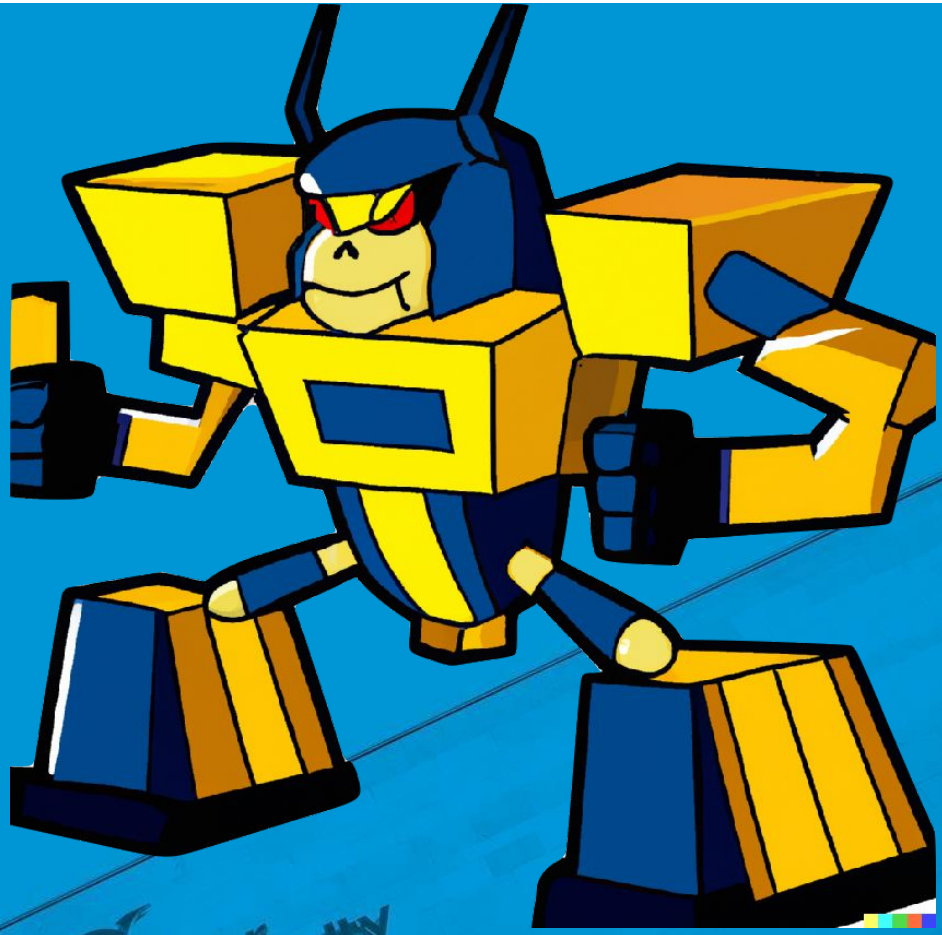
Foundational Models

Muskaan Singh
Lecturer in Data Analytics

CARL, ISRC, SCEIS, Ulster University



Source Transformer cartoon (DALL-E)

“The limits of my language mean the limits of my world.”

