AI CANNIBALISM AND THE LAW

AMY CYPHERT,* SAMUEL J. PERL,** S. SEAN TU, JD, PHD***

INTRODUCTION

Lawyers are already using[1]—and misusing[2]—large language models ("LLMs") like ChatGPT in their daily lives as they practice

* Lecturer in Law, West Virginia University College of Law, Morgantown, WV. This article was supported by a Hodges Fund faculty research grant. The authors are grateful to the editors of the Colorado Technology Law Journal.

** Carnegie Mellon University, Pittsburgh, PA

*** Professor of Law, West Virginia University College of Law, Morgantown, WV

1. *See*, *e.g.*, Chris Stokel-Walker, *Generative AI Is Coming for the Lawyers*, WIRED (Feb. 21, 2023, 10:00 AM), https://www.wired.com/story/chatgpt-generative-ai-is-coming-for-the-lawyers/ [https://perma.cc/6R7R-WE4T] (discussing how law firms are using generative AI tools like large language models in their practices); David Rotman, *ChatGPT Is About to Revolutionize the Economy. We Need to Decide What That Looks Like*, MIT TECH. REV. (Mar. 25, 2023), https://www.technologyreview.com/2023/03/25/1070275/chatgpt-revolutionize-economy-decide-what-looks-like/ [https://perma.cc/6RBZ-QQAJ] (quoting "an MIT labor economist and a leading expert on the impact of technology on jobs" who notes that law firms are using generative AI).

2. *See* Benjamin Weiser, *ChatGPT Lawyers Are Ordered to Consider Seeking Forgiveness*, N.Y. TIMES (June 22, 2023), https://nytimes.com/2023/06/22/nyregion/lawyers-chatgpt-schwartz-loduca.html [https://perma.cc/59UB-YECR] (For example, two New York attorneys were sanctioned by a federal judge for using ChatGPT to draft a brief that included numerous "made up" cases).

law. Despite recent headlines pointing out the very real downsides of misusing the technology,[3] it is all but certain that lawyers will use LLMs with increasing frequency in the coming years.[4] Indeed, many law schools, recognizing that lawyers need to understand LLMs, are scrambling to train students on best practices.[5] However, LLMs are racing toward a potential cliff that could severely undercut their usefulness to lawyers, and potentially even stifle the development of law itself.

As news articles, blog posts, and even works of fiction generated by artificial intelligence ("AI") make up more and more of the internet, those AI-generated outputs will form an ever-larger share of the data training sets of future LLMs. Recent studies suggest this recursive loop is potentially catastrophic for the models' stability and could result in more misinformation and increasing "AI hallucinations."[6] Such a result would lessen the utility of these tools for lawyers.

I. HOW LLMS WORK

ChatGPT, which was introduced to the public in late 2022, has set off an LLM arms race among technology companies.[7] Although the earliest iteration of ChatGPT—GPT-1—was developed more than five years ago, ChatGPT is especially "user-friendly" and easy for non-experts to use and has thus captured the public's attention.[8] It can be difficult for laypersons to understand how machine learning algorithms like LLMs "work." Even for technical experts who understand how the algorithms perform, "explainability"—or

3. *See id.*

4. As one of us argued in early 2022, generative AI is especially likely to be adopted in the practice of law because the tools are "'creation engine[s]' that actually generate[] text," and "because one of the most important 'products lawyers produce is writing (contacts, motions, etc.)" 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, 419 (2021).

5. *See*, *e.g.*, Stephanie Frances Ward, *Can ChatGPT Help Law Students Learn to Write Better?*, ABA J. (Mar. 6, 2023, 8:38 AM), https://www.abajournal.com/web/article/can-chatgpt-help-law-students-learn-to-write-better [https://perma.cc/HV3S-U35P] (discussing the ways various law schools are incorporating ChatGPT into their writing curriculum).

6. *See* Sina Alemohammad et al., *Self-Consuming Generative Models Go MAD*, ARXIV at 3 (July 4, 2023), https://arxiv.org/pdf/2307.01850.pdf [https://perma.cc/N4E3-AXAR].

7. *See* Kevin Roose, *How ChatGPT Kicked Off an A.I. Arms Race*, N.Y. TIMES (Feb. 3, 2023), https://www.nytimes.com/2023/02/03/technology/chatgpt-openai-artificial-intelligence.html [https://perma.cc/VVC7-FW3P] (Since ChatGPT was released, other LLMs have been released by Google, Meta, and others.).

8. *See* Rotman, *supra* note 1 ("For many non-experts, including a growing number of entrepreneurs and businesspeople, the user-friendly chat model—less abstract and more practical than the impressive but often esoteric advances that have been brewing in academia and a handful of high-tech companies over the last few years—is clear evidence that the AI revolution has real potential.")

how the algorithm arrives at its output—is an elusive goal in the field of AI.[9] For the purposes of this article, we will admittedly oversimplify a very complicated process and explain that LLMs are ultimately prediction machines.[10]

LLMs are given a prompt (an input), and then they generate more text that the model predicts is likely to follow that prompt (an output). For example, an LLM that is given the prompt "how are you" will conclude that these words are more likely, based on the texts in its dataset, to be followed by the word "doing" than "cheese." Therefore, when given the prompt "how are you," an LLM will often produce the output of "doing."

II. COMMON LLM PROBLEMS

LLMs represent an extraordinary advancement in the field of AI and have the potential to make many positive contributions. However, numerous problems and potential biases are linked with LLMs. This section reviews four of those issues: (1) data bias, (2) "AI hallucinations," (3) training data limitation, and (4) training on synthetic data. Data bias pertains to biases that exist in training data and thus produce models that generate biased outputs. AI hallucinations involve the generation of "made-up" facts by LLMs. Training data limitation refers to the fact that an LLM is necessarily constrained by that which is in its training data. Finally, dataset changes are problems associated with the type of information consumed by LLMs to assist in generating outputs.

