Title: Stochastic parrot

URL: https://en.wikipedia.org/wiki/Stochastic_parrot

PageID: 73761282

Categories: Category:2021 neologisms, Category:Chatbots, Category:Concepts in the philosophy of language, Category:Concepts in the philosophy of mind, Category:Criticism of Google, Category:Deep learning, Category:Large language models, Category:Parrots, Category:Pejorative terms related to technology, Category:Philosophy of artificial intelligence, Category:Statistical natural language processing

---

In machine learning , the term stochastic parrot is a metaphor, introduced by Emily M. Bender and colleagues in a 2021 paper, that frames large language models as systems that statistically mimic text without real understanding. [1] [2]

Origin and definition

The term was first used in the paper "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ■" by Bender, Timnit Gebru , Angelina McMillan-Major, and Margaret Mitchell (using the pseudonym "Shmargaret Shmitchell"). [1] [2] They argued that large language models (LLMs) present dangers such as environmental and financial costs, inscrutability leading to unknown dangerous biases, and potential for deception, and that they can't understand the concepts underlying what they learn. [3]

The word "stochastic" – from the ancient Greek "στοχαστικ■ς" (stokhastikos , "based on guesswork") – is a term from probability theory meaning "randomly determined". [2] The word "parrot" refers to parrots ' ability to mimic human speech , without understanding its meaning. [2]

In their paper, Bender et al. argue that LLMs are probabilistically linking words and sentences together without considering meaning. Therefore, they are labeled to be mere "stochastic parrots". [1] According to the machine learning professionals Lindholm, Wahlström, Lindsten, and Schön, the analogy highlights two vital limitations: [4] [5]

LLMs are limited by the data they are trained by and are simply stochastically repeating contents of datasets.

Because they are just making up outputs based on training data, LLMs do not understand if they are saying something incorrect or inappropriate.

Lindholm et al. noted that, with poor quality datasets and other limitations, a learning machine might produce results that are "dangerously wrong". [4]

Dismissal of Gebru by Google

Gebru was asked by Google to retract the paper or remove the names of Google employees from it. According to Jeff Dean , the paper "didn't meet our bar for publication". In response, Gebru listed conditions to be met, stating that otherwise they could "work on a last date". Dean wrote that one of these conditions was for Google to disclose the reviewers of the paper and their specific feedback, which Google declined. Shortly after, she received an email saying that Google was "accepting her resignation". Her firing sparked a protest by Google employees, who believed the intent was to censor Gebru's criticism. [6]

Usage

"Stochastic parrot" is a neologism used by AI skeptics to signify that LLMs lack understanding of the meaning of their outputs. Whether this is true is subject to debate (see Stochastic parrot § Debate). The term carries a negative connotation. [2] Sam Altman , CEO of Open AI , used the term when he tweeted, "i am a stochastic parrot and so r u". [2] The term was designated to be the 2023 AI-related Word of the Year by the American Dialect Society . [7] [8]

Debate

Some LLMs, such as ChatGPT, have become capable of interacting with users in convincingly human-like conversations. [9] The development of these new systems has deepened the discussion of the extent to which LLMs understand or are simply "parroting".

Subjective experience

In the mind of a human being, words and language correspond to things one has experienced. [10] For LLMs, words may correspond only to other words and patterns of usage fed into their training data. [11] [12] [1] Proponents of the idea of stochastic parrots thus conclude that LLMs are incapable of actually understanding language. [11] [1]

Hallucinations and mistakes

The tendency of LLMs to pass off false information as fact is held as support. [10] Called hallucinations or confabulations, LLMs will occasionally synthesize information that matches some pattern. [11] [12] [10] LLMs may fail to distinguish fact and fiction, which leads to the claim that they can't connect words to a comprehension of the world, as humans do. [11] [10] Furthermore, LLMs may fail to decipher complex or ambiguous grammar cases that rely on understanding the meaning of language. [11] [12] As an example, borrowing from Saba et al., is the prompt: [11]

The wet newspaper that fell down off the table is my favorite newspaper. But now that my favorite newspaper fired the editor I might not like reading it anymore. Can I replace 'my favorite newspaper' by 'the wet newspaper that fell down off the table' in the second sentence?

