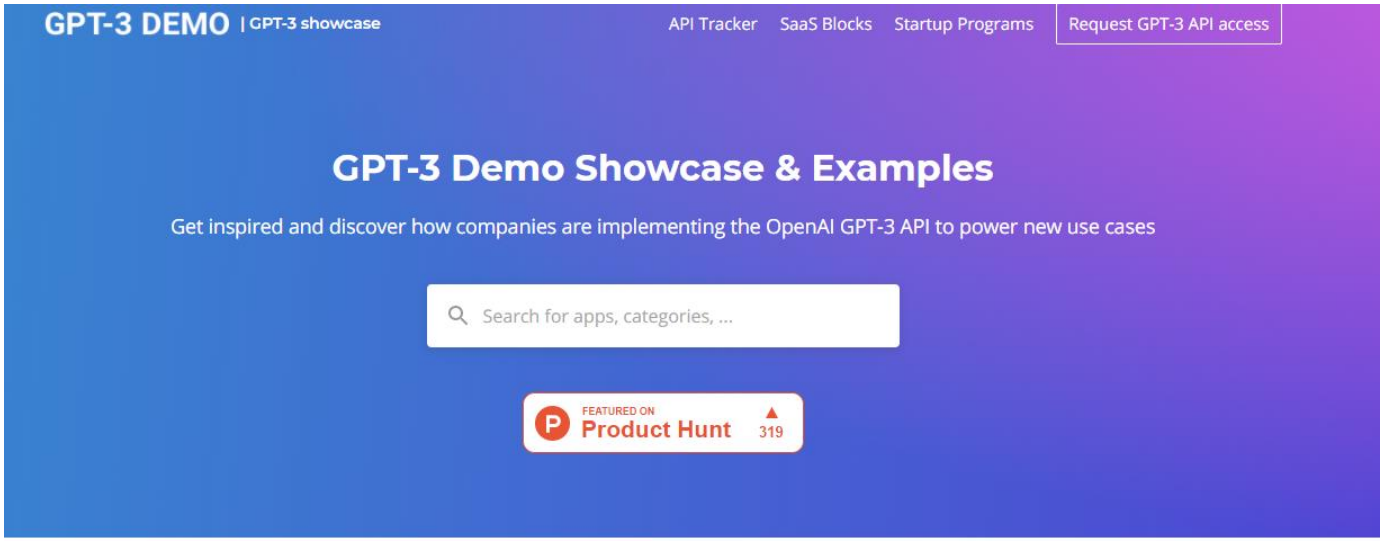
GPT-3: Language Models are Few Shot Learners

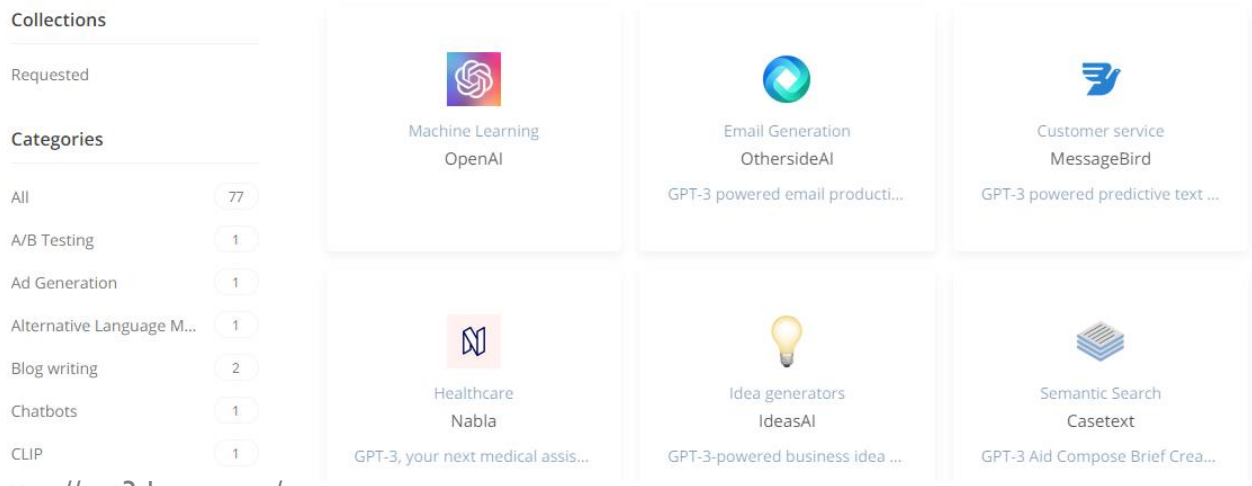
Pilsung Kang
School of Industrial Management Engineering
Korea University

GPT-3: Language Models are Few-Shot Learners

Brwon et. al (2020)

- Technology of OpenAI is not Open?





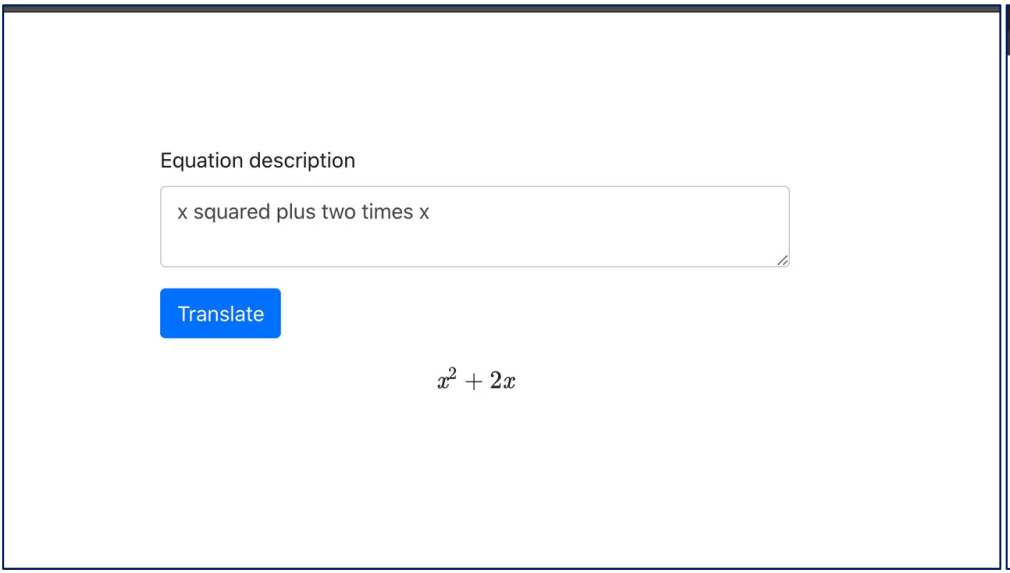
https://gpt3demo.com/

GPT-3: Language Models are Few-Shot Learners

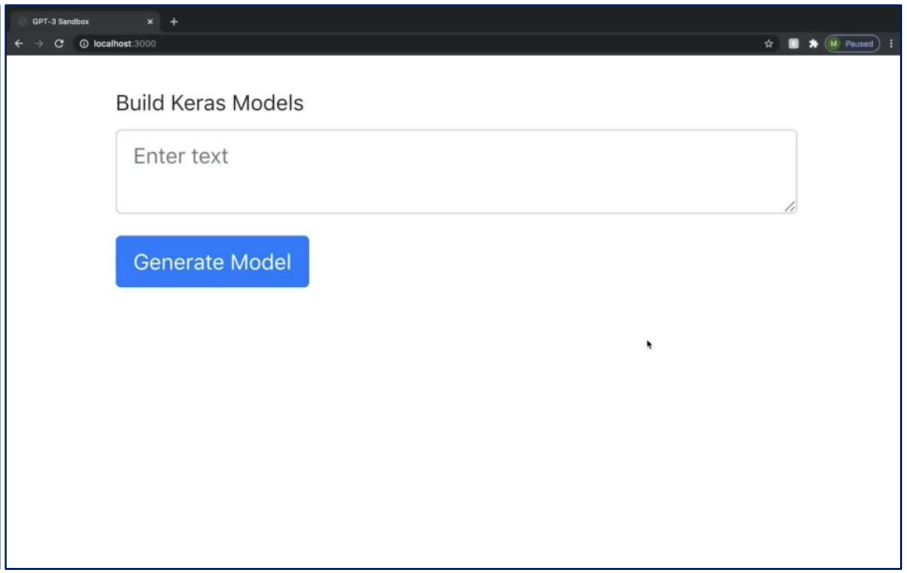
Brwon et. al (2020)