A. Data Bias

Like any predictive machine, an LLM is only as "good" as the underlying data it is based upon.[11] Thus, it is unsurprising that LLMs that are trained on human writing often reflect human

9. AI researchers use terms like "explainability" and "mechanistic interpretability" to discuss the ability of an AI system to be understood by humans. For a general discussion on the difficulty of interpretability in machine learning algorithms, especially in massive ones like LLMs, see *How Generative Models Could Go Wrong*, THE ECONOMIST (Apr. 19, 2023), https://www.economist.com/science-and-technology/2023/04/19/how-generative-models-could-go-wrong [https://perma.cc/X38T-M8ZR]; *see also What Is Explainable AI?*, IBM, https://www.ibm.com/watson/explainable-ai [https://perma.cc/AGR8-XQQH] (last visited Apr. 21, 2024).

10. *See* Cyphert, *supra* note 4, at 407 ("at a basic level, an autoregressive language model like [the precursor to ChatGPT] is one that has been trained to read a series of words and predict what the next word in the 'pattern' should be.").

11. *See*, *e.g.*, Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218, 2224 n.23 (2019) ("The computer-science idiom is 'garbage in, garbage out,' which refers to the fact that algorithmic prediction is only as good as the data on which the algorithm is trained.").

biases.[12] When it introduced GPT-3, the precursor to ChatGPT, OpenAI (the developer of both models) also released a research paper about the model. That research paper acknowledged potential gender, racial, and religious bias with the outputs (using co-occurrence tests) and concluded: "We have presented this preliminary analysis to share some of the biases we found in order to motivate further research[.]"[13] Reports of bias continued when OpenAI launched ChatGPT.[14] In a blog post a few months after the tool was released, the company noted: "Since our launch of ChatGPT, users have shared outputs that they consider politically biased, offensive, or otherwise objectionable. In many cases, we think that the concerns raised have been valid and have uncovered real limitations of our systems which we want to address."[15] In fact, several non-profit organizations, such as the Algorithmic Justice League, were created to promote the equitable and accountable use of AI.[16]

Researchers at the University of Washington, Carnegie Mellon University, and Xi'an Jiaotong University have also discovered that LLMs can produce outputs that reflect political biases.[17] The

12. *See*, *e.g.*, Cyphert, *supra* note 4, at 413–16 (discussing bias in GPT-3, the precursor to ChatGPT); *see also* Mehtab Khan & Alex Hanna, *The Subjects and Stages of AI Dataset Development: A Framework for Dataset Accountability*, 19 OHIO ST. TECH. L.J. 1, 6 (2023) ("large language models have been shown to be biased against certain communities"). For more general information on bias in AI systems see Amy B. Cyphert, *Tinker-ing with Machine Learning: The Legality and Consequences of Online Surveillance of Students*, 20 NEV. L.J. 457, 462–64 (2020) [hereinafter *Tinker-ing with Machine Learning*].

13. TOM B. BROWN ET AL., LANGUAGE MODELS ARE FEW-SHOT LEARNERS 13 (John Hopkins U. 2020).

14. *See*, *e.g.*, Jeremy Baum & John Villasenor, *The Politics of AI: ChatGPT and Political Bias*, BROOKINGS (May 8, 2023), https://www.brookings.edu/articles/the-politics-of-ai-chatgpt-and-political-bias/ [https://perma.cc/8GMQ-QUTV]; Davey Alba, *OpenAI Chatbot Spits Out Biased Musings, Despite Guardrails*, BLOOMBERG (Dec. 8, 2022), https://www.bloomberg.com/news/newsletters/2022-12-08/chatgpt-open-ai-s-chatbot-is-spitting-out-biased-sexist-results [https://perma.cc/FZW2-KKFE].

15. *How Should AI Systems Behave, and Who Should Decide?*, OPENAI (Feb. 16, 2023), https://openai.com/blog/how-should-ai-systems-behave#OpenAI [https://perma.cc/QF5C-D2H3].

16. *See About*, ALGORITHMIC JUST. LEAGUE, https://www.ajl.org/about [https://perma.cc/6CRC-8DU3] (last visited Apr. 21, 2024) (noting that "AI systems can perpetuate racism, sexism, ableism, and other harmful forms of discrimination, therefore, presenting significant threats to our society - from healthcare, to economic opportunity, to our criminal justice system[,]" and that the Algorithmic Justice League's mission is to "raise public awareness about the impacts of AI, equip advocates with resources to bolster campaigns, build the voice and choice of the most impacted communities, and galvanize researchers, policymakers, and industry practitioners to prevent AI harms").