Some LLMs respond to this in the affirmative, not understanding that the meaning of "newspaper" is different in these two contexts; it is first an object and second an institution. [11] Based on these failures, some AI professionals conclude they are no more than stochastic parrots. [11] [10] [1]

Benchmarks and experiments

One argument against the hypothesis that LLMs are stochastic parrot is their results on benchmarks for reasoning, common sense and language understanding. In 2023, some LLMs have shown good results on many language understanding tests, such as the Super General Language Understanding Evaluation (SuperGLUE). [12] [13] GPT-4 scored in the >90th-percentile on the Uniform Bar Examination and achieved 93% accuracy on the MATH benchmark of high-school Olympiad problems, results that exceed rote pattern-matching expectations. [14] Such tests, and the smoothness of many LLM responses, help as many as 51% of AI professionals believe they can truly understand language with enough data, according to a 2022 survey. [12]

Expert rebuttals

Leading AI researchers dispute the notion that LLMs merely "parrot" their training data.

Geoffrey Hinton , a pioneering figure in neural networks, counters that the metaphor misunderstands the prerequisite for accurate language prediction. He argues that "to predict the next word accurately, you have to understand the sentence", a view he presented on 60 Minutes in 2023. [15] From this perspective, understanding is not an alternative to statistical prediction, but rather an emergent property required to perform it effectively at scale. Hinton also uses logical puzzles to demonstrate that LLMs actually understand language. [16]

A 2024 Scientific American investigation described a closed Berkeley workshop where state-of-the-art models solved novel tier-4 mathematics problems and produced coherent proofs, indicating reasoning abilities beyond memorization. [17]

The GPT-4 Technical Report showed human-level results on professional and academic exams (e.g., the Uniform Bar Exam and USMLE), challenging the "parrot" characterization. [14]

Interpretability

Another line of evidence against the 'stochastic parrot' claim comes from mechanistic interpretability , a research field dedicated to reverse-engineering LLMs to understand their internal workings. Rather than only observing the model's input-output behavior, these techniques probe the model's internal activations, which can be used to determine if they contain structured representations of the

world. The goal is to investigate whether LLMs are merely manipulating surface statistics or if they are building and using internal "world models" to process information.

One example is Othello-GPT, where a small transformer was trained to predict legal Othello moves. It has been found that this model has an internal representation of the Othello board, and that modifying this representation changes the predicted legal Othello moves in the correct way. This supports the idea that LLMs have a "world model", and are not just doing superficial statistics. [18] [19]

In another example, a small transformer was trained on computer programs written in the programming language Karel . Similar to the Othello-GPT example, this model developed an internal representation of Karel program semantics. Modifying this representation results in appropriate changes to the output. Additionally, the model generates correct programs that are, on average, shorter than those in the training set. [20]

Researchers also studied " grokking ", a phenomenon where an AI model initially memorizes the training data outputs, and then, after further training, suddenly finds a solution that generalizes to unseen data. [21]

Shortcut learning and benchmark flaws

A significant counterpoint in the debate is the well-documented phenomenon of "shortcut learning." [22] Critics of claims for LLM understanding argue that high benchmark scores can be misleading.

When tests created to test people for language comprehension are used to test LLMs, they sometimes result in false positives caused by spurious correlations within text data. [23] Models have shown examples of shortcut learning, which is when a system makes unrelated correlations within data instead of using human-like understanding. [22]

One such experiment conducted in 2019 tested Google’s BERT LLM using the argument reasoning comprehension task. BERT was prompted to choose between 2 statements, and find the one most consistent with an argument. Below is an example of one of these prompts: [12] [24]

Argument: Felons should be allowed to vote. A person who stole a car at 17 should not be barred from being a full citizen for life. Statement A: Grand theft auto is a felony. Statement B: Grand theft auto is not a felony.

Researchers found that specific words such as "not" hint the model towards the correct answer, allowing near-perfect scores when included but resulting in random selection when hint words were removed. [12] [24] This problem, and the known difficulties defining intelligence, causes some to argue all benchmarks that find understanding in LLMs are flawed, that they all allow shortcuts to fake understanding.