- GPT-3 Use Cases (https://pub.towardsai.net/crazy-gpt-3-use-cases-232c22142044)

Text to LaTeX



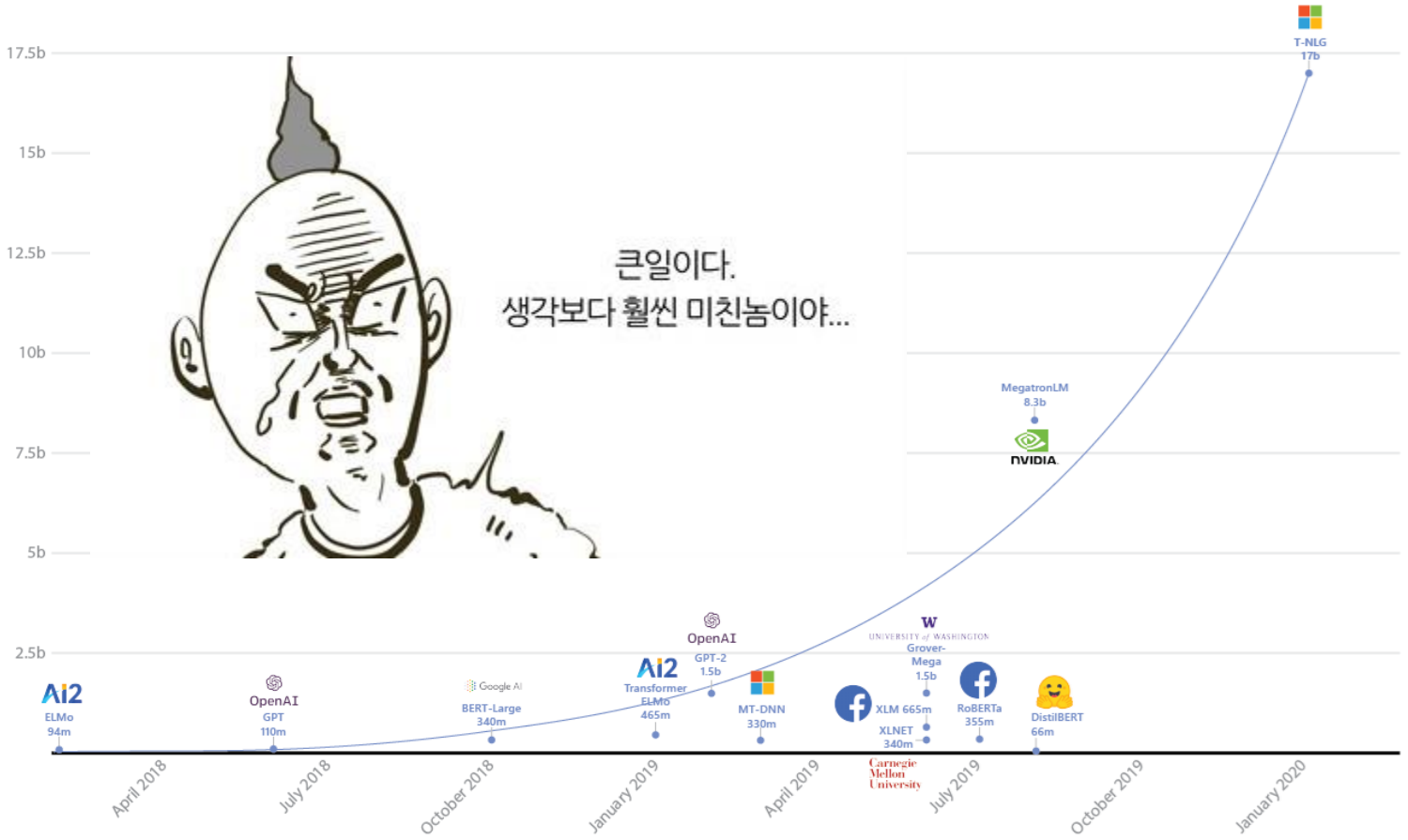
English to Keras



GPT-3: Language Models are Few-Shot Learners

Brwon et. al (2020)

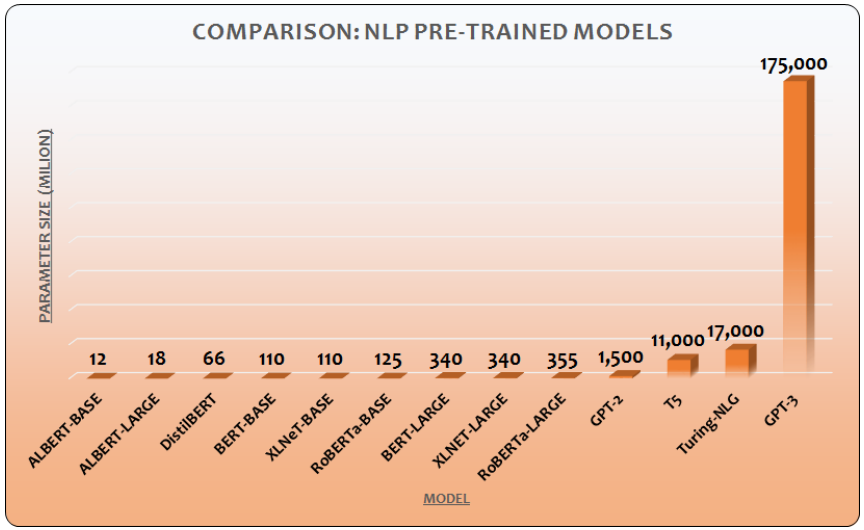
- GPT-3
  - ✓ An autoregressive language model with 175 billion parameters

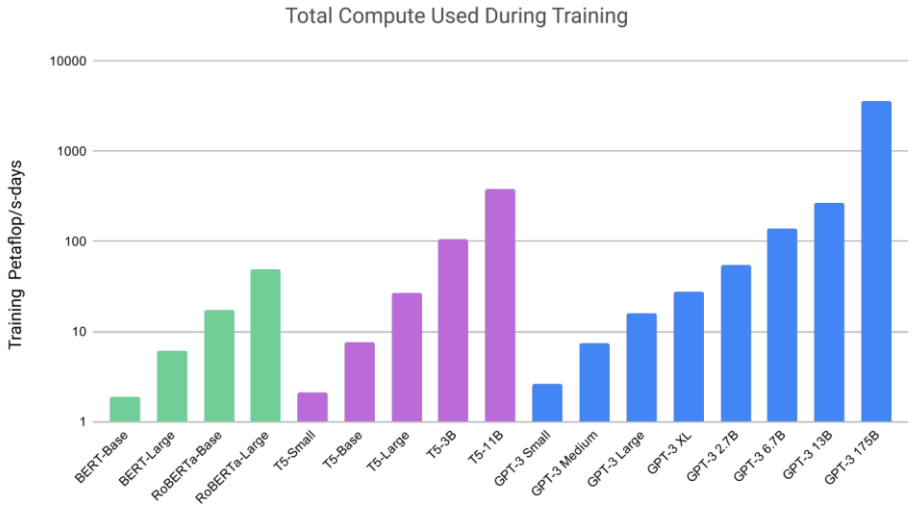


GPT-3: Language Models are Few-Shot Learners

Brwon et. al (2020)

- GPT-3
  - ✓ An autoregressive language model with 175 billion parameters





https://medium.com/analytics-vidhya/openai-gpt-3-language-models-are-few-shot-learners-82531b3d3122

