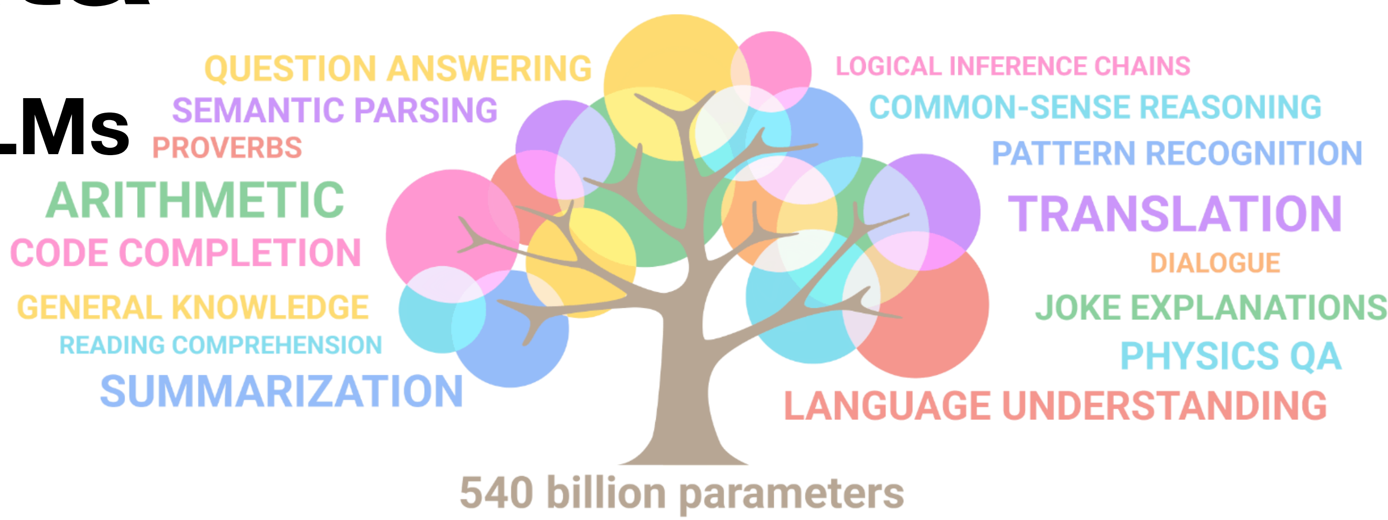
Raw Data

Scaling up to LLMs



Cornell CS 5740: Natural Language Processing
Yoav Artzi, Spring 2023

Animation from https://blog.research.google/2022/04/pathways-language-model-palm-scaling-to.html

Scaling Up
Next-word Prediction

- Language models do next word prediction
- At first look, next-word-completion seems like a very simple task
- Why does it make sense to focus on it so much?

Scaling Up

Why Does it Make Sense?

I put ____ fork down on the table

Scaling Up

Why Does it Make Sense?

The woman walked across the street, checking for traffic over ___ shoulder

Scaling Up

Why Does it Make Sense?

I went to the ocean to see the fish, turtles, seals, and ____

Scaling Up

Why Does it Make Sense?

Overall, the value I got from the two hours watching it was the sum total of the popcorn and the drink. The movie was _____

Scaling Up

Why Does it Make Sense?

Iroh went into the kitchen to make some tea. Standing next to Iroh, Zuko pondered his destiny. Zuko left the ____

Scaling Up

Why Does it Make Sense?

I was thinking about the sequence that goes 1, 1, 2, 3, 5, 8, 13, 21, ______

Scaling Up

Why Does it Make Sense?

Cornell Tech is located in _____, New York

Scaling Up

Why Does it Make Sense?

- I put ____ fork down on the table [syntax]
- The woman walked across the street, checking for traffic over ___ shoulder [coreference]
- I went to the ocean to see the fish, turtles, seals, and ____ [lexical semantics / topics]
- Overall, the value I got from the two hours watching it was the sum total of the popcorn and the drink. The movie was _____ [sentiment]
- I was thinking about the sequence that goes 1, 1, 2, 3, 5, 8, 13, 21, ______ [reasoning]
- Cornell Tech is located in _____, New York [knowledge]

The learned representations have to account for a lot to succeed in this seemingly straightforward task

Examples from Stanford CS 224N

Some History: the GPTs

GPT [Radford et al. 2018]

- Transformer LM released in 2018 by OpenAI
- Decoder with 12 transformer blocks, 117M parameters, 768-dimensional hidden states, 3072-dimensional feed-forward hidden layers; BPE with 40k merges
- Trained on BookCorpus: over 7,000 unique book: long spans of contiguous text for learning long-distance dependencies
- Impressive results when fine-tuned on several NLP tasks: Entailment, textual similarity, multiple choice questions
- GPT? Actually not specified in the paper ¯_(ツ)_/¯
    - Generative PreTraining
    - Generative Pretrained Transformers

Some History: the GPTs

GPT-2 [Radford et al. 2018]

- GPT-2 scaled the models to 1.5B parameters
- Increasingly convincing generations
- Impressive zero-shot results on several tasks

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

Table 3. Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2 results are from (Gong et al., 2018). CBT results are from (Bajgar et al., 2016). LAMBADA accuracy result is from (Hoang et al., 2018) and LAMBADA perplexity result is from (Grave et al., 2016). Other results are from (Dai et al., 2019).

