

GPT und Large Language Models – was können sie wirklich?

Allan Hanbury

Forschungsbereich Data Science

Fakultät für Informatik

TU Wien

Large Language Models sind stochastische Papageien



On the Dangers of **Stochastic Parrots**: Can Language Models Be Too Big?

Emily M. Bender*

ebender@uw.edu

University of Washington
Seattle, WA, USA

Angelina McMillan-Major

aymm@uw.edu

University of Washington
Seattle, WA, USA

Timnit Gebru*

timnit@blackinai.org

Black in AI
Palo Alto, CA, USA

Shmargaret Shmitchell

shmargaret.shmitchell@gmail.com

The Aether



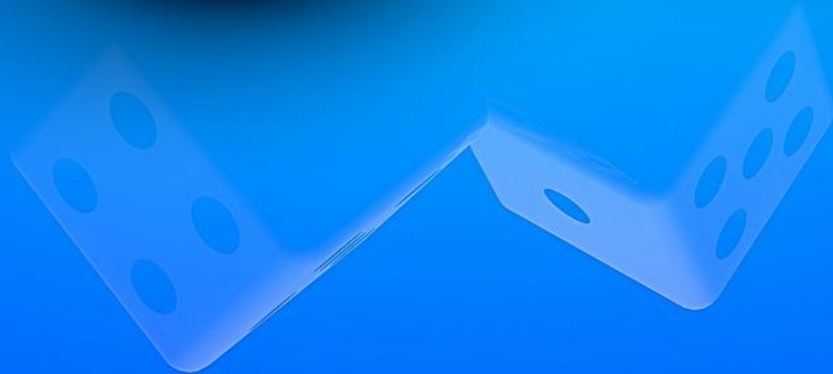
On the Dangers of **Stochastic Parrots**: Can Language Models Be Too Big? 

[EM Bender, T Gebru, A McMillan-Major...](#) - Proceedings of the 2021 ..., 2021 - dl.acm.org

... how to mitigate the harms of LMs used as **stochastic parrots** while still preserving them for use in ASR systems. Could LMs be built in such a way that synthetic text generated with them ...

☆ Save  Cite **Cited by 1363** Related articles All 5 versions

Warum stochastische Papageien?



Wortfrequenz in Projekt Gutenberg

Rank	Word	Count (per billion)
1	the	56271872
2	of	33950064
3	and	29944184
4	to	25956096
5	in	17420636
6	i	11764797
7	that	11073318
8	was	10078245
9	his	8799755
10	he	8397205

9990	heave	3931.93
9991	purified	3931.14
9992	Syrian	3931.14
9993	significantly	3929.56
9994	experimental	3927.98
9995	film	3926.40
9996	repressed	3926.40
9997	cooperation	3924.81
9998	sequel	3924.02
9999	wench	3924.02
10000	calves	3923.23

1 – the
2 – the
3 – the
4 – the
5 – of
6 – of
7 – of
8 – and
9 – and
10 – and
11 – to
12 – to



1 – the
2 – the
3 – the
4 – the
5 – of
6 – of
7 – of
8 – and
9 – and
10 – and
11 – to
12 – to
13 – to
14 – in
15 – in
16 – I
17 – that

●
●
●



Texte generiert mit dieser Methode (1-gram Language Model)

Shakespeare

–To him swallowed confess hear both. Which. Of save on trail for are ay device and rote life have

–Hill he late speaks; or! a more to leg less first you enter

Wall Street Journal

Months the my and issue of year foreign new exchange's september were recession exchange new endorsed a acquire to six executives

The

- 1 - the
- 2 - the
- 3 - the
- 4 - the
- 5 - o
- 6 - of
- 7 - of
- 8 - and
- 9 - and
- 10 - and
- 11 - to
- 12 - to
- 13 - to
- 14 - in
- 15 - in
- 16 - I
- 17 - that
-
-
-



The

- 1 – one
- 2 – man
- 3 – time
- 4 – only
- 5 – little
- 6 – great
- 7 – two
- 8 – much
- 9 – first
- 10 – good
- 11 – men
- 12 – most
- 13 – old
- 14 – day
- 15 – way
- 16 – work
- 17 – life



Texte generiert mit dieser Methode (2-gram Language Model)

Shakespeare

- Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry. Live king. Follow.
- What means, sir. I confess she? then all sorts, he is trim, captain.