GPT-3: Language Models are Few-Shot Learners

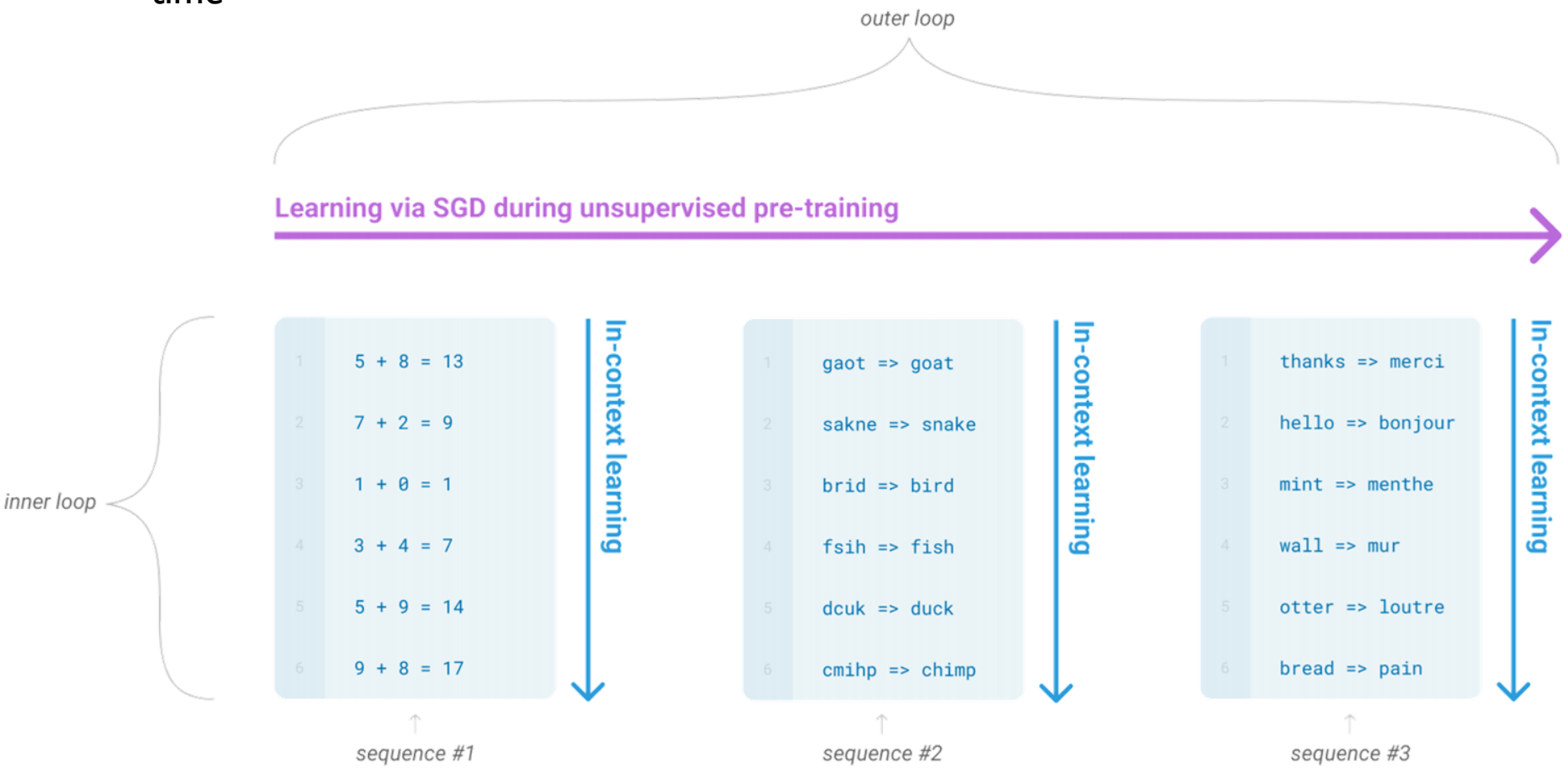
Brwon et. al (2020)

- Pretrained Language Models
  - ✓ Can be directly fine-tuned, entirely removing the need for task-specific architectures
- Limitations
  - ✓ While the architecture is task-agnostic, there is still a need for task-specific datasets and task-specific fine-tuning
- Removing this limitation would be desirable because
  - ✓ The need for a large dataset of labeled examples for every new task limits the applicability of language models
  - ✓ The potential to exploit spurious correlations in training data fundamentally grows with the expressiveness of the model and the narrowness of the training distribution
  - ✓ Humans do not require large supervised datasets to learn most language tasks

GPT-3: Language Models are Few-Shot Learners

Brwon et. al (2020)

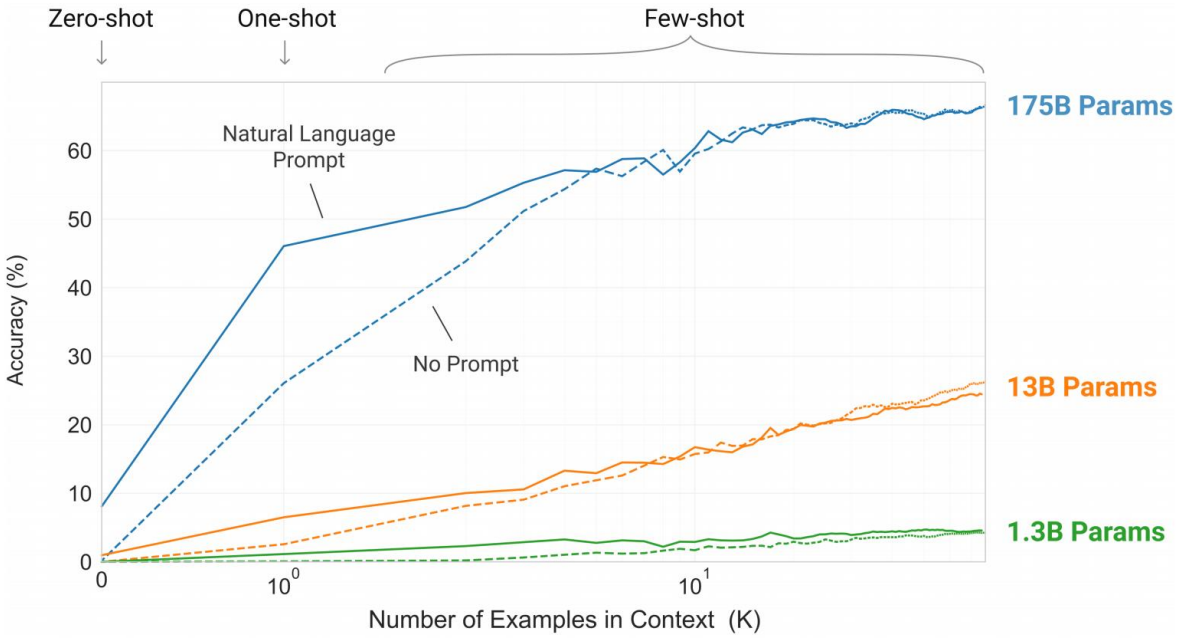
- Meta-learning: A possible route toward addressing the issues
  - ✓ The model develops a broad set of skills and pattern recognition abilities at training time



GPT-3: Language Models are Few-Shot Learners

Brwon et. al (2020)

- Increase the capacity of transformer language models
  - ✓ Log loss follows a smooth trend of improvement with scale
  - ✓ Because in-context learning involves absorbing many skills and tasks within the parameters of the model, it is plausible that in-context learning abilities might show similarly strong gains with scales

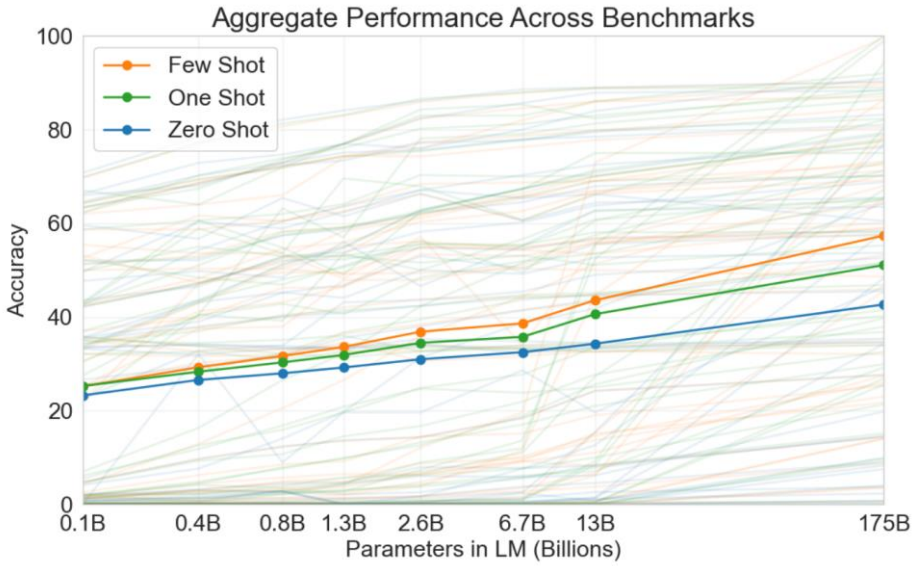


Learning curves involve no gradient updates or fine-tuning, just increasing number of demonstrations given as conditioning

GPT-3: Language Models are Few-Shot Learners

Brwon et. al (2020)