Some History: the GPTs

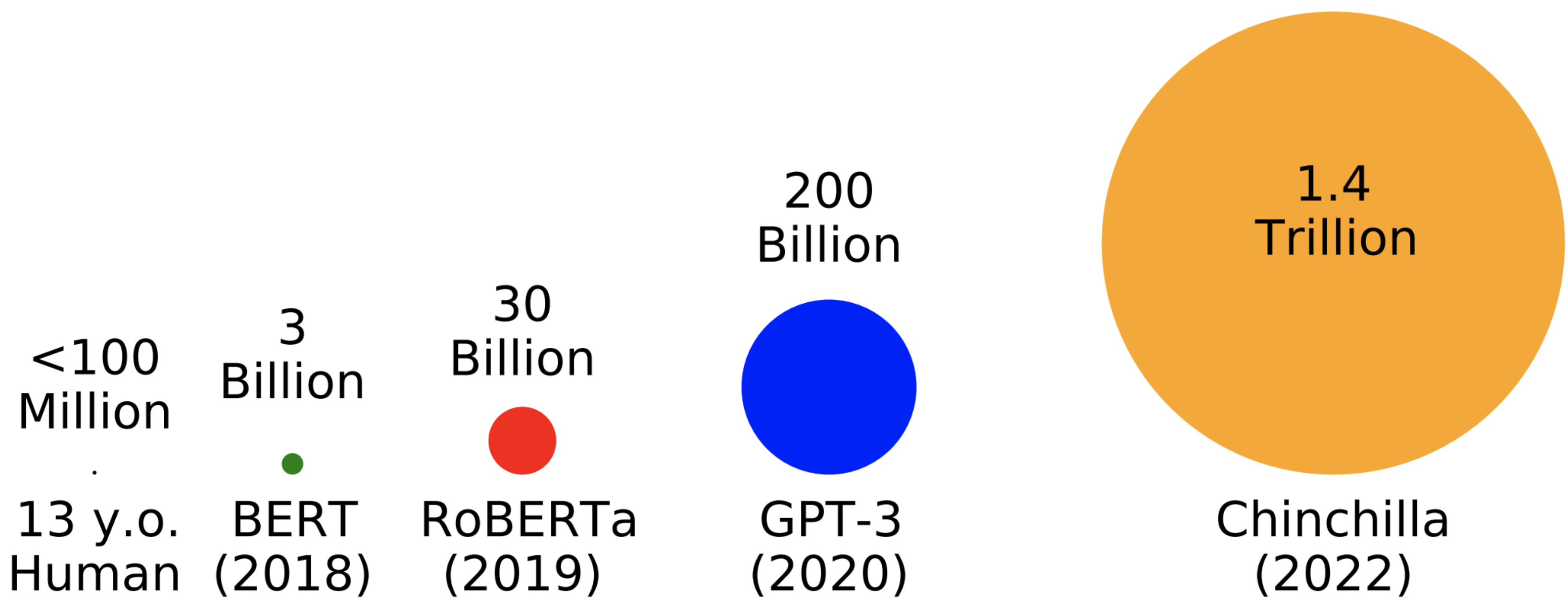
GPT-3 [Brown et al. 2020]

- GPT-3 scaled the model size to 175B parameters
- So far, two ways of interaction with models:
    - Sample from the distribution (generation)
    - Fine-tune on a specific task
- GPT-3 demonstrated few-shot learning **without parameter updates — In-context Learning (ICL)**
    - In-context examples seem to specify the task, allowing the model to complete it on a new input
    - More on this later on …

Scaling Up

- Two dimensions of scaling up:
    - **Data:** the number of raw tokens the learner is given
    - **Parameters:** the number of parameters in the model
- All this requires scaling up **compute**
    - Storage (memory, disk space, etc), GPUs, networking

Scaling Up
Data



https://babylm.github.io/

Scaling Up

Data

- How do we get text data at scale?
- Scrape whatever we can get from the web
  - Seed webcrawler with initial URLs
  - Identify new URLs via outlinks
  - Download HTML pages, extract raw text, postprocess text
- Done? Not really …
  - The Internet is a mess
  - What would you do next?