17. Melissa Heikkilä, *AI Language Models Are Rife with Different Political Biases*, MIT TECH. REV. (Aug. 7, 2023), https://www.technologyreview.com/2023/08/07/1077324/ai-language-models-are-rife-with-political-biases/ [https://perma.cc/RP4R-5THM]; *see also* SHANGBIN FENG ET AL., FROM PRETRAINING DATA TO LANGUAGE MODELS TO DOWNSTREAM TASKS: TRACKING THE TRAILS OF POLITICAL BIASES LEADING TO UNFAIR NLP MODELS 11,737 (2013).

researchers "asked 14 language models to agree or disagree with 62 politically sensitive statements."[18] They were "surprise[d]" to conclude that the AI models they studied "have distinctly different political tendencies."[19] Although the researchers examined ways to potentially reduce political bias in LLMs, they ultimately concluded that the techniques, "while useful in theory, . . . might not be applicable in real-world settings."[20] As one of the researchers put it, "we believe no language model can be entirely free from political biases."[21]

B. AI Hallucinations

Bias is not the only problem users of LLMs must contend with. The tools tend to "hallucinate"—to make up facts and present them as "real."[22] The legal community is not immune from these problems. For example, New York lawyers were sanctioned for submitting a brief that included cases made up by ChatGPT.[23] These lawyers are a cautionary tale for how "believable" AI hallucinations can be to a human who is not aware of them and does not carefully review AI outputs. In their brief, those attorneys incorporated numerous references to cases that ChatGPT fabricated.[24] After opposing counsel asked for copies of the cases, the lawyers went back to ChatGPT, which generated fake texts for these fake cases (all while assuring the lawyers that the cases were, in fact, real).[25] The made-up cases included actual party names and purported to be authored by actual judges. However, the judge who sanctioned the New York lawyers noted that their legal analysis was "gibberish," and that "[t]he summary of the case's procedural

18. Heikkilä, *supra* note 17.

19. *Id.* For example, "[t]he researchers found that BERT models, AI language models developed by Google, were more socially conservative than OpenAI's GPT models."

20. FENG ET AL., *supra* note 17, at 11,745.

21. Heikkilä, *supra* note 17.

22. *See*, *e.g.*, Tate Ryan-Mosley, *Catching Bad Content in the Age of AI*, MIT TECH. REV. (May 15, 2023), https://www.technologyreview.com/2023/05/15/1073019/catching-bad-content-in-the-age-of-ai/ [https://perma.cc/2LV7-LH9Y] (discussing ChatGPT's "propensity to confidently make things up and present them as facts[.]").

23. Weiser, *supra* note 2.

24. *Id.* (The made-up cases included party names that one would expect from cases about airline litigations, including *Martinez v. Delta Air Lines*, *Varghese v. China Southern Airlines*, and *Zicherman v. Korean Air Lines*).

25. At a hearing before the judge, one of the lawyers explained that he thought ChatGPT was a super search engine that had access to legal cases that standard legal research databases did not. *See* Benjamin Weiser & Nate Schweber, *The ChatGPT Lawyer Explains Himself*, N.Y. TIMES (June 8, 2023), https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html [https://perma.cc/DB7M-23K8] (noting that the lawyer "repeatedly tried to explain why he did not conduct further research into the cases that ChatGPT had provided to him," and "that he had believed ChatGPT had greater reach than standard databases").

history is difficult to follow and borders on nonsensical."[26] Additionally, in the summer of 2023, an appeals court in Texas noted that lawyers there had also submitted a brief with cases that appeared to have been made up by generative AI.[27] And finally, in November of 2023, a Colorado attorney had his license suspended for one year and a day for actions that included citing case law he found on ChatGPT in a client brief, including cases that "were either incorrect or fictitious."[28]

C. Training Data Limitation

Today's LLMs were trained on data that often has a "cut off" date of several years in the past. For example, GPT-4, the latest GPT publicly available from OpenAI, "generally lacks knowledge of events that have occurred after the vast majority of its pre-training data cuts off in September 2021."[29] At the risk of stating the obvious, LLMs are therefore necessarily limited in their ability to produce prompts that respond to real-time events such as judicial decisions. Further, because LLMs are predicting words and producing outputs based on documents that already exist (as they must to be included in the training data), they are necessarily biased toward a present orientation: the outputs they create mirror the inputs that already exist.[30] In other words, an LLM arguably does not create anything truly "new."[31] As one commentator put it,

26. Weiser, *supra* note 2.

27. Lauren Berg, *Texas Appeals Court Calls Out Seemingly AI-Generated Cites*, LAW360 (July 26, 2023, 9:38 PM), https://www.law360.com/articles/1704217/texas-appeals-court-calls-out-seemingly-ai-generated-cites [https://perma.cc/5YUV-KM4S].

28. People v. Zachariah C. Crabill, No. 23PDJ067, 2023 WL 8111898, at *1 (Colo. O.P.D.J. Nov. 22, 2023).

29. OPENAI, GPT-4 TECHNICAL REPORT 10 (2023), https://arxiv.org/pdf/2303.08774.pdf [https://perma.cc/QE3C-2T7K].

30. *See* Katharine Miller, *LLMs Aren't Ready for Prime Time. Fixing Them Will Be Hard*, STAN. U. HUMAN-CENTERED A.I. (Oct. 17, 2023), https://hai.stanford.edu/news/llms-arent-ready-prime-time-fixing-them-will-be-hard [https://perma.cc/V5HQ-3SPW] ("Because LLMs merely predict what the next word should be when given an input text, they can only mimic the words and phrases that were used to train them.").

31. Whether AI can "create" something "new" is an ongoing debate that is outside the scope of this article. *See, e.g.*, Richard Moss, *Artificial Intelligence Challenges What It Means to Be Creative*, SCI. NEWS (Feb. 17, 2022, 7:00 AM), https://www.sciencenews.org/article/artificial-intelligence-ai-creativity-art-computer-program [https://perma.cc/D32W-K7RP] ("True creativity is a quest for originality. It is a recombination of disparate ideas in new ways. It is unexpected solutions. It might be music or painting or dance, but also the flash of inspiration that helps lead to advances on the order of light bulbs and airplanes and the periodic table. In the view of many in the computational creativity field, it is not yet attainable by machines."). In any event, even if generative AI is capable of "creative" or "original" legal ideas, an overreliance on the tools can still hamper lawyers' own creativity and weaken their legal arguments, as is addressed below.

