AI

SUB-SECTION ONLY

in “DT-2024”

If you wish to read only the **AI relevant subsection,** then please start on **page 87** of DT-2024. Download DT-2024 **PDF** from:

https://dspace.mit.edu/handle/1721.1/140791

&

https://dspace.mit.edu/handle/1721.1/155442

PS: You may also visit & review the PDFs here - https://dspace.mit.edu/handle/1721.1/104429

www.nature.com/articles/d41586-024-02420-7 www.nature.com/articles/s41586-024-07566-y

nature.com/articles/d41586-024-02420-7

nature

AI ... naturally nonsensical

NEWS | 24 July 2024

AI models fed AI-generated data quickly spew nonsense

Researchers gave successive versions of a large language model information produced by previous generations of the AI – and observed rapid collapse.

By Elizabeth Gibney

www.nature.com/articles/d41586-024-02420-7 www.nature.com/articles/s41586-024-07566-y

Article

AI models collapse when trained on recursively generated data

https://doi.org/10.1038/s41586-024-07566-y
Received: 20 October 2023
Accepted: 14 May 2024
Published online: 24 July 2024
Open access
Check for updates

Ilia Shumailov[1,8], Zakhar Shumaylov[2,8], Yiren Zhao[3], Nicolas Papernot[4,5], Ross Anderson[6,7,9] & Yarin Gal[1]

Stable diffusion revolutionized image creation from descriptive text. GPT-2 (ref. 1), GPT-3(.5) (ref. 2) and GPT-4 (ref. 3) demonstrated high performance across a variety of language tasks. ChatGPT introduced such language models to the public. It is now clear that generative artificial intelligence (AI) such as large language models (LLMs) is here to stay and will substantially change the ecosystem of online text and images. Here we consider what may happen to GPT-{*n*} once LLMs contribute much of the text found online. We find that indiscriminate use of model-generated content in training causes irreversible defects in the resulting models, in which tails of the original content distribution disappear. We refer to this effect as 'model collapse' and show that it can occur in LLMs as well as in variational autoencoders (VAEs) and Gaussian mixture models (GMMs). We build theoretical intuition behind the phenomenon and portray its ubiquity among all learned generative models. We demonstrate that it must be taken seriously if we are to sustain the benefits of training from large-scale data scraped from the web. Indeed, the value of data collected about genuine human interactions with systems will be increasingly valuable in the presence of LLM-generated content in data crawled from the Internet.