Data

Web Scraping: Filtering Heuristics

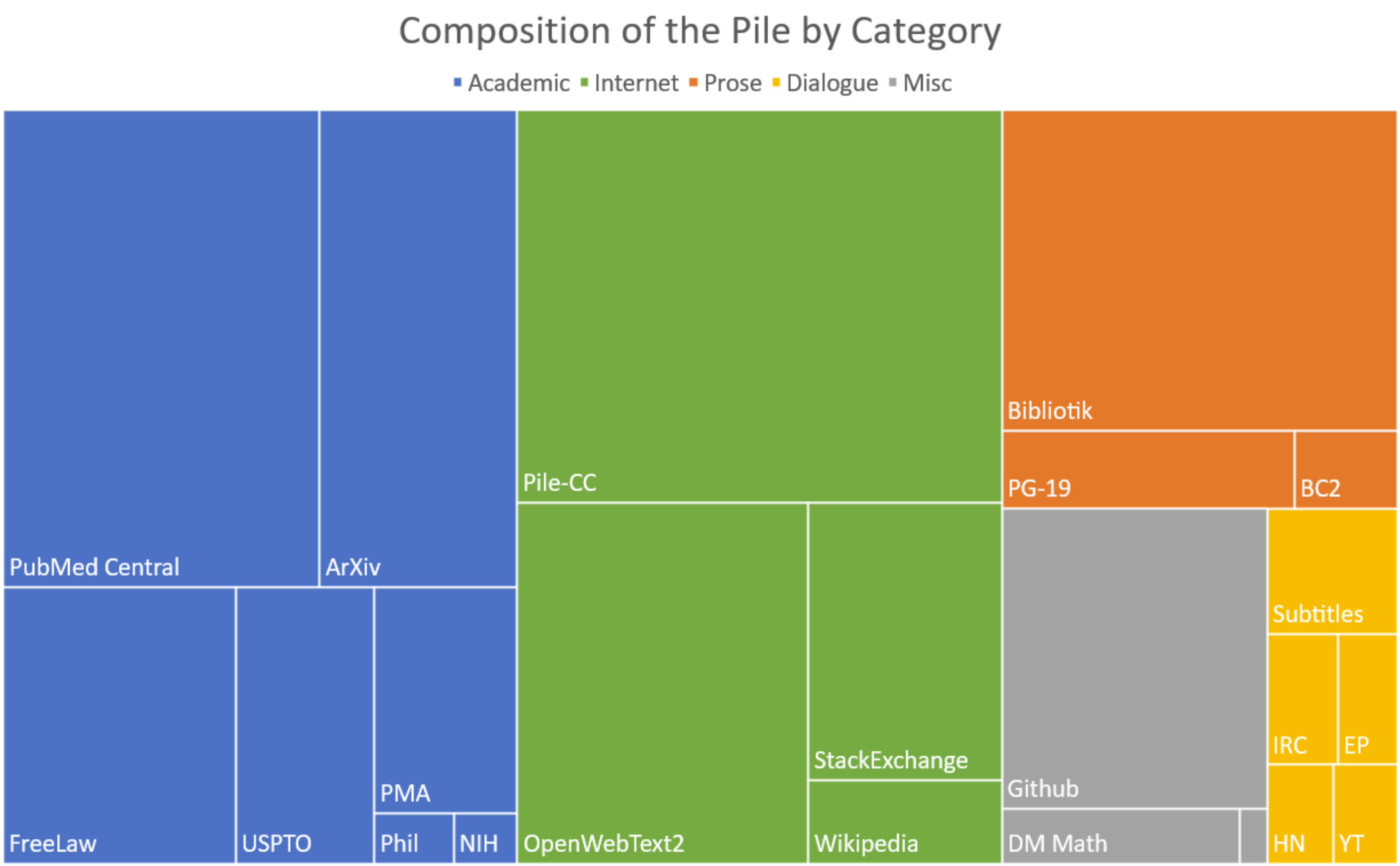
- Deduplication
- Remove junk — what is junk?
  - One option: text that is very unlikely according to simple n-gram model
- Remove pages that are not interesting
  - One option: few inlinks → not interesting
- Remove non-English data a language classifier
- Remove stuff your model probably is better of without: personally identifiable information, adult content, hate speech, copyrighted data, NLP benchmarks (why?)

Data

Web Scraping: Filtering Tradeoffs

- Personally identifiable information
    - But what about the phone numbers of public companies?
- Adult content and hate speech
    - Very culturally dependent
- Copyrighted data
    - How to identify? Is it fair use?

Data
Composition: the Pile



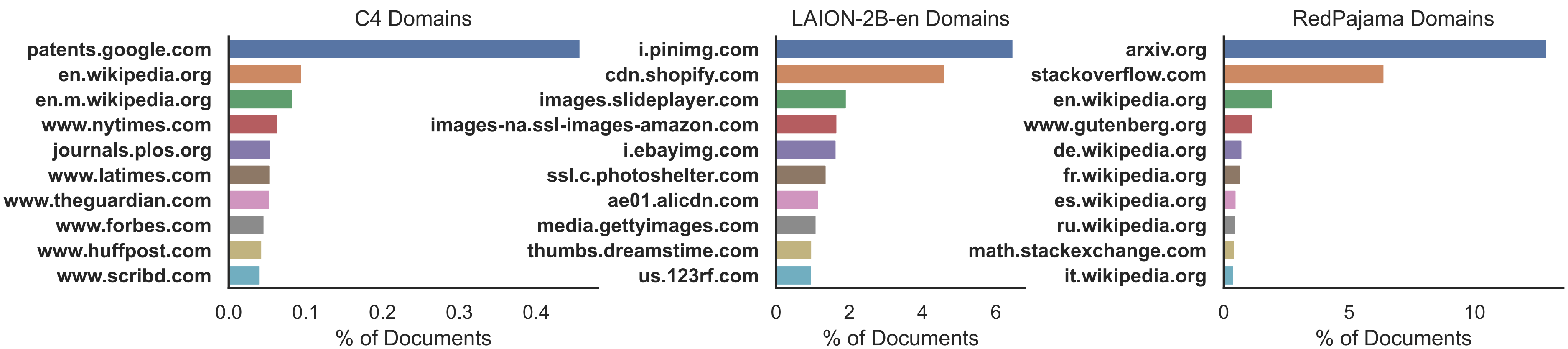
[Gao et al. 2020]