"[n]o matter how impressive a piece of computer-created poetry or artwork might be, it's always built from blocks carved out of the data that's used to train it. In other words, it isn't genuinely capable of what we would call 'original thought' – having new ideas of its own."[32] As is discussed below, this bias that generative AI has toward what is already in existence could have profound impacts on the development of law if lawyers are overly reliant on it.

D. Training on Synthetic Data

LLMs are trained on massive data sets, and today's most well-known LLMs have been trained largely on datasets created by humans. For example, GPT-3, the precursor to ChatGPT, was trained on books, newspaper articles, and online forums such as Reddit.[33] However, as people use these LLMs to draft motions, write articles, and create blogposts, those AI-generated outputs become part of the new datasets that subsequent LLMs will be trained on. LLMs "are already being used to engorge the web with their own machine-made content, which will only continue to proliferate—across TikTok and Instagram, on the sites of media outlets and retailers, and even in academic experiments."[34] Since LLMs have the ability to create content at a scale no human can match, something one commentator termed the "John Henry problem with A.I.,"[35] a significant portion of the internet could be full of AI-generated text in the not-so-distant future. According to Daphne Ippolito, a senior research scientist at Google Brain, "[i]n the future, it's going to get trickier and trickier to find good-quality,

32. As noted in the preceding footnote, the question of whether AI can be truly "creative" is one that has confounded scholars for decades. *See* Bernard Marr, *The Intersection of AI and Human Creativity: Can Machines Really Be Creative?*, FORBES (Mar. 27, 2023, 2:48 AM), https://www.forbes.com/sites/bernardmarr/2023/03/27/the-intersection-of-ai-and-human-creativity-can-machines-really-be-creative/ [https://perma.cc/43W2-5ZB9].

33. *See* Cyphert, *supra* note 4, at 407 ("GPT-3 had an impressively large data training set: it was trained on the Common Crawl dataset, a nearly trillion-word dataset, which includes everything from traditional news sites like the New York Times to sites like Reddit. The Common Crawl dataset represented 60% of GPT-3's training set, and for the remaining 40%, the researchers included sources such as Wikipedia and historical books.").

34. Matteo Wong, *AI Is an Existential Threat to Itself*, THE ATLANTIC (June 21, 2023), https://www.theatlantic.com/technology/archive/2023/06/generative-ai-future-training-models/674478/ [https://perma.cc/SN2G-2QA2].

35. Jon Gertner, *Wikipedia's Moment of Truth*, N.Y. TIMES (Sept. 8, 2023), https://www.nytimes.com/2023/07/18/magazine/wikipedia-ai-chatgpt.html [https://perma.cc/V9KZ-TJ6C] (noting that "[t]he chatbots, unlike their human counterparts, have a formidable ability to churn out language like a steam-driven machine, 24/7.").

guaranteed AI-free training data."[36] As the next section explains, this proliferation of AI-generated content is a problem for the next generation of LLMs.

III. AI CANNIBALISM

The term "AI cannibalism" refers to the phenomenon of AI being trained on AI-generated content.[37] Such training "runs the risk of creating a feedback loop, where AI models 'learn' from content that was itself AI-generated, resulting in a gradual decline in output coherence and quality."[38] A recent study by researchers at Rice University and Stanford University concluded that generative AI image models degrade when trained on their own outputs after very few generations.[39] In that study, the authors approximated what would happen if AI models were trained on ever more increasing amounts of synthetic data (data created by AI). The researchers mixed images created by humans with AI-generated images and used these datasets to train many generations of models. They tested what happened if they changed the balance of real versus synthetic data included in the training dataset and how much synthetic output was selected for the next generation of training input.[40]

The study demonstrated that incorporating real and novel "human" data was crucial for maintaining the AI system's capability of producing accurate results.[41] The researchers used three mix types of real versus synthetic data for training.[42] In "fully synthetic," the models were trained for all generations on fully synthetic data. In "synthetic augmentation," the models had a fixed

36. Melissa Heikkilä, *How AI-generated Text Is Poisoning the Internet*, MIT TECH. REV. (Dec. 20, 2022), https://www.technologyreview.com/2022/12/20/1065667/how-ai-generated-text-is-poisoning-the-internet/ [https://perma.cc/D5F2-ENTK].

37. *See* Christian Guyton, *ChatGPT Use Declines as Users Complain About 'Dumber' Answers, and the Reason Might Be AI's Biggest Threat for the Future*, TECHRADAR (July 14, 2013), https://www.techradar.com/computing/artificial-intelligence/chatgpt-use-declines-as-users-complain-about-dumber-answers-and-the-reason-might-be-ais-biggest-threat-for-the-future [https://perma.cc/U3MD-Z9YJ] (explaining AI cannibalism by noting that "large language models (LLMs) like ChatGPT and Google Bard scrape the public internet for data to be used when generating responses. In recent months, a veritable boom in AI-generated content online - including an unwanted torrent of AI-authored novels on Kindle Unlimited - means that LLMs are increasingly likely to scoop up materials that were already produced by an AI when hunting through the web for information.").

38. *Id.*

39. *See generally* Alemohammad et al., *supra* note 6.

40. A somewhat surprising result in the paper was that adding small amounts of synthetic data improved performance, but when that amount exceeded some threshold the performance of the model degraded. *See id.* at 14.

41. *Id.* at 6.

