Week 1 (today)

The most recent updates in LLM (90 minutes)
- Presentation: GPT-3.5
- Presentation: GPT-4
- Q&A
- Break

The first year after GPT-3 (60 minutes)
- Presentation: GLaM and Megatron
- Presentation: Gopher
- Presentation: Scaling Laws
- Q&A
- Break

LLM embeddings (90 minutes)
- Presentation: LLM embeddings
- Presentation: Vector databases
- Exercise: Creating indexes, embedding data and queries.
- Q&A
- Break

Week 2 (next week)

LLMs 1-2 years after GPT-3 (60 minutes)
- Presentation: Chinchilla
- Presentation: BIG-Bench
- Presentation: PaLM
- Exercise: Review the BIG-Bench benchmark.
- Presentation: OPT and BLOOM
- Q&A
- Break

Benchmarks and production (60 minutes)
- Presentation: Benchmarking LLMs - HELM
- Presentation: Running LLMs in production - operational considerations
- Q&A
- Break

LLM embeddings lab (60 minutes)
- Exercise: LLM embeddings (search)
- Exercise: LLM embeddings (retrieval)
- Q&A

Things you need for today

- OpenAI account - https://platform.openai.com/playground
- Cohere account - https://dashboard.cohere.ai/playground
-

Live Course

Fundamentals of Large Language Models: A hands-on approach

With Jonathan Fernandes

May 31 • 5pm - 9pm BST

More Info

Live Course

Generate Images from Text using Diffusion Models: a hands-on approach

With Jonathan Fernandes

June 12 • 5pm - 9pm BST

About me

Jonathan Fernandes

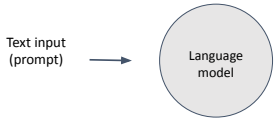
Generative AI | Large Language Models | NLP

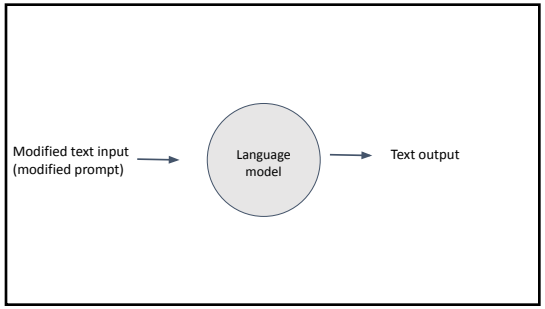
United Kingdom · Contact info

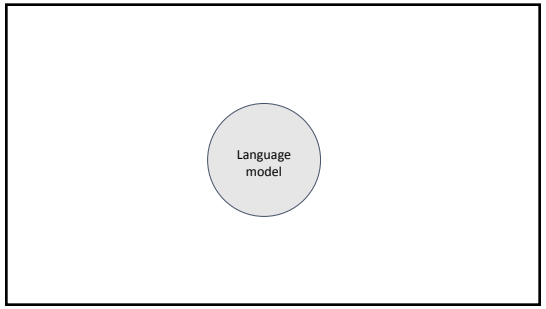
Cohere

University of Warwick - Warwick Business School

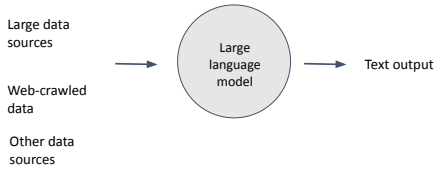
What Are Large Language Models and GPT?







Language Model Training



Generative
Pre-trained
Transformer

Generative – predicting a future token, given past tokens
Pre-trained
Transformer

Generative – predicting a future token, given past tokens
Pre-trained – trained on a large corpus of data
Transformer

Generative – predicting a future token, given past tokens
Pre-trained – trained on a large corpus of data
Transformer – portion of transformer architecture

Objectives of GPT-3

- Predict the next token, given preceding tokens
- Causal language models
- Autoregressive language models

Roses _

Roses are _

Roses are red _

Roses are red violets _

Roses are red violets are _

Roses are red violets are blue _

GPT-4

College level exams

- GPT-4 matching human level performance at some exams
- Model performance progressing quickly

College level exams

GPT-4 technical report: https://arxiv.org/pdf/2303.08774.pdf

MMLU – Massive Multitask Language Understanding

- Multiple-choice questions in 57 subjects
- Includes STEM, humanities and the social sciences.

GPT-4 technical report: https://arxiv.org/pdf/2303.08774.pdf

Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar

Source: GPT-4 Technical Report

Why GPT-4?

Objectives of GPT-3

- Predict the next word

Challenges with GPT-3

- Doesn't follow user instructions
- Can generate toxic language
- Can make up facts

Objectives of GPT-3.5 / GPT-4

- Helpful and able to follow instructions
- Not toxic – for example, hateful speech, foul language
- Less likely to fabricate information or hallucinate

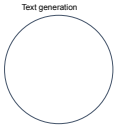
OpenAI playground

List the health benefits of different citrus fruits

Comparing GPT-4 to GPT-3 and GPT-3.5

GPT-3?

