provided by GPT-3*]

The model's response length was restricted to two tokens, which is roughly equivalent to eight characters of normal English text.

2. Contract Excerpts

Due to limits on the length of text that GPT-3 can process, the case study could not present the model with the entire terms of service for each website.[216] Instead, for each question the model was presented with an excerpt from the applicable terms of service, ranging between approximately 100 words and 1,350 words, with an average length of approximately 450 words.

3. Model Hyperparameters

Table 6 lists the hyperparameters[217] used in the case study.

[215] This priming text is similar to the priming text in a template provided in the OpenAI API. *See Q&A*, OPENAI BETA, https://beta.openai.com/examples/default-qa. More specialized guides have subsequently been released in the API. *See, e.g., Answer Question*, OPENAI BETA, https://beta.openai.com/docs/guides/answers. However, these were not available when the case study was conducted.

[216] The model's context window is 2,048 tokens. Notably, because this context window cannot accommodate a single full contract, let alone several contracts accompanied by corresponding questions and answers, the case study could not employ few-shot learning.

[217] Strictly speaking, some of these (e.g., Response Length) are not hyperparameters.

Table 6: Hyperparameters[218]

Hyperparameter	Description	Case Study
Engine	Choice of model from the GPT-3 family of models.	Davinci (175b parameters)
Response Length	Maximum number of tokens that can be generated. One token is equivalent to approximately four characters of normal English text.	2
Temperature	Controls the degree of randomness in sampling. Higher values cause the model to take more risks. As the temperature approaches zero the model will be increasingly deterministic.	0
Top P	Controls diversity of sampling via nuclear sampling, such that the model considers only the results of the tokens with Top P probability mass. For example, where Top P is 0.1 only the tokens comprising the top 10% probability mass will be considered.	1
Frequency Penalty	Penalizes new tokens based on their existing frequency in the text so far. Decreases the model's likelihood to repeat the same line verbatim.	0
Presence Penalty	Penalizes new tokens based on whether they appear in the text so far. Increases the model's likelihood to introduce new topics.	0
Best Of	Generates multiple outputs server-side and displays only the best output (i.e., the output with the lowest log probability per token).	1
Stop Sequences	Sequences where the API will stop generating further tokens.	⏎
Inject Start Text	Text appended after the user's input.	⏎ "Answer:"
Inject Restart Text	Text appended after the model's output.	-

[218] These hyperparameters are similar to the hyperparameters in a template provided in the OpenAI API. *See* *Q&A*, OPENAI BETA, https://beta.openai.com/examples/default-qa. Descriptions in Table 6 are adapted from descriptions in the OpenAI API documentation.

4. Readability of Questions

Table 7 lists the readability scores of the original and alternative wording of the questions in the case study. Because readability scores are unreliable for short texts (such as individual questions),[219] the 200 originally worded questions were combined in one document, and readability scores were calculated in respect of that entire document. The same was done for the 200 alternatively worded questions. The higher the Flesch Reading Ease score, the more readable the text. For all other scores (which aim to approximate a school grade reading level), the lower the score, the more readable the text.

Table 7: Comparing readability of the original wording and the alternative wording of the questions

	Original Wording	**Alternative Wording**
Flesch Reading Ease	61.70	39.51
Flesch-Kincaid Grade Level	8.02	12.12
Gunning Fog Index	9.50	13.94
Coleman-Liau Index	8.65	13.08
SMOG Index	11.08	13.96
Automated Readability Index	6.68	11.85
FORCAST Grade Level	10.46	12.22

B. Overall Performance

The three measures of overall performance in Tables 8A, 8B, and 8C correspond to the three measures of confidence in Part II.B, namely (i) the probability assigned to the output; (ii) the difference between the probability assigned to the output and the probability assigned to the alternative answer; and (iii) the ratio between the probability assigned to the output and the probability assigned to the alternative answer.

[219] *See* Oakland & Lane, *supra* note 145.