42. They termed these as "fully synthetic," "synthetic augmentation," and "fresh data." *Id.* at 3–4.

set of real data but received new synthetic data each generation.[43] In "fresh data," the models received new amounts of real data during training that were combined with synthetic data.[44] In experiments with "fully synthetic" and "synthetic augmentation," the models collapsed and degraded fairly quickly.[45] However, if "fresh data" was available (and used correctly), model collapse could be avoided for much longer periods of time even when mixed with new synthetic data. Having real data—data created, cultivated, and curated by humans—made a big difference.

The authors also tried to choose the outputs of their generative AI model to select certain synthetic samples for use in the next generation of training sets.[46] This tries to simulate the practice AI companies use to increase the quality of their synthetic data by having humans rank the output.[47] This practice can be very labor intensive and expensive, so the authors instead used different statistical distributions to select the amount of synthetic and real data in the training sets, and also to change what samples of synthetic output get "kept" for future training.[48]

The study only includes models for image data, but the authors imply that the argument applies to text (and, as noted below, other studies have found similar results with LLMs).[49] It will be important for other research groups to test those claims with experiments as well. What the study makes very clear is the

43. *Id.* at 4.

44. *Id.*

45. *Id.* (explaining degradation in "fully synthetic" and "synthetic augmentation" models). The term collapse is used by the generative AI community to refer to the problem of models not having enough variance in their output and thus always outputting the same answers. This is also related to the vanishing gradient problem - which can prevent machine learning models from improving during training. *See* Martin Arjovsky & Léon Bottou, *Towards Principled Methods for Training Generative Adversarial Networks*, ARXIV at 6 (Jan. 17, 2017), https://arxiv.org/pdf/1701.04862.pdf [https://perma.cc/C7X5-SLHJ]; *see also* Ishaan Gulrajani et al., *Improved Training of Wasserstein GANs*, ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 6 (2017); *Common Problems*, GOOGLE, https://developers.google.com/machine-learning/gan/problems [https://perma.cc/5Q3E-2ESX] (July 18, 2022).

46. Alemohammad et al., *supra* note 6, at 11.

47. Long Ouyang et al., *Training Language Models to Follow Instructions with Human Feedback*, ARXIV at 2 (Mar. 4, 2023), https://arxiv.org/pdf/2203.02155.pdf [https://perma.cc/5KMQ-GKH5] (describing a model called *InstructGPT* which is combined with GPT-3 to create ChatGPT. OpenAI researchers explain that they use humans in the loop to label the best examples of model outputs. These examples are used to create a new dataset and used to train the next generation of models. ChatGPT is thus trained partly on the best outputs of GPT-3 to a certain set of prompts as ranked by humans. OpenAI calls this technique Reinforcement Learning with Human Feedback (RLHF) and describes the technique as follows: "We then collect a dataset of rankings of model outputs, which we use to further fine-tune this supervised model using reinforcement learning from human Feedback").

48. Alemohammad et al., *supra* note 6, at 13.

49. *Id.* at 3.

importance of datasets that are curated to identify which samples are real versus synthetic, and which samples are of high quality, as rated by humans. However, what will it mean for data to be high-quality? It may be that high quality dataset labels are only valuable for certain model applications and not others. Additionally, it is not clear that everyone will agree on what "high-quality data" means.

Another challenge for researchers and AI companies will be determining applicable metrics for model evaluation. Classic metrics for evaluating the performance of machine learning are widely accepted and allow researchers to compare different models against the same datasets.[50] The best model is the one that performs with the lowest error rates. But what happens when human users or user adoption is the model test? Generative models have long had the problem of researchers disagreeing on the right evaluation metrics.[51]

It is important to note that the Alemohammad et al. study has not yet been peer-reviewed nor has the approach been tested on LLMs. However, other studies specific to LLMs show similar results.[52] A group of researchers used model-generated content in language model training data and found that it caused "irreversible defects in the resulting models."[53] This group investigated the impact of what happens to one language model if increasing portions of its training data included auto-generated text. They found the result is that the output of later model versions moves further away from the original distribution established by the real human-curated data.[54] They termed this phenomenon "model collapse."[55]

Their experiments showed that multiple families of Machine Learning models (typically trained on smaller datasets than LLMs) were susceptible to this problem.[56] However, LLMs are different due to their size. They are so large and expensive to train that most researchers use them by "fine-tuning" an existing LLM toward a

50. TREVOR HASTIE ET AL., THE ELEMENTS OF STATISTICAL LEARNING 219–57 (2001) (discussing approaches to evaluating statistical models, including machine learning); *see also* Jesse Davis & Mark Goadrich, *The Relationship Between Precision-Recall and ROC Curves*, Proc. 23rd Int'l Conf. on Machine Learning 233 (2006) (discussing additional metrics).

51. Shane Barratt & Rishi Sharma, *A Note on the Inception Score*, ARXIV at 3 (June 21, 2018), https://arxiv.org/pdf/1801.01973.pdf [https://perma.cc/G8QR-JY8Q] (noting that a popular metric for comparing the outputs of Generative Adversarial Networks (GANs) called "Inception Score" is fraught with problems).

52. *See generally* Ilia Shumailov et al., *The Curse of Recursion: Training on Generated Data Makes Models Forget*, ARXIV (May 31, 2023), https://arxiv.org/pdf/2305.17493.pdf [https://perma.cc/7JRV-PGXQ].

53. *Id.* at 2.