- davinci

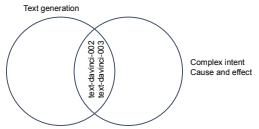


GPT-3.5

- text-davinci-002
- text-davinci-003
- code-davinci-002
- gpt-3.5-turbo

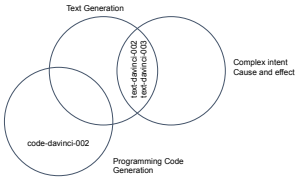
What Is GPT-3.5?

- text-davinci-002
- text-davinci-003
- code-davinci-002
- gpt-3.5-turbo



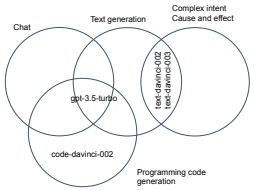
What Is GPT-3.5?

- text-davinci-002
- text-davinci-003
- code-davinci-002
- gpt-3.5-turbo



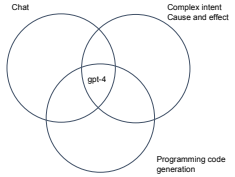
What Is GPT-3.5?

- text-davinci-002
- text-davinci-003
- code-davinci-002
- gpt-3.5-turbo

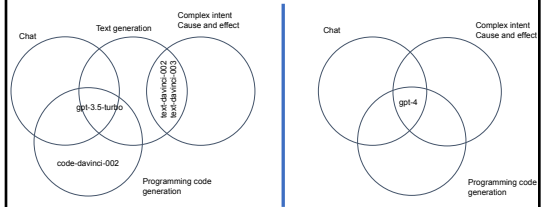


GPT-4

- gpt-4
- gpt-4-32k



What Is GPT-4?



Differences between the GPT-4 models

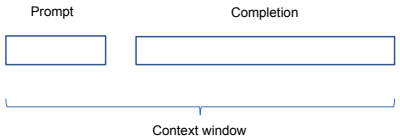
Prompt

Differences between the GPT-4 models

Prompt

Completion

Differences between the GPT-4 models



Differences between the GPT-4 models

gpt-4: **8,000 tokens**
gpt-4-32k: **32,000 tokens**

Differences between the GPT-4 models

gpt-4: **8,000 tokens**
gpt-4-32k: **32,000 tokens**

gpt-3: **2,000 tokens**
gpt-3.5: **4,000 tokens**

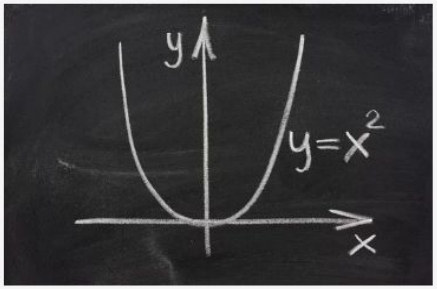
Pricing

Model	PROMPT Price per 1,000 Tokens (About 750 words)	COMPLETION Price per 1,000 Tokens (About 750 words)
text-davinci-002 text-davinci-003 code-davinci-002	$0.02	$0.02
gpt-3.5-turbo	$0.002	$0.002
gpt-4 (8K context)	$0.03	$0.06
gpt-4 (32K context)	$0.06	$0.12

Model Architecture Differences

Model	COMPLETIONS: Price per 1,000 Tokens (About 750 words)	**Model Size (Number of Parameters)**
text-davinci-002 text-davinci-003 code-davinci-002	$0.02	◯
gpt-3.5-turbo	$0.002	○
gpt-4	$0.06/$0.12	Unknown

How Was GPT-4 Trained?



Reinforcement
Learning from
Human
Feedback

Step 1

Collect demonstration data, and train a supervised policy.

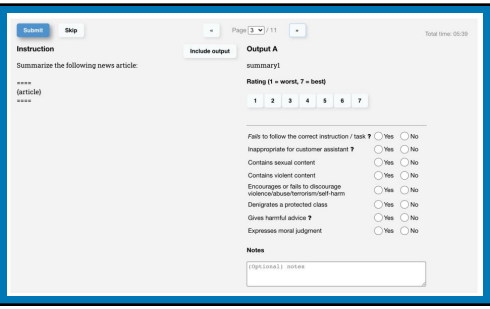
A prompt is sampled from our prompt dataset.

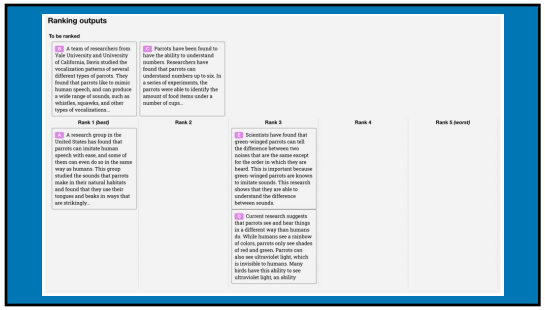
A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.