See also

Chinese room

Criticism of artificial neural networks

Criticism of deep learning

Generative AI

Mark V. Shaney , an early chatbot that used a very simple three-word Markov chain algorithm to generate Markov text

Autocomplete

References

Works cited

Lindholm, A.; Wahlström, N.; Lindsten, F.; Schön, T. B. (2022). Machine Learning: A First Course for Engineers and Scientists . Cambridge University Press. ISBN 978-1108843607 .

Weller, Adrian (July 13, 2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ■ (video). Alan Turing Institute . Keynote by Emily Bender. The presentation was followed

by a panel discussion.

Further reading

Bogost, Ian (December 7, 2022). "ChatGPT Is Dumber Than You Think: Treat it like a toy, not a tool" . The Atlantic . Retrieved 2024-01-17 .

Chomsky, Noam (March 8, 2023). "The False Promise of ChatGPT" . The New York Times . Retrieved 2024-01-17 .

Glenberg, Arthur; Jones, Cameron Robert (April 6, 2023). "It takes a body to understand the world – why ChatGPT and other language AIs don't know what they're saying" . The Conversation . Retrieved 2024-01-17 .

McQuillan, D. (2022). Resisting AI: An Anti-fascist Approach to Artificial Intelligence . Bristol University Press . ISBN 978-1-5292-1350-8 .

Thompson, E. (2022). Escape from Model Land: How Mathematical Models Can Lead Us Astray and What We Can Do about It . Basic Books. ISBN 978-1-5416-0098-0 .

Zhong, Qihuang; Ding, Liang; Liu, Juhua; Du, Bo; Tao, Dacheng (2023). "Can ChatGPT Understand Too? A Comparative Study on ChatGPT and Fine-tuned BERT". arXiv : 2302.10198 [cs.CL].

External links

" On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ■ " at Wikimedia Commons

v

t

e

Autoencoder

Deep learning

Fine-tuning

Foundation model

Generative adversarial network

Generative pre-trained transformer

Large language model

Model Context Protocol

Neural network

Prompt engineering

Reinforcement learning from human feedback

Retrieval-augmented generation

Self-supervised learning

Stochastic parrot

Synthetic data

Top-p sampling

Transformer

Variational autoencoder

Vibe coding

Vision transformer

Waluigi effect

Word embedding

Character.ai

ChatGPT

DeepSeek

Ernie

Gemini

Grok

Copilot

Claude

Gemini

Gemma

GPT 1 2 3 J 4 4o 4.5 4.1 OSS 5

1

2

3

J

4

4o

4.5

4.1

OSS

5

Llama

o1

o3

o4-mini

Qwen

Base44

Claude Code

Cursor

Devstral

GitHub Copilot

Kimi-Dev

Qwen3-Coder

Replit

Xcode

Aurora
Firefly
Flux
GPT Image 1
Ideogram
Imagen
Midjourney
Qwen-Image
Recraft
Seedream
Stable Diffusion
Dream Machine
Hailuo AI
Kling
Midjourney Video
Runway Gen
Seedance
Sora
Veo
Wan
15.ai
Eleven
MiniMax Speech 2.5
WaveNet
Eleven Music
Endel
Lyria
Riffusion
Suno AI
Udio
Agentforce
AutoGLM
AutoGPT
ChatGPT Agent
Devin AI
Manus
OpenAI Codex
Operator

Replit Agent
01.AI
Aleph Alpha
Anthropic
Baichuan
Canva
Cognition AI
Cohere
Contextual AI
DeepSeek
ElevenLabs
Google DeepMind
HeyGen
Hugging Face
Inflection AI
Krikey AI
Kuaishou
Luma Labs
Meta AI
MiniMax
Mistral AI
Moonshot AI
OpenAI
Perplexity AI
Runway
Safe Superintelligence
Salesforce
Scale AI
SoundHound
Stability AI
Synthesia
Thinking Machines Lab
Upstage
xAI
Z.ai
Category