54. *Id.*

55. *Id.*

56. *Id.* at 10.

more specific purpose. For smaller experiments, the team tried up to two hundred future model generations, but it is computationally impractical and cost prohibitive to run experiments where you train an LLM for many iterations.[57] Even a single training run for an LLM can take weeks to months and cost millions of U.S. dollars.[58] Instead, they experimented with taking an LLM and further training it, thus "fine-tuning" it to be more performative on a specific new data set. They used this new model output as training data for new iterations of the fine-tuning process. After a few generations, the models collapsed. The researchers noted: "[L]ater generations start producing samples that would never be produced by the original model *i.e.* they start misperceiving reality based on errors introduced by their ancestors."[59] As to the cost, they reported that the experiments with fine-tuning LLMs alone took weeks to run.[60]

These research teams were not the first to suggest that the outputs of generative AI would "degrade" over time as the training datasets included more and more AI-generated images or text.[61] Indeed, a very similar thing happened to early internet keyword searches. These searches worked well due in large part to the limited number of websites.[62] For example, in 1993 there were only about 130 websites on the internet, but just three years later, there were over one hundred thousand.[63] Results from keyword searches became less useful as the web increased in size.[64] Additionally, "Search Engine Optimization" companies worked to get their clients' content prominently positioned in search results. This led to an arms race between internet search engines, who tried to

57. Lennart Heim, *Estimating PaLM's Training Cost*, BLOG.HEIM.XYZ (Apr. 5, 2022), https://blog.heim.xyz/palm-training-cost/ [https://perma.cc/KB4R-EQM2].

58. *See* Aakanksha Chowdhery et al., *PaLM: Scaling Language Modeling with Pathways*, ARXIV at 66 (Oct. 5, 2022), https://arxiv.org/pdf/2204.02311.pdf [https://perma.cc/6VZ9-8RZH] (discussing how Google reported training "PaLM-540B on 6144 TPU v4 chips for 1200 hours and 3072 TPU v4 chips for 336 hours including some downtime and repeated steps"); *see also* Lennart Heim, *Estimating PaLM's Training Cost*, BLOG.HEIM.XYZ (Apr. 5, 2022), https://blog.heim.xyz/palm-training-cost/ [https://perma.cc/S6R6-NS3V] (estimating that the cost of training Google's PaLM LLM to be between $9 million and $23 million).

59. Shumailov et al., *supra* note 52, at 12.

60. *Id.* at 11 n.4 (noting that "just the language experiments described in the paper took weeks to run").

61. *See* Lasha Maden & Adam Rogers, *Search and Ye Might Find*, 99% INVISIBLE (Sept. 13, 2022), https://99percentinvisible.org/episode/search-and-ye-might-find/ [https://perma.cc/NQ85-9UW5]; *see also id.* at 3.

62. *See* Maden & Rogers, *supra* note 61.

63. *See id.*

64. Shumailov et al., *supra* note 52, at 12–13 (noting that "[t]he negative effect these poisoning attacks had on search results led to changes in search algorithms: e.g., Google downgraded farmed articles, putting more emphasis on content produced by trustworthy sources e.g. education domains, while DuckDuckGo removed them altogether.").

improve their algorithms to avoid artificially inflated results, and spammers, who tried to work around these algorithms.[65]

Although the developers of LLMs do "curate" the datasets they use to train the models,[66] it will not be easy for the developers of the next generation of LLMs to efficiently "weed out" AI-generated content from data training sets. This concern will intensify as AI continues to generate a larger volume of information that it then utilizes as inputs. As a threshold problem, it is becoming increasingly difficult for humans to distinguish between content produced by other humans and content produced by AI.[67] Further, even if we could easily distinguish this content, each successive generation of LLMs has needed exponentially larger datasets than their predecessors to sustain the explosive growth in the utility of the tools. Each new generation of LLMs is "trained on ever more data, and the number of parameters—the variables in the models that get tweaked—is rising dramatically."[68] The growth from one generation to the next is truly exponential. For example, GPT-2 had 1.5 billion parameters and GPT-3 had 175 billion parameters.[69] To continue this exponential growth, the datasets will have to be massive, and excluding AI-generated content simply may not be possible while sustaining dataset size. Developers of AI are racing to address the issues caused by AI cannibalism,[70] but unless and until they find a sustainable solution, it is a problem to be mindful of.

IV. THE IMPACT OF AI ON LAWYERS AND LAW

The biases and problems outlined in Section III will have a dramatic impact on the use of AI for legal services. For example, as noted above, AI hallucinations have already appeared in AI-generated legal briefs, resulting in documents filled with fabricated

65. *See*, *e.g.*, Shumailov, *supra* note 52, at 12 (describing "the creation of *click*, *content*, *and troll* farms . . . whose job is to misguide social networks and search algorithms," as well as Google's response of downgrading "farmed" articles in its search results).

66. *See*, *e.g.*, Wong, *supra* note 34 (noting that "[f]iltering is a whole research area right now").

67. *See*, *e.g.*, Gil Press, *Is It an AI Chatbot or a Human? 32% Can't Tell*, FORBES (June 1, 2023, 8:19 AM), https://www.forbes.com/sites/gilpress/2023/06/01/is-it-an-ai-chatbot-or-a-human-32-cant-tell/ [https://perma.cc/778C-H2JZ] (describing an experiment wherein 32% of research participants could not correctly identify, after a 2-minute-long conversation, whether the "person" they were talking to was a human or a chatbot).

68. Rotman, *supra* note 1.

69. *See* Cyphert, *supra* note 4, at 407–08.

70. *See, e.g.*, Cade Metz et al., *How Tech Giants Cut Corners to Harvest Data for A.I.*, N.Y. TIMES, https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html [https://perma.cc/RR8W-FER9] (Apr. 8, 2024) (describing how OpenAI and other developers of generative AI are attempting to create AI-generated synthetic data that would not cause model collapse problems).

cases resulting in attorney and law firm sanctions and discipline. AI can also mirror and exacerbate the implicit biases already found in our legal system. AI might amplify these biases if left unchecked and allowed to cannibalize data that is already biased. Finally, the limitations on LLM dataset training, when applied to the legal field, have the potential to render the law stagnant and hinder the emergence of innovative new legal theories that could better reflect future values.

