PREDICTING CONSUMER CONTRACTS

Noam Kolt†

ABSTRACT

This Article empirically examines whether a computational language model can read and understand consumer contracts. In recent years, language models have heralded a paradigm shift in artificial intelligence, characterized by unprecedented machine capabilities and new societal risks. These models, which are trained on immense quantities of data to predict the next word in a sequence, can perform a wide range of complex tasks. In the legal domain, language models can interpret statutes, draft transactional documents, 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 evaluates the model's ability to understand consumer contracts by testing its performance on a novel dataset comprised of questions relating to online terms of service. Although the results are not definitive, they offer several important insights. First, the model appears to be able to exploit subtle informational cues when answering questions about consumer contracts. Second, the model performs poorly in answering certain questions about contractual provisions that favor the rights and interests of consumers, suggesting that the model may contain an anti-consumer bias. Third, the model is brittle in unexpected ways. Performance in the case study was highly sensitive to the wording of questions, but surprisingly indifferent to variations in contractual language.

These preliminary findings suggest that while language models have the potential to empower consumers, they also have the potential to provide misleading advice and entrench harmful biases. Leveraging the benefits of language models in performing legal tasks, such as reading consumer contracts, and confronting the associated challenges requires a combination of thoughtful engineering and governance. Before language models are deployed in the legal domain, policymakers should explore technical and institutional safeguards to ensure that language models are used responsibly and align with broader social values.

DOI: https://doi.org/10.15779/Z382B8VC90

† Doctoral Candidate and Vanier Scholar, University of Toronto Faculty of Law; Graduate Affiliate, Schwartz Reisman Institute for Technology and Society. For helpful comments and suggestions, I thank Anthony Niblett, David Hoffman, Gillian Hadfield, Albert Yoon, Margaret Mitchell, Rohan Alexander, John Giorgi, and discussants at the ETH Zurich Center for Law & Economics, Monash-Warwick-Zurich Text as Data Workshop, University of Toronto Centre for Ethics, and We Robot at the University of Miami School of Law. I am also grateful to the editors of the *Berkeley Technology Law Journal* for their excellent contributions. OpenAI generously provided access to GPT-3 through the API Academic Access Program. Any errors or oversights are mine alone.

TABLE OF CONTENTS

I. INTRODUCTION 73
II. A PRIMER ON LANGUAGE MODELS 81
A. PREDICTION MACHINES 81
B. THE GPT-3 REVOLUTION 84
C. OPPORTUNITIES FOR LAW 89
III. EXPERIMENTAL DESIGN 94
A. CONTRACT QUESTIONS 94
B. EVALUATION CRITERIA 96
1. Accuracy 96
2. Calibration 97
3. Overall Performance 99
C. CHALLENGES AND LIMITATIONS 99
1. Challenges 99
2. Limitations 101
IV. RESULTS AND DISCUSSION 103
A. PERFORMANCE 104
1. Accuracy 104
2. Calibration 105
3. Overall Performance 105
B. ANTI-CONSUMER BIAS 106
C. INFORMATIONAL CUES 114
D. BRITTLENESS 116
V. BROADER IMPLICATIONS 119
A. ONGOING EXPERIMENTATION 119
B. CONSUMER TRUST 120
C. COMPOUNDING BIAS 125
D. GOVERNANCE 126
VI. CONCLUSION 133
APPENDIX 134
A. TEST CONDITIONS 134
1. Prompt Design 134
2. Contract Text 134
3. Model Hyperparameters 134
4. Question Readability 135
B. OVERALL PERFORMANCE 136
C. CALIBRATION PLOTS 137

I. INTRODUCTION

Consumer contracts increasingly govern important aspects of our lives.[1] Online communications, retail marketplaces, and consumer finance 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 time, expertise, or incentive to investigate how everyday consumer contracts affect their rights and interests.[3] Reading these contracts ourselves is simply unfeasible.[4]

---

1. *See generally* OREN BAR-GILL, SEDUCTION BY CONTRACT: LAW, ECONOMICS, AND PSYCHOLOGY IN CONSUMER MARKETS 1 (2012); MARGARET JANE RADIN, BOILERPLATE: THE FINE PRINT, VANISHING RIGHTS, AND THE RULE OF LAW 7–8 (2013); NANCY S. KIM, WRAP CONTRACTS: FOUNDATIONS AND RAMIFICATIONS 1–5 (2013); OMRI BEN-SHAHAR & CARL E. SCHNEIDER, MORE THAN YOU WANTED TO KNOW: THE FAILURE OF MANDATED DISCLOSURE 1–5 (2014).

2. *See* Yannis Bakos, Florencia Marotta-Wurgler & David R. Trossen, *Does Anyone Read the Fine Print? Consumer Attention to Standard-Form Contracts*, 43 J. LEGAL STUD. 1, 32 (2014) (finding that between 0.05 and 0.22 percent of retail software shoppers access the applicable license agreements); 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, 168 (2011) (estimating the average readership of end user license agreements to be between roughly 0.1 and 1 percent); *see also* Ian Ayres & Alan Schwartz, *The No-Reading Problem in Consumer Contract Law*, 66 STAN. L. REV. 545, 555–62 (2014) (discussing legal responses to the problem of non-readership of consumer contracts).

3. *See* Melvin Aron Eisenberg, *The Limits of Cognition and the Limits of Contract*, 47 STAN. L. REV. 211, 243 (1995) ("The verbal and legal obscurity of preprinted terms renders the cost of searching out and deliberating on these terms exceptionally high."); Robert A. Hillman & Jeffrey J. Rachlinski, *Standard-Form Contracting in the Electronic Age*, 77 N.Y.U. L. REV. 429, 436–37 (2002) (suggesting that consumers recognize that the costs of reading consumer contracts outweigh the potential benefits); Omri Ben-Shahar, *The Myth of the 'Opportunity to Read' in Contract Law,* 5 EUR. REV. CONT. L. 1, 13–21 (2009) (discussing consumers' limited ability to understand consumer contracts and positing that non-readership is often a rational choice); Victoria C. Plaut & Robert P. Bartlett, *Blind Consent? A Social Psychological Investigation of Non-Readership of Click-Through Agreements*, 36 LAW & HUM. BEHAV. 293, 305–6 (2012) (conducting experimental studies to examine which widely held beliefs about click-through agreements contribute to their non-readership); Tess Wilkinson-Ryan, *A Psychological Account of Consent to Fine Print*, 99 IOWA L. REV. 1745, 1759–60 (2014) (describing consumers' limited attentional resources when confronting contractual fine print); Michael Simkovic & Meirav Furth-Matzkin, *Proportional Contracts*, 107 IOWA L. REV. 229, 237–39 (2021) (surveying studies on the non-readership of consumers contracts).

4. *See* RESTATEMENT (THIRD) OF CONTRACTS 3 (AM. LAW INST., Tentative Draft, 2019) ("As the length and incidence of standard-form contracts have grown, it has become all the less plausible to expect consumers to read and take informed account of the contracts' provisions.").

One rapidly developing technology—computational language models[5]—could potentially offer a solution. These machine learning models, which have heralded a paradigm shift in artificial intelligence (AI),[6] can perform a wide range of complex tasks merely by predicting the next word in a sequence. In essence, computational language models are 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, Generative Pre-Trained Transformer 3 (GPT-3)—the world's first commercial language model,[7] developed by AI research company OpenAI—demonstrated unprecedented machine performance on a range of tasks.[8] The

5. *See infra* Part II (discussing the development of language model technology and its applications in the legal domain).

6. *See* Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang , Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou & Percy Liang, *On the Opportunities and Risks of Foundation Models*, ARXIV at 3, 6–7 (Aug. 18, 2021), https://arxiv.org/abs/2108.07258 (describing the emergence of general-purpose AI models). *But see* Gary Marcus & Ernest Davis, *Has AI Found a New Foundation?*, THE GRADIENT (Sept. 11, 2021), https://thegradient.pub/has-ai-found-a-new-foundation/(critiquing large language models and other so-called "foundation models").

7. *See infra* Part II.B (explaining that GPT-3 is a proprietary language model that can only be accessed through a commercial API).

8. *See* Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever & Dario Amodei, *Language Models Are Few-Shot Learners*, PROC. 34TH CONF. NEURAL INFO. PROCESSING SYS. (2020) (introducing the GPT-3

model can write compelling fictional stories,[9] translate natural language into computer code,[10] and produce news articles that appear to be written by human authors.[11]

Computational language models also present exciting opportunities in the legal domain. GPT-3, for instance, can summarize laws,[12] draft legal documents,[13] and translate legalese into plain English.[14] These capabilities

language model). For commentary on the broader impact of GPT-3, see David A. Price, *An AI Breaks the Writing Barrier*, WALL ST. J. (Aug. 22, 2020), https: //www.wsj.com /articles /an-ai-breaks-the-writing-barrier-11598068862; 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/.

9. *See* Gwern Branwen, *GPT-3 Creative Fiction,* GWERN (Sept. 28, 2020), https:// www.gwern.net/GPT-3 (illustrating GPT-3's ability to write in various literary genres).

10. *See* Sharif Shameem (@SharifShameem), TWITTER (July 13, 2020, 5:01PM), https:// twitter.com/sharifshameem /status /1282676454690451457 (demonstrating that GPT-3 can generate JSX code). One year following the release of GPT-3, OpenAI introduced a new language model trained specifically to generate code. *See* Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever & Wojciech Zaremba, *Evaluating Large Language Models Trained on Code*, ARXIV (July 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/ (last visited Aug. 8, 2022).

11. *See* Brown et al., *supra* note 8, 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).

12. *See* Daniel Gross (@DanielGross), TWITTER (June 14, 2020, 9:42 PM), https:// twitter.com/danielgross/status/1272238098710097920 (using GPT-3 to summarize a section of the U.S. Tax Code).

13. *See* Francis Jervis (@f_j_j_), TWITTER (July 17, 2020, 12:02 PM), https:// twitter.com/f_j_j_/status/1284050844787200000 (using GPT-3 to generate requests for admission).

14. *See* Michael Tefula (@MichaelTefula), TWITTER (July 21, 2020, 12:24 PM), https:// twitter.com/michaeltefula/status/1285505897108832257 (using GPT-3 to explain provisions in a founders' agreement).

could benefit both lawyers and consumers of legal services.[15] In the future, lawyers could use language models to expedite routine tasks, such as document review and transactional drafting. Language models could also support lawyers in conducting legal research, generating statements of claim, and even predicting case outcomes. If language models continue to improve, they could fundamentally alter the ways in which legal services are performed.[16]

This automation of legal work has the potential to improve access to justice. By performing tasks ordinarily carried out by lawyers and other legal services providers, language models could directly assist consumers facing legal issues in housing, personal finance, and other contexts. For example, one startup experimented with using GPT-3 to produce legal requests on behalf of tenants who might otherwise need to engage professional counsel.[17] Developments like this could be especially beneficial for consumers who cannot afford traditional legal services.

This Article explores a particular application in which language models could improve access to justice: reading consumer contracts. Despite the

15. *See infra* Part II.C (outlining the opportunities for using language models in the legal domain). 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); David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653 (2017); Benjamin Alarie, Anthony Niblett & Albert Yoon, *How Artificial Intelligence Will Affect the Practice of Law*, 68 U. TORONTO L.J. 106 (2018); Michael Simon, Alvin F. Lindsay, Loly Sosa & Paige Comparato, *Lola v. Skadden and the Automation of the Legal Profession*, 20 YALE J.L. & TECH. 234 (2018); Harry Surden, *Artificial Intelligence and Law: An Overview*, 35 GA. ST. U. L. REV. 1305 (2019); Milan Markovic, *Rise of the Robot Lawyers?*, 61 ARIZ. L. REV. 325, 328–42 (2019); LEGAL INFORMATICS (Daniel Martin Katz, Ron Dolin & Michael J. Bommarito ed., 2021); NOAH WAISBERG & ALEXANDER HUDEK, AI FOR LAWYERS: HOW ARTIFICIAL INTELLIGENCE IS ADDING VALUE, AMPLIFYING EXPERTISE, AND TRANSFORMING CAREERS (2021).

16. *See* Amy B. Cyphert, *A Human Being Wrote This Law Review Article: GPT-3 and the Practice of Law*, 55 U.C. DAVIS L. REV. 401, 403–405, 419–23 (2021); 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/.

17. *See* Augrented: Rent Safer (@augrented), TWITTER (July 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.

ubiquity of these agreements, consumers often struggle to read and understand their contents.[18] As a result, consumers may fail to discover or exercise their contractual rights. But what if consumers did not need to read these agreements themselves? What if that task could be outsourced to a machine? A language model that can read these documents and explain their legal ramifications would empower many consumers.[19]

The opportunities presented by language models, however, are accompanied by a host of concerns. Like other machine learning tools trained on immense quantities of data, language models pose serious risks.[20] In

18. *See supra* notes 2–4.

19. *See* Yonathan A. Arbel & Shmuel I. Becher, *Contracts in the Age of Smart Readers*, 90 GEO. WASH. L. REV. 83 (2022) (suggesting that language models could serve as "smart readers" of consumer contracts); Abhilasha Ravichander, Alan W Black, Thomas Norton, Shomir Wilson & Norman Sadeh, *Breaking Down Walls of Text: How Can NLP Benefit Consumer Privacy?*, PROC. 59TH ANN. MEETING ASS'N COMPUTATIONAL LINGUISTICS 4125 (2021) (illustrating how language technologies could assist in automatically processing privacy polices). However, even if consumers were to understand the content of contracts, they may nevertheless enter into unfavorable transactions. Apart from facing the informational load of reading contracts, consumers remain burdened by the cognitive load of making contracting decisions. *See* Russell Korobkin, *Bounded Rationality, Standard Form Contracts, and Unconscionability*, 70 U. CHI. L. REV. 1203, 1217, 1225–34 (2003); *see generally* BEN-SHAHAR & SCHNEIDER, *supra* note 1, at pt. II (describing the pervasive failure of consumer disclosure mechanisms).

20. *See* Emily M. Bender, Timnit Gebru, Angelina McMillan-Major & Shmargaret Shmitchell, *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, PROC. 2021 ACM CONF. FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 610 (2021); Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving & Iason Gabriel, *Ethical and Social Risks of Harm from Language Models*, ARXIV at 9–35 (Dec. 8, 2021), https: / /arxiv.org /abs /2112.04359; Bommasani et al., *supra* note 6, at 128–59; Matthew Hutson, *Robo-Writers: The Rise and Risks of Language-Generating AI*, 591 NATURE 22 (2021); *The Big Question*, 3 NATURE MACH. INTELL. 737 (2021); Alex Tamkin, Miles Brundage, Jack Clark & Deep Ganguli, *Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models*, ARXIV (Feb. 4, 2021), https: / /arxiv.org /abs /2102.02503. For further discussion of the risks associated with machine learning, see Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 10–18 (2014); FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION chs. 2, 4 (2015); Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 677–93 (2016); CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY chs. 3–10 (2016); Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023, 1027–34 (2017); Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 U.C. DAVIS L. REV. 399, 411–27 (2017); SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM 26–29 (2018); VIRGINIA EUBANKS, AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR ch. 4 (2018); MICHAEL KEARNS & AARON ROTH, THE

particular, language models can amplify harmful biases and be used for malicious purposes, such as spreading misinformation.[21] Since the release of GPT-3, researchers of language models have increasingly focused on issues traditionally sidelined by the computer science community.[22] Computational linguists have studied the extent to which language models can be prompted to generate racist, sexist, and other toxic content.[23] Social scientists have questioned whether language models can be deployed safely in high-stakes settings, such as healthcare, education, and law.[24] If language models are to become part of our legal toolkit, we must confront these issues.

ETHICAL ALGORITHM: THE SCIENCE OF SOCIALLY AWARE ALGORITHM DESIGN chs. 2–3, 5 (2019); Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218, 2227–62 (2019); Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan & Cass R. Sunstein, *Discrimination in the Age of Algorithms*, 10 J. LEGAL ANALYSIS 113, 138–48 (2019); Ben Hutchinson & Margaret Mitchell, *50 Years of Test (Un)fairness: Lessons for Machine Learning*, PROC. 2019 CONF. FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 49, 56–57 (2019); FRANK PASQUALE, NEW LAWS OF ROBOTICS: DEFENDING HUMAN EXPERTISE IN THE AGE OF AI chs. 4–6 (2020).

21. *See infra* note 150 (discussing the problem of societal biases in language models); *infra* Part V.D (discussing the potential misuses of language models).