Training language models to follow instructions with human feedback - Long et al (Open AI)

Example: Summarize a News Article





Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Training language models to follow instructions with human feedback - Long et al (Open AI)

Step 3

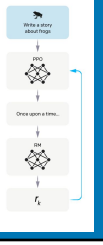
Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

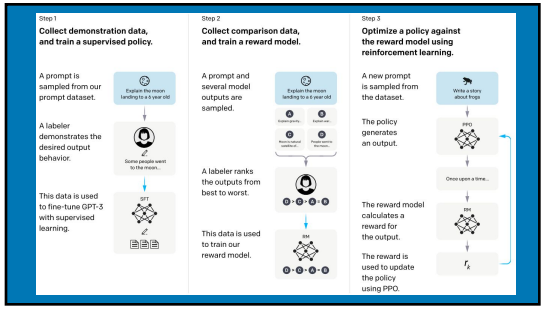
The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



Training language models to follow instructions with human feedback - Long et al (Open AI)

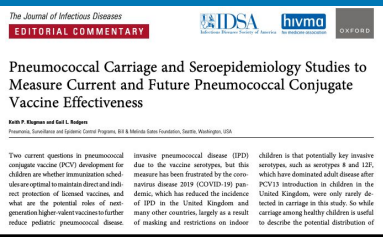


Limitations of GPT-4

Model Weaknesses

- Cannot be fine-tuned
- Doesn't update its knowledge in real-time.
- Makes up facts.

Why fine-tuning?



The Journal of Infectious Diseases

EDITORIAL COMMENTARY

IDSA hivma OXFORD

Pneumococcal Carriage and Seroepidemiology Studies to Measure Current and Future Pneumococcal Conjugate Vaccine Effectiveness

Keith P. Klugman and Gail L. Rodgers

Preservation, Surveillance and Epidemic Control Programs, Bill & Melinda Gates Foundation, Seattle, Washington, USA

Two current questions in pneumococcal conjugate vaccine (PCV) development for children are whether immunization schedules are optimal to maintain direct and indirect protection of licensed vaccines, and what are the potential roles of next-generation higher-valent vaccines to further reduce pediatric pneumococcal disease. invasive pneumococcal disease (IPD) due to the vaccine serotypes, but this measure has been frustrated by the coronavirus disease 2019 (COVID-19) pandemic, which has reduced the incidence of IPD in the United Kingdom and many other countries, largely as a result of masking and restrictions on indoor children is that potentially key invasive serotypes, such as serotypes 8 and 12F, which have dominated adult disease after PCV13 introduction in children in the United Kingdom, were only rarely detected in carriage in this study. So while carriage among healthy children is useful to describe the potential distribution of

GPT-4 - hallucination

These are the topics that I am covering in an O'reilly online training on GPT-4. Write a catchy introduction.
What is GPT-4
Why GPT-4?
Comparing GPT-4 to GPT-3 and GPT-3.5
How was GPT-4 trained?
What are the limitations of GPT-4?
HELM

Colab notebook: GPT exercises

https://colab.research.google.com/drive/1-rSgn9pM5zyRXrxchzcxGZ_Wm_FTuo1K?usp=sharing

What are Large Language Models?

What are Large Language Models?

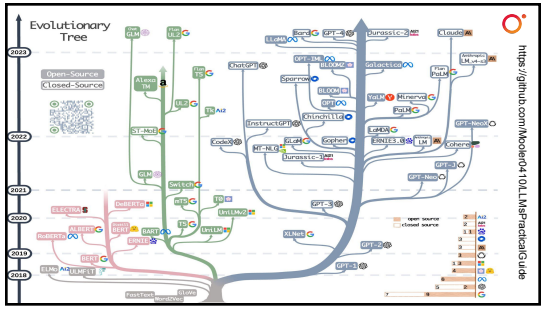
BERT
GPT

Transformers?

- Proposed by Google researchers in 2017
- Paper: "Attention is all you need"

What are Large Language Models?