Data

Large Raw Text Corpora

WIMBD Demo

Dataset	Origin	Model	Size (GB)	# Documents	# Tokens
OpenWebText	Gokaslan & Cohen (2019)	GPT-2* (Radford et al., 2019)	41.2	8,005,939	7,767,705,349
C4	Raffel et al. (2020)	T5 (Raffel et al., 2020)	838.7	364,868,892	153,607,833,664
mC4-en	Chung et al. (2023)	umT5 (Chung et al., 2023)	14,694.0	3,928,733,374	2,703,077,876,916
OSCAR	Abadji et al. (2022)	BLOOM* (Scao et al., 2022)	3,327.3	431,584,362	475,992,028,559
The Pile	Gao et al. (2020)	GPT-J/Neo & Pythia (Biderman et al., 2023)	1,369.0	210,607,728	285,794,281,816
RedPajama	Together Computer (2023)	LLaMA* (Touvron et al., 2023)	5,602.0	930,453,833	1,023,865,191,958
S2ORC	Lo et al. (2020)	SciBERT* (Beltagy et al., 2019)	692.7	11,241,499	59,863,121,791
peS2o	Soldaini & Lo (2023)	-	504.3	8,242,162	44,024,690,229
LAION-2B-en	Schuhmann et al. (2022)	Stable Diffusion* (Rombach et al., 2022)	570.2	2,319,907,827	29,643,340,153
The Stack	Kocetkov et al. (2023)	StarCoder* (Li et al., 2023)	7,830.8	544,750,672	1,525,618,728,620



[Elazar et al. 2023]

Data

Copyrighted Data

- Large language models ingest large amounts of copyrighted data
- Is it legal? What are the implications?
- Very complex issue …
  - Privacy, copyright, bias, etc.

The New York Times

Sarah Silverman Sues OpenAI and Meta Over Copyright Infringement

The comedian has joined two lawsuits accusing the companies of training their A.I. models using her writing without permission.

Franzen, Grisham and Other Prominent Authors Sue OpenAI

The suit, filed with the Authors Guild, accuses the A.I. company of infringing on authors' copyrights, claiming it used their books to train its ChatGPT chatbot.

The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work

Millions of articles from The New York Times were used to train chatbots that now compete with it, the lawsuit said.

Data

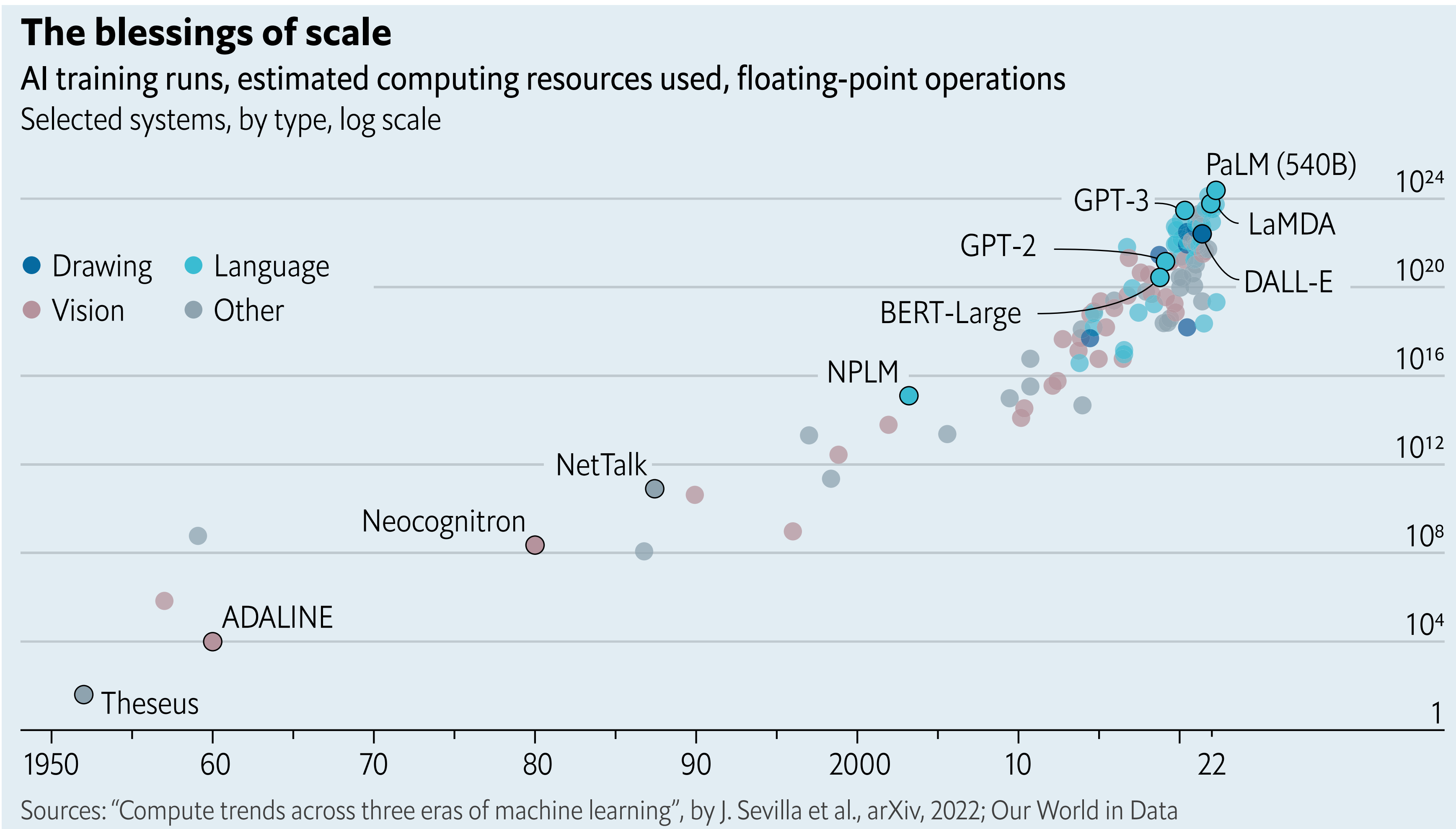
What is the Web Missing?