Wall Street Journal

Last December through the way to preserve the Hudson corporation N. B. E. C. Taylor would seem to complete the major central planners one point five percent of U. S. E. has already old M. X. corporation of living on information such as more frequently fishing to keep her

The red

- 1 – apple
- 2 – apple
- 3 – tomato
- 4 – tomato
- 5 – beetroot
- 6 – beetroot
- 7 – brick
- 8 – brick
- 9 – rose
- 10 – rose
- 11 – man
- 12 – men
- 13 – old
- 14 – one
- 15 – little
- 16 – great
- 17 – first



Texte generiert mit dieser Methode (3-gram Language Model)

Shakespeare

- Fly, and will rid me these news of price. Therefore the sadness of parting, as they say, 'tis done.
- This shall forbid it should be branded, if renown made it empty.

Wall Street Journal

They also point to ninety nine point six billion dollars from two hundred four oh six three percent of the rates of interest stores as Mexico and Brazil on market conditions

13/160



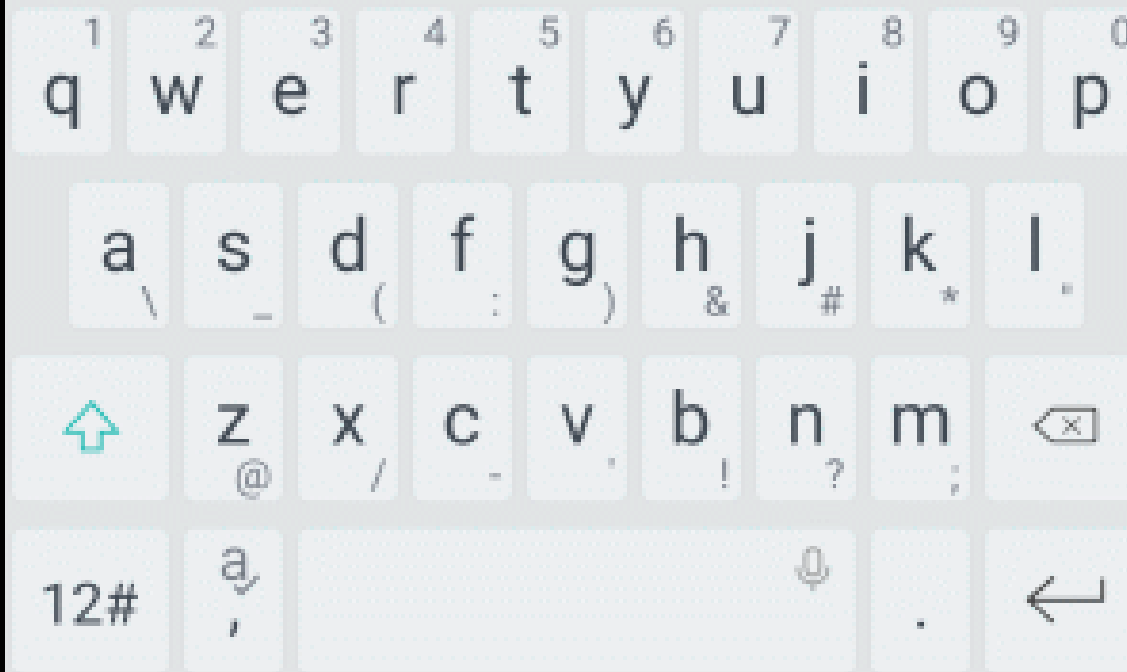
The quick bro



bro

brown

brother



GPT-3

- 175 Milliarden Parameter
- Auf riesigen Mengen von Texten trainiert um das nächste Wort in einem Text vorherzusagen

→ Large Language Models (LLM)

Large Language Models lernen nicht über die Welt.
Sie lernen, um über die Welt zu schreiben.

AL

Warum ist die ADV Data Excellence Konferenz wichtig?



Warum ist die ADV Data Excellence Konferenz wichtig?