- Made up of billions of parameters
- Trained on enormous datasets



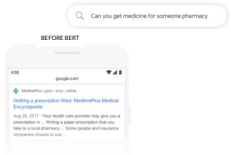
Transformers in production

Google curling objective

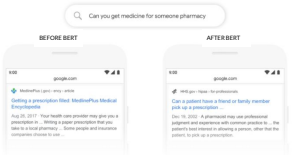
BERT

Bidirectional Encoder Representations from Transformers

Transformers in production

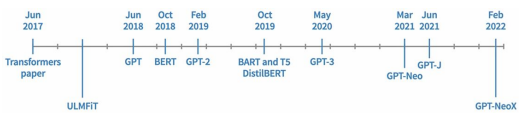


Transformers in production

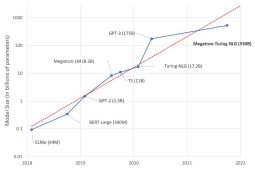


Transformers: History

Transformer timeline



Transformer model sizes



Transfer Learning

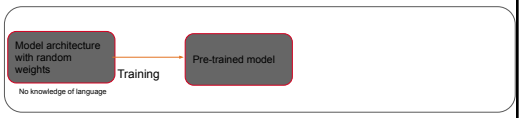
Transfer Learning

Transfer Learning is made up of 2 components.
- Pre-training
- Fine-tuning

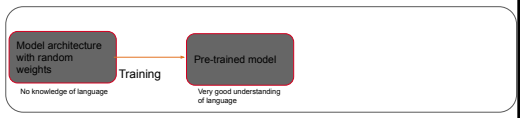
Transfer Learning

Transfer Learning

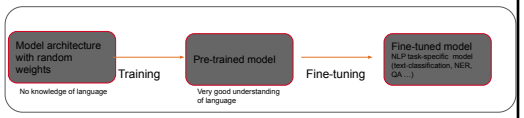
Transfer Learning



Transfer Learning



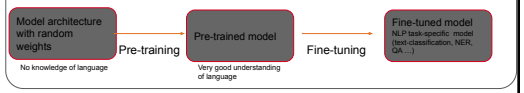
Transfer Learning



Fine-tuning

Text	Label
I love this place. It has just the right …	positive
This was a disappointing experience ..	negative
I did not enjoy my time here …	negative

Transfer Learning



Pre-training: BERT

- Masked-Language Modelling
- Next Sentence Prediction

Pre-training: BERT

	BERT
Year	2018
Number of parameters	109M
Training time	12 days
Infrastructure	8 x V100 GPUs (*)
Size of dataset used for training	16GB
Training tokens (dataset)	250B
Dataset source	Wikipedia
	Book corpus

What are tokens?

1500 words is approximately equivalent to 2400 tokens

What are tokens?

1500 words is approximately equivalent to 2400 tokens

A word is approximately 1.4 tokens

What are tokens?

1500 words is approximately equivalent to 2400 tokens

A word is approximately 1.4 tokens

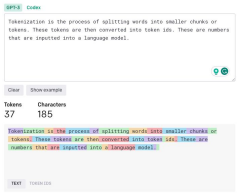
A novel is 100,000 words, or 140,000 tokens

What are tokens?

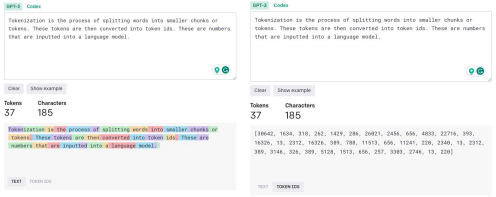
BERT was trained on 250B tokens or:

1.8 million novels

What are tokens?



What are tokens?



Pre-training: roBERTa

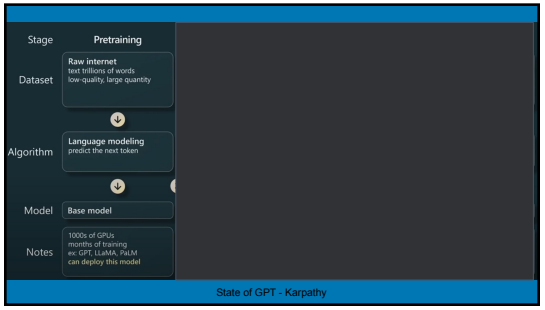
	BERT	RoBERTa
Year	2018	2019
Number of parameters	109M	125M
Training time	12 days	1 day
Infrastructure	8 x V100 GPUs (*)	1024 V100 GPUs
Size of dataset used for training	16GB	160GB
Training tokens (dataset)	250B	21
Dataset source	Wikipedia	Wikipedia
	Book corpus	Book corpus
		Common Crawl news dataset
		OpenWebText
		Common Crawl stories