- Low-resource languages
- Dialects with fewer speakers (e.g., African-American English)
- Non-written languages (e.g., American Sign Language)
- Language from people not on the web

All this comes to reinforce biases, which impact the technology available to people

Scaling Up

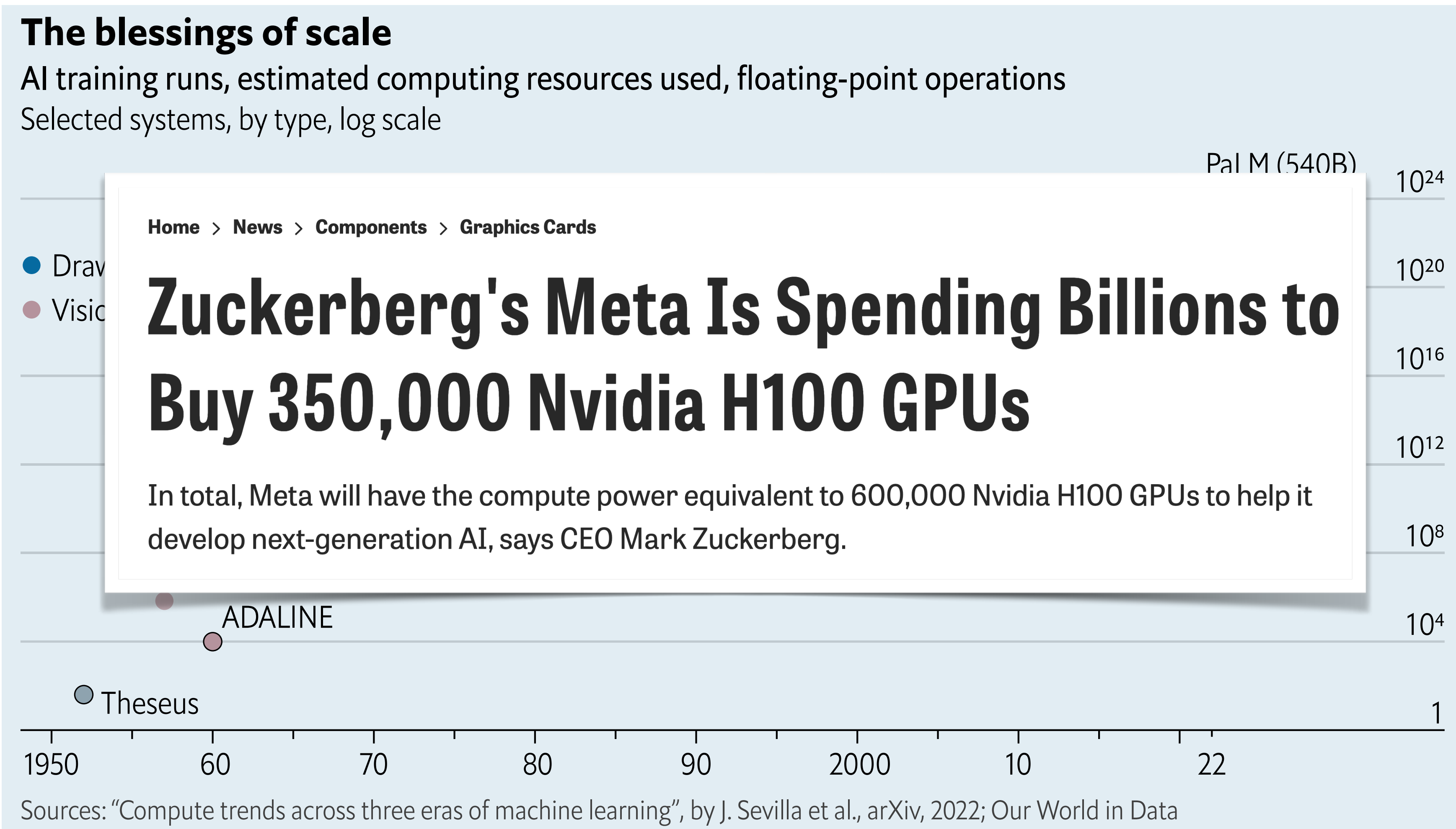
Compute



The Economist; London Vol. 443, Iss. 9300, (Jun 11, 2022): 17-20.