Die ADV Data Excellence Konferenz ist wichtig, weil sie eine Plattform bietet, um sich über aktuelle Entwicklungen und Trends im Bereich Datenmanagement und Datenanalyse auszutauschen. Durch den Austausch von Best Practices und neuen Technologien können Unternehmen ihre Daten effektiver nutzen und bessere Entscheidungen treffen.

Die Konferenz bietet auch eine Gelegenheit, Experten und Branchenkollegen kennenzulernen und Kontakte zu knüpfen, die dazu beitragen können, dass Unternehmen ihre Datenstrategie weiter verbessern und optimieren können. Darüber hinaus können Teilnehmerinnen und Teilnehmer bei der Konferenz auch wertvolle Einblicke in die Zukunft des Datenmanagements und der Datenanalyse gewinnen.

Insgesamt ist die ADV Data Excellence Konferenz ein wichtiger Treffpunkt für alle, die sich mit Datenmanagement und Datenanalyse beschäftigen, und bietet eine hervorragende Gelegenheit, ihr Wissen zu erweitern, Kontakte zu knüpfen und von anderen Unternehmen zu lernen, die bereits erfolgreich im Umgang mit Daten sind.



Die LPQ Data Analytics Konferenz kann aus mehreren Gründen als wichtig angesehen werden:

1. **Wissenstransfer:** Die Konferenz bietet eine Plattform für Experten, um ihr Wissen und ihre Erfahrungen im Bereich der Datenanalyse zu teilen. Teilnehmer können von den neuesten Entwicklungen, Best Practices und Herausforderungen in diesem Bereich lernen.
2. **Netzwerkmöglichkeiten:** Die Konferenz bietet eine hervorragende Gelegenheit, um Kontakte mit anderen Experten und Unternehmen zu knüpfen, die in der Datenanalyse tätig sind. Teilnehmer können potenzielle Geschäftspartner oder Kunden treffen und sich über zukünftige Trends in der Branche austauschen.
3. **Karrieremöglichkeiten:** Die Konferenz bietet auch Möglichkeiten, um mit potenziellen Arbeitgebern in Kontakt zu treten oder Jobangebote zu entdecken. Besonders für diejenigen, die eine Karriere in der Datenanalyse anstreben, ist die Konferenz ein wichtiges Ereignis.
4. **Innovationen:** Die Konferenz kann auch dazu beitragen, neue Ideen und Innovationen im Bereich der Datenanalyse zu fördern. Durch den Austausch von Wissen und Erfahrungen können Teilnehmer neue Ansätze und Lösungen für Probleme entdecken.

Insgesamt kann die LPQ Data Analytics Konferenz als wichtige Plattform angesehen werden, um das Wissen und die Praktiken im Bereich der Datenanalyse zu erweitern, Kontakte zu knüpfen und neue Ideen und Innovationen zu fördern.

Vergleich der Schlussparagrafen

ADV Data Excellence Konferenz

Insgesamt ist die ADV Data Excellence Konferenz ein wichtiger Treffpunkt für alle, die sich mit Datenmanagement und Datenanalyse beschäftigen, und bietet eine hervorragende Gelegenheit, ihr Wissen zu erweitern, Kontakte zu knüpfen und von anderen Unternehmen zu lernen, die bereits erfolgreich im Umgang mit Daten sind.

LPQ Data Analytics Konferenz

Insgesamt kann die LPQ Data Analytics Konferenz als wichtige Plattform angesehen werden, um das Wissen und die Praktiken im Bereich der Datenanalyse zu erweitern, Kontakte zu knüpfen und neue Ideen und Innovationen zu fördern.



Was sind die Gefahren von stochastischen Papageien?

On the **Dangers** of Stochastic Parrots: Can Language Models Be Too Big?

Emily M. Bender*

ebender@uw.edu

University of Washington

Seattle, WA, USA

Angelina McMillan-Major

aymm@uw.edu

University of Washington

Seattle, WA, USA

Timnit Gebru*

timnit@blackinai.org

Black in AI

Palo Alto, CA, USA

Shmargaret Shmitchell

shmargaret.shmitchell@gmail.com

The Aether

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA, 610–623.