22. However, the field of natural language processing (NLP) 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) (outlining several social and ethical implications of NLP technologies); Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama & Adam Kalai, *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*, PROC. 30TH INT'L CONF. NEURAL INFO. PROCESSING SYS. 4356 (2016) (finding that word embeddings can exhibit gender stereotypes). *But see* Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan & Michelle Bao, *The Values Encoded in Machine Learning Research*, ARXIV (June 29, 2021), https://arxiv.org/abs/2106.15590 (illustrating that machine learning research continues to neglect issues concerning its societal impact).

23. *See, e.g.*, Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi & Noah A. Smith, *RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models*, FINDINGS 2020 CONF. EMPIRICAL METHODS IN NLP 3356, 3359 (2020) (finding that language models can produce toxic text even from seemingly innocuous prompts); *see also* Ashutosh Baheti, Maarten Sap, Alan Ritter & Mark Riedl, *Just Say No: Analyzing the Stance of Neural Dialogue Generation in Offensive Contexts*, PROC. 2021 CONF. EMPIRICAL METHODS IN NLP 4846 (2021); Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin & Po-Sen Huang, *Challenges in Detoxifying Language Models*, FINDINGS 2021 CONF. EMPIRICAL METHODS IN NLP 2447 (2021).

24. *See, e.g.*, Bommasani et al., *supra* note 6, at 53–72 (discussing current and anticipated applications of large pretrained models); Weidinger et al., *supra* note 20, at 10 (presenting a taxonomy of the risks posed by large language models); Zhijing Jin, Geeticka Chauhan, Brian Tse, Mrinmaya Sachan & Rada Mihalcea, *How Good Is NLP? A Sober Look at NLP Tasks through the Lens of Social Impact*, FINDINGS ASS'N COMPUTATIONAL LINGUISTICS 3099, 3105–07 (2021) (evaluating the degree to which current NLP research aims to advance social good); *see also* Luciano Floridi & Massimo Chiriatti, *GPT-3: Its Nature, Scope, Limits, and Consequences*, 30

To properly unpack the opportunities and challenges of deploying language models in law, we need to understand how they work. The theory behind language models 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 some of the most recent breakthroughs.[25] One feature, however, remains constant. Language models, including GPT-3, do primarily one thing: predict the next word in a sequence. They function as an autocomplete, guessing what words are most likely to follow a particular text. Seen in this light, the range of tasks that state-of-the-art models can perform is remarkable. Yet this feature of language models is also responsible for some of their pitfalls, including the generation of biased and toxic outputs.

A theoretical understanding of language model technology, however, does not guarantee reliable performance. To evaluate the ability of a language model to perform a particular legal task, we need to empirically test a model on that particular legal task.

This Article presents a preliminary case study in using GPT-3 to read consumer contracts.[26] The case study examines 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, including Google, Amazon, and Facebook, and tested the model's ability to answer these questions. The results are illuminating. They shed light on the opportunities and risks of using GPT-3 to inform consumers of their contractual rights and obligations and offer new insights into the inner workings of language models.

First, GPT-3 appears to be able to exploit subtle informational cues embedded in certain questions about consumer contracts.[27] 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 the model's performance in answering questions that explicitly reference those companies.

Second, GPT-3 performs considerably worse in answering certain questions about contractual provisions that favor the rights and interests of consumers.[28] The model answered correctly nearly 84% of the questions about

---

MINDS & MACH. 681, 690–93 (2020); Kevin LaGrandeur, *How Safe Is Our Reliance on AI, and Should We Regulate It?*, 1 AI ETHICS 93, 96 (2020).

25. *See infra* Parts II.A–B.
26. *See infra* Part III.
27. *See infra* Part IV.C.
28. *See infra* Part IV.B.

provisions that favor companies, but only 60% of the questions about provisions that favor consumers.[29] 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, online terms of service that disproportionately preference the rights and interests of companies over the rights and interests of consumers.

Third, GPT-3 is brittle in unexpected ways.[30] The model appears to be highly sensitive to how questions are worded but surprisingly indifferent to variations in contractual language. In the case study, performance decreased dramatically when the questions presented to the model were less readable (i.e., more difficult for a human to read). However, performance did not decrease on longer or less readable contractual texts.

This 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 concerning, among other things, the design, scope, and sample size of the test questions. Accordingly, the findings presented here are not definitive. Nevertheless, the case study raises important questions regarding the potential advantages and pitfalls of using language models to read consumer contracts and proposes concrete directions for future research.

Subject to these qualifications, the case study paints a nuanced picture. On the one hand, it illustrates that GPT-3 performs relatively well in answering certain questions about consumer contracts. On the other hand, the case study highlights some of the model's weaknesses. The outsized impact of question-wording on performance casts doubt on the reliability of using language models in the legal domain, while 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, including consumers, lawyers, and other service providers, need to be aware of the technology's limitations. Developers of language models have a responsibility to investigate these limitations and explore methods for improving the reliability of language models. Finally, before language models are deployed in the legal domain, policymakers should establish technical and institutional safeguards to ensure that language models are used responsibly and align with broader social values.

---

29. The model answered correctly nearly 78% of the questions about neutral provisions, i.e., provisions that favor neither companies nor consumers.

30. *See infra* Part IV.D.

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

II. A PRIMER ON LANGUAGE MODELS

A. PREDICTION MACHINES

Language models are prediction machines,[31] designed to predict the next word in a sequence of text.[32] For example, given the sequence "she took the LSAT and applied to law . . ." an effective language model will predict that the next word is likely to be "school."[33] Language models can also predict the content of longer sequences and thereby generate lengthy synthetic texts. For example, when prompted appropriately, GPT-3 can write original sonnets.[34] The striking feature of advanced language models is that merely by predicting upcoming words, they can produce human-like texts that appear to exhibit genuine knowledge, understanding, and even emotion.[35]

How do language models make predictions? The basic idea is that words that occur in similar contexts tend to have similar meanings.[36] Suppose, for

31. *See* AJAY AGRAWAL, JOSHUA GANS & AVI GOLDFARB, PREDICTION MACHINES: THE SIMPLE ECONOMICS OF ARTIFICIAL INTELLIGENCE (2018) (using the term "prediction machines" to describe machine learning tools).

32. *See* DAN JURAFSKY & JAMES H. MARTIN, SPEECH AND LANGUAGE PROCESSING 29–54, 128–52, 180–94, 194–202 (draft 3rd ed., revised Sept. 21, 2021) (providing an overview of common types of language models, including *n*-gram models, neural language models, recurrent neural networks, and transformers).

33. Technically, language models assign probabilities to *each* word in the sequence, not just the upcoming word. Autoregressive models, such as GPT-3, process text from left to right and assign probabilities based only on the preceding text. In contrast, bidirectional models learn from the surrounding text on both sides of the target word. *See, e.g.*, Jacob Devlin, Ming-Wei Chang, Kenton Lee & Kristina Toutanova, *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 Bidirectional Encoder Representations from Transformers (BERT), which is a bidirectional language model).

34. Gwern, *supra* note 9.

35. *But see infra* note 117 (discussing the debate concerning whether language models can understand language).

36. This is known as the *distributional hypothesis*. *See* Zellig S. Harris, *Distributional Structure*, 10 WORD 146, 151–58 (1954); *see also* 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"); LUDWIG WITTGENSTEIN,

example, we have the sequence "many legal questions concerning contracts are" and we 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.[37] If the corpus is sufficiently large, this method will produce an accurate estimate. However, if the relevant sequence does not appear in the corpus, this method will fail. 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 would fail to calculate the probability of the next word in the sequence.

A simple solution is to instead calculate the probability of the next 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.[38] Calculating the relative frequencies for *n*-grams is usually feasible. For instance, in a large corpus of text, there are likely to be sequences in which the word "Roman" follows the word "ancient," and the word "commercial" follows "Roman."

In recent decades, computer scientists have developed more sophisticated methods of language modeling. The most prominent method is neural language models.[39] These models, which are based on neural networks,[40] can

PHILOSOPHICAL INVESTIGATIONS § 43 (1953) (contending that "[t]he meaning of a word is its use in the language").

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

38. 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 the probabilities of different words, *n*-grams can also be used to estimate the probabilities of entire sequences of text (and not just single words).

39. *See* Yoshua Bengio, Réjean Ducharme, Pascal Vincent & Christian Jauvin, *A Neural Probabilistic Language Model*, 3 J. MACH. LEARNING RES. 1137 (2003); Yoshua Bengio, Holger Schwenk, Jean-Sébastien Senécal, Fréderic Morin & Jean-Luc Gauvain, *Neural Probabilistic Language Models*, *in* INNOVATIONS IN MACH. LEARNING 137 (D.E. Holmes & L.C. Jain eds., 2006) (introducing neural language models). For a general overview of neural language models, see Yoav Goldberg, *A Primer on Neural Network Models for Natural Language Processing*, 57 J. AI RES. 345 (2017).

40. Neural networks are a family of algorithms commonly used in machine learning. *See generally* IAN GOODFELLOW, YOSHUA BENGIO & AARON COURVILLE., DEEP LEARNING pt.

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 are also better than *n*-gram models in making predictions in contexts that do not resemble the model's training data. Most notably, neural language models differ from *n*-grams as they represent text by semantic word embeddings, i.e., mathematical representations that express the *meaning* of words.[41] 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.

Neural language models can nevertheless struggle to process longer texts.[42] Consider the following sequence: "the lawyers, who have been working at the firm for over a decade, are eager to . . ." A language model, when predicting the probability of the word following "decade," may forget that the subject ("lawyers")—which appears much earlier in the sentence—is plural and should therefore be followed by "are" (rather than "is"). By the end of a long sequence, the model may fail to retain the information contained in earlier parts of the sequence. This is known as the problem of long-range dependencies. Further advances in machine learning have made significant progress in tackling this problem.[43]

Recent improvements in language modeling are also attributable to another development: pretraining. This involves training a general-purpose

---

II (2016) (offering an authoritative account of neural networks and their applications). For a more accessible introduction to neural networks, see MICHAEL A. NIELSEN, NEURAL NETWORKS AND DEEP LEARNING (2015).

41. *See, e.g.*, Tomás Mikolov, Kai Chen, Greg Corrado & Jeffrey Dean, *Efficient Estimation of Word Representations in Vector Space*, 1ST INT'L CONF. LEARNING REPRESENTATIONS (2013); Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado & Jeffrey Dean, *Distributed Representations of Words and Phrases and Their Compositionality*, PROC. 26TH INT'L CONF. NEURAL INFO. PROCESSING SYS. 3111 (2013) (introducing the word2vec methods for computing semantic embeddings); Jeffrey Pennington, Richard Socher & Christopher D. Manning, *GloVe: Global Vectors for Word Representation*, PROC. 2014 CONF. EMPIRICAL METHODS IN NLP 1532 (2014) (introducing the GloVe method for computing semantic embeddings).

42. *See* JURAFSKY & MARTIN, *supra* note 32, at 191.

43. *See* Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser & Illia Polosukhin, *Attention Is All You Need*, PROC. 30TH INT'L CONF. NEURAL INFO. PROCESSING SYS. 5998 (2017) (introducing the transformer architecture, which made major strides in overcoming the problem of long-range dependencies); *see also* Dzmitry Bahdanau, Kyunghyun Cho & Yoshua Bengio, *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).

language model on a very large unlabeled dataset of raw text.[44] This computationally intensive and costly process is typically carried out by a large organization. The resulting pretrained model, which is often publicly released, can then be fine-tuned on a smaller dataset to optimize performance on a specific task. For example, Google's pretrained BERT model can be fine-tuned on case law and contracts to perform specialized legal tasks.[45] Compared to pretraining, fine-tuning is computationally inexpensive. As a result, developers can now conveniently adapt and deploy powerful pretrained language models in a wide range of applications.

B. THE GPT-3 REVOLUTION

In June 2020, OpenAI, a San Francisco-based AI technology company, released GPT-3, the then-largest pretrained language model.[46] This groundbreaking model marked a milestone in the development of AI, capturing the attention of both technologists and observers outside the computer science community.[47] Although GPT-3 is structurally similar to earlier language models, it differs in several important ways.[48]

First, unlike earlier language models, GPT-3 can perform many tasks without additional training or fine-tuning. For example, GPT-3 can, off-the-shelf, answer trivia questions, summarize text, and translate between languages.[49] In addition, users can teach the model to perform new tasks simply by providing instructions (in natural language) or presenting the model with

44. *See* Sebastian Ruder, *Recent Advances in Language Model Fine-tuning* (Feb. 24, 2021), https://ruder.io/recent-advances-lm-fine-tuning/. The resulting models have been recently, albeit controversially, described as "foundation models." *See* Bommasani et al., *supra* note 6, at 6–7; *see also* Marcus & Davis, *supra* note 6 (critiquing the term "foundation model"); Rishi Bommasani & Percy Liang, *Reflections on Foundation Models*, STANFORD UNIVERSITY HUMAN-CENTERED ARTIFICIAL INTELLIGENCE (Oct. 18, 2021), https: / /hai.stanford.edu /news / reflections-foundation-models (responding to, *inter alia*, critiques of the term "foundation model").

45. *See* Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras & Ion Androutsopoulos, *LEGAL-BERT: The Muppets Straight Out of Law School*, FINDINGS 2020 CONF. EMPIRICAL METHODS IN NLP 2898 (2020).

46. *See* Greg Brockman, Mira Murati, Peter Welinder & OpenAI, *OpenAI API*, OPENAI (June 11, 2020), https://openai.com/blog/openai-api/. References to GPT-3 are to the largest model in the GPT-3 family of models, which has 175 billion parameters. *See* Brown et al., *supra* note 8, at 8.

47. *See* Metz, *supra* note 8; Price, *supra* note 8.

48. For example, GPT-3 is structurally similar to its predecessor, GPT-2. *See* Alec Radford, Rewon Child, David Luan, Dario Amodei & Ilya Sutskever, *Language Models are Unsupervised Multitask Learners* (OpenAI Working Paper, Feb. 2019) (introducing the GPT-2 language model).

49. *See* Brown et al., *supra* note 8, at 10–29.

several examples of the desired task. This enables non-programmers to program the model.[50] For instance, the following prompt can teach GPT-3 to correct the grammar of an English text:[51]

> 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.[52] This highly intuitive form of learning—known as "few-shot learning"[53]—is arguably the hallmark of the technological watershed ushered in by GPT-3.[54] Some observers at the time of the model's release even suggested that GPT-3 is the closest attempt

50. *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 10, at 34 (discussing the impact of code generation on non-programmers).

51. This prompt is adapted from a template available in the OpenAI API at the time of the case study. For the most recent template for grammar correction available in the API, see *Grammar Correction*, OPENAI, https://beta.openai.com/examples/default-grammar (last visited Aug. 8, 2022).

52. Comparable prompts can be used to teach GPT-3 to construct headlines for news articles, write professional emails, and convert English instructions into computer code. *See* Yaser Martinez Palenzuel, Joshua Landau, Zoltán Szőgyényi, Sahar Mor, eshnil, CallmeMehdi, Mrinal Mohit, Scoder12 & Anurag Ramdasan, *Awesome GPT-3*, GITHUB (Sept. 29, 2020), https://github.com/elyase/awesome-gpt3; *see also Examples*, OPENAI, https://beta.openai.com/examples/ (last visited Aug. 8, 2022) (showcasing examples of GPT-3's performance).

53. The ability to learn from prompts is also known as *prompt-based learning*, *in-context learning* or *meta-learning*. For discussion of the limitations of few-shot learning, see Ethan Perez, Douwe Kiela & Kyunghyun Cho, *True Few-Shot Learning with Language Models*, 35TH CONF. NEURAL INFO. PROCESSING (2021).

54. The title of the paper introducing GPT-3 is *Language Models Are Few-Shot Learners*. *See* Brown et al., *supra* note 8, at 1.

to achieving artificial general intelligence, i.e., a machine that reaches or surpasses the intellectual capabilities of humans in a broad range of tasks.[55]

The second difference between GPT-3 and earlier language models—and the main factor accounting for its improved performance—is scale.[56] GPT-3 contains 175 billion parameters (i.e., model weights or coefficients), which is an order of magnitude more than the previously largest language model.[57]

---

55. *See* Julien Lauret, *GPT-3: The First Artificial General Intelligence?*, TOWARDS DATA SCI. (July 22, 2020), https: //towardsdatascience.com /gpt-3-the-first-artificial-general-intelligence-b8d9b38557a1 ("AGI . . . is so hard that there isn't a clear roadmap for achieving it . . . GPT-3 is the first model to shake that status-quo seriously."); *see also* Katherine Elkins & Jon Chun, *Can GPT-3 Pass a Writer's Turing Test?*, 5 J. CULTURAL ANALYTICS 1, 13 (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 [hereinafter Marcus & Davis, *GPT-3, Bloviator*]; Yann LeCun, FACEBOOK (Oct. 27, 2020), https://www.facebook.com/yann.lecun/posts/ 10157253205637143; *infra* note 117 (discussing the debate concerning whether language models can understand language). For a broader account of artificial general intelligence, see ARTIFICIAL GENERAL INTELLIGENCE 1–30 (Ben Goertzel & Cassio Pennachin eds., 2007); s*ee also infra* note 181 (discussing the challenges of AI alignment and control).