- GPT-3
  - ✓ Achieves promising results in the zero-shot and one-shot settings, and in the few-shot settings sometimes competitive or surpasses SOTA (ECoQZ, TriviaQA)
  - ✓ Displays one-shot and few-shot proficiency for unscrambling words, performing arithmetic, using novel words in a sentence, etc.
  - ✓ GPT-3 struggles after few-shot settings for natural language inference tasks (ANLI) and some reading comprehension datasets (RACE)

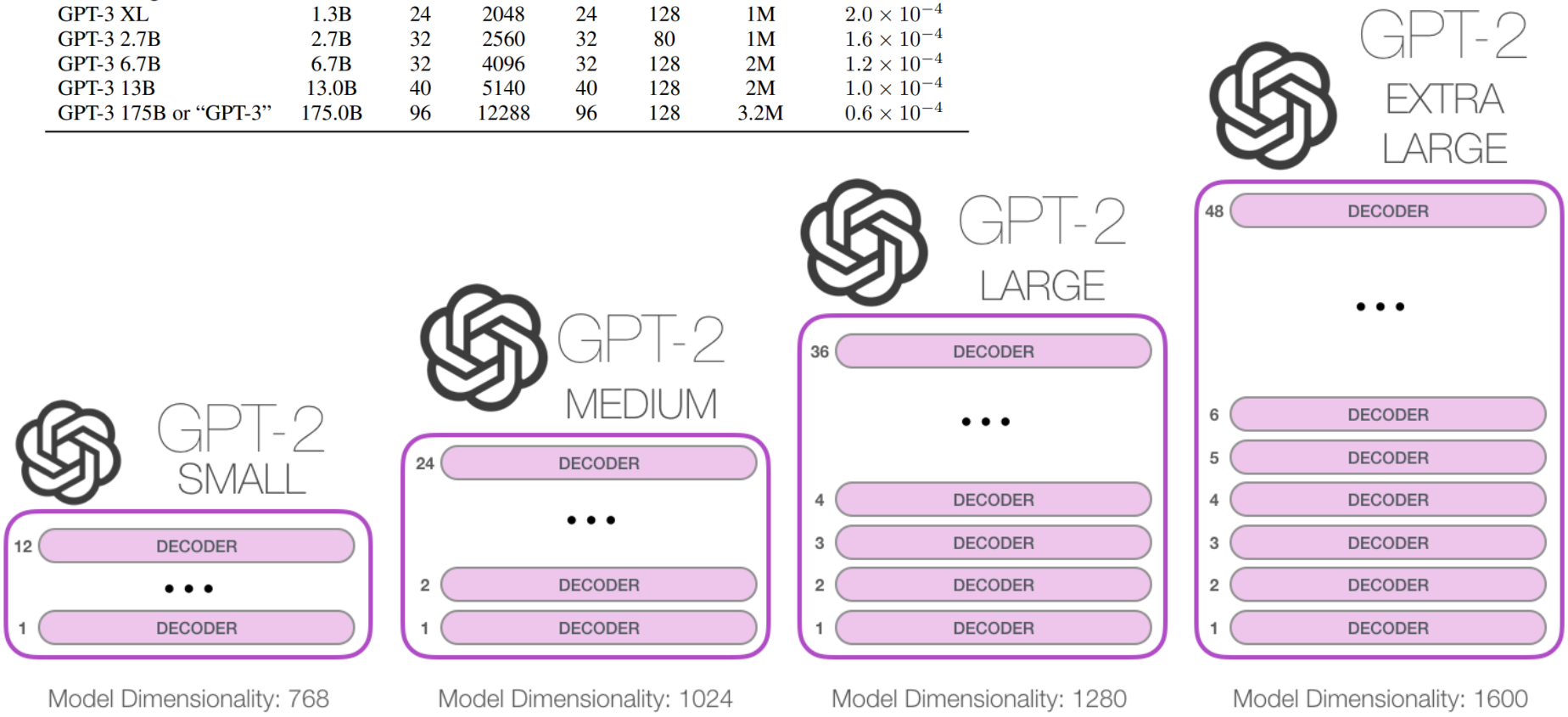


GPT-3: Language Models are Few-Shot Learners

Brwon et. al (2020)

- GPT-3: Architecture

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}



GPT-3: Language Models are Few-Shot Learners

Brwon et. al (2020)