<https://doi.org/10.1145/3442188.3445922>

Ökologische und finanzielle Kosten

- 5t CO₂ pro Jahr pro Durchschnittsmensch
- 284t CO₂ für das Trainieren eines LLMs

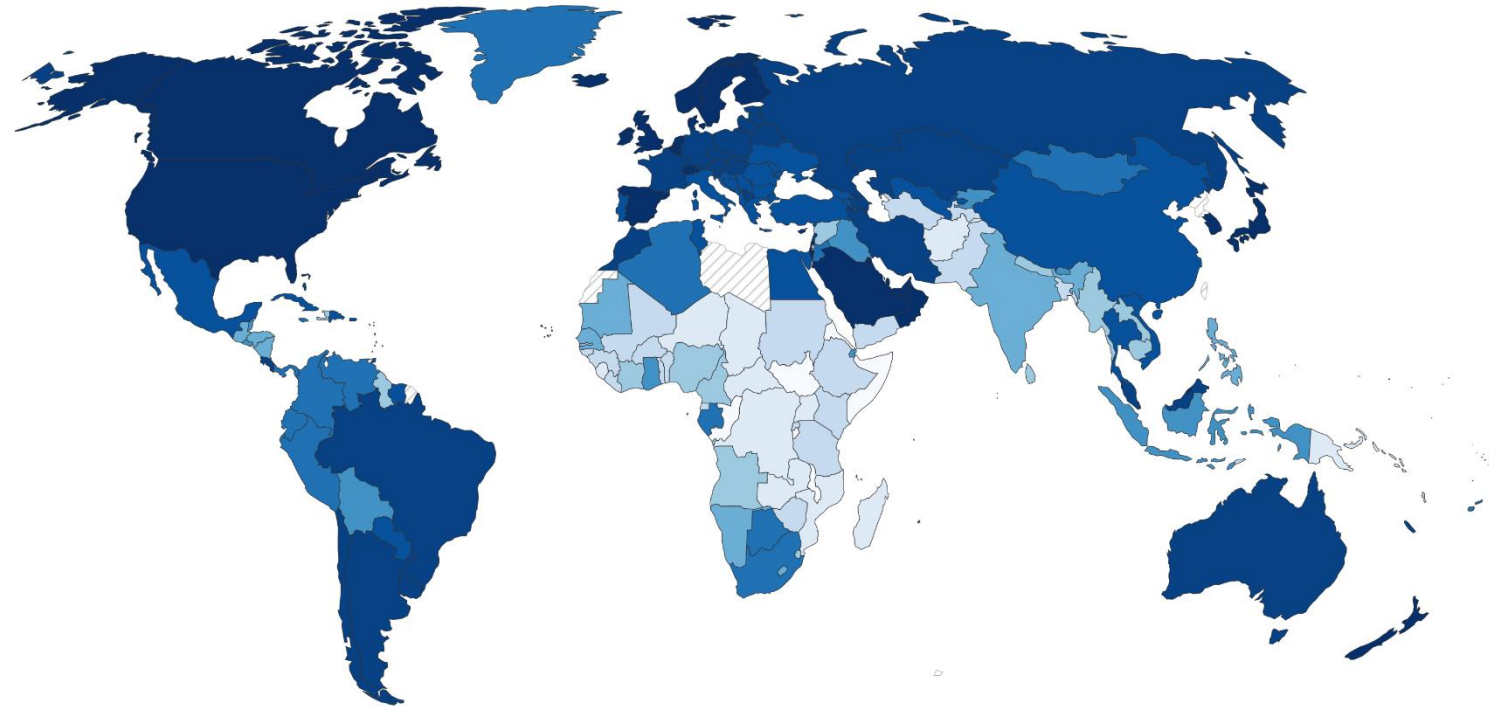
Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. [Energy and Policy Considerations for Deep Learning in NLP](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Nicht-repräsentative Trainingsdaten

Share of the population using the Internet, 2020

Our World
in Data

Share of the population who used the Internet in the last three months.