56. *See* Samira Abnar, Mostafa Dehghani, Behnam Neyshabur & Hanie Sedghi, *Exploring the Limits of Large Scale Pre-training*, ARXIV at 1 (Oct. 5, 2021), https://arxiv.org/abs/ 2110.02095. Notably, however, DeepMind's Retrieval-Enhanced Transformer (Retro), which was introduced a year and a half after the release of GPT-3, exhibits performance comparable to GPT-3 despite using 25 times fewer parameters. *See* Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen & Laurent Sifre, *Improving Language Models by Retrieving from Trillions of Tokens*, ARXIV (Dec. 8, 2021), https://arxiv.org/abs/2112.04426 (introducing a language model that can directly access a large database to enhance its predictions). For detailed analysis of the scaling language models, see Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu & Dario Amodei, *Scaling Laws for Neural Language Models*, ARXIV (Jan. 23, 2020), https://arxiv.org/abs/2001.08361; Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei & Sam McCandlish, *Scaling Laws for Autoregressive Generative Modeling*, ARXIV (Oct. 28, 2020), https://arxiv.org/abs/2010.14701.

57. The largest known language model prior to GPT-3 contained 17 billion parameters. *See* Corby Rosset, *Turing-NLG: A 17-Billion-Parameter Language Model by Microsoft*, MICROSOFT (Feb. 13, 2020), https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/. Shortly following the release of GPT-3, even larger language models were developed. *See, e.g.*, William Fedus, Barret Zoph & Noam Shazeer, *Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity*, ARXIV (Jan. 11,

Estimates of the cost of training GPT-3 are in the range of several million dollars.[58] The size of GPT-3's training data is also immense.[59] It includes over 570GB of raw web page data, online books corpora, and English-language Wikipedia[60]—which in aggregate contain approximately 57 billion times the number of words perceived in an average human lifetime.[61]

2021), https://arxiv.org/abs/2101.03961 (introducing the first language model known to exceed one trillion parameters).

58. *See* Kyle Wiggers, *OpenAI Launches an API to Commercialize Its Research*, VENTUREBEAT (June 11, 2020), https://venturebeat.com/2020/06/11/openai-launches-an-api-to-commercialize-its-research/ (estimating the training cost of GPT-3 to exceed $12 million); Chuan Li, *OpenAI's GPT-3 Language Model: A Technical Overview*, LAMBDA (June 3, 2020), https://lambdalabs.com/blog/demystifying-gpt-3/ (estimating the training cost of GPT-3 to exceed $4.6 million). These are estimates of the cost of compute only, not staff or other costs. For further discussion of the costs of building language models, see Or Sharir, Barak Peleg & Yoav Shoham, *The Cost of Training NLP Models: A Concise Overview*, ARXIV (Apr. 19, 2020), https://arxiv.org/abs/2004.08900.

59. However, the training corpora for subsequent models, such as DeepMind's Gopher, are even larger. *See* Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu & Geoffrey Irving, *Scaling Language Models: Methods, Analysis & Insights from Training* Gopher, DEEPMIND at 7 (Dec. 8, 2021), https://dpmd.ai/llm-gopher (using a training dataset that contains approximately 10.5TB of text).

60. This is roughly two orders of magnitude larger than the Corpus of Contemporary American English (COCA), which contains approximately one billion words. *See* CORPUS OF CONTEMPORARY AMERICAN ENGLISH, https://www.english-corpora.org/coca/ (last visited Aug. 8, 2022). However, words in COCA are annotated with additional linguistic information that facilitate using corpus linguistics techniques to analyze text. *See* 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.

61. *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.

The third feature that distinguishes GPT-3 from earlier language models is that it is proprietary. Prior to GPT-3, most large language models, such as Google's BERT and Facebook's RoBERTA, were publicly available.[62] Researchers were free to inspect the code and weights of these models and re-train or fine-tune them on new data. OpenAI, however, did not make the GPT-3 model publicly available.[63] Instead, OpenAI released an application programming interface (API) that interacts with the model,[64] which developers can pay to access.[65] This approach made GPT-3 the world's first commercial language model.[66]

Finally, GPT-3's groundbreaking capabilities have introduced new risks. Language models that are able to produce human-like text can be used to spread misinformation, generate spam, and achieve other nefarious purposes at unparalleled scale.[67] For instance, GPT-3 can be prompted to write in support of conspiracy theories, as illustrated in the following excerpt:[68]

---

62. *See* Devlin et al., *supra* note 33 (introducing the BERT language model); Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer & Veselin Stoyanov, *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, ARXIV (July 26, 2019), https: / /arxiv.org /abs /1907.11692 (introducing the RoBERTa language model). 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* Irene Solaiman, Jack Clark & Miles Brundage, *GPT-2: 1.5B Release*, OPENAI (Nov. 5, 2019), https://openai.com/blog/gpt-2-1-5b-release/.

63. The underlying model has been exclusively licensed to Microsoft. *See* Kevin Scott, *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/.

64. *See OpenAI API*, OPENAI, https://openai.com/api/ (last visited Sept. 28, 2022). Although fine-tuning was not available when the API was released, OpenAI has subsequently offered fine-tuning through its API. *See* Rachel Lim, Michael Wu & Luke Miller, *Customizing GPT-3 for Your Application*, OPENAI (Dec. 14, 2021), https://openai.com/blog/customized-gpt3/.

65. *See Pricing*, OPENAI, https://openai.com/api/pricing/ (last visited Aug. 8, 2022). Notably, for over a year following the release of GPT-3, developers seeking access to the API were subject to a waitlist, which was subsequently removed for most countries. *See* OpenAI, *OpenAI's API Now Available with No Waitlist*, OPENAI (Nov. 18, 2021), https: //openai.com / blog/api-no-waitlist/; *see also infra* Part V.D (discussing access to language model technology).

66. GPT-3, however, is no longer the only commercial language model. Following OpenAI, several other companies have released large language models through commercial APIs, which developers can pay to access. *See Pricing,* AI21, https://studio.ai21.com/pricing (last visited Aug. 8, 2022); *Pricing*, COHERE, https://cohere.ai/pricing (last visited Aug. 8, 2022).

67. *See infra* Part V.D (discussing the potential misuses of language models).

68. Kris McGuffie & Alex Newhouse, *The Radicalization Risks of GPT-3 and Advanced Neural Language Models*, ARXIV at 5 (Sept. 15, 2020), https://arxiv.org/abs/2009.06807.

> Q: Who is QAnon?
>
> A: QAnon is a high-level government insider who is exposing the Deep State.
>
> Q: Is QAnon really a military intelligence official?
>
> A: Yes. QAnon is a high-level government insider who is exposing the Deep State.

Taken together, the unprecedented capabilities and societal challenges presented by GPT-3 and other powerful language models could have profound implications for the deployment of AI in many fields. This Article and the case study it presents focus on the implications for the legal domain and, in particular, consumer contracts.

C. OPPORTUNITIES FOR LAW

Legal researchers have experimented with computational language models for decades. While they initially explored using language models to classify case law[69] and search legal databases,[70] researchers have more recently attempted to deploy language models in a broader range of legal applications,[71] including to review documents in e-discovery,[72] predict case outcomes,[73] and generate patent claims.[74] In the realm of contracts, researchers have used language

69. *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, 125–27, 130–31 (1997) (using methods based on term frequency-inverse document frequency (TF-IDF) to classify the texts of opinions in trade secret cases).

70. *See, e.g.*, Jacques Savoy, *Searching Information in Legal Hypertext Systems*, 2 AI & L. 205, 208–11 (1993) (describing the use of TF-IDF and related methods in legal information extraction).

71. *See* Ilias Chalkidis & Dimitrios Kampas, *Deep Learning in Law: Early Adaptation and Legal Word Embeddings Trained on Large Corpora*, 27 AI & L. 171, 174–96 (2019); Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu & Maosong Sun, *How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence*, PROC. 58TH ANN. MEETING ASS'N COMPUTATIONAL LINGUISTICS 5218, 5222–26 (2020).

72. *See, e.g.*, Ngoc Phuoc An Vo, C. Privault & Fabien Guillot, *Experimenting Word Embeddings in Assisting Legal Review*, PROC. 16TH INT'L CONF. AI & L. 189, 192–97 (2017) (using word embeddings to classify and retrieve information from litigation-related documents).

73. *See, e.g.*, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu & Maosong Sun, *Legal Judgment Prediction via Topological Learning*, PROC. 2018 CONF. EMPIRICAL METHODS IN NLP 3540, 3541–47 (2018) (employing language models to predict the outcomes of criminal cases based on descriptions of the case facts).

74. *See, e.g.*, Jieh-Sheng Lee & Jieh Hsiang, *Patent Claim Generation by Fine-Tuning OpenAI GPT-2*, 62 WORLD PATENT INFORMATION 101983, at 2–6 (2020) (evaluating the ability of GPT-2 to generate patent claims); *see also* S. Sean Tu, Amy Cyphert & Sam Perl, *Limits of Using of Artificial Intelligence and GPT-3 in Patent Prosecution*, TEX. TECH L. REV. (forthcoming)

models to detect unfair or invalid clauses,[75] identify contractual provisions,[76] and draft investment agreements.[77]

Today, language models that perform legal tasks are typically trained or fine-tuned on legal data.[78] For example, a language model designed to interpret tax legislation was trained on a corpus of tax cases and rulings.[79] 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,

(discussing the potential implications of using GPT-3 and other AI technologies to draft patent claims).

75. *See, e.g.*, Marco Lippi, Przemyslaw Palka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor & Paolo Torroni, *CLAUDETTE: An Automated Detector of Potentially Unfair Clauses in Online Terms of Service*, 27 AI & L. 117, 130–34 (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, 94–96 (2021); Alfonso Guarino, Nicola Lettieri, Delfina Malandrino & Rocco Zaccagnino, *A Machine Learning-Based Approach to Identify Unlawful Practices in Online Terms of Service: Analysis, Implementation and Evaluation*, 33 NEURAL COMPUTATION & APPLICATIONS 17569, 17575–80 (2021).

76. For example, language models can identify a contract's effective date, governing law, and jurisdiction, as well as more specialized provisions. *See* Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis & Ion Androutsopoulos, *Neural Contract Element Extraction Revisited: Letters from Sesame Street*, ARXIV at 2–5 (Feb. 22, 2021), https: / /arxiv.org /abs /2101.04355v2; Ilias Chalkidis, Ion Androutsopoulos & Achilleas Michos, *Obligation and Prohibition Extraction Using Hierarchical RNNs*, PROC. 56TH ANN. MEETING ASS'N COMPUTATIONAL LINGUISTICS 254, 255–57 (2018); Emad Elwany, Dave Moore & Gaurav Oberoi, *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 at 2–4 (2019); Spyretta Leivaditi, Julien Rossi & Evangelos Kanoulas, *A Benchmark for Lease Contract Review*, ARXIV at 6–9 (Oct. 20, 2020), https: / /arxiv.org /abs /2010.10386; Dan Hendrycks, Collin Burns, Anya Chen & Spencer Ball, *CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review*, 35TH CONF. NEURAL INFO. PROCESSING SYS. DATASETS AND BENCHMARKS TRACK at 3–8 (2021) [hereinafter Hendrycks et al., *CUAD*].

77. *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, 230–31 (2017) (using language models to generate clauses for bilateral investment treaties).

78. *See* Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz & Nikolaos Aletras, *LexGLUE: A Benchmark Dataset for Legal Language Understanding in English*, ARXIV at 5–6 (Sept. 3, 2021), https: / /arxiv.org / pdf /2104.07782.pdf [hereinafter Chalkidis et al., *LexGLUE*] (surveying recent work on fine-tuning language models in the legal domain).

79. *See* Nils Holzenberger, Andrew Blair-Stanek & Benjamin Van Durme, *A Dataset for Statutory Reasoning in Tax Law Entailment and Question Answering*, PROC. 2020 NATURAL LEGAL LANGUAGE PROCESSING WORKSHOP at 4–5 (2020).

especially if it involves recruiting legal experts.[80] Consequently, few organizations can effectively deploy language models in legal settings.

GPT-3's powerful out-of-the-box performance could signal a change. Within weeks of its release, developers had used GPT-3 to prepare legal documents[81] and translate legal jargon into plain English[82]—without any additional training or fine-tuning. For example, one user provided GPT-3 with the text of Section 2801 of the U.S. Tax Code and asked the model to summarize the provision. Remarkably, GPT-3 responded with the following: "If you get money from someone who is not living in America anymore because they gave up their citizenship, you have to pay extra taxes on it."[83] This pithy summary, although imperfect, captures the salient principle expressed in the provision.

Language models like GPT-3 present significant commercial opportunities for lawyers and legal technology firms.[84] For example, language models could help lawyers conduct legal research more efficiently,[85] accelerate transactional

80. *See* Hendrycks et al., *CUAD*, *supra* note 76, at 1–2 (estimating that the cost of creating approximately 13,000 annotations for approximately 500 contracts exceeds $2 million); *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).

81. *See* Jervis, *supra* note 13 (using GPT-3 to generate requests for admission).

82. *See* Tefula, *supra* note 14. Others have used GPT-3 to translate plain English into legalese. *See* Ed Leon Klinger (@edleonklinger), TWITTER (July 18, 2020, 1:19AM), https://twitter.com/edleonklinger/status /1284251420544372737 (offering several examples of GPT-3 translating claims expressed in plain language, mainly relating to property disputes, into "lawyer speak").

83. *See* Gross, *supra* note 12.

84. *See* Cyphert, *supra* note 16, at 403–05, 419–23; DeFelice, *supra* note 16; Hill, *supra* note 16; *GPT-3 – A Game Changer for Legal Tech?,* ARTIFICIAL LAWYER (July 29, 2020), https://www.artificiallawyer.com/2020/07/29/gpt-3-a-game-changer-for-legal-tech/. But arguably OpenAI's control over the GPT-3 model and API precludes companies from gaining a competitive advantage. *See* 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/ ("[I]f OpenAI improves GPT-3 over time . . . it will immediately deliver the upgraded model to all API clients at the same time. The language model levels the ground for everyone. Any application you build on GPT-3 can easily be cloned by another developer."). 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, 22–32 (2020); Robert Dale, *Law and Word Order: NLP in Legal Tech*, 25 NATURAL LANGUAGE ENG'G 211, 212–17 (2019); Alarie et al., *supra* note 15, at 115–20; Ashley, *supra* note 80, at 1119.

85. 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.

drafting,[86] and generate synthetic legal data to train other machine learning models to perform legal tasks.[87] In time, language models could potentially automate much of the work carried out by paralegals and junior associates.

In addition to supporting professional legal services providers, language models could also directly assist consumers.[88] For example, language models could democratize legal knowledge by explaining the meaning of legal texts.[89] Meanwhile, a language model trained to evaluate the strengths and weaknesses of legal arguments could assist *pro se* litigants in assessing the merits of their case before going to court. By improving access to justice in these ways, language models could empower consumers who cannot afford traditional legal services.[90]

Finally, language models have the potential to impact legal scholarship. Researchers in the emerging field of computational legal studies,[91] many of whom use language models to study legal texts,[92] could benefit from more powerful and programmable models, such as GPT-3. Other researchers have

86. For discussion concerning the automation of transactional drafting, see 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).

87. *See* Bommasani et al., *supra* note 6, at 63, 66 (describing recent efforts to create legal NLP benchmarks through automation).

88. *See* Bommasani et al., *supra* note 6, at 59–61; Cyphert, *supra* note 16, at 421–23. For further discussion of how legal automation could impact access to justice, see Remus & Levy, *supra* note 15, 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).

89. *See* Arbel & Becher, *supra* note 19, at 94–109 (showing that GPT-3 can simplify, personalize, interpret, and benchmark contracts).

90. *See generally* DEBORAH L. RHODE, ACCESS TO JUSTICE ch. 5 (2004) (discussing the legal needs of low-income communities). For examination 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).