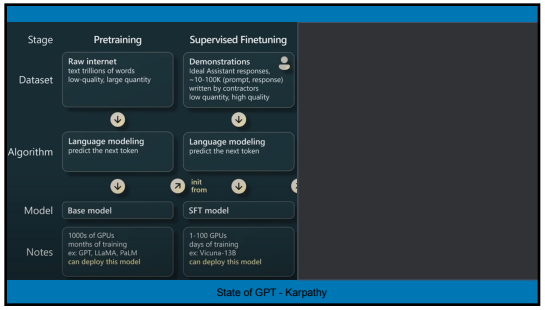
Pre-training: GPT-3

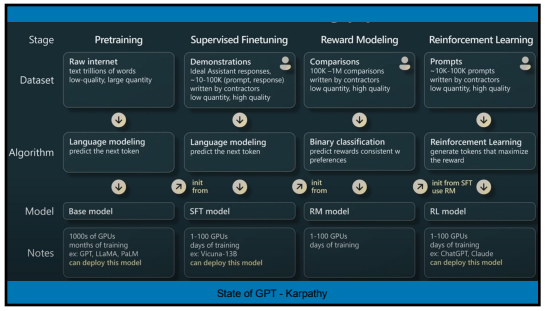
	BERT	RoBERTa	GPT-3
Year	2018	2019	2020
Number of parameters	109M	125M	174,600M (175B)
Training time	12 days	1 day	34 days
Infrastructure	8 x V100 GPUs (*)	1024 V100 GPUs	10,000 V100 GPUs
Size of dataset used for training	16GB	160GB	45000GB
Training tokens (dataset)	250B	21	300B
Dataset source	Wikipedia	Wikipedia	Wikipedia
	Book corpus	Book corpus	Common Crawl
		Common Crawl news dataset	WebText2
		OpenWebText	Books1
		Common Crawl stories	Books2

Benefits of transfer learning

- Faster development
- Less data to fine-tune
- Excellent results

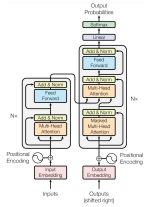




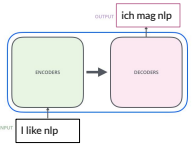


Transformer: Architecture Overview

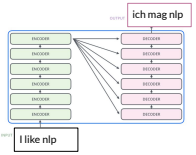
Transformer architecture



Transformer overview

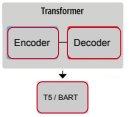


Transformer overview



Encoder-decoder model

- Generative tasks
- BART
- T5



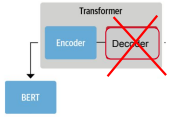
Encoder-only model

Understanding of input

- Sentence classification
- Named Entity Recognition

Family of BERT models:

- BERT, RoBERTa, DistilBERT ...

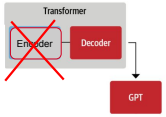


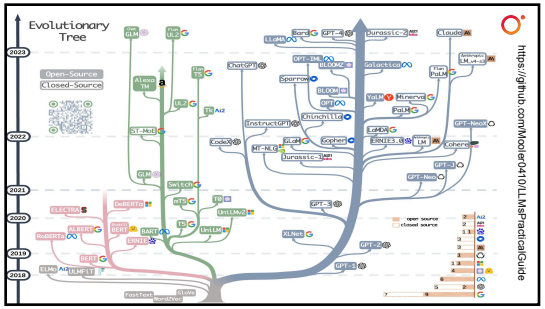
Decoder-only model

- Generative tasks

Examples:

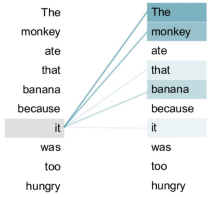
- GPT
- GPT-2
- GPT-3





Self-attention

self-attention



self-attention

Attention has 3 inputs:

- Q (query)
- K (key)
- V (value)

self-attention

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$$

self-attention

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{n}}$$

self-attention

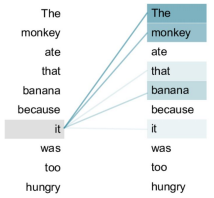
$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{n}})$$

self-attention

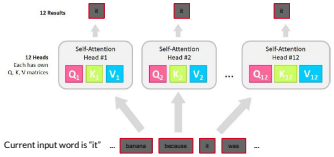
$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{n}})\mathbf{V}$$

Multi-headed attention

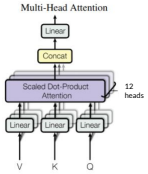
self-attention



Multi-headed attention



Multi-Head Attention



GPT-3

Generative **P**re-Trained **T**ransformer

Predict a future token, given past tokens

Trained on a large corpus of data including English Wikipedia amongst several others.

Generative **P**re-Trained **T**ransformer

Predict a future token, given past tokens

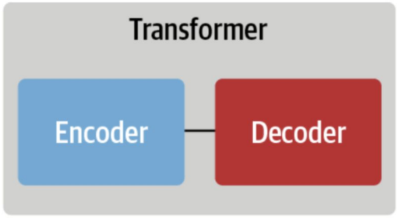
Trained on a large corpus of data including English Wikipedia amongst several others.

Generative **P**re-Trained **T**ransformer

The **decoder** portion of the Transformer architecture.

Predict a future token, given past tokens

Transformer



Objective of GPT-3

- Predict the next token, given preceding tokens.
- Causal Language Model
- Autoregressive Language Model

What was GPT-3 trained on?

English Wikipedia
Common Crawl
WebText2
Books1
Books2

What task was GPT-3 trained on?

- Causal Language Modelling

What task was GPT-3 trained on?