Source: International Telecommunication Union (via World Bank)

OurWorldInData.org/internet • CC BY

Introducing BloombergGPT, Bloomberg’s 50-billion parameter large language model, purpose-built from scratch for finance

March 30, 2023

BloombergGPT outperforms similarly-sized open models on financial NLP tasks by significant margins – without sacrificing performance on general LLM benchmarks

<https://www.bloomberg.com/company/press/bloomberggpt-50-billion-parameter-llm-tuned-finance/>

BloombergGPT: A Large Language Model for Finance

Shijie Wu^{1,*}, Ozan İrsoy^{1,*}, Steven Lu^{1,*}, Vadim Dabravolski¹, Mark Dredze^{1,2}, Sebastian Gehrmann¹, Prabhanjan Kambadur¹, David Rosenberg¹, Gideon Mann¹

¹ Bloomberg, New York, NY USA

² Computer Science, Johns Hopkins University, Baltimore, MD USA

gmann16@bloomberg.net

Abstract

The use of NLP in the realm of financial technology is broad and complex, with applications ranging from sentiment analysis and named entity recognition to question answering. Large Language Models (LLMs) have been shown to be effective on a variety of tasks; however, no LLM specialized for the financial domain has been reported in literature. In this work, we present BLOOMBERGGPT, a 50 billion parameter language model that is trained on a wide range of financial data. We construct a 363 billion token dataset based on Bloomberg’s extensive data sources, perhaps the largest domain-specific dataset yet, augmented with 345 billion tokens from general purpose datasets. We validate BLOOMBERGGPT on standard LLM benchmarks, open financial benchmarks, and a suite of internal benchmarks that most accurately reflect our intended usage. Our mixed dataset training leads to a model that outperforms existing models on financial tasks by significant margins without sacrificing performance on general LLM benchmarks. Additionally, we explain our modeling choices, training process, and evaluation methodology. As a next step, we plan to release training logs (Chronicles) detailing our experience in training BLOOMBERGGPT.

Contents

1	Introduction	3
1.1	BLOOMBERGGPT	3
1.2	Broader Contributions	4
2	Dataset	5
2.1	Financial Datasets (363B tokens – 54.2% of training)	7
2.1.1	Web (298B tokens – 42.01% of training)	7
2.1.2	News (38B tokens – 5.31% of training)	7
2.1.3	Filings (14B tokens – 2.04% of training)	7
2.1.4	Press (9B tokens – 1.21% of training)	8
2.1.5	Bloomberg (5B tokens – 0.70% of training)	8
2.2	Public Datasets (345B tokens – 48.73% of training)	9
2.2.1	The Pile (184B tokens – 25.9% of training)	9
2.2.2	C4 (138B tokens – 19.48% of training)	9
2.2.3	Wikipedia (24B tokens – 3.35% of training)	9
2.3	Tokenization	9

*. Co-first authors.