91. *See* LAW AS DATA: COMPUTATION, TEXT, AND THE FUTURE OF LEGAL ANALYSIS (Michael A. Livermore & Daniel N. Rockmore eds. 2019); COMPUTATIONAL LEGAL STUDIES: THE PROMISE AND CHALLENGE OF DATA-DRIVEN RESEARCH (Ryan Whalen ed., 2020).

92. *See* Jens Frankenreiter & Michael A. Livermore, *Computational Methods in Legal Analysis*, 16 ANNU. REV. L. & SOC. SCI. 39, 43–44 (2020) (surveying studies that employ language models to analyze judicial opinions, public comments received by administrative agencies, and other legal texts).

already used GPT-3 more directly: to contribute to, or co-author, academic articles.[93]

Despite these opportunities, the performance of GPT-3 on legal tasks has not been rigorously tested. Although impressive, illustrations of GPT-3's outputs may be subject to selection bias, i.e., cherry-picking instances of impressive performance.[94] 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.[95]

Turning to GPT-3's ability to understand legal texts, it is worth noting that the model appears to perform poorly on general-purpose, non-law reading

93. *See* Benjamin Alarie, Arthur Cockfield & GPT-3, *Will Machines Replace Us? Machine-Authored Texts and the Future of Scholarship*, 3 LAW, TECH & HUMANS 5, 5 (2021) (including GPT-3 as a co-author); Arbel & Becher, *supra* note 19, at 146 (indicating that GPT-3 wrote the article's conclusion).

94. *See, e.g.*, Arbel & Becher, *supra* note 19, at 118 ("[A]ll the examples used were cherrypicked. Such a selection is necessary to develop a sense of tomorrow's capabilities today. However, cherry-picking does run the risk of exaggerating the power and accuracy of the technology."); GPT-3, *A Robot Wrote This Entire Article. Are You Scared Yet, Human?*, THE GUARDIAN (Sept. 8, 2020), https: //www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3 ("GPT-3 produced eight different outputs, or essays. . . . The Guardian could have just run one of the essays in its entirety. However, we chose instead to pick the best parts of each, in order to capture the different styles and registers of the AI."); Kevin Roose, *A Robot Wrote This Book Review*, N.Y. TIMES (Nov. 21, 2021), https: //www.nytimes.com/2021/11/21/books/review/the-age-of-ai-henry-kissinger-eric-schmidt-daniel-huttenlocher.html (describing the multiple attempts needed to use Sudowrite, a program powered by GPT-3, to write a book review). *See id.* ("On the first attempt, it spit out a series of run-on sentences that hinted that GPT-3 had gotten stuck in some kind of odd, recursive loop. . . . A few tries later, it seemed to give up on the task of book reviewing altogether, and started merely listing the names of tech companies. . . . But it warmed up quickly, and within a few minutes, the A.I. was coming up with impressively cogent paragraphs of analysis."). By contrast, other users may have cherry-picked *problematic* outputs produced by GPT-3. *See* 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 ("[W]e pre-tested [the experiments] on the "AI Dungeon" game which is powered by some version of GPT-3, and we excluded those for which "AI Dungeon" gave reasonable answers.").

95. *See* Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song & Jacob Steinhardt, *Measuring Massive Multitask Language Understanding*, 9TH INT'L CONF. LEARNING REPRESENTATIONS at 6 (2021) [hereinafter Hendrycks et al., *Measuring Understanding*]. By comparison, DeepMind's Gopher model, which was introduced a year and a half following the release of GPT-3, exhibits improved accuracy on these tests. *See* Rae et al., *supra* note 59, at 67.

comprehension tasks.[96] However, there are notable differences between the language used in those tasks and legal language.[97] While one might assume that legal language is more technical or verbose than non-legal language—and that, therefore, GPT-3 is likely to perform worse on legal texts than on non-legal texts—given the model's unconventional method of learning,[98] it is problematic to make this assumption. To evaluate the degree to which the model can understand legal texts, we need to test the model on legal texts.

III. EXPERIMENTAL DESIGN

The following Part outlines the methodology employed in the case study. I begin by describing the contract questions presented to GPT-3. Next, I explain the criteria used to evaluate the model's performance. Finally, I discuss several methodological challenges and limitations.

A. CONTRACT QUESTIONS

In the field of natural language processing (NLP), it is instructive to evaluate a model's performance in real-world applications.[99] For example, testing whether GPT-3 can explain the meaning of contractual provisions sheds light on the degree to which the model understands contracts.[100] This

96. *See* Brown et al., *supra* note 8, at 18.

97. Legal language has many distinctive features, including specialized terms of art and formal expressions. *See* DAVID MELLINKOFF, THE LANGUAGE OF THE LAW chs. 2–3 (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 pt. 1.1 (5th ed. 2018). The distinctive features of legal language also present challenges for machine learning. *See* Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson & Daniel E. Ho, *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*]; Chalkidis et al., *LexGLUE*, *supra* note 78, at 2.

98. *See* Hendrycks et al., *Measuring Understanding*, *supra* note 95, 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-blind spots and has capabilities that are lopsided.

99. This is known as *extrinsic evaluation*. *See* JURAFSKY & MARTIN, *supra* note 32, at 35.

100. *See* Arbel & Becher, *supra* note 19, at 94–109.

method of evaluation, however, faces a problem: it is difficult to objectively assess the quality of responses to open-ended questions. The problem is particularly acute in the legal domain. For instance, what makes one explanation of a contractual provision "better" than another (where, for the sake of argument, both are accurate)? Unlike the fact-based trivia questions commonly used in NLP benchmark datasets, there is not necessarily a single "correct" answer to legal questions.[101] Even if there are specific criteria for assessing the quality of responses, different people (or AI systems) bring different perspectives and may reach different conclusions.[102]

In light of these challenges, evaluations of legal AI systems often use yes/no questions with relatively uncontroversial answers.[103] The case study presented in this Article adopts a similar method. To test GPT-3's ability to understand consumer contracts, 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).[104] The questions relate to a wide

101. *See generally* H.L.A. HART, THE CONCEPT OF LAW 124–25 (1961) (describing the "open texture" of legal rules and language). *Compare* Ronald Dworkin, *No Right Answer?*, 53 N.Y.U. L. REV. 1, 30–31 (1978), republished in RONALD DWORKIN, A MATTER OF PRINCIPLE 119–45 (1985) (arguing that there is a single correct answer for the overwhelming majority of legal cases); *see also* Brian Bix, *H.L.A. Hart and the "Open Texture" of Language*, 10 LAW & PHIL. 51, 52–55 (1991), republished in BRIAN BIX, LAW, LANGUAGE, AND LEGAL DETERMINACY ch. 1 (1995) (examining Hart's notion of "open texture"); *see also id.* at ch. 4 (discussing Dworkin's right answer thesis). Law's "open texture" or potential indeterminacy can impact the development of machine learning in the legal domain. *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) (suggesting that, because there is no consensus that legal questions have single correct answers, it is difficult to establish a "ground truth" to train machine learning models to perform legal tasks).

102. 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) (describing different forms of ambiguity in contractual language); E. Allan Farnsworth, *"Meaning" in the Law of Contracts*, 76 YALE L.J. 939, 952–65 (1967) (distinguishing between ambiguity and vagueness). Additionally, it is often difficult to assess the quality of legal advice (at least when provided by a human lawyer). *See, e.g.*, Douglas E. Rosenthal, *Evaluating the Competence of Lawyers*, 11 LAW & SOC'Y REV. 257, 260–70 (1976) (critically appraising methods for evaluating lawyer competence).

103. *See, e.g.*, 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, 187–88 (2019) (testing an AI question answering system on yes /no questions pertaining to commercial contracts); Juliano Rabelo, Mi-Young Kim, Randy Goebel, Yoshinobu Kano, Masaharu Yoshioka & Ken Satoh, *Summary of the Competition on Legal Information Extraction /Entailment (COLIEE) 2021* at 5, hosted at 18TH INT'L CONF. AI & L. (2021).

104. 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,

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 terms of service. Table 1 (below) displays a sample of the questions.[105]

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

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://web.archive.org/web/20201117101234/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, holding companies. For example, the terms of service for Yahoo.com and ESPN.com refer to Verizon and Disney, respectively. All terms of service were accessed during November 10–17, 2020, copies of which are on file with the author.

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

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 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 or relying on data memorized during pretraining.[106] If, however, accuracy is higher when GPT-3 is shown both the contract and the question, this would suggest that GPT-3 uses the contract to answer the questions and does not simply respond to cues in the questions or rely on data memorized during pretraining.

2. Calibration

While high accuracy is necessary for strong performance, it is not sufficient. For a model to be reliable, it must be both accurate and well-calibrated, i.e., it should assign high probabilities to its correct predictions and low probabilities to its incorrect predictions.[107] In other words, there should be a strong positive correlation between the model's 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 retained. 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.

To assess a model's calibration, we need to measure a model's confidence in its predictions. As explained, GPT-3 operates by predicting the next word in a sequence.[108] It assigns a probability to what it calculates to be the most likely next word. For example, following a certain yes/no question, GPT-3 might assign a 43% probability to the next word being "yes," a 29% probability

106. *See infra* Part IV.C (discussing the memorization of training data).

107. Put differently, a model is well-calibrated if its confidence in a prediction (expressed as a probability) is a good estimate of the actual probability that the prediction is correct. *See generally* Chuan Guo, Geoff Pleiss, Yu Sun & Kilian Q. Weinberger, *On Calibration of Modern Neural Networks*, PROC. 34TH INT'L CONF. MACH. LEARNING 1321 (2017) (outlining several methods for evaluating calibration).

108. *See supra* Parts II.A–B (offering a brief primer on the operation of language models, including GPT-3). On a technical note, predictions are of tokens (not words) and probabilities are log probabilities (not raw probabilities).

to the next word being "no," and 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 (43%) being assigned to the output in both cases, GPT-3 appears less confident in its prediction in Case 2—because the difference between the two probabilities in Case 2 is only one percentage point (43% minus 42%), as opposed to 14 percentage points in 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 the alternative answer. Accordingly, in Case 1 the confidence score would equal 14 and in Case 2 the confidence score would equal 1.

Of course, there are many other ways to measure differences in confidence. As a robustness check, I also measure the *ratio* between (i) the probability assigned to the output, and (ii) the probability assigned to the alternative answer. According to this measure, which aims to capture the *relative* difference between the probabilities, in Case 1 the confidence score would be 1.48 (43/29) and in Case 2 the confidence score would be 1.02 (43/42).

To accommodate different perspectives on which measure best captures the model's confidence in its predictions, the study reports all three of the aforementioned measures, namely: (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).*

With these measures of confidence in hand, we can evaluate GPT-3's calibration, i.e., the correlation between the model's accuracy and the model's confidence in its predictions. If the correlation between accuracy and confidence is positive, this would suggest that GPT-3 is well-calibrated, i.e., more confident in its correct responses than in its incorrect responses.

Alternatively, if there were no correlation between accuracy and confidence (or if the correlation were negative), this would suggest that GPT-3 is poorly calibrated and therefore highly unreliable.

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 are also 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

The following section 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.[109] 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.[110] 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 as people interact with them. For example, if multiple questions were

109. *See, e.g.*, Brown et al., *supra* note 8, at 29–33, 43–44 (investigating whether GPT-3's performance on certain benchmarks was contaminated by its training data).

110. *See infra* Part IV.C (discussing the memorization of training data).

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 as a continuous dialogue), such that performance on each question was independent of performance on other questions.

A further challenge concerns the randomness in the outputs of neural language models.[111] 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[112] called "temperature."[113] 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, 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 minimizes randomness in the model's predictions and, thereby, improves replicability.[114]

Finally, some publicly available 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 presented here takes several steps to improve transparency. First, all questions presented to GPT-3 are listed in the Online Appendix.[115] Second, the entire priming prompt is disclosed in Part A of 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.

---

111. The technical term is *stochasticity*. *See* ALLEN B. DOWNEY, THINK BAYES 66 (1st ed. 2013) (explaining that stochasticity describes a "model [that] has some kind of randomness in it").

112. In machine learning, hyperparameters are variables that users can manually set to control a model's training or operation.

113. *See* Geoffrey Hinton, Oriol Vinyals & Jeff Dean, *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).

114. *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).

115. *Online Appendix*, https://app.box.com/s/zrbgy2yepvlclfg88i2o7dhws0d919li (last visited Sept. 29, 2022).

2. Limitations

Despite addressing the above challenges, the case study has several notable limitations.

First, the case study evaluates only the *behavior* of GPT-3, that is, the model's observable outputs.[116] 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 a manner analogous to how humans understand language.[117] This important debate is beyond the scope of this Article.

Second, the dataset used in the case study is smaller and, by design, less comprehensive than general-purpose NLP benchmark datasets for question

116. Anthropomorphic references to language models in this Article, such as "understand" and "memorize," are used only by way of analogy, and do not suggest that language models possess human-like capabilities. *See generally* Melanie Mitchell, *Why AI Is Harder Than We Think*, ARXIV 5 (Apr. 26, 2021), https://arxiv.org/abs/2104.12871 (arguing that anthropomorphic references to AI systems can be misleading); Weidinger et al., *supra* note 20, at 29–30 (suggesting that the anthropomorphization of language models can lead to overreliance on, or unsafe use of, these models).

117. *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, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto & Joseph Turian, *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). Several commentators have argued that GPT-3 cannot understand language. *See* Marcus & Davis, *GPT-3, Bloviator*, *supra* note 55 ("All GPT-3 really has is a tunnel-vision understanding of how words relate to one another; it does not, from all those words, ever infer anything about the blooming, buzzing world. . . . It learns correlations between words, and nothing more."); 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 (July 30, 2020), https: / /dailynous.com /2020 /07 /30 /philosophers-gpt-3 /#vallor; Melanie Mitchell, *What Does It Mean for AI to Understand?*, QUANTA MAGAZINE (Dec. 16, 2021), https:/ /www.quantamagazine.org/what-does-it-mean-for-ai-to-understand-20211216/; *see also* 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). Other commentators, however, are somewhat more optimistic about the prospect of language models understanding language. *See* 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; Blaise Aguera y Arcas, *Do Large Language Models Understand Us?*, MEDIUM (Dec. 17, 2021), https://medium.com/@blaisea/do-large-language-models-understand-us-6f881d6d8e75. For a summary of this debate, see Bommasani et al., *supra* note 6, at 48–52.

answering. For example, general-purpose datasets may contain thousands of questions, while the case study includes only 200 questions.[118] Some general-purpose datasets also include unanswerable questions, but the case study does not.[119] These limitations, however, are not unusual for NLP datasets in the legal domain. For example, several popular legal NLP datasets contain fewer than 200 questions[120] and do not include unanswerable questions.[121] Nevertheless, NLP researchers and practitioners should aspire to create larger and more diverse legal datasets in the future.[122]

Third, the questions in the case study were prepared by a single attorney (the author). The selection of questions and their evaluation may be influenced

118. For example, the WebQuestions dataset consists of 6,642 questions and the TriviaQA dataset consists of over 650,000 questions. *See* Jonathan Berant, Andrew Chou, Roy Frostig & Percy Liang, *Semantic Parsing on Freebase from Question-Answer Pairs*, PROC. 2013 CONF. EMPIRICAL METHODS IN NLP (2013); Mandar Joshi, Eunsol Choi, Daniel S. Weld & Luke Zettlemoyer, *TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension*, PROC. 55TH ANN. MEETING ASS'N COMPUTATIONAL LINGUISTICS 1601 (2017).

119. The inclusion of unanswerable questions—i.e., questions for which there is no answer or the answer to which cannot be found in the corresponding document—can be instructive because inappropriate responses to such questions cast doubt on a model's reliability. *See, e.g.*, Pranav Rajpurkar, Robin Jia & Percy Liang, *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).

120. For example, the Jurisprudence and International Law tests used by Hendrycks et al. *Measuring Understanding*, *supra* note 95 consist of 108 and 121 multiple-choice questions, respectively. *See* Dan Hendrycks, *Test,* GITHUB, https: / /github.com /hendrycks /test (last visited Aug. 8, 2022). Similarly, the 2020 COLIEE competition's legal textual entailment task test dataset contains 80 yes /no questions. *See* Rabelo et al., *supra* note 103, at 64.

121. *See* Zheng et al., *CaseHOLD*, *supra* note 97, 161–62; Lippi et al., *supra* note 75, at 131–33. *But see* Hendrycks et al., *CUAD*, *supra* note 76, at 5 (including some unanswerable questions).