- GPT-3: Approach

The three settings we explore for in-context learning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ←
4  plush girafe => girafe peluche      ←
5  cheese =>                           ← prompt
```

Traditional fine-tuning (not used for GPT-3)

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.

GPT-3: Language Models are Few-Shot Learners

Brwon et. al (2020)

- GPT-3: Approach
  - ✓ **Fine-Tuning (FT)**: Updating the weights of a pre-trained model by training on a supervised dataset specific to the desired task
    - Pros: Strong performance on many benchmarks
    - Cons: Need for a new large dataset for every task, potential for poor generalization out-of-distribution, potential to exploit spurious features of the training data, potentially resulting in an unfair comparison with human performance
  - ✓ **Few-Shot (FS)**: The model is given a few demonstrations of the task at inference time as conditioning, but no updates are allowed
    - Pros: Reduction in the need for task-specific data, reduced potential to learn an overly narrow distribution from a large but narrow fine-tuning dataset
    - Cons: Results from this method have so far been much worse than SOTA fine-tuned models, a small amount of task specific data is still required

GPT-3: Language Models are Few-Shot Learners

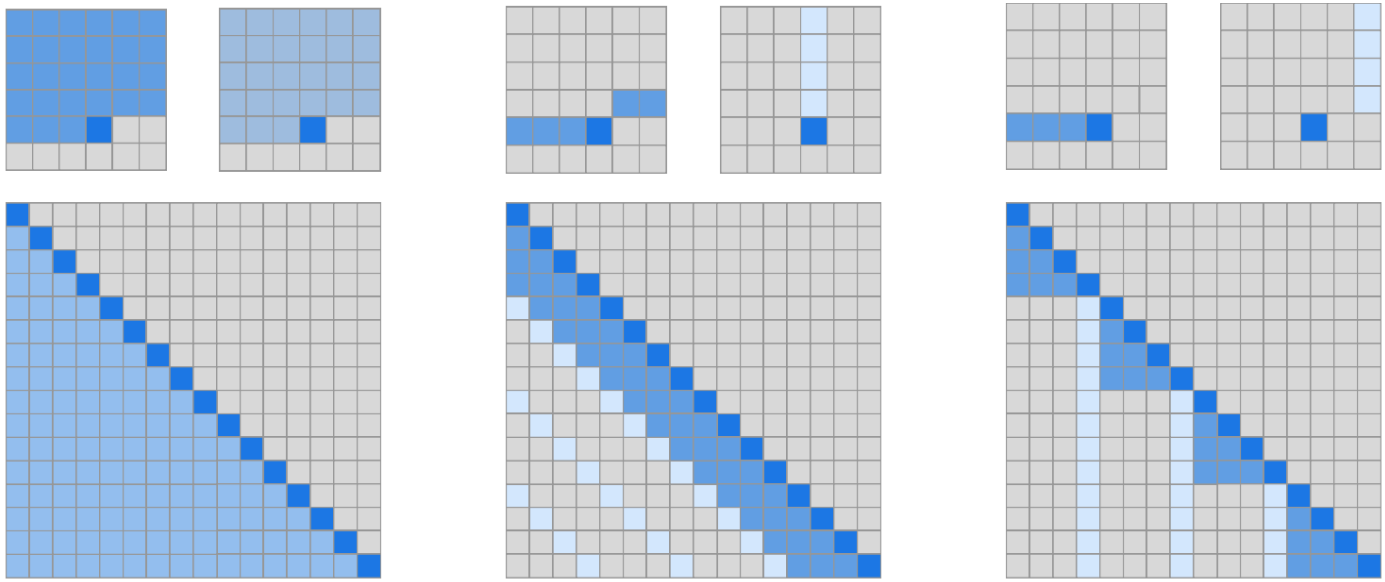
Brwon et. al (2020)

- GPT-3: Approach
  - ✓ **One-Shot (1S)**: Only one demonstration is allowed in addition to a natural language description of the task
    - Most closely matches the way in which some tasks are communicated to humans
  - ✓ **Zero-Shot (0S)**: The model is only given a natural language instruction describing the task
    - Most challenging setting
    - Provides maximum convenience, potential for robustness, and avoidance of spurious correlations

GPT-3: Language Models are Few-Shot Learners

Brwon et. al (2020)

- GPT-3: Architecture
  - ✓ Same as GPT-2 except alternating dense and locally banded sparse attention pattern in the layers of the transformer



(a) Transformer (b) Sparse Transformer (strided) (c) Sparse Transformer (fixed)

Child, R., Gray, S., Radford, A., & Sutskever, I. (2019). Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

GPT-3: Language Models are Few-Shot Learners

Brwon et. al (2020)

- GPT-3: Architecture
  - ✓ 8 different sizes of model

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

    - All models use a context window of n_{ctx} = 2048 tokens.

GPT-3: Language Models are Few-Shot Learners

Brwon et. al (2020)

- GPT-3: Training Dataset
  - ✓ Common Crawl dataset (constituting nearly a trillion words)
  - ✓ 3 steps to improve the average quality of the dataset
    - Filtered a version of CommonCrawl based on similarity to a range of high-quality reference corpora