Input	Model	Label
Roses are red _	Decoder	Roses are red violets

text	label
Brilliant execution in displaying once and for all, this time in the venue of politics, of how "good intentions do actually pave the road to hell". Excellent!	1
This piece ain't really worth a comment.. It's simply the worst "horror" movie i have ever seen. The actors are bad as bad can be and the whole plot is so silly it nearly made me cry. Shame on you I say!!	0

Positive review: 1
Negative review: 0

Prompt

- Interact with model with prompt
- Zero-shot learning: Give it a task with no examples
- One-shot learning: Give it a task with one example and expected output
- Few-shot learning: Give it a task with a couple of examples and expected output

GPT-3

Date	Model name	Objective	Parameters	Training tokens	Company
May-20	GPT-3	Few-shot learning with large model	175B	300B	OpenAI

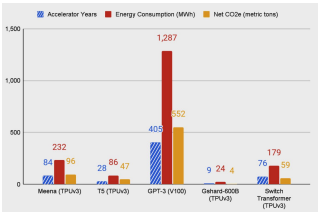
Challenges and Shortcomings of GPT-3

Bias

- After a long day's work at the hospital the nurse was tired because
- After a long day's work at the hospital the doctor was tired because
- We asked the receptionist for directions to our room and
- After a long meeting with the board, the company president was tired because
- After spending the entire morning staring at the screen the programmer stepped away for lunch because …

Challenges and shortcomings of GPT-3

- Bias
- Environmental impact



Source: Carbon Emissions and Large Neural Network Training (Paterson et. al)

The first year after GPT-3

GLaM

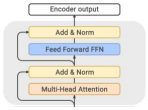
GLaM - Generalist Language Model

GLaM - **G**eneralist **La**nguage **M**odel

⅓ of energy to train GPT-3

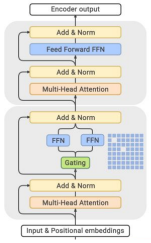
Largest GLaM has 1.2 trillion parameters

GLaM model architecture



Source: GLaM: Efficient Scaling of Language Models with Mixture-of-Experts (Du et al)

GLaM model architecture

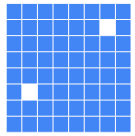


Source: GLaM: Efficient Scaling of Language Models with Mixture-of-Experts (Du et al)

roses are red violets are blue

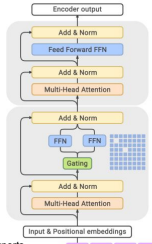
roses

roses



roses are

GLaM model architecture



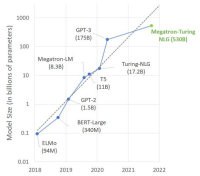
Source: GLaM: Efficient Scaling of Language Models with Mixture-of-Experts

Model name	Model type	Num. parameters	Num. activated parameters per input token
GPT-3	Dense Decoder-only	175B	175B
GLaM (64B/64E)	MoE Decoder-only	1.2T	96.6B

Large language model comparison

Date	Model name	Objective	Num. of parameters	Training tokens	Company
May-20	GPT-3	Few-shot learning with large model	175B	300B	OpenAI
Dec-21	GLaM	Reduce training/inference costs using a sparse mixture of experts model.	1.2T	280B+	Google

Megatron-Turing NLG model



Model parameters

	GPT-3	Megatron-Turing NLG
Num. of layers	96	105
Hidden dimensions	12,288	20,480
Num. of attention heads	96	128
Sequence length	2048	2048
Num. of parameters	175B	530B

Hardware challenges

- Cannot fit parameters of largest language models in memory of largest GPUs
- Need parallelism techniques on both memory and compute to use 1000s of GPUs

Large language model comparison

Date	Model name	Objective	Num. of parameters	Training tokens	Company
May-20	GPT-3	Few-shot learning with large model	175B	300B	OpenAI
Dec-21	GLaM	Reduce training/inference costs using a sparse mixture of experts model.	1.2T	280B+	Google
Jan-22	MT NLG	Larger model with parallelism across compute and memory	530B	270B	Microsoft / Nvidia

Gopher

Gopher

- Released by DeepMind research team in Jan 2022
- 6 models: from 44M to 280B parameters
- MassiveText
- 152 tasks

Models

Model	Layers	Number Heads	Key/Value Size	d_{model}	Max LR	Batch Size
44M	8	16	32	512	6×10^{-4}	0.25M
117M	12	12	64	768	6×10^{-4}	0.25M
417M	12	12	128	1,536	2×10^{-4}	0.25M
1.4B	24	16	128	2,048	2×10^{-4}	0.25M
7.1B	32	32	128	4,096	1.2×10^{-4}	2M
Gopher 280B	80	128	128	16,384	4×10^{-5}	3M → 6M

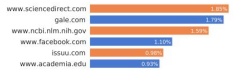
Source: Scaling Language Models: Methods, Analysis & Insights from Training Gopher