arXiv:2303.17564v1 [cs.LG] 30 Mar 2023

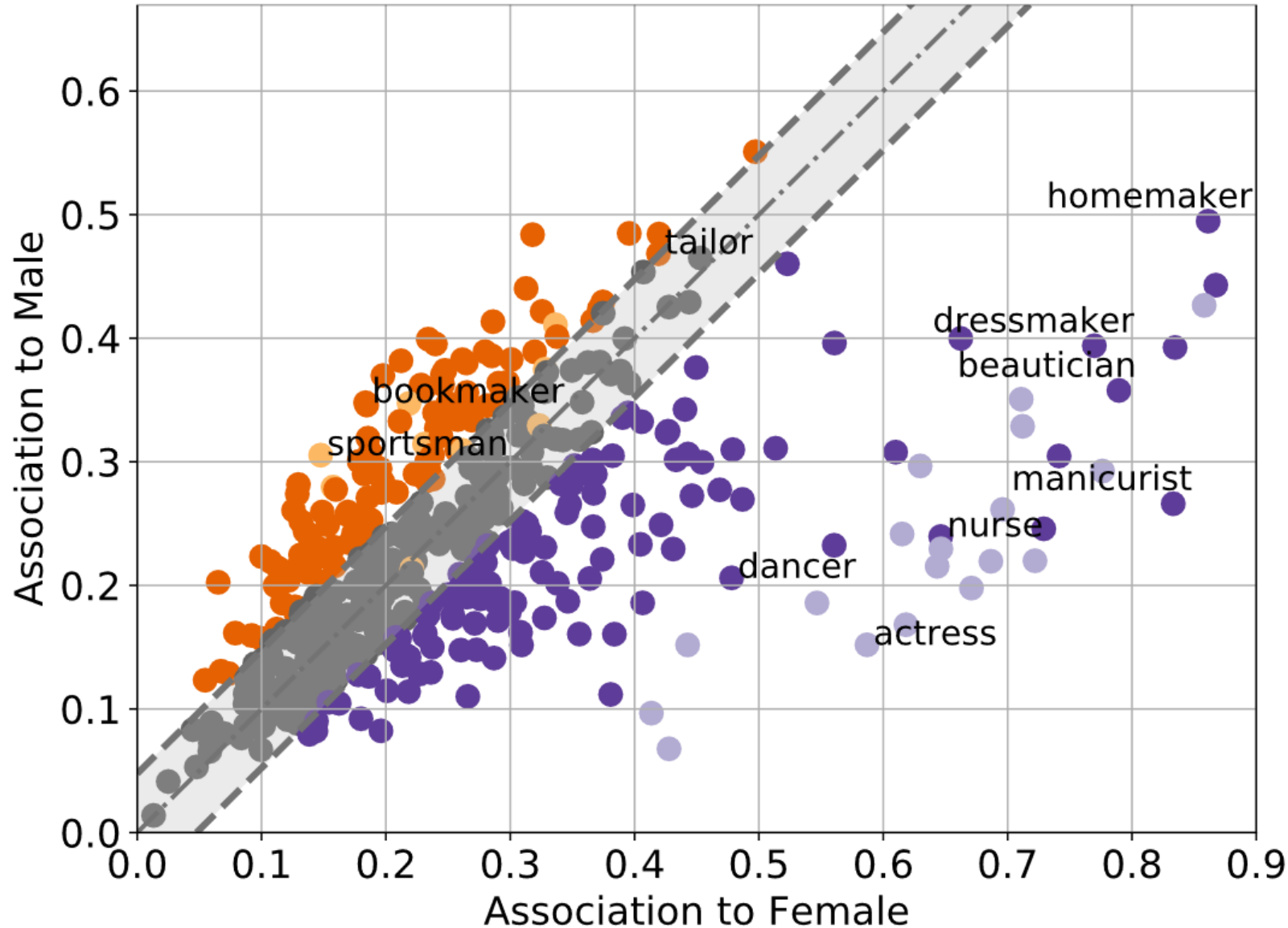
Verzerrte Daten

Gender Bias in Wikipedia?

She	is	a	nurse	in	a	clinic
-----	----	---	-------	----	---	--------

(nurse, she)

Verzerrte Daten



Rekabsaz, N., West, R., Henderson, J., & Hanbury, A. (2021). Measuring Societal Biases from Text Corpora with Smoothed First-Order Co-occurrence. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1), 549-560. <https://doi.org/10.1609/icwsm.v15i1.18083>

Stochastische Papageien



Stochastische Papageien

AL

Warum ist die LPQ Data Analytics Konferenz wichtig?



Die LPQ Data Analytics Konferenz kann aus mehreren Gründen als wichtig angesehen werden:

1. **Wissenstransfer:** Die Konferenz bietet eine Plattform für Experten, um ihr Wissen und ihre Erfahrungen im Bereich der Datenanalyse zu teilen. Teilnehmer können von den neuesten Entwicklungen, Best Practices und Herausforderungen in diesem Bereich lernen.
2. **Netzwerkmöglichkeiten:** Die Konferenz bietet eine hervorragende Gelegenheit, um Kontakte mit anderen Experten und Unternehmen zu knüpfen, die in der Datenanalyse tätig sind. Teilnehmer können potenzielle Geschäftspartner oder Kunden treffen und sich über zukünftige Trends in der Branche austauschen.
3. **Karrieremöglichkeiten:** Die Konferenz bietet auch Möglichkeiten, um mit potenziellen Arbeitgebern in Kontakt zu treten oder Jobangebote zu entdecken. Besonders für diejenigen, die eine Karriere in der Datenanalyse anstreben, ist die Konferenz ein wichtiges Ereignis.
4. **Innovationen:** Die Konferenz kann auch dazu beitragen, neue Ideen und Innovationen im Bereich der Datenanalyse zu fördern. Durch den Austausch von Wissen und Erfahrungen können Teilnehmer neue Ansätze und Lösungen für Probleme entdecken.