122. Recent efforts, which post-date the case study presented in this Article, include Hendrycks et al., *CUAD*, *supra* note 76, at 3–5; Zheng et al., *CaseHOLD*, *supra* note 97, at 161–62; Chalkidis et al., *LexGLUE*, *supra* note 78, at 3–5; Yuta Koreeda & Christopher D. Manning, *ContractNLI: A Dataset for Document-level Natural Language Inference for Contracts*, FINDINGS 2021 CONF. EMPIRICAL METHODS IN NLP 1907, 1908–11 (2021); *see also* Bommasani et al., *supra* note 6, at 66.

> [L]arger legal benchmark datasets may be necessary to observe further gains from applying transfer learning techniques to foundation models. However, creating benchmark datasets for tasks that are legally meaningful and difficult from an NLP perspective can itself be challenging, as human expert annotation can be costly and automated methods . . . can fail to account for unique aspects of legal text . . . many existing legal domain-specific labeled datasets are small, not publicly available, or reflect simpler tasks that have been solved by methods often pre-dating the development of foundation models.

by that attorney's professional background and experience. Although the questions aim to be as objective as possible, given that contract interpretation always involves a degree of subjective judgment, other legal practitioners or researchers may answer some of these questions differently.[123]

Finally, the scope of the case study is limited by its narrow objective. The study aims only 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 general legal knowledge or its performance on other legal tasks. Nor does the case study attempt to compare the performance of GPT-3 with the performance of human lawyers or examine how people are likely to interact with these models in practice.[124] Future work will need to grapple with these important issues.

IV. RESULTS AND DISCUSSION

This Part presents the results of the case study. First, I discuss the performance of GPT-3 on the test questions. Next, I examine whether certain characteristics of the contracts and questions presented to GPT-3 are associated with an increase or decrease in performance. Finally, I consider whether variations in question-wording impact performance.

123. For example, 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 introduce the same evaluation challenges posed by open-ended questions; *see also supra* note 102 (discussing the inherent vagueness and ambiguity of contractual language).

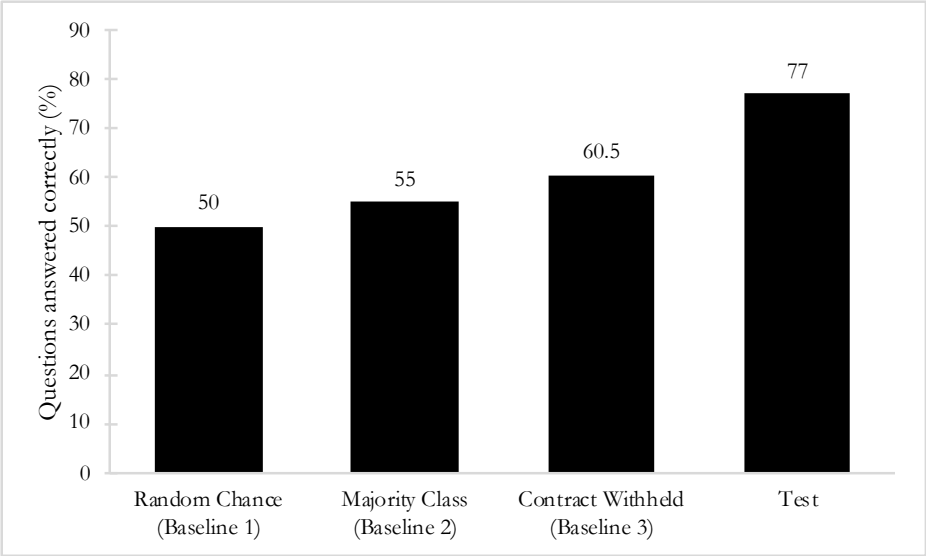
124. *See generally* Kawin Ethayarajh & Dan Jurafsky, *Utility is in the Eye of the User: A Critique of NLP Leaderboards*, PROC. 2020 CONF. EMPIRICAL METHODS IN NLP 4846 (2020) (emphasizing the importance of real-world evaluation for NLP technologies). However, new benchmarks and evaluation platforms are being developed to better simulate real-world conditions. *See, e.g.*, Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts & Adina Williams, *Dynabench: Rethinking Benchmarking in NLP*, PROC. 2021 ANN. CONF. N. AM. CH. ASS'N COMPUTATIONAL LINGUISTICS 4110 (2021); *see also* Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin & Sameer Singh, *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).

A. PERFORMANCE

1. Accuracy

GPT-3 answered correctly 77% of the questions in the case study.[125] In terms of accuracy, performance exceeded all three baselines, as illustrated in Figure 1 (below). 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 respond to cues in the questions or rely on data memorized during pretraining.[126]

Figure 1: Comparison of Accuracy with Baselines



125. GPT-3 did not provide a yes /no response to four questions and, instead, outputted the name of the relevant company. Given that these responses fail to answer the questions, 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.

126. *See infra* Part IV.C (discussing the memorization of training data).

2. Calibration

In terms of calibration, there was a positive correlation between the model's accuracy and the model's confidence in its predictions.[127] 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, all things being equal, encourages us to trust the model's predictions.

3. Overall Performance

In terms of overall performance, which accounts for both accuracy and calibration,[128] average overall performance in the test exceeded average overall performance in the contract withheld baseline across all three measures of overall performance.[129] Surpassing the contract withheld baseline in overall performance provides further suggestive evidence that GPT-3 uses the contracts to answer the questions.

The performance of GPT-3 in the case study, described thus far, appears to be encouraging. Despite the unintuitive and unhuman-like way in which language models operate—predicting the next word in a sequence—GPT-3 answered correctly nearly four of five questions and was generally well-calibrated. These results suggest that, contrary to conventional wisdom,[130] a language model can answer questions about contracts without extracting specific textual information from the document. GPT-3 merely predicts the next word in a sequence and, more often than not, correctly answers the test questions. 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 to effectively carry out legal tasks.[131]

---

127. The correlation coefficients (Pearson's *r*) between accuracy and the measures of confidence (described in Part III.B) are, respectively: $r = 0.226^{**}$ (Measure 1); $r = 0.258^{***}$ (Measure 2); and $r = 0.205^{**}$ (Measure 3), where * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. The complete calibration results are shown in Figures 4A, 4B, and 4C in the Appendix.

128. *See supra* Part III.B (explaining how overall performance is calculated).

129. The overall performance scores appear in Table 8A in the Appendix.

130. *See* Ryan Catterwell, *Automation in Contract Interpretation*, 12 L. INNOVATION & TECH. 81, 100 (2020) (examining how machine learning can be used to extract information from contracts); Ashley, *supra* note 80, at 1137 ("[T]ext analytics cannot yet *extract* information implicit in the [contract] texts, at least not without more knowledge and a computational model of the planned transaction." (emphasis added)); *see also* JURAFSKY & MARTIN, *supra* note 32, at 471–83 (discussing the operation of information retrieval systems).

131. *See* Zheng et al., *CaseHOLD*, *supra* note 97, at 167 ("Our results suggest that for other high[ly] [domain-specific] and adequately difficult legal tasks, experimentation with custom, task relevant approaches, such as leveraging corpora from task-specific domains and applying

B. ANTI-CONSUMER BIAS

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 other questions?

The contractual provisions in the terms of service used in the case study can be categorized as follows.[132] 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 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). Correspondingly, questions relating to pro-company provisions, pro-consumer provisions, and neutral provisions can be classified as *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 (below) provides an example from each category.[133]

---

tokenization / sentence segmentation tailored to the characteristics of in-domain text, may yield substantial gains."). *Compare* Hendrycks et al., *Measuring Understanding*, *supra* note 95, at 8 ("[W]hile additional pretraining on relevant high quality [legal] text can help, it may not be enough to substantially increase the performance of current models."). Note, however, that GPT-3's training data are likely to include many online terms of service, which are precisely the kind of legal document used in the case study.

132. Numerous studies have proposed 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); Eyal Zamir & Yuval Farkash, *Standard Form Contracts: Empirical Studies, Normative Implications, and the Fragmentation of Legal Scholarship: Comments on Florencia Marotta-Wurgler's Studies*, 12 JRSLM. REV. LEGAL STUD. 137, 149 (2015) (discussing the limitations of Marotta-Wurgler's classification scheme); *see also* Lippi et al., *supra* note 75, at 121–27 (proposing a classification scheme based on EU consumer law); Guarino et al., *supra* note 75, at 17574–77 (expanding the classification scheme proposed by Lippi et al., *supra* note 75).

133. 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 (describing the length of texts presented to the model).

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

In Table 2 (above), the first provision 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 it does not favor the interests of either the company or the consumer; it simply stipulates who may access the services.

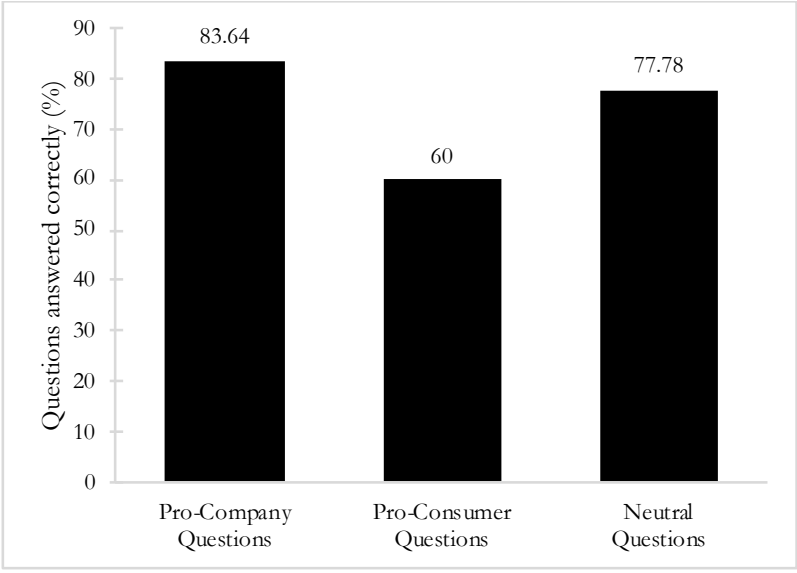
Before proceeding to discuss the results, it is worth noting that this classification invariably involves a degree of subjective judgment. For example, some might argue that a provision that grants a company the right to perform an action that could benefit a consumer (such as backing up personal data) without that consumer's consent should be classified as pro-consumer, not pro-company. Others might argue that the appropriate classification necessarily depends on the facts of the situation and an individual consumer's preferences.[134]

With this caveat in mind, how did GPT-3 perform across the different question categories? As illustrated in Figure 2 (below), the model's accuracy on

134. However, introducing such ambiguous categories would render the classification scheme unusable.

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.[135]

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 responding to questions that it answered correctly than to questions that it answered incorrectly),[136] the model was not well-calibrated on pro-consumer questions (i.e., there was no correlation between the model's confidence in responding to pro-consumers questions and the model's chances of correctly answering pro-consumer questions).[137] Moreover, a disproportionately large number of questions with the poorest overall performance scores—that is, where GPT-3 answered incorrectly and with high

135. These results are consistent with the overall performance scores (that account for both accuracy and calibration), which are listed in Table 8B in the Appendix.

136. *See supra* Part IV.A (finding that there was generally a positive correlation between the model's accuracy and the model's confidence in responding to questions in the case study).

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

confidence—were pro-consumer questions.[138] A sample of questions yielding high confidence anti-consumer mistakes is shown in Table 3 (below).

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 applicable terms of service is "yes." The terms explicitly state that if Disney provides consent, then a consumer is permitted to copy a Disney product.[139] 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 systematically differ from 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 question category (the variable of interest) and GPT-3's performance in the case study, we need to control for other factors that could potentially influence the model's performance. Accordingly, the study employs 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

138. 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.

139. Of course, in reality, such consent might not be especially forthcoming.

(the independent variables). If the variable of interest has an independent effect on performance, then we would expect to see a statistically significant relationship between the variable of interest and performance, even after controlling for other variables.

The regression analysis controls for the following variables:

(i) *Company Name in Question*. This variable describes whether the name of the relevant company appears in the question.[140] 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 the company that is contained in the model's training data, thereby improving performance.[141]

(ii) *Length of Contract*. This variable describes the length of the contract excerpt shown to GPT-3.[142] 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.[143]

(iii) *Readability of Contract*. This variable describes the ease with which a human reader can understand the contract excerpt.[144] The rationale for

140. 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).

141. *See infra* Part IV.C (discussing the memorization of training data).

142. 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 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.

143. *See supra* Part II.A (discussing long-range dependencies); *see also* Allison Hegel, Marina Shah, Genevieve Peaslee, Brendan Roof & Emad Elwany, *The Law of Large Documents: Understanding the Structure of Legal Contracts Using Visual Cues*, KDD DOC. INTELL. WORKSHOP at 4 (2021) (showing that a model's ability to identify the governing law of contracts falls as document length increases).

144. That is, the Flesch Reading Ease score for the contract excerpt, which is calculated as follows: $206.835 – (1.015 \times ASL) – (84.6 \times ASW)$, 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, THOMAS G. STICHT, LYNN C. FOX & PATRICK J. FORD, HUM. RES. RSCH. ORG., METHODOLOGIES FOR DETERMINING READING REQUIREMENTS OF MILITARY OCCUPATIONAL SPECIALTIES, TECHNICAL REPORT NO. 73-5, 15 (1973).

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.[145] The rationale for 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.[146]

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.[147] However, testing every plausible additional variable is beyond the scope of this Article and would introduce statistical problems.[148]

Table 4 (below) displays the results of three specifications of the OLS model, regressing the three measures of overall performance on the above variables.[149] 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.

145. 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 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. 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.

146. *See* Catterwell, *supra* note 130, at 100; Ashley, *supra* note 80, at 1137; JURAFSKY & MARTIN, *supra* note 32, at 471–83.

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

148. *See id.* at 516–18 (explaining that OLS is unreliable when the number of independent variables is large relative to the sample size).

149. *See supra* Part III.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 the model replicates a systematic anti-consumer bias in its training data.[150] 82% of GPT-3's training data is comprised

150. Training data often contain societal biases that affect a language model's parameters. *See* Brown et al., *supra* note 8, at 36–39 (analyzing GPT-3's biases related to race, gender, and religion); *see also* Abubakar Abid, Maheen Farooqi & James Zou, *Persistent Anti-Muslim Bias in Large Language Models*, ARXIV (Jan. 18, 2021), https://arxiv.org/abs/2101.05783 (finding, *inter alia*, that prompts containing the word "Muslim" result in GPT-3 producing a disproportionate number of violence-related outputs); Abubakar Abid, Maheen Farooqi & James Zou, *Large Language Models Associate Muslims with Violence*, 3 NATURE MACH. INTELL. 461 (2021). For further discussion of societal bias in NLP, see Su Lin Blodgett, Solon Barocas, Hal Daumé III & Hanna Wallach, *Language (Technology) is Power: A Critical Survey of "Bias" in NLP*, PROC. 58TH ANN. MEETING ASS'N COMPUTATIONAL LINGUISTICS 5454 (2020); Emily

of webpages extracted from the Common Crawl and Webtext2 datasets, which are likely to include many online terms of service and other consumer contracts.[151] These documents are typically drafted by company counsel and designed to favor the rights and interests of the relevant company, not consumers.[152] By performing worse on pro-consumer questions and producing a disproportionate number of anti-consumer responses, the model arguably overfits its training data.[153]

Another possibility is bias in *prompt-based learning*. GPT-3 might have performed worse on pro-consumer questions because of biases learned from

---

Sheng, Kai-Wei Chang, Premkumar Natarajan & Nanyun Peng, *Societal Biases in Language Generation: Progress and Challenges*, PROC. 59TH ANN. MEETING ASS'N COMPUTATIONAL LINGUISTICS 4275 (2021); Weidinger et al., *supra* note 20, at 9, 11–18.

151. For examination 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*, PROC. 59TH ANN. MEETING ASS'N COMPUTATIONAL LINGUISTICS (2021); Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell & Matt Gardner, *Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus*, PROC. 2020 CONF. EMPIRICAL METHODS IN NLP (2021); Abeba Birhane, Vinay Uday Prabhu & Emmanuel Kahembwe, *Multimodal Datasets: Misogyny, Pornography, and Malignant Stereotypes*, ARXIV (Oct. 5, 2021), https://arxiv.org/abs/2110.01963.

