Generative AI Data and Training Issues

Bias, Privacy, Intellectual Property

Derek Harter

Department of Computer Science
East Texas A&M University

Spring 2024

Divisions of Artificial Intelligence

01

ARTIFICIAL INTELLIGENCE

Artificial Intelligence is the mechanism to incorporate human intelligence into machines through a set of rules(algorithms).

02

MACHINE LEARNING

Machine Learning is an application of AI that provides systems the ability to automatically learn, predict, and improve from experience without being explicitly programmed.

03

DEEP LEARNING

Deep Learning is a subset of ML that uses Neural Networks(similar to the neurons working in our brain) to mimic human brain-like behavior.

04

GENERATIVE AI

Generative AI, also known as generative modeling or generative deep learning, refers to the branch of deep learning that focuses on creating new content or data that resembles a

EAST TEXAS A&M

Generative AI Data and Training

- Large Language Models (LLM)
  - ChatGPT (3, 3.5, 4 OpenAI)
  - Claude (Anthropic)
  - LLaMa (Meta)
- Image Generators
  - DallE (2, 3 OpenAI)
  - Stable Difussion / DreamStudio (Stability AI)
  - Gemini (Google)

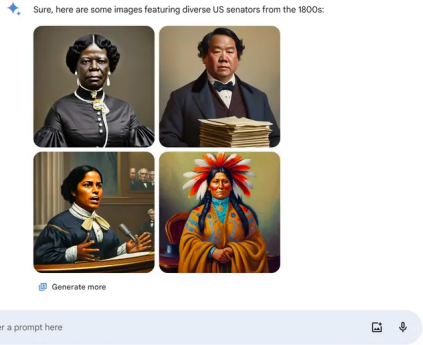


Figure 1: Google Gemini results for prompt to generate senator from the 1800s

EAST TEXAS A&M

Generative AI Data and Training

ChatGPT Facts:

- ChatGPT-3 was trained on a massive corpus of text data, around 570GB of datasets, including web pages, books, and other sources.
- GPT-3 has also been criticized for its lack of common sense knowledge and susceptibility to producing biased or misleading responses.
- ChatGPT generates biased or inaccurate responses, particularly when the model has not been fine-tuned on specific domain.
- ChatGPT 3.5 and 4 are not reliable as they hallucinate facts and make reasoning errors.
- Open AI is facing several lawsuits for copyright infringement in ChatGPT's training process. Some of the plaintiffs are The New York Times and book authors like George R.R. Martin and John Grisham.

EAST TEXAS A&M

Issue: Bias and Fairness

Figure 2: Survey of representation in image generation for

- Training data reflects the biases present in society.
- Results can range from mild problems (Ali et al., 2024) Figure 5, to more serious like medical applications (Hastings, 2024)

Bias

An ambitious study by Travis Zack and Eric Lehman and colleagues in The Lancet Digital Health comprehensively shows that GPT-4 exhibits racial and gender bias across clinically relevant tasks, including the generation of cases for medical education, support for differential diagnostic reasoning, medical plan recommendation, and subjective assessments of patients.

EAST TEXAS A&M

Issue: Intellectual Property and Copyright

Copyright is not an effective solution to the challenges faced by creators in the age of generative AI:

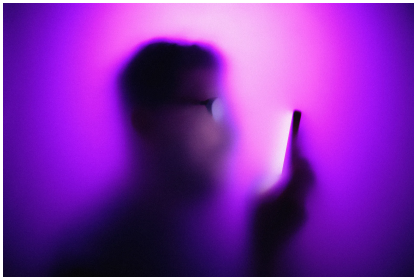
- Despite expansion of copyright laws, creators share of profits has declined (while media companies have increased).
  1. Creators are not compensated for work used in training from which products are generated.
  2. Consumers/employers may use Generative AI instead of employing creators.

1. Training Datasets: The use of copyrighted materials in training data raises concerns about potential infringement and unauthorized production.
2. Output Generation: relationship between outputs and pre-existing materials.
3. Ownership and Protection: Are Generative AI products eligible for copyright protection?

EAST TEXAS A&M

Issue: Data Privacy

- Included in ChatGPT training data are millions of pages scraped from the web, Reddit posts, Twitter
  - With large amounts of personal information people share about themselves online.
- This is getting OpenAI into trouble.
- Generative AI fine-tuning has been unreliable in preventing private information from leaking back out.



EAST TEXAS A&M

Bibliography

- Ali, R., Tang, O. Y., Connolly, I. D., Abdulrazeq, H. F., Mirza, F. N., Lim, R. K., Johnston, B. R., Groff, M. W., Williamson, T., Svokos, K., et al. (2024). Demographic representation in 3 leading artificial intelligence text-to-image generators. *JAMA surgery*, *159*(1), 87–95.
- Doctorow, C. (2023). Copyright won't solve creators generative AI problem. *Pluralistic*.
- Hastings, J. (2024). Preventing harm from non-conscious bias in medical generative AI. *The Lancet Digital Health*, *6*(1), e2–e3.
- Novelli, C., Casolari, F., Hacker, P., Spedicato, G., & Floridi, L. (2024). Generative AI in EU law: Liability, privacy, intellectual property, and cybersecurity. *arXiv preprint arXiv:2401.07348*.
- Tokayev, K.-J. (2023). Ethical implications of large language models a multidimensional exploration of societal, economic, and technical concerns. *International Journal of Social Analytics*, *8*(9), 17–33.
- Wei, A., Haghtalab, N., & Steinhardt, J. (2024). Jailbroken: How does LLM safety training fail? *Advances in Neural Information Processing Systems*, *36*.

EAST TEXAS A&M