A. AI Hallucinations

The potential impacts outlined above—more disinformation, increasing AI hallucinations, etc.—obviously pose a problem for any user of an LLM, but we argue here that these potential impacts are especially acute for lawyers. Lawyers are, of course, "the most highly paid rhetoricians in the world"[71] and writing is one of our most important products. Thus, as one of us has argued elsewhere, lawyers might be especially drawn to LLMs.[72] As LLMs produce more AI hallucinations and more misinformation, these tools will be of limited utility to lawyers and could subject them to malpractice and professional discipline. Indeed, judges across the nation are beginning to issue orders requiring the lawyers who appear before them to disclose if they have used generative AI in drafting the briefs they file, and sometimes even requiring them to confirm that a human has in fact checked that the cases cited are real.[73] For example, Judge Brantley Starr of the Northern District of Texas requires attorneys appearing before him to file "a certificate attesting either that no portion of any filing will be drafted by generative artificial intelligence (such as ChatGPT, Harvey.AI, or Google Bard) or that any language drafted by generative artificial intelligence will be checked for accuracy, using print reporters or traditional legal databases, by a human being."[74]

71. Bryan A. Garner, *Why Lawyers Can't Write*, ABA J. (Mar. 1, 2013, 9:00 AM), https://www.abajournal.com/magazine/article/why_lawyers_cant_write [https://perma.cc/R2XN-EGX4].

72. *See* Cyphert, *supra* note 4, at 408 (noting that lawyers' ability to customize GPT-3, a precursor to ChatGPT, on legal documents "is one of the reasons GPT-3 is well-poised for wide adoption in the legal field").

73. This is one way that judges will address regulation of AI before Congress takes up the topic. *See* Melissa Heikkilä, *How Judges, Not Politicians, Could Dictate America's AI Rules*, MIT TECH. REV. (July 17, 2023), https://www.technologyreview.com/2023/07/17/1076416/judges-lawsuits-dictate-ai-rules/ [https://perma.cc/GZN5-9323] ("It's becoming increasingly clear that courts, not politicians, will be the first to determine the limits on how AI is developed and used in the US.").

74. *See Judge Starr's Mandatory Certification Regarding Generative Artificial Intelligence*, U.S. DIST. CT. DIST. OF TEX., https://www.txnd.uscourts.gov/judge/judge-

Judge Michael Baylson of the Eastern District of Pennsylvania has a similar requirement,[75] as does Judge Stephen Vaden of the Court of International Trade.[76] As of January 2024, the Fifth Circuit is considering a rule that would require "attorneys to verify they checked the accuracy of any generative artificial intelligence material they file with the court."[77] Courts have certainly taken notice of AI hallucinations, and lawyers will have to be mindful of the phenomenon, at least until AI developers develop better tools to combat it. For example, OpenAI reports that its GPT-4 model produces significantly fewer hallucinations than ChatGPT, but "it still is not fully reliable" because "it 'hallucinates' facts and makes reasoning errors."[78]

B. Bias

As is discussed above in Part III.A, AI systems' ability to produce biased outputs has been well-documented.[79] Some have

brantley-starr [https://perma.cc/396K-5Q45] (last visited Apr. 21, 2024) (Judge Starr's statement also discusses bias in generative AI).

75. *See Standing Order re: Artificial Intelligence ("AI") in Cases Assigned to Judge Baylson*, U.S. DIST. CT. E. DIST. OF PA. (June 6, 2023), https://www.paed.uscourts.gov/sites/paed/files/documents/Standing%20Order%20Re%20Artificial%20Intelligence%206.6.pdf [https://perma.cc/RY38-2LWH] ("If any attorney for a party, or a pro se party, has used Artificial Intelligence ("AI") in the preparation of any complaint, answer, motion, brief, or other paper, filed with the Court, and assigned to Judge Michael M. Baylson, **MUST**, in a clear and plain factual statement, disclose that AI has been used in any way in the preparation of the filing, and **CERTIFY**, that each and every citation to the law or the record in the paper, has been verified as accurate." (emphasis in original)).

76. *See Order on Artificial Intelligence*, U.S. CT. OF INT'L TRADE (June 8, 2023), https://www.cit.uscourts.gov/sites/cit/files/Order%20on%20Artificial%20Intelligence.pdf [https://perma.cc/AE5X-CN8C] (Judge Vaden's order places great emphasis on privacy and confidentiality, and on the ways in which lawyers using generative AI may compromise their clients' confidential information. Thus, his order requires that any submission that was created "with the assistance of a generative artificial intelligence program on the basis of natural language prompts, including but not limited to ChatGPT and Google Bard, must be accompanied by: (1) A disclosure notice that identifies the program used and the specific portions of text that have been so drafted; (2) A certification that the use of such program has not resulted in the disclosure of any confidential or business proprietary information to any unauthorized party[.]").

77. Jacqueline Thomsen, *Lawyers Must Certify AI Review Under Fifth Circuit Proposal*, BLOOMBERG L. (Nov. 21, 2023, 4:26 PM), https://news.bloomberglaw.com/us-law-week/lawyers-must-certify-ai-review-under-fifth-circuit-proposal [https://perma.cc/Y2XK-CRTC]; *see also Notice of Proposed Amendment to 5TH CIR. R. 32.3*, U.S. CT. OF APP. FOR THE FIFTH CIR., https://www.ca5.uscourts.gov/docs/default-source/default-document-library/public-comment-local-rule-32-3-and-form-6 [https://perma.cc/D8SS-MF93] (last visited Apr. 21, 2024) (the comment period on the proposed rule closed on January 4, 2024).