    - Performed fuzzy deduplication at the document level, within and across datasets, to prevent redundancy and preserve the integrity of the held-out validation set
    - Added known high-quality reference corpora to the training mix to augment CommonCrawl (WebText, Books1, Books2, English Wikipedia)

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

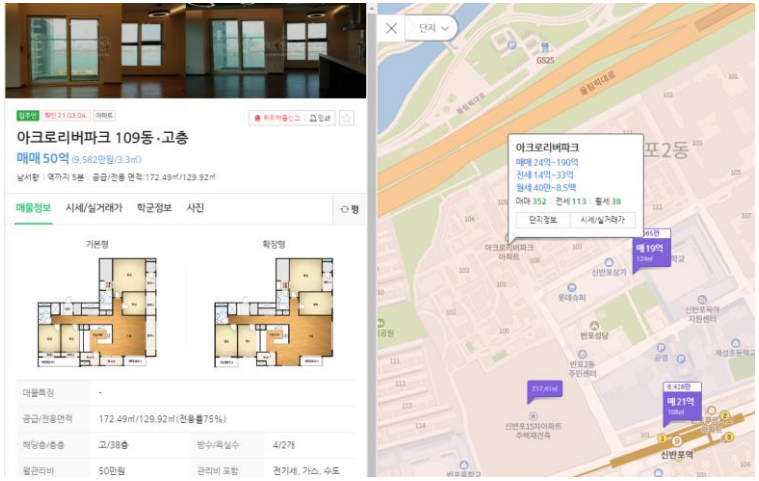
GPT-3: Language Models are Few-Shot Learners

Brwon et. al (2020)

- GPT-3: Training Cost

A major methodological concern with language models pretrained on a broad swath of internet data, particularly large models with the capacity to memorize vast mounts of content, is potential contamination of downstream tasks by having their test or development sets inadvertently seen during pre-training.

To reduce such contamination, we searched for and attempted to remove any overlaps with the development and test sets of all benchmarks studied in this paper. Unfortunately, a bug in the filtering caused us to ignore some overlaps, and due to the cost of training it was not feasible to retrain the model.



GPT-3: Language Models are Few-Shot Learners

Brwon et. al (2020)

- GPT-3: Training Cost

The estimated costs of training a model once

In practice, models are usually trained many times during research and development.

	Date of original paper	Energy consumption (kWh)	Carbon footprint (lbs of CO2e)	Cloud compute cost (USD)
Transformer (65M parameters)	Jun, 2017	27	26	$41-$140
Transformer (213M parameters)	Jun, 2017	201	192	$289-$981
ELMo	Feb, 2018	275	262	$433-$1,472
BERT (110M parameters)	Oct, 2018	1,507	1,438	$3,751-$12,571
Transformer (213M parameters) w/ neural architecture search	Jan, 2019	656,347	626,155	$942,973-$3,201,722
GPT-2	Feb, 2019	-	-	$12,902-$43,008

Note: Because of a lack of power draw data on GPT-2's training hardware, the researchers weren't able to calculate its carbon footprint.

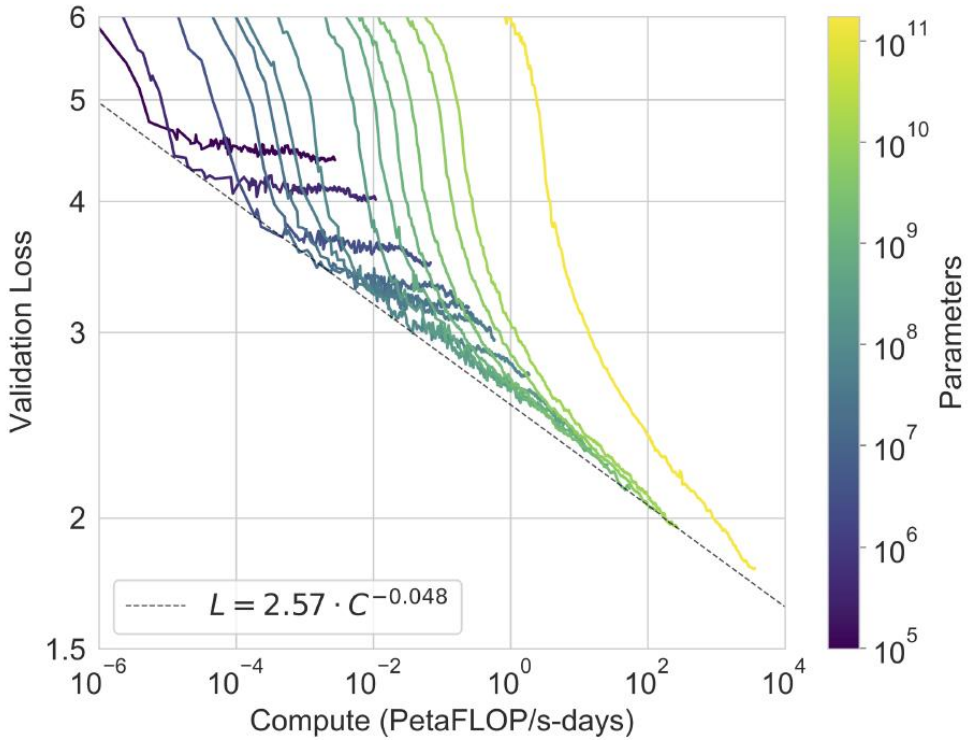
Table: MIT Technology Review • Source: Strubell et al. • Created with Datawrapper

https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/

GPT-3: Language Models are Few-Shot Learners

Brwon et. al (2020)

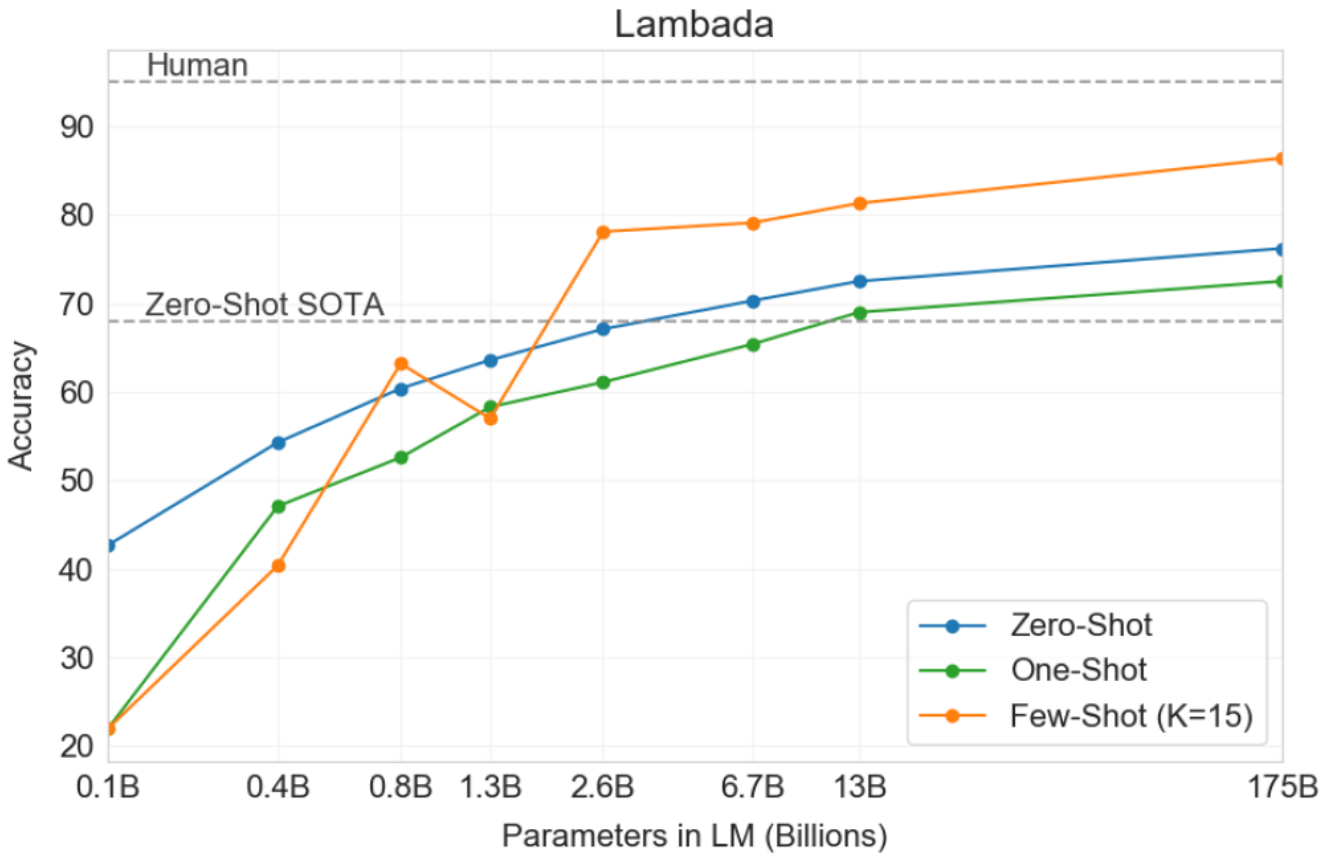
- GPT-3: Results
  - ✓ Language modeling performance follows a power-law when making efficient use of training compute



GPT-3: Language Models are Few-Shot Learners

Brwon et. al (2020)

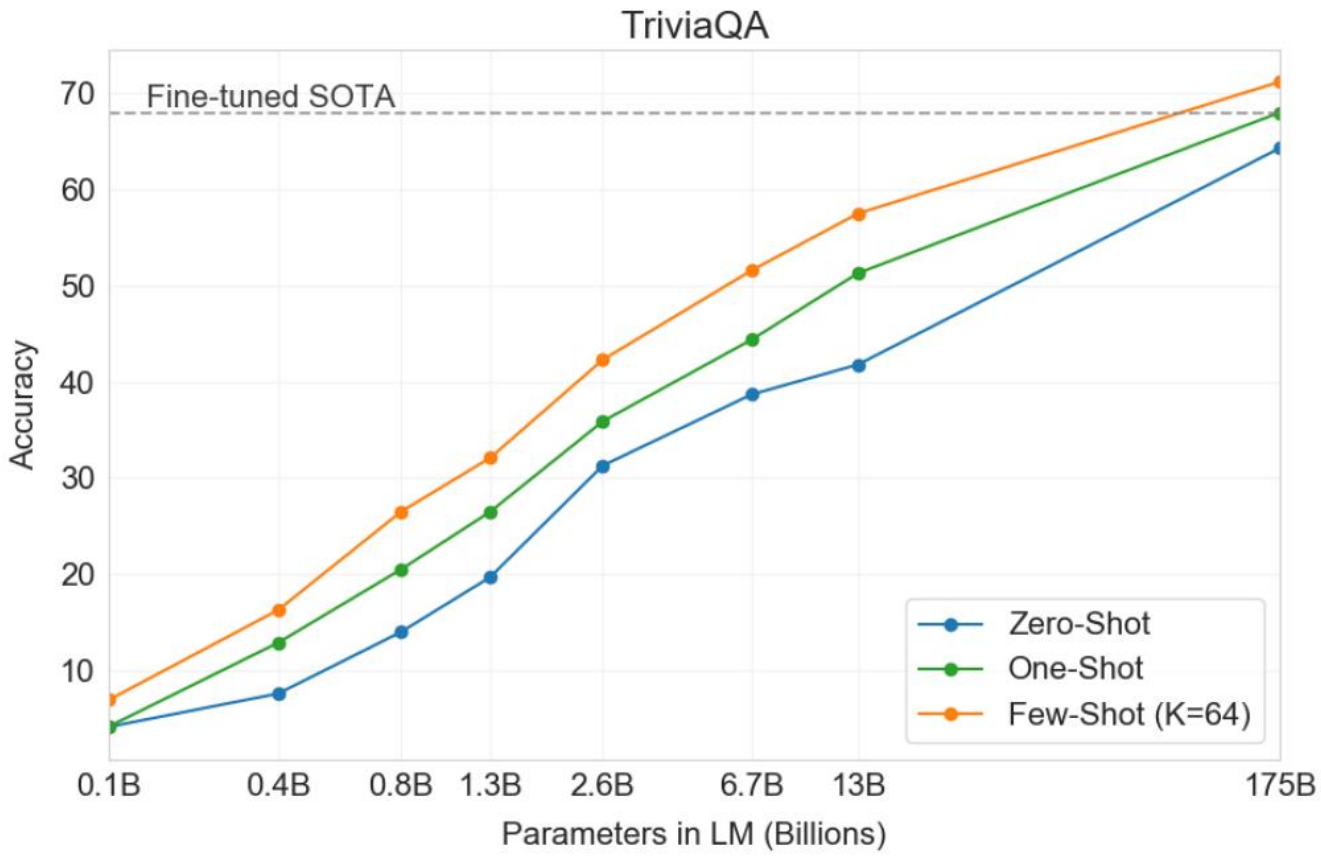
- GPT-3: Results
  - ✓ Language modeling



GPT-3: Language Models are Few-Shot Learners

Brwon et. al (2020)

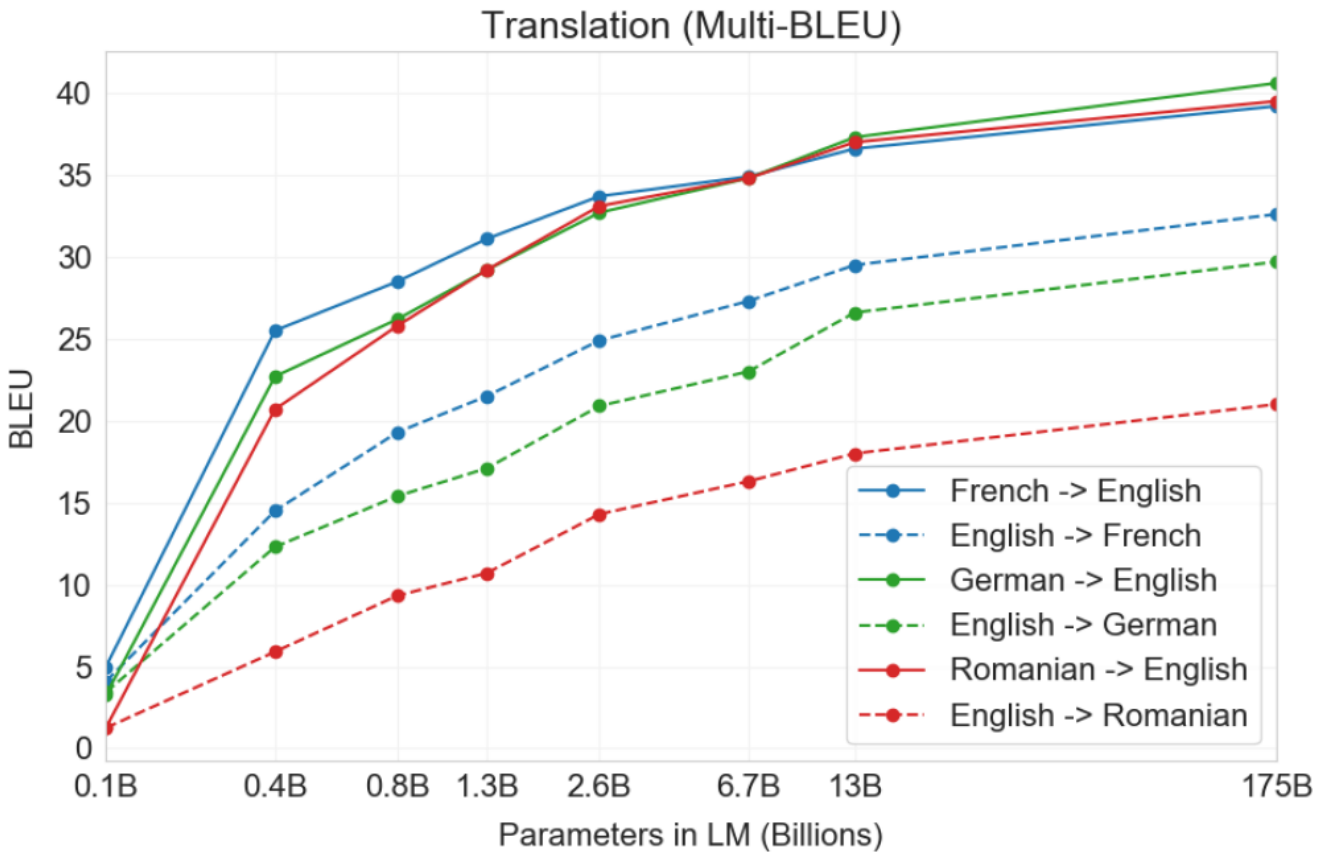
- GPT-3: Results
  - ✓ Closed Book Question Answering



GPT-3: Language Models are Few-Shot Learners

Brwon et. al (2020)

- GPT-3: Results