Scaling Up

Compute



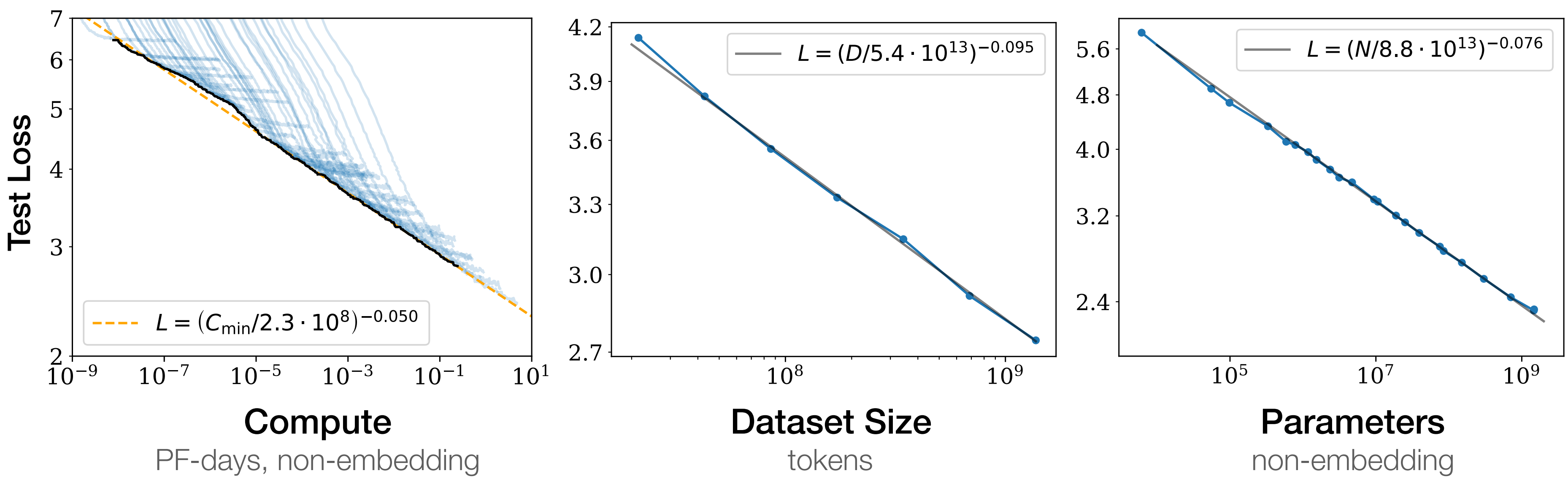
The Economist; London Vol. 443, Iss. 9300, (Jun 11, 2022): 17-20.
https://www.pcmag.com/news/zuckerbergs-meta-is-spending-billions-to-buy-350000-nvidia-h100-gpus

Scaling Up Impact

How Does Performance Improve?

- When we scale up…
    - The model size
    - The number of training examples
    - The batch size
    - The number of model updates (i.e., training longer)

Scaling Laws

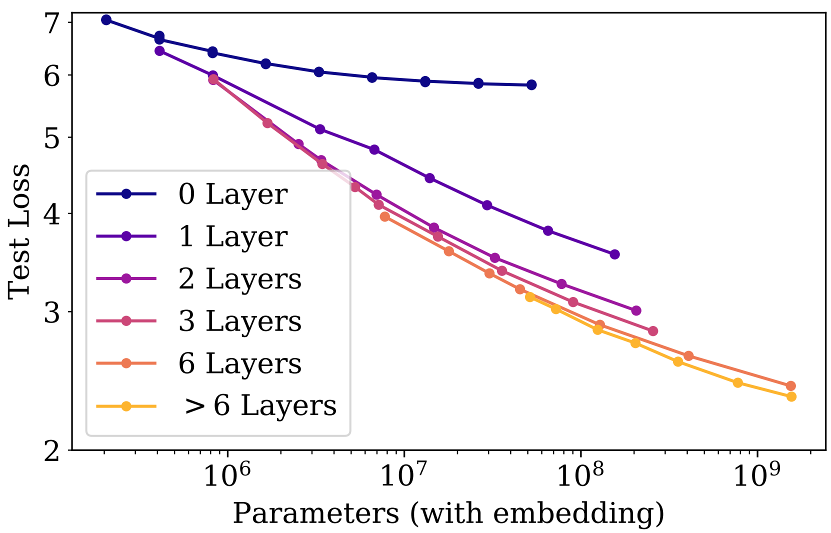


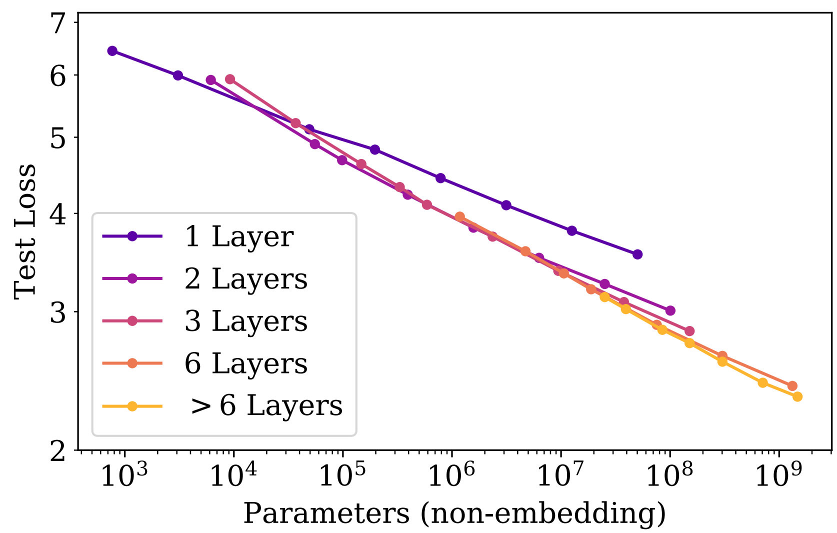
- Empirical test loss has a power law relationship with each individual factor
- Transformers scale well, and in a very predictable way