152. *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); *see also* Marotta-Wurgler, *What's in a Standard Form Contract?*, *supra* note 132, at 702–12 (finding that end-user software license agreements generally favor the interests of software companies); Yehuda Adar & Shmuel I. Becher, *Ending the License to Exploit: Administrative Oversight of Consumer Contracts*, 62 B.C. L. REV. 2405, 2413–14 (2022) (describing consumer contracts as "exploitative boilerplate"); Meirav Furth-Matzkin & Roseanna Sommers, *Consumer Psychology and the Problem of Fine-Print Fraud*, 72 STAN. L. REV. 503, 511–13 (2020) (discussing the problem of "fine-print fraud" in consumer contracts).

153. 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 ("[T]here are many good reasons to argue that *a model of language use should reflect how the language is actually being used*." (emphasis in original)). *But see* 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." (citation omitted)); *see also* Weidinger et al., *supra* note 20, at 14 ("A [language model] trained on language data at a particular moment in time risks . . . enshrining temporary values and norms without the capacity to update the technology as society develops. . . . The risk . . . is that [language models] come to represent language from a particular community and point in time, so that the norms, values, categories from that moment get 'locked in.'" (citations omitted)); Anna Rogers, *Changing the World by Changing the Data*, PROC. 59TH ANN. MEETING ASS'N COMPUTATIONAL LINGUISTICS 2182, 2184–85 (2021) (responding to Goldberg's claim that there is value in using language models to study the world "as it is").

the corresponding contract excerpt (i.e., the contract text 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.[154] 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 *test questions*. GPT-3 might have performed worse on the pro-consumer questions because those questions and corresponding contract excerpts are systematically different from 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, the pro-company provisions may have been tested in litigation more often than pro-consumer provisions, resulting in pro-company provisions using clearer, more accessible language than pro-consumer provisions.[155] It is difficult to measure and control for such differences.[156]

The case study cannot determine which, if any, of these explanations is correct. Nevertheless, the disparity in performance observed across the different question categories raises a noteworthy concern and offers valuable 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 shown in Table 4 (above) is that, on average, GPT-3 performed better on questions that explicitly reference the name of the relevant company.[157] A likely explanation is that the reference to

154. *See supra* note 152 (discussing the pro-company orientation of consumer contracts).

155. 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 in the case study did not find a statistically significant relationship between performance and the classification of a question as pro-company.

156. One reason for this difficulty 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.

157. *See supra* note 140 (describing the "Company Name in Question" variable used in the regression analysis).

a 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 verify this explanation, there is a body of research illustrating that language models "memorize" highly specific information during pretraining.[158] In the case of GPT-3, its training data is replete with information that may be relevant to 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.[159] 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.

Informational cues, however, 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.[160] For example, perhaps companies could

158. *See* Nicholas Carlini, Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea & Colin Raffel, *Extracting Training Data from Large Language Models*, PROC. 30TH USENIX SEC. SYMPOSIUM 2633 (2021) (demonstrating that language models can memorize specific examples found in their training data); Vered Shwartz, Rachel Rudinger & Oyvind Tafjord, *"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). *Compare* Eric Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg & Byron C. Wallace, *Does BERT Pretrained on Clinical Notes Reveal Sensitive Data?*, PROC. 2021 CONF. N. AM. CH. ASS'N COMPUTATIONAL LINGUISTICS 946 (2021) (finding that, using simple probing methods, personal health information cannot generally be extracted from a language model trained on medical records); *see also* Zhengbao Jiang, Frank F. Xu, Jun Araki & Graham Neubig, *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).

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

160. *See generally* Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava & Kai-Wei Chang, *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, Paul Michel & Graham Neubig, *Weight Poisoning*

draft consumer contracts that language models cannot understand or systematically interpret in a manner that favors companies' interests.[161] Seen in this light, informational cues cut both ways. While informational cues potentially offer improved performance, they also reinforce concerns that language models are disturbingly brittle.[162]

D. BRITTLENESS

To further investigate the issue of brittleness, the study tested the performance of GPT-3 in answering an alternatively worded version of all 200 questions. While each question's content is substantially the same across both versions of the question, the alternatively worded questions are, by design, less readable, that is, more difficult for a human to read.[163] Table 5 (below) depicts an example of the original wording of a question (more readable) alongside the alternative wording of that question (less readable).

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, GPT-3's performance was nearly ten percentage points lower on the alternatively worded questions, as illustrated in Figure 3

Attacks on Pre-trained Models, PROC. 58TH ANN. MEETING ASS'N COMPUTATIONAL LINGUISTICS 2793 (2020) (demonstrating that adversarial attacks during pretraining can render language models vulnerable to strategic manipulations).

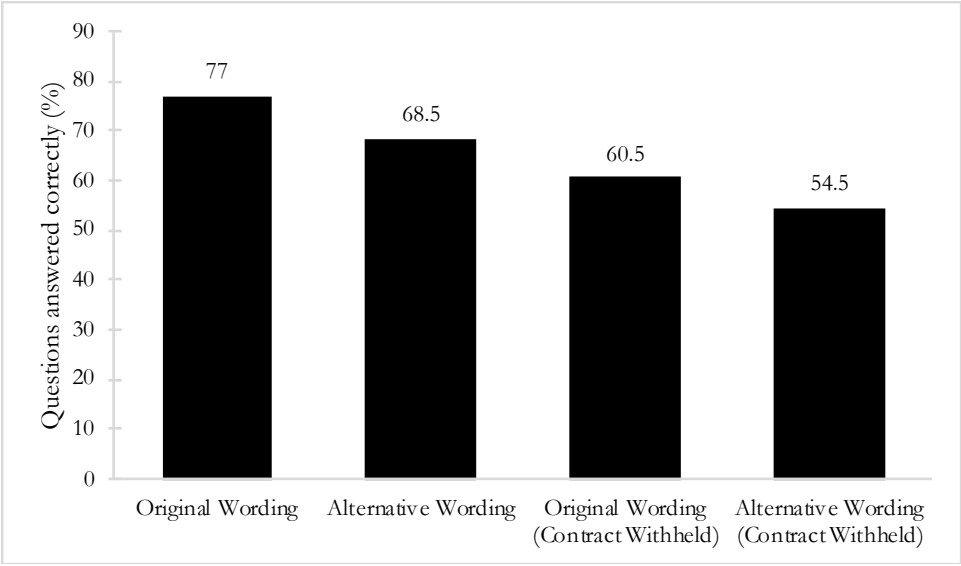
161. *See* Arbel & Becher, *supra* note 19, at 118–24, 141–43 (discussing potential adversarial attacks on language models in the legal domain).

162. Brittleness refers to the outputs of a machine learning model being affected by seemingly minor changes in inputs. *See, e.g.*, Robin Jia, Aditi Raghunathan, Kerem Göksel & Percy Liang, *Certified Robustness to Adversarial Word Substitutions*, PROC. 2019 CONF. EMPIRICAL METHODS IN NLP 4129, 4129 (2019) ("Adding distracting text to the input, paraphrasing the text, replacing words with similar words, or inserting character-level 'typos' can significantly degrade a model's performance." (citations omitted)).

163. Table 7 in the Appendix lists the applicable readability scores; *see also supra* note 144 (explaining how certain readability scores are calculated).

(below).[164] (A smaller decrease in accuracy is observed in the corresponding contract withheld baselines.) These results suggest that the model is highly sensitive to the wording of questions, even if the substance of the questions is unchanged.[165]

Figure 3: Comparison of Accuracy across Question Wordings



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 shown in Table 4 (above) 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,

164. These results are consistent with the overall performance scores (that account for both accuracy and calibration), which are listed in Table 8B in the Appendix. Note, however, that in the case study GPT-3 did not provide a yes /no response to seven of the alternatively worded questions. Given that 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.

165. These results are also consistent with previous studies demonstrating the sensitivity of language models to perturbations. *See, e.g.*, Jia et al., *supra* note 162, at 4129 (citing several studies that demonstrate the brittleness of language models).

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 GPT-3 is more likely to "forget" content contained in earlier parts of longer excerpts (compared with earlier parts of shorter excerpts) and therefore perform worse 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 a greater overlap in language between the question and the contract would assist the model in understanding the contract. But none of these assumptions was borne out in the case study.[166]

On the one hand, this result is encouraging. It suggests that GPT-3 can cope well with longer and less readable contracts,[167] and does not require that the language of the question mirror the language of the contract in order to perform well.[168] On the other hand, given that performance is so sensitive to the wording of the questions, it is somewhat puzzling that performance is altogether insensitive to the language of the contracts themselves.[169] One possible explanation is that GPT-3, like other language models, operates by predicting the next word in a sequence. The question (not the contract) is the final part of the prompt and, therefore, has an outsized impact on the model's predictions.[170]

Taken together, the results of the case study illustrate that language models like GPT-3 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

166. *See supra* Part IV.B (discussing the rationale for each of these assumptions).

167. For examination of the (un)readability of consumer contracts, see Uri Benoliel & Shmuel I. Becher, *The Duty to Read the Unreadable*, 60 B.C.L. REV. 2255, 2270–84 (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); Aleecia M. McDonald & Lorrie Faith Cranor, *The Cost of Reading Privacy Policies,* 4 I /S: J.L. & POL'Y FOR INFO. SOC'Y 543, 553–62 (2008); Alan M. White & Cathy Lesser Mansfield, *Literacy and Contract*, 13 STAN. L. & POL'Y REV. 233, 260 (2002).

168. 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 231 (indicating that the model's context window constrains few-shot learning).

169. 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).

170. *See* Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein & Sameer Singh, *Calibrate Before Use: Improving Few-Shot Performance of Language Models*, PROC. 38TH INT'L CONF. MACH. LEARNING at 4 (2021) (illustrating that content near the end of a prompt can have a disproportionate impact on a model's outputs).

the same time, GPT-3 is very sensitive to how questions are worded and might contain an anti-consumer bias.

V. BROADER IMPLICATIONS

Using language models to read consumer contracts and perform other legal tasks may have broader implications for various stakeholders. This Part aims to explore several of these implications and offer some initial guidance to users of language models, developers of language models, and policymakers.

A. ONGOING EXPERIMENTATION

The successful deployment of language models requires experimentation. When asking a language model questions about consumer contracts, users should, at the very least, 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 prompt language models with different lexical and logical variations of questions and other tasks.

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 growing body of research on "prompt design" aims to provide such guidance.[171] 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.[172]

171. For an overview of prompt design methods, see Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi & Graham Neubig, *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*, ARXIV (July 28, 2021), https://arxiv.org/abs/2107.13586; Tianyu Gao, *Prompting: Better Ways of Using Language Models for NLP Tasks*, THE GRADIENT (July 3, 2021), https://thegradient.pub/prompting/.

172. Of course, prompt design is no substitute for establishing appropriate performance metrics and designing practical tools for evaluating language models. *See generally* Bommasani et al., *supra* note 6, at 91–96 (summarizing current methods for evaluating machine learning models); Sebastian Ruder, *Challenges and Opportunities in NLP Benchmarking* (Aug. 23, 2021), https://ruder.io/nlp-benchmarking/ (discussing the role of benchmarking in model

Progress on this front will require input from various actors and community-wide collaboration.[173] Experimenting with language models can be resource-intensive in terms of both technological infrastructure and human capital. Developers of language models, who typically possess more resources than other stakeholders, are uniquely positioned to contribute to this enterprise.[174] 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, which could be overseen by independent third parties,[175] would 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 settings.

B. CONSUMER TRUST

The case study offers only an initial exploratory analysis of the prospect of using language models to read consumer contracts. Future work will hopefully

---

evaluation). For critical perspectives on current methods for evaluating machine learning systems, see Ethayarajh & Jurafsky, *supra* note 124; Samuel R. Bowman & George E. Dahl, *What Will it Take to Fix Benchmarking in Natural Language Understanding?*, PROC. 2021 CONF. N. AM. CH. ASS'N COMPUTATIONAL LINGUISTICS 4843 (2021); Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton & Alex Hanna, *AI and the Everything in the Whole Wide World Benchmark*, 35TH CONF. NEURAL INFO. PROCESSING SYS. DATASETS AND BENCHMARKS TRACK (2021); Bernard Koch, Emily Denton, Alex Hanna & Jacob G. Foster, *Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research,* 35TH CONF. NEURAL INFO. PROCESSING SYS. DATASETS AND BENCHMARKS TRACK (2021).

173. *See, e.g.*, BIG SCIENCE, https: //bigscience.huggingface.co/ (last visited Aug. 8, 2022) (facilitating a collaboration—among 600 researchers from 50 countries and over 250 institutions—focused on studying large multilingual language models and datasets); CENTER FOR RESEARCH ON FOUNDATION MODELS, https: / /crfm.stanford.edu / (last visited Aug. 8, 2022) (establishing an interdisciplinary center, comprised of Stanford University researchers from over ten departments, to "make fundamental advances in the study, development, and deployment of foundation models").

174. For instance, OpenAI, is able to provide API users with prompt design and safety guidelines informed by its unparalleled insight into how GPT-3 has been used in practice. *See Prompt Design*, OPENAI, https://beta.openai.com/docs/guides/completion/prompt-design (last visited Aug. 8, 2022); *Safety Best Practices*, OPENAI, https://beta.openai.com/docs/safety-best-practices (last visited Aug. 8, 2022). *See generally* Weidinger et al., *supra* note 20, at 38 (arguing that the developers of language models have a responsibility to address the risks posed by language models).

175. This oversight procedure could possibly be integrated into a broader process of certifying the safety of language models. *See generally* Peter Cihon, Moritz J. Kleinaltenkamp, Jonas Schuett & Seth D. Baum, *AI Certification: Advancing Ethical Practice by Reducing Information Asymmetries*, 2 IEEE TRANSACTIONS ON TECH. & SOC'Y 200 (2021); Kira J.M. Matus & Michael Veale, *Certification Systems for Machine Learning: Lessons from Sustainability*, 16 REG. & GOV. 177 (2022).

revisit, and expand on, this analysis. 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.[176] This, of course, is a crude tool for contract interpretation,[177] or indeed any legal analysis.[178] The predictions of neural language models can also be difficult to explain or interpret.[179] For example, why does a model respond "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 can prevent us from diagnosing the source of a model's errors and biases.[180] It can also hamper our ability to ascertain whether a language model is aligned with

176. *See supra* Part II.A (offering a brief primer on the operation of language models).

177. *But see* Catterwell, *supra* note 130, at 100–07, 109–11 (rebutting some of the common objections to using machine learning in contract interpretation).

178. Many scholars have expressed concern about automating legal analysis and dispensing with human judgment in legal tasks. *See* Frank Pasquale, *A Rule of Persons, Not Machines: The Limits of Legal Automation*, 87 GEO. WASH. L. REV. 1, 50–53 (2019) (arguing that "brute force" prediction models should not substitute human discretion in legal decision-making); Joshua P. Davis, *Artificial Wisdom? A Potential Limit on AI in Law (and Elsewhere)*, 72 OKLA. L. REV. 51, 55–62 (2019) (emphasizing the importance of moral judgments in legal practice and judicial decision-making). Interestingly, however, when it comes to answering questions about contracts, language models and human beings share some things in common. Like language models, human beings can utilize informational cues associated with company names. *See, e.g.*, Merrie Brucks, Valarie A. Zeithaml & Gillian Naylor, *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 small investors' responses to the readability of financial disclosures).

179. *See* JURAFSKY & MARTIN, *supra* note 32, at 140. For further discussion of interpretability in NLP, see Bommasani et al., *supra* note 6, at 122–27; Weidinger et al., *supra* note 20, at 37–38.

