Generative AI Meets Responsible AI: Practical Challenges and Opportunities

Krishnaram Kenthapadi
Fiddler AI
krishnaram@fiddler.ai

Himabindu Lakkaraju
Harvard University
hlakkaraju@hbs.edu

Nazneen Rajani
Hugging Face
nazneen@huggingface.co

ABSTRACT

Generative AI models and applications are being rapidly developed and deployed across a wide spectrum of industries and applications ranging from writing and email assistants to graphic design and art generation to educational assistants to coding to drug discovery [12]. However, there are several ethical and social considerations associated with generative AI models and applications. These concerns include lack of interpretability, bias and discrimination, privacy, lack of model robustness, fake and misleading content, copyright implications, plagiarism, and environmental impact associated with training and inference of generative AI models.

In this tutorial, we first motivate the need for adopting responsible AI principles when developing and deploying large language models (LLMs) and other generative AI models, as part of a broader AI model governance and responsible AI framework, from societal, legal, user, and model developer perspectives, and provide a roadmap for thinking about responsible AI for generative AI in practice. We provide a brief technical overview of text and image generation models, and highlight the key responsible AI desiderata associated with these models. We then describe the technical considerations and challenges associated with realizing the above desiderata in practice. We focus on real-world generative AI use cases spanning domains such as media generation, writing assistants, copywriting, code generation, and conversational assistants, present practical solution approaches / guidelines for applying responsible AI techniques effectively, discuss lessons learned from deploying responsible AI approaches for generative AI applications in practice, and highlight the key open research problems. We hope that our tutorial will inform both researchers and practitioners, stimulate further research on responsible AI in the context of generative AI, and pave the way for building more reliable and trustworthy generative AI applications in the future.

ACM Reference Format:
Krishnaram Kenthapadi, Himabindu Lakkaraju, and Nazneen Rajani. 2023. Generative AI Meets Responsible AI: Practical Challenges and Opportunities. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3580305.3599557

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08.
https://doi.org/10.1145/3580305.3599557

1 OUTLINE OF THE TUTORIAL

The tutorial will consist of two parts: (1) technical deepdive into generative AI landscape including advances, challenges, and opportunities (90 minutes); (2) ethical considerations including privacy, consent, and responsible release, along with approaches for mitigating harms and long term planning (90 minutes).

Introduction and overview of the generative AI landscape (15 minutes). Give an overview of the generative AI landscape in ML and motivate the topic with some questions. What constitutes generative AI? Why is generative AI an important topic? What are the origins of the research field?

Technical overview of LLMs and other generative AI models (75 minutes).
(1) Generative AI models for different domains such as image generation (e.g., Stable diffusion [23], CLIP [22]), text generation (e.g., BLOOM [25], InstructGPT [20], OPT [33]), dialog agents (e.g., ChatGPT, LaMDA, Sparrow, Claude, BlenderBot 3), code generation (e.g., Codex, AlphaCode, CodeWhisperer), video generation (e.g., Make-a-video), and audio generation (e.g., AudioLM).
(2) Applications of generative AI - images, music, text, code, video.
(3) Model training: (a) Pretraining method and datasets; (b) Diffusion approach; (c) Supervised fine-tuning [9]; (d) Instruction datasets – Self-instruct [31], Supernatural Instructions [30]; (e) Reinforcement Learning with Human Feedback (RLHF) [8, 16, 34]; (f) Compute costs and infrastructure [21, 27]
(4) Model evaluation [5] and auditing [18] including (a) metrics, datasets, and benchmarks; (b) Automated vs. human evaluations; (c) Red-teaming [10] and evaluations on toxicity/harmfulness.
(5) Model Access.

Technical and ethical challenges with generative AI and solution approaches (90 minutes) We will highlight the following challenges [2, 4]:
(1) Trust and lack of interpretability: are significant concerns for LLMs and other generative AI models especially due to their large size and opaque behavior. Often, such models exhibit emergent behavior, and demonstrate capabilities not intended as part of the architectural design and not anticipated by the model developers [14]. A lack of transparency, lineage, and trustworthiness prevents users from validating and citing the responses generated by search and information retrieval mechanisms powered by LLMs [17, 26]. Further, LLMs and other generative AI models could be used to generate fake and misleading content (including deepfakes) and spread misinformation with serious social and political consequences.
(2) Bias and discrimination: Generative AI models are often trained on large corpuses of data, making it difficult to audit the training data for different types of biases [2]. For example, many LLMs have

been shown to exhibit different types of biases such as gender stereotypes [3, 15], undesirable biases towards mentions of disability [13], and religious stereotypes [1]. Similarly, contrastive language-vision AI models (such as Stable Diffusion) trained on automatically collected web scraped data have been shown to learn biases of sexual objectification, which can propagate to downstream applications [32]. Further, generative AI models are typically trained on data crawled from the internet, and consequently the models often reflect the practices of the wealthiest communities and countries [2].
(3) Privacy and copyright implications: LLMs have been shown to memorize personally identifiable information occurring just once in the training data and reproduce such data, raising potential privacy concerns [7, 11]. Further, image diffusion models such as DALL-E 2, Imagen, and Stable Diffusion have been shown to memorize individual images from their training data and emit them at generation time, with potential privacy as well as copyright implications [6].
(4) Model robustness and security: LLMs often lack the ability to provide uncertainty estimates [24]. Without knowledge of the extent of confidence (or uncertainty) of the model, it becomes difficult for users to decide when the model's output can be trusted [19]. Model security is a key concern for generative AI models, especially since several applications may be derived from the same underlying foundation model. LLMs have been shown to be vulnerable to data poisoning attacks [29].
We will discuss practical solution approaches such as watermarking, release norms [28], red-teaming [10], and confidence building measures (CBMs).