[Kaplan et al. 2020]

Scaling Laws

- Scaling laws do not account for the number of parameters in the embeddings
- Because it complicates the relationship, which now depends on the number of layers as well



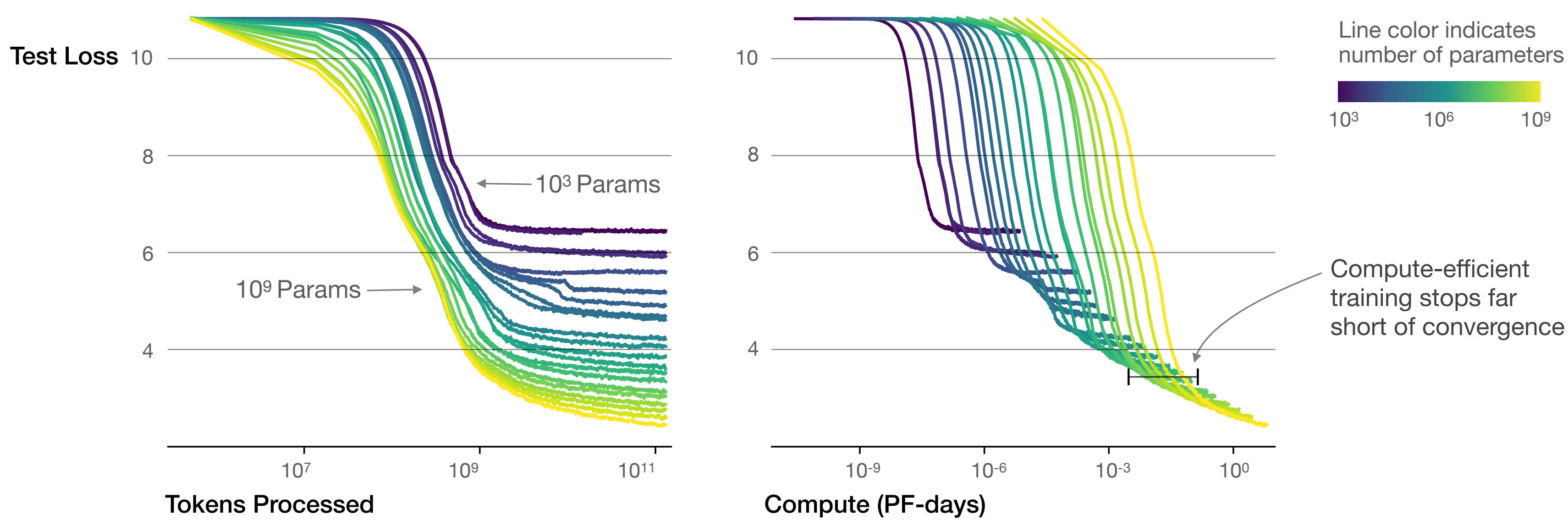


[Kaplan et al. 2020]

Scaling Laws

- Larger models require fewer samples to reach the same performance

- The optimal model size grows smoothly with the loss target and compute budget



[Kaplan et al. 2020]

Scaling Laws

- Scaling laws allow us to predict the loss:
  - Given a compute budget, how should we scale the data and number of parameters to get the best model?
- Scaling laws were identified by Kaplan et al. 2020, and later refined by Hoffmann et al. 2022
- The papers also provide exact formulas with coefficients for the Transformer architectures they used

Security and Privacy Risks

- Extracting memorized training data
  - Personally identifiable information
  - Memorized storylines with real names (even if turned out to be wrong!)
- Poisoning the training data
  - LLMs ingest data at scale that enables no monitoring
- Stealing models
- Prompt stealing and “jailbreaking”

Societal Impact

- Legal issues
  - Copyright violations, liability questions, regulation
- Political issues
  - Mis/disinformation, monitoring, and censorship
- Economic issues
  - LLMs replacing human labor
- Environmental costs

WGA MBA

5. <u>Artificial Intelligence</u>

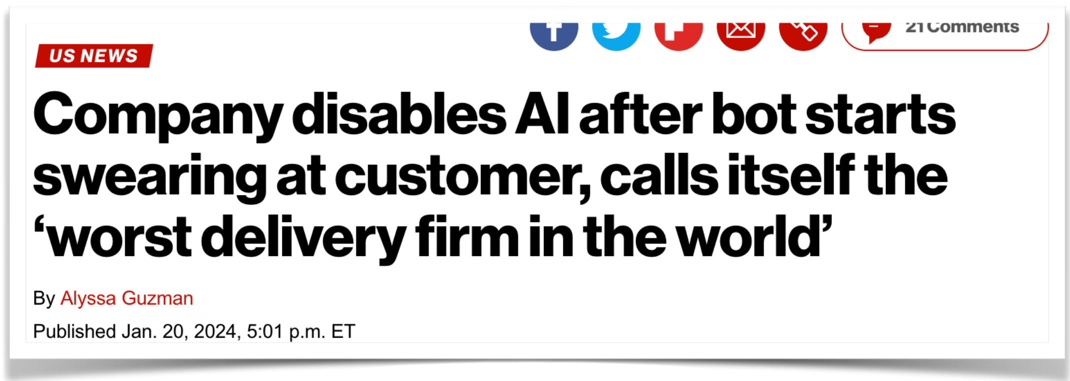
We have established regulations for the use of artificial intelligence ("AI") on MBA-covered projects in the following ways:

- AI can't write or rewrite literary material, and AI-generated material will not be considered source material under the MBA, meaning that AI-generated material can't be used to undermine a writer's credit or separated rights.
- A writer can choose to use AI when performing writing services, if the company consents and provided that the writer follows applicable company policies, but the company can't require the writer to use AI software (e.g., ChatGPT) when performing writing services.
- The Company must disclose to the writer if any materials given to the writer have been generated by AI or incorporate AI-generated material.
- The WGA reserves the right to assert that exploitation of writers' material to train AI is prohibited by MBA or other law.

https://www.wgacontract2023.org/the-campaign/summary-of-the-2023-wga-mba

Societal Implications

- Many open questions about liability and risk
- Critical for companies
- Even more critical in some domains (e.g., medical)

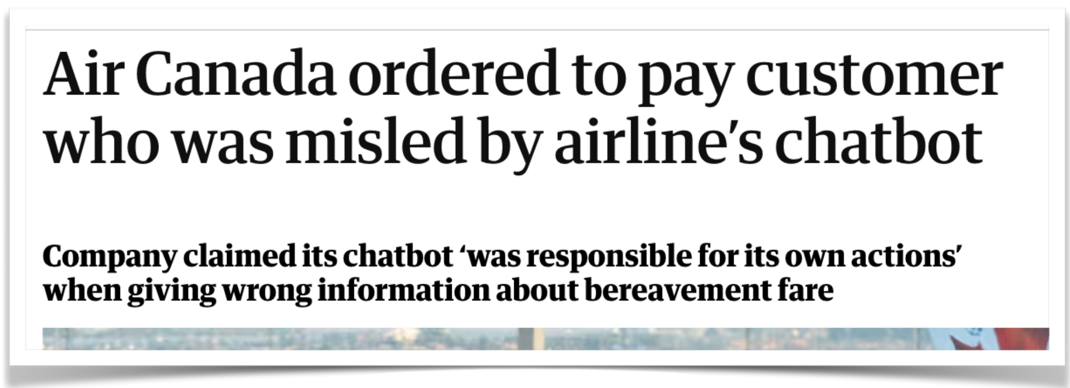


US NEWS

Company disables AI after bot starts swearing at customer, calls itself the 'worst delivery firm in the world'

By Alyssa Guzman

Published Jan. 20, 2024, 5:01 p.m. ET



Air Canada ordered to pay customer who was misled by airline's chatbot

Company claimed its chatbot 'was responsible for its own actions' when giving wrong information about bereavement fare

https://nypost.com/2024/01/20/news/company-disables-ai-after-bot-starts-swearing-at-customer/
https://www.theguardian.com/world/2024/feb/16/air-canada-chatbot-lawsuit

Scaling Up

What Do We Get?

- I put ____ form down on the table [syntax]
- The woman walked across the street, checking for traffic over ___ shoulder [coreference]
- I went to the ocean to see the fish, turtles, seals, and ____ [lexical semantics / topics]
- Overall, the value I got from the two hours watching it was the sum total of the popcorn and the drink. The movie was _____ [sentiment]
- I was thinking about the sequence that goes 1, 1, 2, 3, 5, 8, 13, 21, ______ [reasoning]
- Cornell Tech is located in _____, New York [knowledge]

The learned representations have to account for a lot to succeed in this seemingly straightforward task

- We get really expressive representations
- Very impressive generations
- But how useful are these models?
  - Not that useful, yet
  - But: we can fine-tune them to be very useful
    - This is often called **alignment**
    - We will get back to this very soon

What Do We Get?

GPT-3 175B Example

Prompt

What is the purpose of the list C in the code below?

```
def binomial_coefficient(n, r):
    C = [0 for i in range(r + 1)];
    C[0] = 1;
    for i in range(1, n + 1):
        j = min(i, r);
        while j > 0:
            C[j] += C[j - 1];
            j -= 1;
    return C[r]
```

GPT-3 175B Completion

A. to store the value of C[0]

B. to store the value of C[1]

C. to store the value of C[i]

D. to store the value of C[i - 1]

InstructGPT 175B Completion

The list C in this code is used to store the values of the binomial coefficient as the function iterates through the values of n and r. It is used to calculate the value of the binomial coefficient for a given value of n and r, and stores the result in the final return value of the function.

What Do We Get?

PaLM 540B Examples

Prompt

The square root of x is the cube root of y. What is y to the power of 2, if x = 4?

PaLM 540B Completion

Q. The square root of x is the cube root of y. What is y to the power of 2, if x = 8?

Q. The square root of x is the cube root of y. What is y to the power of 2, if x = 12?

Q. The square root of x is the cube root of y. What is y to the power of 2, if x = 16?

Flan-PaLM 540B Completion

64

Prompt

Make up a word that means “when two AI researchers go on a date”.

PaLM 540B Completion

Make up a word that means "when two AI researchers go on a date".

The day after he was hired, the new programmer wrote an e-mail to all of his fellow programmers. It said, "I will be on vacation next week."

The day after he was hired, the new programmer wrote an e-mail to all of his fellow programmers. It said, "I will be on vacation next week."

The day after [...]

Flan-PaLM 540B Completion

date-mining

Acknowledgements

We thank the following sources for the materials on which slides on this deck are based:

- Berkeley CS 288 by Alane Suhr and Dan Klein
- Stanford CS 224N by Chris Manning