78. OPENAI, *supra* note 29, at 10.

79. *See*, *e.g.*, *Tinker-ing with Machine Learning*, *supra* note 12, at 462–64 (discussing "the various ways that algorithms and machine learning can be inadvertent tools for

suggested that AI cannibalism could further exacerbate the problem of bias already present in LLMs.[80] As one of us wrote previously, "[b]ecause discrimination by lawyers 'undermine[s] confidence in the legal profession and the legal system,' the Model Rules deem it professional misconduct for lawyers to engage in conduct that is harassment or discriminatory while practicing law," and "[i]t is not hard to imagine a scenario involving a lawyer using [an LLM] that would implicate" that model rule.[81] An increase in the bias of the outputs of future models only makes this issue even more acute for lawyers. Judge Starr addresses the issue of bias in his order on AI. He notes that AI systems are prone to bias, and explains why this creates special concerns for lawyers:

> While attorneys swear an oath to set aside their personal prejudices, biases, and beliefs to faithfully uphold the law and represent their clients, generative artificial intelligence is the product of programming devised by humans who did not have to swear such an oath. As such, these systems hold no allegiance to any client, the rule of law, or the laws and Constitution of the United States (or, as addressed above, the truth). Unbound by any sense of duty, honor, or justice, such programs act according to computer code rather than conviction, based on programming rather than principle.[82]

If AI cannibalizes too much of its own information, its output could act as a magnifying glass of unjust results. Marginalized communities will only become more marginalized as AI algorithms become more adept at amplifying the current implicitly biased inputs. These feed-forward loops will make it much harder to break the cycle of unjust laws. It will be crucial for lawyers who use generative AI to be aware of the propensity toward biased outputs and to carefully review and supervise all generated text.

C. Impact of AI on the Development of Law

Of course, even if developers are somehow able to limit AI cannibalism, and thus the outputs of LLMs remain relatively useful, there are still potential problems for the field of law. If lawyers overly rely on generative AI, the very development of the

deepening inequality[.]"); *see also* Drew Simshaw, *Access to A.I. Justice: Avoiding an Inequitable Two-Tiered System of Legal Services*, 24 YALE J.L. & TECH. 150, 200 (2022) ("Bias can manifest in virtually any AI-driven legal process.").

80. *See* Wong, *supra* note 34 ("Of greater concern is the compounding of smaller, hard-to-detect biases and misperceptions—especially as machine-made content becomes harder, if not impossible, to distinguish from human creations.").

81. Cyphert, *supra* note 4, at 434–35.

82. U.S. DIST. CT. DIST. OF TEX., *supra* note 74.

field of law could be altered. For example, take appellate lawyers, who are often advocating for important changes to the existing state of law. An appellate lawyer who relies on generative AI to draft an appellate brief may be less likely to advance a novel way of approaching law because of the generative AI training data limitation discussed above. If lawyers rely on an LLM trained on the existing corpus of law to craft legal arguments, they may not offer as many novel arguments in favor of the expansion or change of existing law.[83] This could cause stagnation in the development of law. This is especially true if the AI the lawyers are using is trained on too much synthetic data, as the outcome could make the law more static and frozen in one moment in time. Lawyers play an important role in helping ensure that law evolves and develops in ways that benefit and mirror society.

"Legal formalism" is a school of thought that posits that the outcome of a court case is either correct or incorrect, in the same way a math problem is correct or incorrect.[84] In contrast, "legal realists" argue that the result of a court case both is and should be based on public policy and our current political values.[85] Overuse of AI has the potential to upend our current legal system by moving toward a formalistic approach and away from legal realism. AI output is only as good as its input. If AI cannibalizes its own output, it creates a recursive loop of information. When applied to legal cases, this means that AI tools will only reflect the current state of the law and will have a harder time adjusting to novel legal arguments that may better reflect future societal values. This may lead to stagnation in the law resulting in a turn toward formalism.

CONCLUSION

Generative AI has the potential to help lawyers be faster, more efficient, and better at writing. But it also has the potential to embed bias, embarrass lawyers, subject them to professional discipline, and negatively impact creativity in the practice and the very development of law itself. Whether the tools ultimately prove

83. *See, e.g.*, Cameron Shackell, *Will AI Kill Our Creativity? It Could – If We Don't Start to Value and Protect the Traits That Make Us Human*, THE CONVERSATION (Sept. 27, 2023, 4:04 PM), https://theconversation.com/will-ai-kill-our-creativity-it-could-if-we-dont-start-to-value-and-protect-the-traits-that-make-us-human-214149#:~:text=The%20danger%20here%20is%20that,deviations%20from%20the%20status%20quo [https://perma.cc/UQ6A-B3U2] ("AI models don't contain reality. They rely on the complex statistical abstraction of digital data. This limits their real-world creative significance and their capacity to produce 'eureka' moments.").

84. Richard A. Posner, *Legal Formalism, Legal Realism and the Interpretation of Statues and the Constitution*, 37 CASE W. RSRV. L. REV. 179, 181 (1986).

85. *Id.*

to be more help than hindrance will depend in part on how the tools themselves develop and whether AI cannibalism reduces their utility for all. It will also depend on how ethically and professionally lawyers use them and whether they can resist the "easy out" of overusing them for drafting tasks and instead use them to supplement, augment, and improve their own writing.