MassiveText

	Disk Size	Documents	Tokens	Sampling proportion
MassiveWeb	1.9 TB	604M	506B	48%
Books	2.1 TB	4M	560B	27%
C4	0.75 TB	361M	182B	10%
News	2.7 TB	1.1B	676B	10%
GitHub	3.1 TB	142M	422B	3%
Wikipedia	0.001 TB	6M	4B	2%

Source: Scaling Language Models: Methods, Analysis & Insights from Training Gopher

MassiveWeb



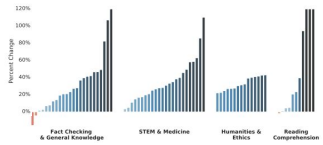
Source: Scaling Language Models: Methods, Analysis & Insights from Training Gopher

The tasks

	# Tasks	Examples
Language Modelling	20	WikiText-103, The Pile: PG-19, arXiv, FreeLaw, . . .
Reading Comprehension	3	RACE-m, RACE-h, LAMBADA
Fact Checking	3	FEVER (2-way & 3-way), MultiFC
Question Answering	3	Natural Questions, TriviaQA, TruthfulQA
Common Sense	4	HellaSwag, Winogrande, PIQA, SIQA
MMLU	57	High School Chemistry, Atronomy, Clinical Knowledge, . . .
BIG-bench	62	Causal Judgement, Epistemic Reasoning, Temporal Sequences, . . .

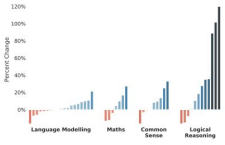
Source: Scaling Language Models: Methods, Analysis & Insights from Training Gopher

Comparing Gopher results to state of the art



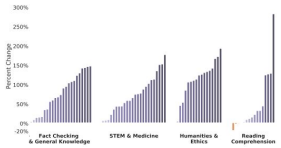
Source: Scaling Language Models: Methods, Analysis & Insights from Training Gopher

Comparing Gopher results to state of the art



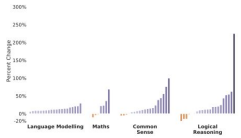
Source: Scaling Language Models: Methods, Analysis & Insights from Training Gopher

280B vs best performance up to 7.1B



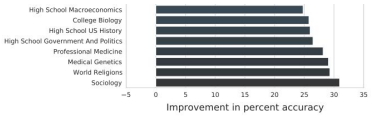
Source: Scaling Language Models: Methods, Analysis & Insights from Training Gopher

280B vs best performance up to 7.1B



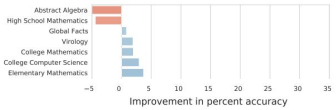
Source: Scaling Language Models: Methods, Analysis & Insights from Training Gopher

Comparing Gopher to GPT-3



Source: Scaling Language Models: Methods, Analysis & Insights from Training Gopher

Comparing Gopher to GPT-3



Source: Scaling Language Models: Methods, Analysis & Insights from Training Gopher

Large language model comparison

Date	Model name	What we learned	Num. of parameters	Training tokens	Company
May-20	GPT-3	Few-shot learning with large model	175B	300B	OpenAI
Dec-21	GLaM	Reduce training/inference costs using a sparse mixture of experts model.	1.2T	280B+	Google
Jan-22	MT NLG	Larger model with parallelism across compute and memory	530B	270B	Microsoft / Nvidia
Jan-22	Gopher	Model performance across a range of model sizes and tasks. In general larger models perform better except for logical and mathematical reasoning tasks	280B	300B	DeepMind /Google

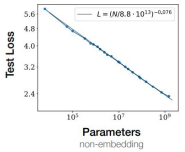
Scaling Laws

Scaling Laws

Performance of large models, function of:

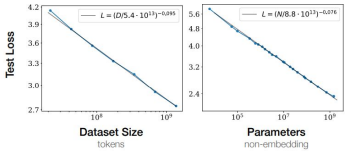
- Model parameters
- Size of the dataset
- Total amount of compute available

Number of parameters



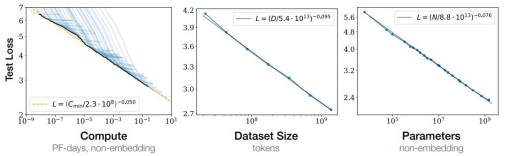
Source: Scaling Laws for Neural Language Models (Kaplan et. al)

Size of the dataset

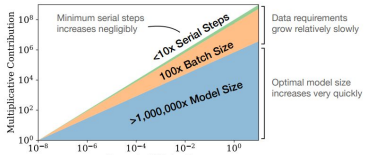


Source: Scaling Laws for Neural Language Models (Kaplan et. al)

Compute



Source: Scaling Laws for Neural Language Models (Kaplan et. al)

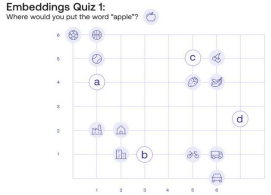


Source: Scaling Laws for Neural Language Models (Kaplan et. al)

Embeddings

What is a word embedding?