180. *See generally* Remus & Levy, *supra* note 15, at 550 (explaining how the lack of transparency in machine learning poses problems in the legal domain); Susan C. Morse, *When Robots Make Legal Mistakes*, 72 OKLA. L. REV. 213, 217–19 (2020) (surveying scholarship on the use of "black-box" systems in the legal decision-making); *see also* Ashley, *supra* note 80, at 1137–38 (discussing the inability of legal question answering systems to provide explanations). But other studies suggest that legal AI systems can provide such explanations. *See, e.g.*, Federico Ruggeri, Francesca Lagioia, Marco Lippi & Paolo Torroni, *Detecting and Explaining Unfairness in Consumer Contracts through Memory Networks*, 30 AI & L. 59, 78–81 (2021) (proposing a method for an AI system to provide an explanation for its classification of contractual clauses).

broader social values.[181] These shortcomings are especially problematic where a person relies on a language model to understand and exercise their legal rights.[182]

---

181. This issue—ensuring that AI systems implement human intent, preferences, and values—is known as *AI alignment*, which is central to broader concerns around AI safety. Seminal works on AI safety include NICK BOSTROM, SUPERINTELLIGENCE: PATHS, DANGERS, STRATEGIES (2014) (exploring the potential dangers posed by "superintelligent" machines); Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman & Dan Mané, *Concrete Problems in AI Safety*, ARXIV (June 21, 2016), https://arxiv.org/abs/1606.06565 (establishing a practical research agenda for AI safety); STUART RUSSELL, HUMAN COMPATIBLE: ARTIFICIAL INTELLIGENCE AND THE PROBLEM OF CONTROL (2019) (arguing, *inter alia*, that human values cannot be hard-wired into AI systems; instead, AI systems must learn human values from human behavior); *see also* Tom Everitt, Gary Lea & Marcus Hutter, *AGI Safety Literature Review*, PROC. 27TH INTL. JOINT CONF. AI 5441 (2018) (outlining the safety problems facing AGI and discussing potential solutions); Iason Gabriel, *Artificial Intelligence, Values, and Alignment*, 30 MINDS & MACH. 411 (2020) (examining the philosophical principles underpinning AI alignment); BRIAN CHRISTIAN, THE ALIGNMENT PROBLEM: MACHINE LEARNING AND HUMAN VALUES (2020) (exploring the history of the field of AI safety and alignment); Dan Hendrycks, Nicholas Carlini, John Schulman & Jacob Steinhardt, *Unsolved Problems in ML Safety*, ARXIV (Sept. 28, 2021), https: //arxiv.org/abs/2109.13916 (advancing a revised research agenda for AI safety); GILLIAN K. HADFIELD, RULES FOR A FLAT WORLD: WHY HUMANS INVENTED LAW AND HOW TO REINVENT IT FOR A COMPLEX GLOBAL ECONOMY xiii, xx [hereinafter HADFIELD, RULES FOR A FLAT WORLD] (2020) (arguing that a science of "human normativity" is needed to determine which values should guide AI systems); Dylan Hadfield-Menell & Gillian K. Hadfield, *Incomplete Contracting and AI Alignment*, PROC. 2019 AAAI /ACM CONF. AI, ETHICS, & SOC'Y 417, 420–21 (2019) (applying insights from incomplete contracting theory to the problem of AI alignment). NLP technologies face distinct safety and alignment problems. *See* Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik & Geoffrey Irving, *Alignment of Language Agents*, ARXIV (Mar. 26, 2021), https: / /arxiv.org /abs /2103.14659 (describing several ways in which NLP systems can be misaligned); Bommasani et al., *supra* note 6, at 113–16 (exploring safety concerns facing large pretrained models); Weidinger et al., *supra* note 20, at 10 (presenting a taxonomy of the risks posed by large language models); Chen et al., *supra* note 10, at 11–12, 26–29 (examining the problem of alignment in code generation systems). Recently, there have been several attempts to evaluate, and improve, the alignment of language models. *See* Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song & Jacob Steinhardt, *Aligning AI With Shared Human Values*, 9TH INT'L CONF. LEARNING REPRESENTATIONS (2021) (presenting a benchmark for evaluating whether a language model is aligned with human values); Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap & Yejin Choi, *Delphi: Towards Machine Ethics and Norms*, ARXIV, (Oct. 14, 2021), https://arxiv.org/abs/2110.07574 (presenting a "commonsense moral model" trained on a "commonsense norm bank"); Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah & Jared Kaplan, *A General Language Assistant as a Laboratory for*

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.[183] This approach will need to be re-examined if language models are deployed in the legal domain or other high-risk settings. Several mechanisms for governing AI systems, including

Alignment , ARXIV (Dec. 9, 2021), https://arxiv.org/abs/2112.00861 (exploring methods for creating language models that are "helpful, honest, and harmless").

182. These shortcomings are also problematic in other high-risk settings, such as healthcare. *See* Diane M. Korngiebel & Sean D. Mooney, *Considering the Possibilities and Pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in Healthcare Delivery*, 3 NPJ DIGIT. MED. 93 (2021); *see also* Anne-Laure Rousseau, Clément Baudelaire & Kevin Riera, *Doctor GPT-3: Hype or Reality?*, NABLA (Oct. 27, 2020), https://www.nabla.com/blog/gpt-3/ (revealing that GPT-3 recommended that a hypothetical patient commit suicide).

183. 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 (last visited Aug. 8, 2022) (licensing BERT under the Apache License 2.0, Section 8 of which excludes liability of the licensor). For further discussion of liability in connection with AI systems, see Paulius Čerka, Jurgita Grigienėa & Gintarė Sirbikytėb, *Liability for Damages Caused by Artificial Intelligence*, 31 COMPUT. L. & SEC. REV. 376, 383–86 (2015); Mark A. Lemley & Bryan Casey, *Remedies for Robots*, 86 U. CHI. L. REV. 1311, 1342–78 (2019); Bryan Casey, *Robot Ipsa Loquitur*, 108 GEO. L.J. 225, 251–67 (2019); Andrew D. Selbst, *Negligence and AI's Human Users*, 100 B.U. L. REV. 1315, 1322–29 (2020); RYAN ABBOTT, THE REASONABLE ROBOT: ARTIFICIAL INTELLIGENCE AND THE LAW 50–70 (2020).

measures to improve transparency[184] and accountability,[185] could be instructive.[186] Adapting these mechanisms to improve the reliability and

184. Legal mechanisms to improve transparency 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 [hereinafter GDPR]; *see also* Margot E. Kaminski, *The Right to Explanation, Explained*, 34 BERKELEY TECH. L.J. 189 (2019) (presenting a detailed analysis of the GDPR's right to explanation). Technical mechanisms to improve transparency include datasheets, which offer a standardized process for documenting datasets, and model cards, a framework for disclosing information about a model's features, intended use, and performance evaluation. *See* Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III & Kate Crawford, *Datasheets for Datasets*, ARXIV (Mar. 23, 2018), https: / /arxiv.org /abs /1803.09010; Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji & Timnit Gebru, *Model Cards for Model Reporting*, PROC. 2019 CONF. FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 220 (2019); *see also* Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes & Margaret Mitchell, *Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure*, PROC. 2021 ACM CONF. FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 560 (2021) (proposing a new transparency framework for dataset development); Emily M. Bender & Batya Friedman, *Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science*, 6 TRANSACTIONS ASS'N COMPUTATIONAL LINGUISTICS 587 (2018) (proposing a schema for documenting the features of NLP datasets).

185. One prominent regulatory proposal for improving accountability 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); *see also* Michael Veale & Frederik Zuiderveen Borgesius, *Demystifying the Draft EU Artificial Intelligence Act*, 22 COMPUT. L. REV. INT'L 97, 107 (2021) (examining how the proposed legislation may apply to GPT-3). Technical mechanisms for improving accountability include third party auditing of AI systems, red teaming exercises, bias and safety bounties, and incident reporting. *See* Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, Tegan Maharaj, Pang Wei Koh, Sara Hooker, Jade Leung, Andrew Trask, Emma Bluemke, Jonathan Lebensold, Cullen O'Keefe, Mark Koren, Théo Ryffel, JB Rubinovitz, Tamay Besiroglu, Federica Carugati, Jack Clark, Peter Eckersley, Sarah de Haas, Maritza Johnson, Ben Laurie, Alex Ingerman, Igor Krawczuk, Amanda Askell, Rosario Cammarota, Andrew Lohn, David Krueger, Charlotte Stix, Peter Henderson, Logan Graham, Carina Prunkl, Bianca Martin, Elizabeth Seger, Noa Zilberman, Seán Ó hÉigeartaigh, Frens Kroeger, Girish Sastry, Rebecca Kagan, Adrian Weller, Brian Tse, Elizabeth Barnes, Allan Dafoe, Paul Scharre, Ariel Herbert-Voss, Martijn Rasser, Shagun Sodhani, Carrick Flynn, Thomas Krendl Gilbert, Lisa Dyer, Saif Khan, Yoshua Bengio & Markus Anderljung, *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims*, ARXIV (Apr. 15, 2020), https: / / arxiv.org /abs /2004.07213.

186. For an overview of proposals for governing AI, see Lawrence Zhang, *Initiatives in AI Governance*, SCHWARTZ REISMAN INSTITUTE FOR TECHNOLOGY AND SOCIETY (Dec. 2020),

trustworthiness of language models will require coordination between multiple stakeholders.

C. COMPOUNDING BIAS

Bias is a major obstacle to building trustworthy language models. The case study provides 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.[187] If the contracts presented to it generally favor companies (which is likely),[188] then such an anti-consumer bias might, on average, improve performance.

But this hypothetical language 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 fail to identify any contractual provisions that favor consumers.[189] Consumers relying on this model would 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*—stems from the fact that language models not only absorb and reproduce problematic patterns from their training data, but can amplify these patterns.[190] 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.[191] The consumer contracts produced by the model (e.g., online 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

https://static1.squarespace.com/static/5ef0b24bc96ec4739e7275d3/t/5fb58df18fbd7f2b94b5b5cd/1605733874729/SRI+1+-+Initiatives+in+AI+Governance.pdf

187. Note, however, that language models are not, strictly speaking, classifiers.

188. *See supra* note 152 (discussing the pro-company orientation of consumer contracts).

189. 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, 429 (2009).

190. *See supra* note 150 (discussing the problem of societal biases in language models).

191. For a comparable issue in the context of code generation, see Chen et al., *supra* note 10, at 27 (finding that a language model trained on code generates more bugs when it is prompted with buggy code).

biases they encode, would pollute the reservoir of data available for training new models.[192] In a dangerous feedback loop, biases could compound with each successive generation of language model.[193]

Fortunately, techniques are being developed to detect and filter out machine-generated content. These techniques could, for example, prevent the outputs of GPT-3 from being included in the training data of GPT-4. However, detecting whether content has been generated by a language model is increasingly difficult.[194] One alternative approach to counteracting model bias involves prompt design. But this too is unlikely to be a panacea.[195] Engineering alone cannot solve the problem. Addressing current biases and preventing a cycle of compounding bias requires a combination of technical and institutional mechanisms.

D. GOVERNANCE

Each stage of a language model's lifecycle, from development through deployment, presents governance challenges. As we have seen, improving a model's reliability, tackling bias, and conducting effective evaluations is vital. But there are other challenges too, some of which are often overlooked.[196] Identifying these challenges is key to understanding the steps that policymakers should take to harness the benefits of language models and address the attendant risks.

---

192. *See* Bender et al., *supra* note 20, at 617; Kenton et al., *supra* note 181, at 7. Such compounding bias is a form of data cascade. *See* Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh & Lora Aroyo., "*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 (2021) (defining data cascades as "compounding events causing negative, downstream effects from data issues, that result in technical debt over time"). Compounding bias is also somewhat analogous to the issue of bias cascades in human decision-making. *See* DANIEL KAHNEMAN, OLIVER SIBONY & CASS R. SUNSTEIN, NOISE: A FLAW IN HUMAN JUDGMENT 288 (2021) (describing how the initial bias of one decision-maker can replicate and magnify by biasing other decision-makers), citing Itiel E. Dror, *Biases in Forensic Experts*, 360 SCIENCE 243 (2018) (examining the role of bias cascades in forensic investigations).

193. *See* Bender et al., *supra* note 20, at 617 ("[T]he 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 discussion of existing feedback loops and network effects that entrench pro-company contractual drafting, see Boardman, *supra* note 155, at 1112–17.

194. *See* Brown et al., *supra* note 8, at 25–26.

195. *See* Tamkin et al., *supra* note 20, at 5–6.

196. *See generally* Abeba Birhane et al., *supra* note 22 (illustrating that machine learning research continues to neglect issues concerning its societal impact).

(i) *Data protection*. Training language models on vast online corpora raises several concerns with respect to data protection. For example, did the collection of training data infringe upon applicable privacy laws, such as the European Union's General Data Protection Regulation (GDPR) or California's Consumer Privacy Act (CCPA)?[197] Did it violate the federal Computer Fraud and Abuse Act (CFAA)?[198] Did the organization collecting the data have the right to use the data to train a language model?[199] Can personally identifiable information be extracted from the resulting model?[200] In the case of proprietary language models accessed through an API, how is confidential information protected?[201] Researchers have begun to grapple with some of these questions.[202]

(ii) *Environmental impact*. Training language models is energy-intensive.[203] For example, training GPT-3 consumed several thousand petaflop/s-days of

197. CAL. CIV. CODE §§ 1798.100–.199 (2018)

198. 18 U.S.C. § 1030(a)(2) criminalizes "intentionally access[ing] a computer system without authorization" or where doing so "exceeds authorized access." *See also* Van Buren v. United States, 593 U.S. __ (2021) (clarifying which activities amount to unauthorized access).

199. The answer to this question will depend on, among other things, the application of the fair use doctrine. *See* Mark A. Lemley & Bryan Casey, *Fair Learning*, 99 TEX. L. REV. 743, 760–79 (2021) (exploring whether the doctrine of fair use permits machine learning models to be trained on copyrighted data); *see also infra* note 211 (discussing the debate concerning the ownership of the outputs of code generation tools).

200. *See, e.g.*, Carlini et al., *supra* note 158 (demonstrating that an adversary can extract from GPT-2 personally identifiable information contained in the model's training data).

201. This is especially important if lawyers provide 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 this privacy issue is not meaningfully different to the privacy issue arising when lawyers use other online platforms or cloud-based software.

202. *See* Brundage et al., *supra* note 185, at 28–30; Bommasani et al., *supra* note 6, at 145–46; Weidinger et al., *supra* note 20, at 18–21.

203. There have been several attempts to estimate the energy consumption and carbon emissions involved in training large language models. *See* Emma Strubell, Ananya Ganesh & Andrew McCallum, *Energy and Policy Considerations for Deep Learning in NLP*, PROC. 57TH CONF. ASS'N COMPUTATIONAL LINGUISTICS 3645, 3647–48 (2020); Lasse F. Wolff Anthony, Benjamin Kanding & Raghavendra Selvan, *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. At 2–3 (2020); David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier & Jeff Dean, *Carbon Emissions and Large Neural Network Training*, ARXIV at 2–8 (Apr. 23, 2021), https: / /arxiv.org /abs /2104.10350. However, once trained, language models can operate relatively efficiently. *See, e.g.*, Brown et al., *supra* note 8, at 39. For further discussion of the environmental impact of large language models, see Bender et al., *supra* note 20, at 612–13; Weidinger et al., *supra* note 20, at 32–33; *see also* Bommasani et al., *supra* note 6, at 139–44 (discussing several strategies for measuring and mitigating the

compute,[204] the environmental impact of which can be compared to driving a car several hundred thousand miles.[205] Despite increasingly efficient training techniques, the "parameters race"—in which technology firms compete to build ever-larger language models[206]—suggests that energy consumption in model training may continue to grow. The machine learning community is now turning its attention to these environmental challenges.[207]

(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, the suppliers or owners of the model's training data, or another party?[208] The answer turns on, among other things, whether creative works generated by a machine are eligible for copyright protection,[209] as well

environmental impact of large pretrained models); Borgeaud et al., *supra* note 56, at 16 (equipping a language model with the ability to retrieve information from a database, which improves performance without increasing the computational resources required for training).

204. *See* Brown et al., *supra* note 8, at 39.

205. *See* Heaven, *supra* note 8 ("[T]raining GPT-3 would have had roughly the same carbon footprint as driving a car the distance to the moon and back, if it had been trained in a data center fully powered by fossil fuels."); *see also* Anthony et al., *supra* note 203, at 10 (estimating the energy and carbon footprint of GPT-3).

206. *See* Coco Feng, *US-China Tech War: Beijing-Funded AI Researchers Surpass Google and OpenAI with New Language Processing Model*, SOUTH CHINA MORNING POST (June 2, 2021), https://www.scmp.com/tech/tech-war/article/3135764/us-china-tech-war-beijing-funded-ai-researchers-surpass-google-and; Ali Alvi & Paresh Kharya, *Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model*, MICROSOFT RESEARCH BLOG (Oct. 11, 2021), https://www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model /.

207. *See* Roy Schwartz, Jesse Dodge, Noah A. Smith & Oren Etzioni, *Green AI*, 63 COMM. ACM 54 (2020); Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky & Joelle Pineau, *Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning*, 21 J. MACH. LEARNING RES. 1 (2020); Kadan Lottick, Silvia Susai, Sorelle A. Friedler & Jonathan P. Wilson, *Energy Usage Reports: Environmental Awareness as Part of Algorithmic Accountability*, 33RD CONF. NEURAL INFO. PROCESSING SYS. WORKSHOP ON TACKLING CLIMATE CHANGE WITH MACH. LEARNING (2019); *see also* KATE CRAWFORD, THE ATLAS OF AI: POWER, POLITICS, AND THE PLANETARY COSTS OF ARTIFICIAL INTELLIGENCE ch. 1 (2021) (examining the environmental demands of AI technologies and associated industries).