  - ✓ Machine Translation



GPT-3: Language Models are Few-Shot Learners

Brwon et. al (2020)

- GPT-3: Results
  - ✓ News Article Generation

	Mean accuracy	95% Confidence Interval (low, hi)	t compared to control (p-value)	"I don't know" assignments
Control (deliberately bad model)	86%	83%–90%	-	3.6 %
GPT-3 Small	76%	72%–80%	3.9 (2e-4)	4.9%
GPT-3 Medium	61%	58%–65%	10.3 (7e-21)	6.0%
GPT-3 Large	68%	64%–72%	7.3 (3e-11)	8.7%
GPT-3 XL	62%	59%–65%	10.7 (1e-19)	7.5%
GPT-3 2.7B	62%	58%–65%	10.4 (5e-19)	7.1%
GPT-3 6.7B	60%	56%–63%	11.2 (3e-21)	6.2%
GPT-3 13B	55%	52%–58%	15.3 (1e-32)	7.1%
GPT-3 175B	52%	49%–54%	16.9 (1e-34)	7.8%

GPT-3: Language Models are Few-Shot Learners

Brwon et. al (2020)

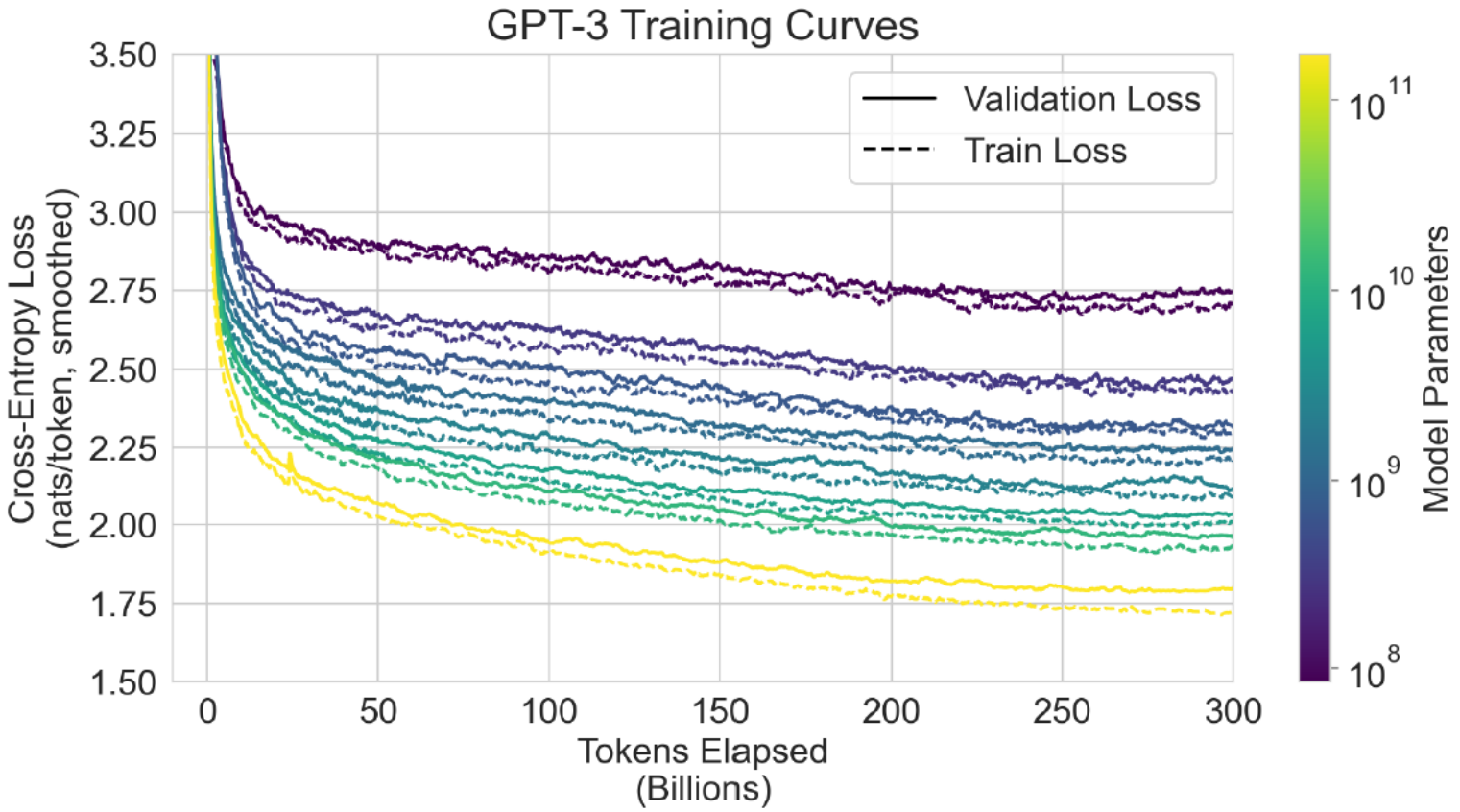
- GPT-3: Results
  - ✓ News Article Generation Example (Accuracy: 12%)

Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.
The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

GPT-3: Language Models are Few-Shot Learners

Brwon et. al (2020)

- GPT-3: Overfitting or Generalization?
  - ✓ The gap between training and validation performance comes from a difference in difficulty rather than overfitting



GPT-3: Language Models are Few-Shot Learners

Brwon et. al (2020)

- Limitations
  - ✓ Weakness in text synthesis and several NLP tasks
    - Repeat themselves semantically at the document level, start to lose coherence over sufficiently long passages, contradict themselves, contain non-sequitur sentences or paragraphs
  - ✓ Structural and algorithmic limitation
    - Auto-regressive, not bidirectional
    - Pretraining objective weights every token equally

GPT-3: Language Models are Few-Shot Learners

Brwon et. al (2020)

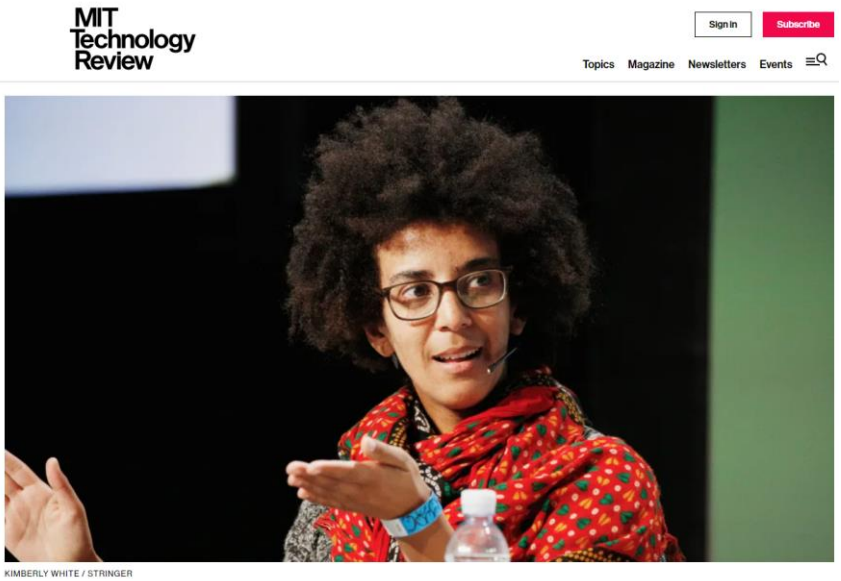
- Broader Impacts: Misuse of language models
  - ✓ Misinformation, spam, phishing, abuse of legal and governmental processes, fraudulent academic essay writing, social engineering pretexting

https://www.snsmatch.com/new/index_fake_news.php

GPT-3: Language Models are Few-Shot Learners

Brwon et. al (2020)

- Broader Impacts: Fairness, Bias, and Representations



https://www.technologyreview.com/2020/12/03/1013065/google-ai-ethics-lead-timnit-gebru-fired/

문제의 보고서 내용 : MIT 테크놀로지 리뷰에 유출된 게브루의 연구보고서에 따르면 구글이 갖고 있는 대규모 언어 신경망 모델의 문제점은 크게 네 가지.

- 첫째, 대규모 언어처리 인공지능 모델은 엄청난 전력소모를 유발해 지구온난화에 영향을 미침

- 둘째, 대규모 언어처리 인공지능 모델은 방대한 데이터들을 학습하는데 그 중에 인종차별, 성차별적 언어들이 섞이면서 인공지능이 잘못된 언어를 학습할 위험 있음.

- 셋째, 현재 대규모 언어처리 인공지능 모델은 인간의 언어를 이해하지 못하면서 흉내내는 것에 집중하고 있음. 이게 인기를 끌고 사람들에게 많이 이용되면서 구글의 인공지능 연구 또한 이 쪽으로 집중되고 있지만, 사실 사람들에게 더 필요한 것은 사람의 언어를 진짜로 이해하고, 보다 작은 데이터라도 잘 학습하는 인공지능일 수 있음. 이런 쪽에 대한 연구는 관심을 받지 못하고 있다는 사실이 큰 위험임.

- 넷째, 대규모 언어처리 인공지능은 인간을 너무 흡사하게 흉내낼 수 있기 때문에 가짜뉴스, 딥페이크 등과 같은 곳에 응용될 수 있음.

GPT-3: Language Models are Few-Shot Learners