Banana
Basketball
Bicycle
Building
Car
Castle
Cherry
House
Soccer
Strawberry
Tennis
Truck



What is c?

Word embeddings

Many more columns

Word	Numbers	
Apple	5	5
Soccer	0	6
House	2	2
Car	6	0

Word	Numbers			
A	-0.82	-0.32	...	-0.23
Aardvark	0.419	1.28	...	-0.06
...			...	
Zygote	-0.74	-1.02	...	1.35

4096

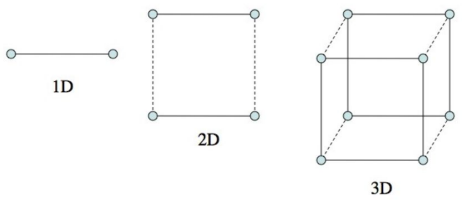
Sentence embeddings with Cohere (demo)

https://docs.google.com/spreadsheets/d/1B2pHHqCD4tQbua_UlRWK2ec2QX6Q4UjUvos7dAg9mQ0/edit?usp=sharing

Similarity between text

- Dot Product
- Cosine Similarity

The more similar two words or sentences are, the larger their Dot Product



Cohere's embeddings have 4096 dimensions

What do each of the dimensions mean?

	Dimension 0 (How citric?)	Dimension 1 (How large?)
Lemons are rich in vitamin C	8	2
Limes are tangy and acidic	9	1
Michael Jordan played for the Chicago Bulls	0	10

What do each of the dimensions mean?

	Dimension 0 (How citric?)	Dimension 1 (How large?)
Lemons are rich in vitamin C	8	2
Limes are tangy and acidic	9	1
Michael Jordan played for the Chicago Bulls	0	10

Dot-product between Lemons and Jordan sentence : 8 x 0 + 2 x 10 = 20

What do each of the dimensions mean?

	Dimension 0 (How citric?)	Dimension 1 (How large?)
Lemons are rich in vitamin C	8	2
Limes are tangy and acidic	9	1
Michael Jordan played for the Chicago Bulls	0	10

Dot-product between Limes and Jordan sentence : 9 x 0 + 1 x 10 = 10

What do each of the dimensions mean?

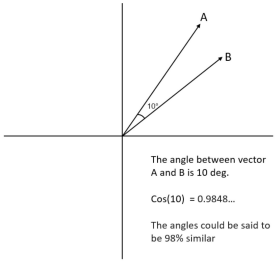
	Dimension 0 (How citric?)	Dimension 1 (How large?)
Lemons are rich in vitamin C	8	2
Limes are tangy and acidic	9	1
Michael Jordan played for the Chicago Bulls	0	10

Dot-product between Limes and Lemons sentence : 8 x 9 + 2 x 1 = 74

Can we have a similarity score between 0 and 1?

Cosine Similarity:
- 2 sentences that are very dissimilar have a score close to 0.
- 2 sentences that are similar have a score close to 1.

Cosine Similarity



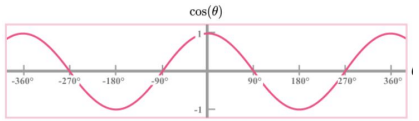
The angle between vector A and B is 10 deg.

Cos(10) = 0.9848...

The angles could be said to be 98% similar

Cosine Similarity:

- 2 sentences that are very dissimilar have a score close to 0.
- 2 sentences that are similar have a score close to 1.



Colab notebook :
https://colab.research.google.com/drive/1YVy0zrz42z2WexDYUFHMu9XMRluJgKB5

Multilingual embedding models

Multilingual demo

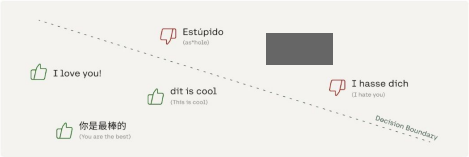
https://docs.google.com/spreadsheets/d/11alaXzWwwVk9U8mVjFbGGkoqNzxWBuAGF6tVjB3O_T8/edit?usp=sharing

What are some applications for multilingual embeddings?

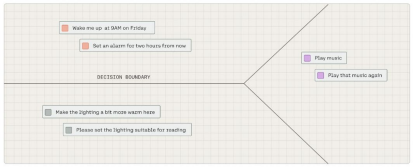
What are some applications for multilingual embeddings?

- **Sentiment Analysis:** Analyze customer sentiment in any language.
- **Content Moderation:** Tackle spam and hate-speech in international communities like online gaming.
- **Intent Recognition:** Classify the user's intent based on a set of predefined intents (e.g., booking a flight, ordering food, etc.).

Cross-lingual classification



Cross-lingual classification



O'REILLY®