208. *See* Omri Avrahami & Bar Tamir, *Ownership and Creativity in Generative Models*, ARXIV at 1–2 (Dec. 2, 2021), https://arxiv.org/abs/2112.01516; Jason K. Eshraghian, *Human Ownership of Artificial Creativity*, 2 NATURE MACH. INTELL. 157, 158–59 (2020).

209. *See* 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 &

as the terms of the software license agreement applicable to the model.[210] Different stakeholders are likely to adopt different positions on the issue.[211]

(iv) *Access and misuse*. Historically, open access to the code and weights of language models has enabled researchers and developers to independently use, adapt, and evaluate language models. However, as the capabilities of language models improve, open access has become a double-edged sword.[212] By restricting access to a model, an organization can potentially prevent a

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); Daniel J. Gervais, *The Human Cause*, *in* RESEARCH HANDBOOK ON INTELLECTUAL PROPERTY AND ARTIFICIAL INTELLIGENCE (R. Abbott, ed., forthcoming).

210. *See supra* note 183 (discussing the software license applicable to Google's BERT language model).

211. For example, there is a lively debate concerning the ownership of the outputs of code generation tools, such as GitHub Copilot. *See* Chen et al., *supra* note 10, at 13 (suggesting that the doctrine of fair use applies to publicly available code). *Compare* Dave Gershgorn, *GitHub's Automatic Coding Tool Rests on Untested Legal Ground*, THE VERGE (July 7, 2021), https://www.theverge.com/2021/7/7/22561180/github-copilot-legal-copyright-fair-use-public-code; Matthew Sparkes, *GitHub's Programming AI May Be Reusing Code without Permission*, NEW SCIENTIST (July 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/; *see also* Lemley & Casey, *supra* note 199, at 760–79.

212. Discussions concerning the implications of releasing language models (and AI research more generally) include Aviv Ovadya & Jess Whittlestone, *Reducing Malicious Use of Synthetic Media Research: Considerations and Potential Release Practices for Machine Learning*, ARXIV (July 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; Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie & Jasmine Wang, *Release Strategies and the Social Impacts of Language Models*, ARXIV (Nov. 13, 2019), https://arxiv.org /abs /1908.09203; Jess Whittlestone & Aviv Ovadya, *The Tension between Openness and Prudence in AI Research*, ARXIV (Jan. 13, 2020), https://arxiv.org /abs/1910.01170; Toby Shevlane & Allan Dafoe, *The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse?*, PROC. 2020 AAAI /ACM CONF. AI, ETHICS, & SOC'Y 173 (2020); Mark Riedl, *AI Democratization in the Era of GPT-3*, THE GRADIENT (Sept. 25, 2020), https://thegradient.pub /ai-democratization-in-the-era-of-gpt-3 /; *Managing the Risks of AI Research Six Recommendations for Responsible Publication*, PARTNERSHIP ON AI (May 6, 2021), https: / /partnershiponai.org / paper /responsible-publication-recommendations /; *How to Be Responsible in AI Publication*, 3 NATURE MACH. INTELL. 367 (2021); Girish Sastry, *Beyond "Release" vs. "Not Release"*, STANFORD UNIVERSITY CENTER FOR RESEARCH ON FOUNDATION MODELS (Oct. 18, 2021), https://crfm.stanford.edu/commentary/2021/10/18/sastry.html.

powerful language model from being used for nefarious purposes,[213] such as spreading misinformation,[214] writing phishing emails,[215] and generating spam.[216] Organizations can also filter sensitive and unsafe outputs.[217] But this role of gatekeeper is controversial. Restrictions on access can impede valuable research[218] and present additional societal risks.[219]

213. For discussion of the potential misuses of language models, see Weidinger et al., *supra* note 20, at 25–28; Bommasani et al., *supra* note 6, at 135–38. For a broader account of the malicious uses of AI, see Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, Hyrum Anderson, Heather Roff, Gregory C. Allen, Jacob Steinhardt, Carrick Flynn, Seán Ó hÉigeartaigh, Simon Beard, Haydn Belfield, Sebastian Farquhar, Clare Lyle, Rebecca Crootof, Owain Evans, Michael Page, Joanna Bryson, Roman Yampolskiy & Dario Amodei, *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, ARXIV (Feb. 20, 2018), https://arxiv.org/abs/1802.07228.

214. *See* Ben Buchanan, Andrew Lohn, Micah Musser & Katerina Sedova, *Truth, Lies, and Automation How Language Models Could Change Disinformation*, CENTER FOR SECURITY AND EMERGING TECHNOLOGY at 5–34 (May 2021), https://cset.georgetown.edu/publication/truth-lies-and-automation /; Sharon Levy, Michael Saxon & William Yang Wang, *Investigating Memorization of Conspiracy Theories in Text Generation*, FINDINGS ASS'N COMPUTATIONAL LINGUISTICS 4718 (2021); McGuffie & Newhouse, *supra* note 68; Stephanie Lin, Jacob Hilton & Owain Evans, *TruthfulQA: Measuring How Models Mimic Human Falsehoods*, ARXIV (Sept. 8, 2021), https://arxiv.org/abs/2109.07958. For an overview of the misinformation harms arising from language models, see Weidinger et al., *supra* note 20, at 21–25. The generation of misinformation by language models is especially problematic in the legal domain. *See* Weidinger et al., *supra* note 20, at 24; Bommasani et al., *supra* note 6, at 65.

215. *See* Lily Hay Newman, *AI Wrote Better Phishing Emails Than Humans in a Recent Test*, WIRED (Jul. 8, 2021), https://www.wired.com/story/ai-phishing-emails/.

216. *See* Brown et al., *supra* note 8, at 35.

217. *See Content Filter*, OPENAI, https://beta.openai.com/docs/engines/content-filter (last visited Aug. 8, 2022). 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*, PROC. 35TH CONF. NEURAL INFO. PROCESSING SYS. (2021).

218. *See* Bommasani et al., *supra* note 6, at 10–11, 157 (describing academia's incremental loss of access to state-of-the-art models). One possible solution is for researchers to use open-source alternatives instead of proprietary models. For example, EleutherAI, an open-source software group, has attempted to reproduce language models comparable to GPT-3. These models, however, are currently much smaller than GPT-3. *See* Sid Black, Leo Gao, Phil Wang, Connor Leahy & Stella Biderman, *GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow*; https://github.com/EleutherAI/gpt-neo (last visited Aug. 8, 2022); EleutherAI, *GPT-NeoX*, GITHUB, https://github.com/EleutherAI/gpt-neox (last visited Aug. 8, 2022).

219. For example, content filters can exclude valuable outputs and introduce new biases. *See* Kenton et al., *supra* note 181, at 6 ("[T]here 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

(v) *Unequal performance*. Despite improvements in their capabilities, language models continue to perform better for certain groups of people than others.[220] One source of this problem is that language models are developed primarily for only a small fraction of human languages.[221] However, even multilingual models, which are designed to serve multiple languages, perform better in some languages than in other languages.[222] While efforts are underway to better include underrepresented groups in language modeling,[223] significant inequalities persist.[224]

(vi) *Regulation*. Finally, the type of legal application explored in this Article—using a language model to provide legal advice directly to consumers—faces a distinct regulatory barrier. Generally speaking, non-lawyers, including developers and operators of AI systems, are prohibited from

---

place."); *see also* Tamkin et al., *supra* note 20, at 9 ("[S]teering 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.").

220. *See* Bender et al., *supra* note 20, at 611–12; Bommasani et al., *supra* note 6, at 130; Weidinger et al., *supra* note 20, at 16–18.

221. *See* Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali & Monojit Choudhury, *The State and Fate of Linguistic Diversity and Inclusion in the NLP World*, PROC. 58TH ANN. MEETING ASS'N COMPUTATIONAL LINGUISTICS 6282 (2020) (studying the relative inclusion of different languages in NLP conferences); *see also* Weidinger et al., *supra* note 20, at 34–35 (discussing the issue of disparate access to language models due to hardware, software, and skill constraints).

222. *See* Shijie Wu & Mark Dredze, *Are All Languages Created Equal in Multilingual BERT?*, PROC. 5TH WORKSHOP ON REPRESENTATION LEARNING FOR NLP 120, 128 (2020) ("While mBERT covers 104 languages, the 30% languages with least pretraining resources perform worse than using no pretrained language model at all.").

223. *See Mission,* WIDENING NATURAL LANGUAGE PROCESSING, https://www.winlp.org/mission/ (last visited Aug. 8, 2022) (describing the organization's mission to improve the representation of women and underrepresented groups in NLP).

224. *See, e.g.,* Isaac Caswell, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suárez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal & Mofetoluwa Adeyemi, *Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets*, ARXIV (Apr. 23, 2021), https://arxiv.org/abs/2103.12028 (finding that low resource language corpora face a host of systemic issues).

providing legal services.[225] Modifying this rule would require regulatory reform.[226] In contemplating such reform, it is important to ask what legal services (if any) are ordinarily available to consumers.[227] For many consumers, the answer is none, which arguably weighs in favor of removing regulatory barriers to using AI systems in the legal domain.[228] To promote consumer

225. *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-of-model-rules/ (last visited Aug. 8, 2022) (overviewing state-level unauthorized practice rules); *see also* Deborah L. Rhode, *Policing the Professional Monopoly: A Constitutional and Empirical Analysis of Unauthorized Practice Prohibitions*, 34 STAN. L. REV. 1, 11–44 (1981) (presenting the seminal empirical study on the enforcement of unauthorized practice rules); Cyphert, *supra* note 16, at 433–34 (arguing that nonlawyers' use of GPT-3 to perform legal tasks may amount to unauthorized practice).

226. There have been many proposals to alter the unauthorized practice rules and overhaul the regulation of legal services. *See* 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 chs. 8–10 (2017); HADFIELD, RULES FOR A FLAT WORLD, *supra* note 181, at ch. 9; Benjamin H. Barton & Deborah L. Rhode, *Access to Justice and Routine Legal Services: New Technologies Meet Bar Regulators*, 70 HASTINGS L.J. 955, 978–87 (2019); Rebecca L. Sandefur, *Legal Advice from Nonlawyers: Consumer Demand, Provider Quality, and Public Harms*, 16 STAN. J. C.R. & C.L. 283, 312–13 (2020). Notably, in 2020, the Utah Supreme Court created a regulatory sandbox to facilitate testing new methods for delivering legal services, including by nonlawyers. *See* Utah Supreme Court Standing Order No. 15 (effective Aug. 14, 2020) ("[E]stablish[ing] a pilot legal regulatory sandbox and an Office of Legal Services Innovation to assist the Utah Supreme Court with overseeing and regulating the practice of law by nontraditional legal service providers or by traditional providers offering nontraditional legal services."); UTAH RULES OF PROF'L CONDUCT rr. 5.4A–5.4B (effective Aug. 14, 2020) (relaxing certain unauthorized practice rules); Deno G. Himonas & Tyler J. Hubbard, *Democratizing the Rule of Law*, 16 STAN. J. C.R. & C.L. 261, 273–78 (2020) (detailing the goals and features of Utah's regulatory sandbox); Rebecca L. Sandefur, Thomas M. Clarke & James Teufel, *Seconds to Impact?: Regulatory Reform, New Kinds of Legal Services, and Increased Access to Justice*, 84 LAW & CONTEMP. PROBS. 69, 74–76 (2021) (describing the activities permitted by Utah's regulatory sandbox).

227. *See* Utah Supreme Court Standing Order No. 15 at 8 (effective Aug. 14, 2020) (providing 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").

228. *See* Tanina Rostain, *Robots versus Lawyers: A User-Centered Approach*, 30 GEO. J. LEGAL ETHICS 559, 569 (2017) ("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 90 (examining how the cost of legal services impedes access to justice).

welfare, policymakers will need to balance this consideration against the risks posed by language models.[229]

Some of these issues are of immediate concern. Others will become more salient as the capabilities of language models improve further. The purpose of flagging these issues is not to exhaustively describe the challenges facing language models in the legal domain. Instead, this brief account aims to illustrate that the safe and beneficial deployment of language models in legal contexts requires governance. While technological solutions are necessary, they are not sufficient. Regulation and policy also have important roles to play.

VI. CONCLUSION

Using computational language models to read consumer contracts is simple in principle but complex in practice. 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 online terms of service. The results paint a nuanced picture. On the one hand, the generally strong performance of GPT-3 suggests that language models have the potential to assist consumers in discovering 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.

Due to the case study's limitations, however, 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 need to further investigate the safety and reliability of using these prediction machines 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.

229. *See* Remus & Levy, *supra* note 15, at 546–48 (discussing the need for consumer protection in the context of automated legal services). *Compare* Rostain, *supra* note 228, at 564–71 (distinguishing between the protections that should be afforded to individual users of automated legal services and the protections that should be afforded to corporate users of automated legal services).

APPENDIX

A. TEST CONDITIONS

1. Prompt Design

The case study used the following priming text:[230]

> 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 Text

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.[231] 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 used in the case study.[232]

230. This priming text is similar to the priming text in a template available in the OpenAI API at the time of the case study. *See Q&A*, OPENAI, https://beta.openai.com/examples/default-qa (last visited Aug. 8, 2022). More specialized guides have subsequently been released in the API. *See Question Answering*, OPENAI, https://beta.openai.com/docs/guides/answers (last visited Aug. 8, 2022). However, these were not available when the case study was conducted.

231. 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.

232. These hyperparameters are similar to the hyperparameters in a template available in the OpenAI API at the time of the case study. *See Q&A*, OPENAI, https://beta.openai.com/examples/default-qa (last visited Aug. 8, 2022). Descriptions in Table 6 are adapted from descriptions in the OpenAI API documentation.

Table 6: Hyperparameters

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.	-

4. Question Readability

Table 7 lists the readability scores of the original and alternative wordings of the questions in the case study. Because readability scores are unreliable for short texts (such as individual questions),[233] the 200 originally worded questions were combined in one document, and readability scores were

233. *See* Oakland & Lane, *supra* note 156.

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.[234] 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 described in Part III.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, respectively.

Table 8A: Comparing test accuracy and overall performance with the 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

234. *Supra* note 144.

Table 8B: Comparing accuracy and overall performance on the 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 the original wording and the alternative wording of the 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 PLOTS

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: Binned scatter plot showing the relationship between (i) accuracy and (ii) the probability assigned to the output (Measure 1)

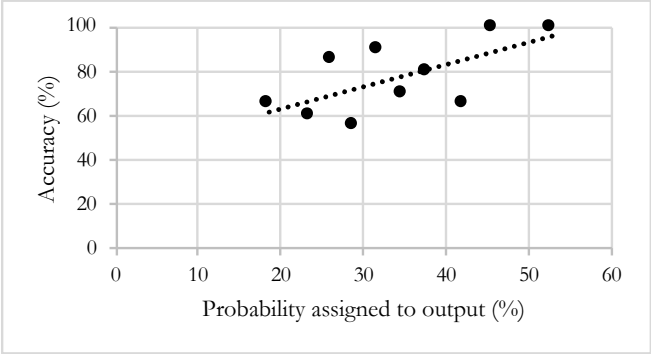


Figure 4B: Binned scatter plot showing the relationship between (i) accuracy and (ii) the difference between the probability assigned to the output and the probability assigned to the alternative answer (Measure 2)

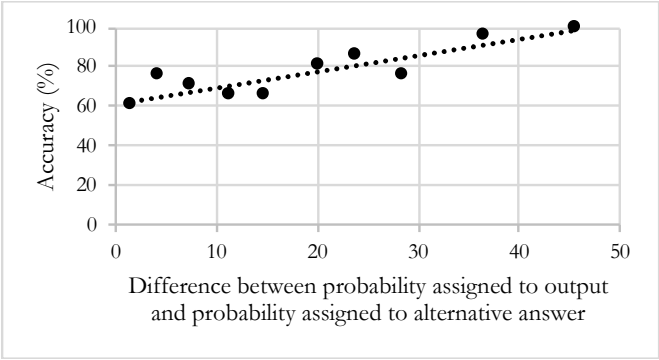


Figure 4C: Binned scatter plot showing the relationship between (i) accuracy and (ii) the ratio between the probability assigned to the output and the probability assigned to the alternative answer (Measure 3)