2 CONCLUSION

In light of the increasing role played by AI based systems in our daily lives and the disruptive impact of generative AI models and systems, responsible AI techniques need to be incorporated when developing and deploying LLMs and other generative AI models to help build trust into such systems and applications. Our tutorial is a step towards helping data scientists and ML developers build generative AI systems that are secure, privacy-preserving, transparent, explainable, fair, and accountable – to avoid unintended consequences and compliance challenges that can be harmful to individuals, businesses, and society. By emphasizing the need for responsible AI as well as key challenges, we aspire to stimulate further research on responsible AI in the context of generative AI, and thereby pave the way for developing more reliable and trustworthy generative AI applications in the future.

REFERENCES

[1] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Large language models associate Muslims with violence. *Nature Machine Intelligence* 3, 6 (2021), 461–463.
[2] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Margaret Mitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *ACM Conference on Fairness, Accountability, and Transparency*.
[3] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *NeurIPS*.
[4] Rishi Bommasani et al. 2021. On the Opportunities and Risks of Foundation Models. *ArXiv* (2021). https://crfm.stanford.edu/assets/report.pdf
[5] Rishi Bommasani, Daniel Zhang, Tony Lee, and Percy Liang. 2023. Improving Transparency in AI Language Models: A Holistic Evaluation. *Stanford HAI Policy Brief* (2023).
[6] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. 2023. Extracting training data from diffusion models. *arXiv preprint arXiv:2301.13188* (2023).
[7] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting Training Data from Large Language Models. In *USENIX Security Symposium*, Vol. 6.
[8] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *NeurIPS*.
[9] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416* (2022).
[10] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858* (2022).
[11] Jie Huang, Hanyin Shao, and Kevin Chang. 2022. Are Large Pre-Trained Language Models Leaking Your Personal Information?. In *ICML Workshop on Knowledge Retrieval and Language Models*.
[12] Sonya Huang, Pat Grady, and GPT-3. 2022. Generative AI: A Creative New World. https://www.sequoiacap.com/article/generative-ai-a-creative-new-world/
[13] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social Biases in NLP Models as Barriers for Persons with Disabilities. In *ACL*. 5491–5501.
[14] Subbarao Kambhampati. 2022. Changing the nature of AI research. *Commun. ACM* 65, 9 (2022).
[15] Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*. 48–55.
[16] James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, Guan Wang, David L Roberts, Matthew E Taylor, and Michael L Littman. 2017. Interactive learning from policy-dependent human feedback. In *ICML*.
[17] Donald Metzler, Yi Tay, Dara Bahri, and Marc Najork. 2021. Rethinking search: Making domain experts out of dilettantes. In *ACM SIGIR Forum*, Vol. 55.
[18] Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2023. Auditing large language models: A three-layered approach. *arXiv preprint arXiv:2302.08500* (2023).
[19] Andrew Ng. 2022. ChatGPT Mania!, Crypto Fiasco Defunds AI Safety, Alexa Tells Bedtime Stories. https://www.deeplearning.ai/the-batch/issue-174/ The Batch — Deeplearning.ai newsletter.
[20] Long Ouyang et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155* (2022).
[21] David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350* (2021).
[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.
[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.
[24] Karthik Abinav Sankararaman, Sinong Wang, and Han Fang. 2022. BayesFormer: Transformer with Uncertainty Estimation. *arXiv preprint arXiv:2206.00826* (2022).
[25] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv preprint arXiv:2211.05100* (2022).
[26] Chirag Shah and Emily M Bender. 2022. Situating search. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*. 221–232.
[27] Or Sharir, Barak Peleg, and Yoav Shoham. 2020. The cost of training NLP models: A concise overview. *arXiv preprint arXiv:2004.08900* (2020).
[28] Irene Solaiman. 2023. The Gradient of Generative AI Release: Methods and Considerations. *arXiv preprint arXiv:2302.04844* (2023).
[29] Eric Wallace, Tony Zhao, Shi Feng, and Sameer Singh. 2021. Concealed Data Poisoning Attacks on NLP Models. In *NAACL-HLT*.
[30] Yizhong Wang et al. 2022. Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks. In *EMNLP*.
[31] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-Instruct: Aligning Language Model with Self Generated Instructions. *arXiv preprint arXiv:2212.10560* (2022).
[32] Robert Wolfe, Yiwei Yang, Bill Howe, and Aylin Caliskan. 2022. Contrastive Language-Vision AI Models Pretrained on Web-Scraped Multimodal Data Exhibit Sexual Objectification Bias. *arXiv preprint arXiv:2212.11261* (2022).
[33] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068* (2022).
[34] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593* (2019).