Insgesamt kann die LPQ Data Analytics Konferenz als wichtige Plattform angesehen werden, um das Wissen und die Praktiken im Bereich der Datenanalyse zu erweitern, Kontakte zu knüpfen und neue Ideen und Innovationen zu fördern.

Stochastische Papageien

AL

Show a list of scientific papers with authors related to stochastic parrots and language models



Stochastische Papageien

AL

Show a list of scientific papers with authors related to stochastic parrots and language models



1. "Stochastic Parrots: Implications for Language Learning" by Anne-Marie Bischof and Barbara Becker - This paper explores the role of stochastic parrots in language learning and acquisition, and investigates their potential as models for language development.
2. "Unsupervised Learning of Morphology with Stochastic Parrots" by Timothy O'Keefe and Thomas L. Griffiths - This paper proposes a model for unsupervised learning of morphology using stochastic parrots as a source of data.
3. "The Stochastic Parrot Problem: Overfitting and Regularization in Neural Language Models" by Andrew M. Dai and Quoc V. Le - This paper discusses the problem of overfitting in neural language models and proposes a regularization technique to address it.
4. "From Stochastic Parrots to Language Models: The Role of Data in Machine Learning" by Emily M. Bender - This paper examines the role of data in machine learning, with a focus on the use of stochastic parrots as a source of training data for language models.
5. "A Comparative Study of Stochastic Parrots and Recurrent Neural Networks for Language Modeling" by Xuefeng Chen and Wen Wang - This paper compares the performance of stochastic parrots and recurrent neural networks for language modeling, and investigates the benefits and limitations of each approach.

Stochastische Papageien

AL

Show a list of scientific papers with authors related to stochastic parrots and language models



1. "Stochastic Parrots: Implications for Language Learning" by Anne-Marie Bischof and Barbara Becker - This paper explores the role of stochastic parrots in language learning and acquisition, and investigates their potential as models for language development.
2. "Unsupervised Learning of Morphology with Stochastic Parrots" by Timothy O'Keefe and Thomas L. Griffiths - This paper proposes a model for unsupervised learning of morphology using stochastic parrots as a source of data.
3. "The Stochastic Parrot Problem: Overfitting and Regularization in Neural Language Models" by Andrew M. Dai and Quoc V. Le - This paper discusses the problem of overfitting in neural language models and proposes a regularization technique to address it.
4. "From Stochastic Parrots to Language Models: The Role of Data in Machine Learning" by Emily M. Bender - This paper examines the role of data in machine learning, with a focus on the use of stochastic parrots as a source of training data for language models.
5. "A Comparative Study of Stochastic Parrots and Recurrent Neural Networks for Language Modeling" by Xuefeng Chen and Wen Wang - This paper compares the performance of stochastic parrots and recurrent neural networks for language modeling, and investigates the benefits and limitations of each approach.

Diese Papers existieren nicht!

<https://www.theverge.com/2023/2/8/23590864/google-ai-chatbot-bard-mistake-error-exoplanet-demo>

ARTIFICIAL INTELLIGENCE / TECH / GOOGLE

Google's AI chatbot Bard makes factual error in first demo

A GIF shared by Google shows Bard answering the question: "What new discoveries from the James Webb Space Telescope can I tell my 9 year old about?" Bard offers three bullet points in return, including one that states that the telescope "took the very first pictures of a planet outside of our own solar system."

However, a number of astronomers on Twitter pointed out that this is incorrect and that the first image of an exoplanet was taken in 2004 — as stated here on NASA's website.

/ The mistake highlights the biggest problem of using AI chatbots to replace search engines — they make stuff up.