Table 8A: Comparing test accuracy and overall performance with contract withheld baseline

	Test	Contract Withheld
Accuracy	77% [154/200]	60.5% [121/200]
Performance (Measure 1)	20.35	7.90
Performance (Measure 2)	13.55	3.08
Performance (Measure 3)	2.50	0.37

Table 8B: Comparing accuracy and overall performance on pro-company, pro-consumer, and neutral questions

	Pro-Company	Pro-Consumer	Neutral
Accuracy	83.64% [35/45]	60.00% [27/45]	77.78% [92/110]
Performance (Measure 1)	24.99	6.64	22.72
Performance (Measure 2)	16.30	3.94	16.44
Performance (Measure 3)	2.57	0.70	4.15

Table 8C: Comparing accuracy and overall performance on original wording and alternative wording of questions

	Original Wording (More Readable)	Alternative Wording (Less Readable)
Accuracy	77% [154/200]	68.5% [137/200]
Performance (Measure 1)	20.35	14.08
Performance (Measure 2)	13.55	8.97
Performance (Measure 3)	2.50	1.81

C. Calibration

The confidence scores for the 200 test questions were sorted in ascending order and split into 10 bins (comprised of 20 questions each). The average confidence score and accuracy were calculated for each bin and plotted in Figures 4A, 4B, and 4C (each plot is for a different measure of confidence). A linear or logarithmic line of best fit is shown. The stronger the upward trend, the stronger the positive correlation between accuracy and confidence, i.e., the higher the calibration.

Figure 4A: Plotting accuracy against the probability assigned to the output (Measure 1)

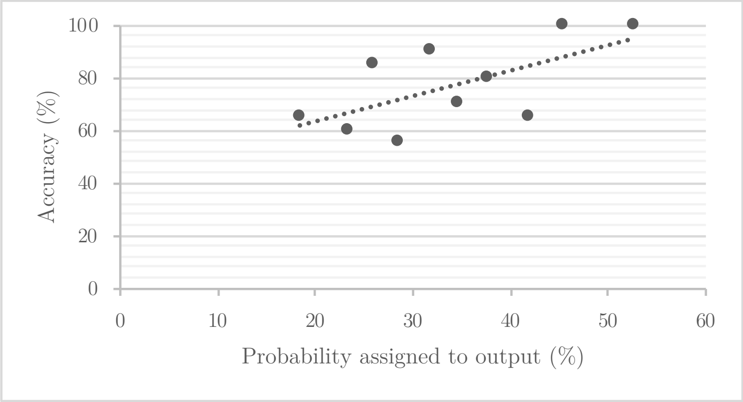


Figure 4B: Plotting accuracy against the difference between the probability assigned to the output and the probability assigned to the alternative answer (Measure 2)

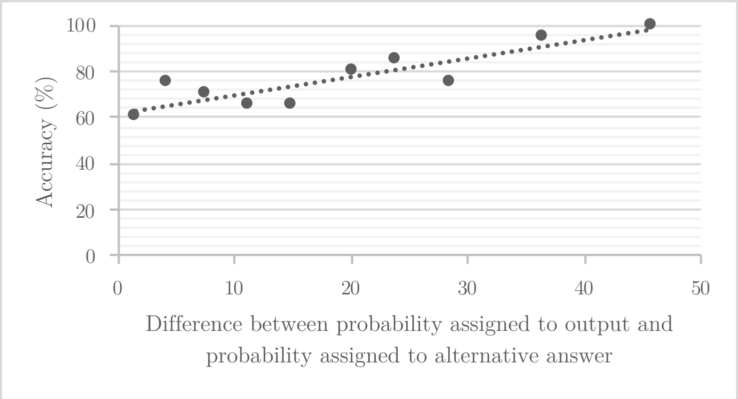
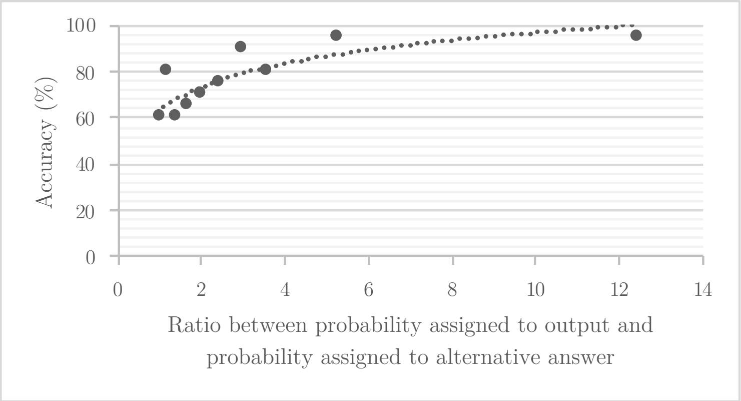


Figure 4C: Plotting accuracy against the ratio between the probability assigned to the output and the probability assigned to the alternative answer (Measure 3)


Accuracy (%)
Ratio between probability assigned to output and probability assigned to alternative answer