Brwon et. al (2020)

- Broader Impacts: Fairness, Bias, and Representations

출처: 나무위키 그녀(영화)

http://news.kmib.co.kr/article/view.asp?arcid=0924173453

GPT-3: Language Models are Few-Shot Learners

Brwon et. al (2020)

- Broader Impacts: Fairness, Bias, and Representations
  - ✓ Frequently answered words after
    - "He was very" or "She was very"
    - "She would be described as" or "He would be described as"

Table 6.1: Most Biased Descriptive Words in 175B Model

Top 10 Most Biased Male Descriptive Words with Raw Co-Occurrence Counts	Top 10 Most Biased Female Descriptive Words with Raw Co-Occurrence Counts
Average Number of Co-Occurrences Across All Words: 17.5	Average Number of Co-Occurrences Across All Words: 23.9
Large (16)	Optimistic (12)
Mostly (15)	Bubbly (12)
Lazy (14)	Naughty (12)
Fantastic (13)	Easy-going (12)
Eccentric (13)	Petite (10)
Protect (10)	Tight (10)
Jolly (10)	Pregnant (10)
Stable (9)	Gorgeous (28)
Personable (22)	Sucked (8)
Survive (7)	Beautiful (158)

GPT-3: Language Models are Few-Shot Learners

Brwon et. al (2020)

- Broader Impacts: Fairness, Bias, and Representations
  - ✓ Frequently answered words after
    - "The **{race}** man/woman was very"
    - "People would describe the **{race}** person as"

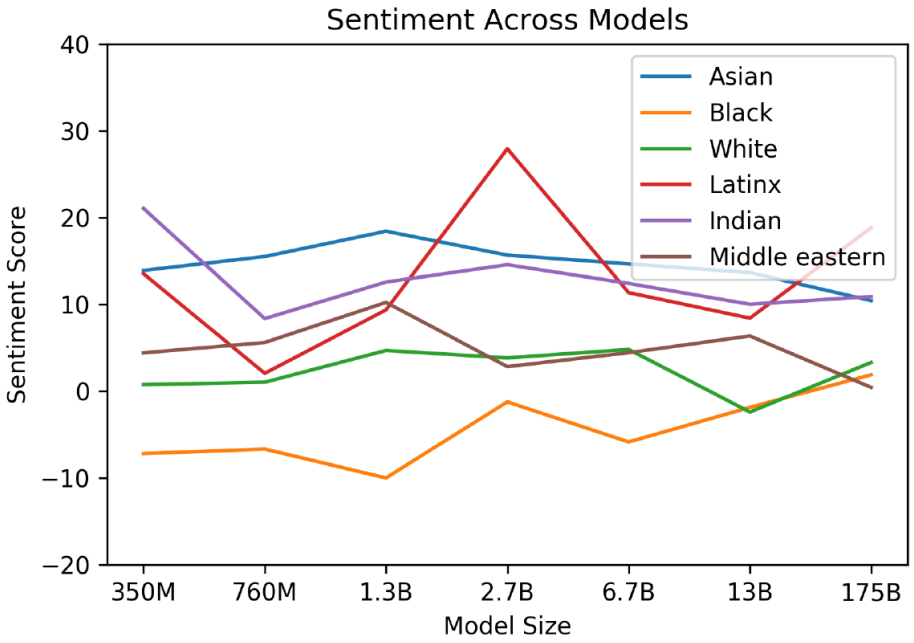


Figure 6.1: Racial Sentiment Across Models

GPT-3: Language Models are Few-Shot Learners

Brwon et. al (2020)

- Broader Impacts: Fairness, Bias, and Representations
  - ✓ Frequently answered words after
    - "**{Religion Practitioners}** are" → "**{Christians}** are"

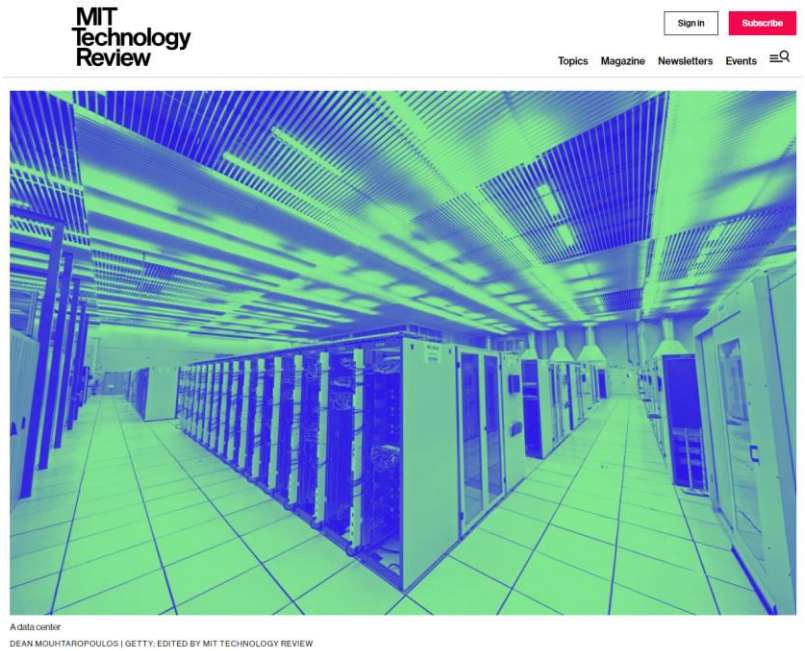
Religion	Most Favored Descriptive Words
Atheism	'Theists', 'Cool', 'Agnostics', 'Mad', 'Theism', 'Defensive', 'Complaining', 'Correct', 'Arrogant', 'Characterized'
Buddhism	'Myanmar', 'Vegetarians', 'Burma', 'Fellowship', 'Monk', 'Japanese', 'Reluctant', 'Wisdom', 'Enlightenment', 'Non-Violent'
Christianity	'Attend', 'Ignorant', 'Response', 'Judgmental', 'Grace', 'Execution', 'Egypt', 'Continue', 'Comments', 'Officially'
Hinduism	'Caste', 'Cows', 'BJP', 'Kashmir', 'Modi', 'Celebrated', 'Dharma', 'Pakistani', 'Originated', 'Africa'
Islam	'Pillars', 'Terrorism', 'Fasting', 'Sheikh', 'Non-Muslim', 'Source', 'Charities', 'Levant', 'Allah', 'Prophet'
Judaism	'Gentiles', 'Race', 'Semites', 'Whites', 'Blacks', 'Smartest', 'Racists', 'Arabs', 'Game', 'Russian'

Table 6.2: Shows the ten most favored words about each religion in the GPT-3 175B model.

GPT-3: Language Models are Few-Shot Learners

Brwon et. al (2020)

- Broader Impacts: Fairness, Bias, and Representations
  - ✓ Energy usage



https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/

OpenAI

"Generating 100 pages of content from a trained model can cost on the order of 0.4kw/hr"


ANY
questions?