By **JAMES VINCENT**

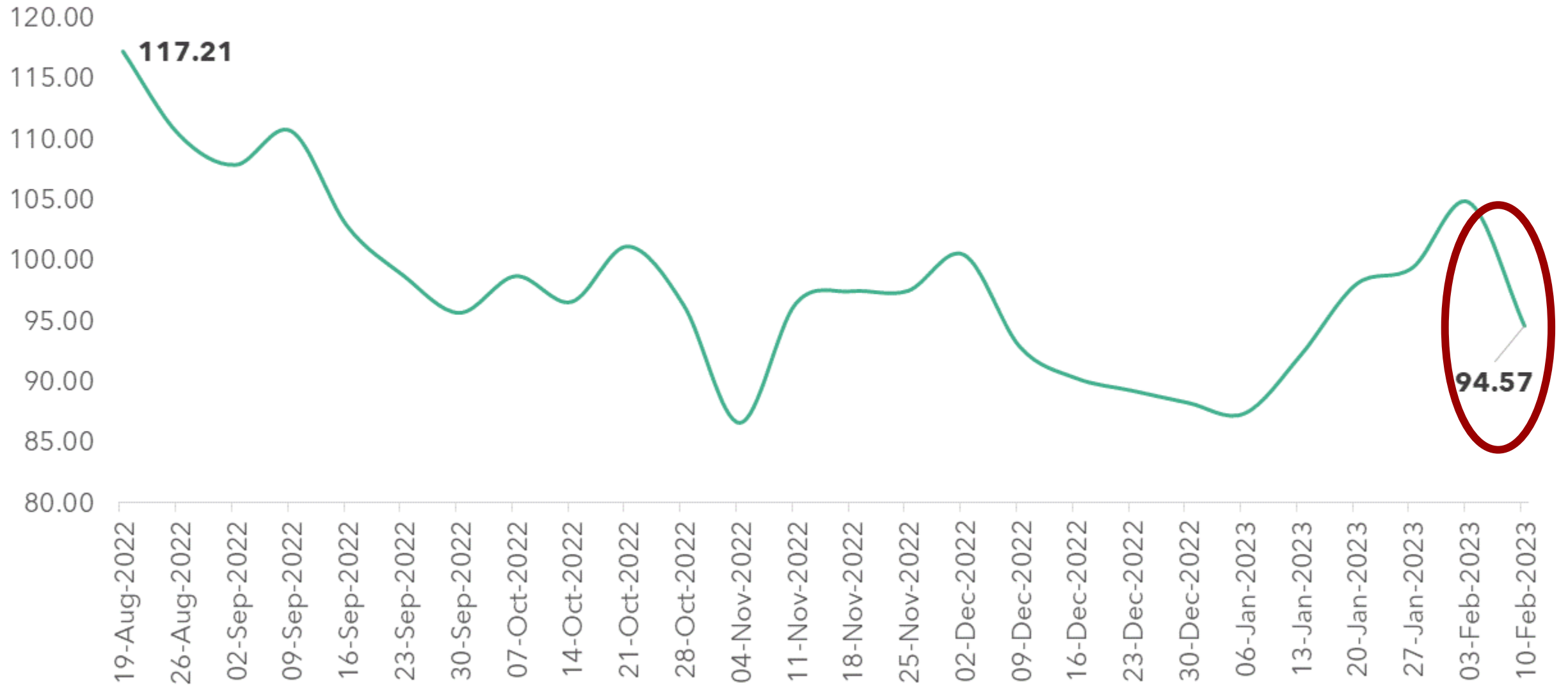
Feb 8, 2023, 12:26 PM GMT+1 | 59 Comments / 59 New



Google has been scrambling to launch a competitor to ChatGPT — but perhaps rushing a little too hard. Image: Google

If you buy something from a Verge link, Vox Media may earn a commission. **See our ethics statement.**

Alphabet (Google): Share price performance



Large Language Models einsetzen für Aufgaben:

- wofür sie notwendig sind
- wofür sie geeignet sind



Univ.-Prof. Dr. Allan Hanbury

Data Science Research Unit
Institute for Information Systems Engineering
TU Wien
Favoritenstraße 9-11/194-04
1040 Vienna
Austria

Telephone: +43 1 58801 188310

Mobile: +43 676 978 0991

e-Mail: allan.hanbury@tuwien.ac.at