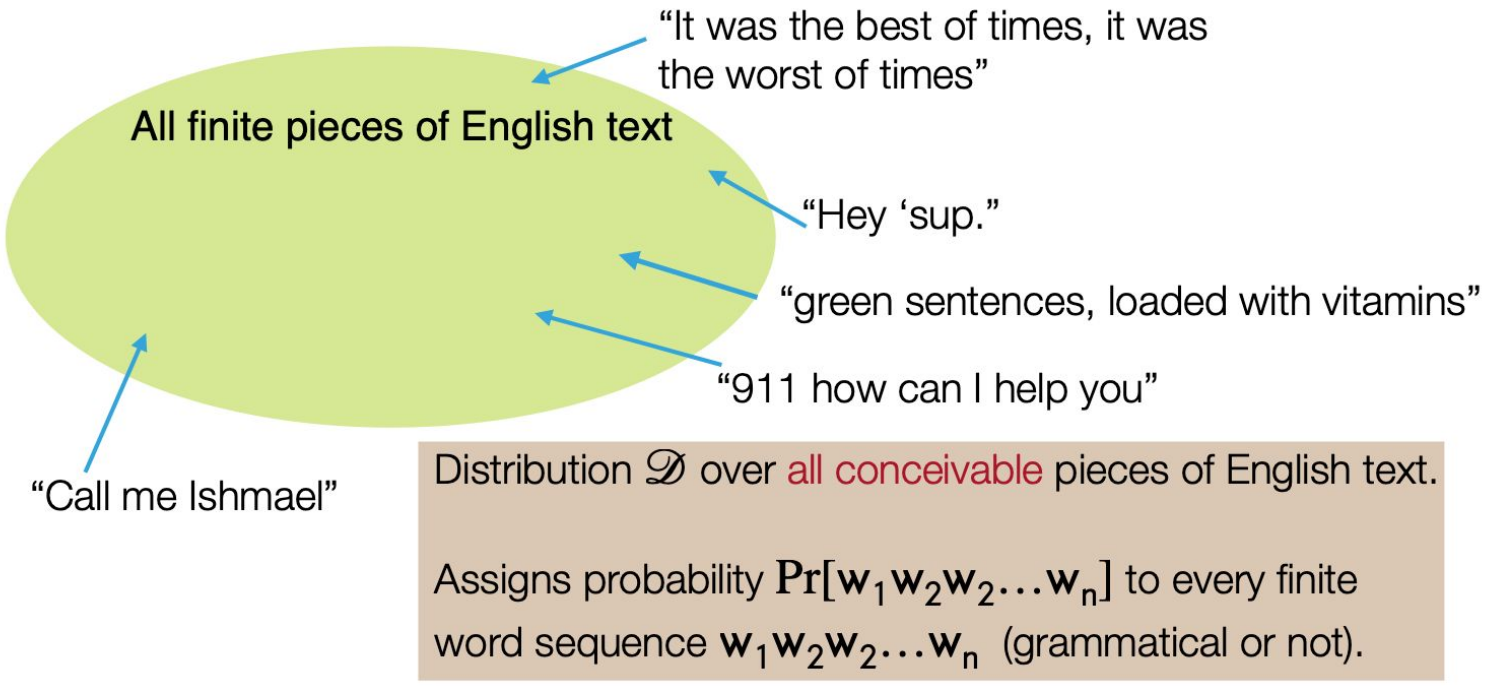
—Ludwig Wittgenstein

An Old Analogy

What are language models (LMs)?

Language Model (LM)

- A probabilistic model that assigns a probability *P*[*w*1, *w*2, ..., *wn*] to every finite sequence *w*1, ..., *wn* (grammatical or not) to predict the probabilities of future tokens.



GPT-3 works on same principle but with very large neural network model of 175-billion parameters!

Source: COS324

A human wants to submerge himself in water, what should he use?

Humans *select* options

(a) Coffee cup

(b) Whirlpool bath

(c) Cup

(d) Puddle

Language Models assign probability to *every possible string*

OK = right concept, wrong surface form

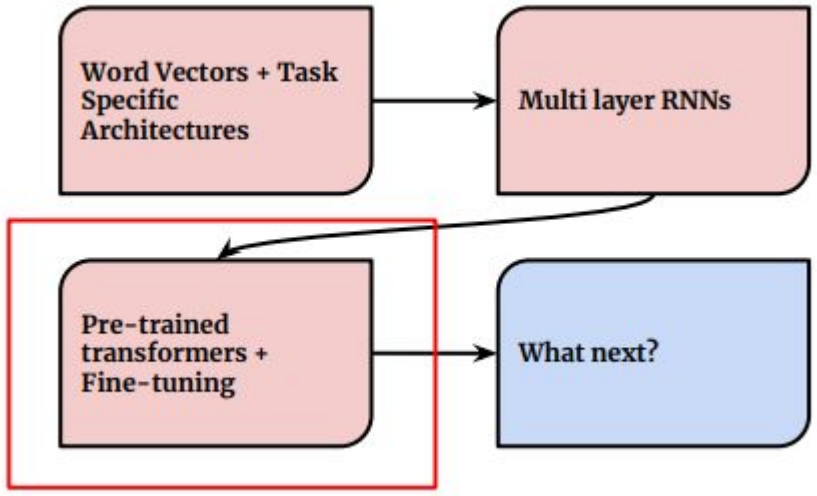
Competes for probability mass

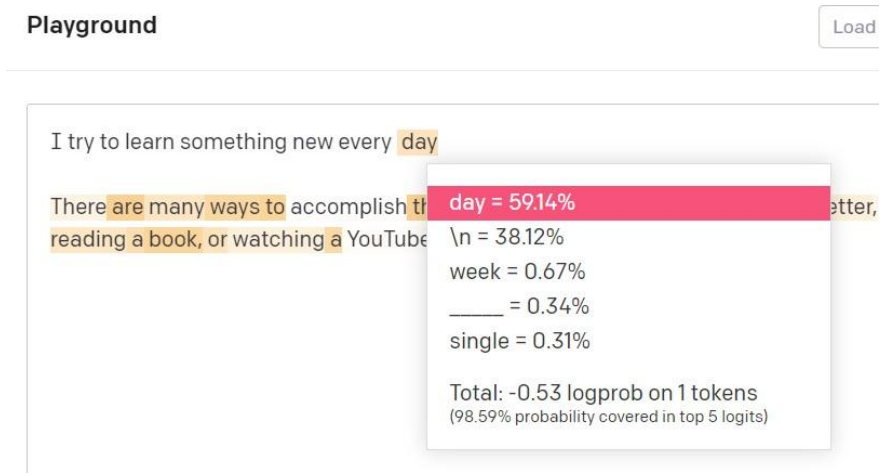
Generic output always assigned high probability

Every correct string is assigned lower score than expected

Language Models: Stages

- Statistical Language Models (Markov Decision)
- Neural Language Models (RNN, CNN)
- Pretrained language Models /Foundation Models (ELMo, GPT-2, BART)
- Large Language Models- (GPT-3, ChatGPT, GPT-4)



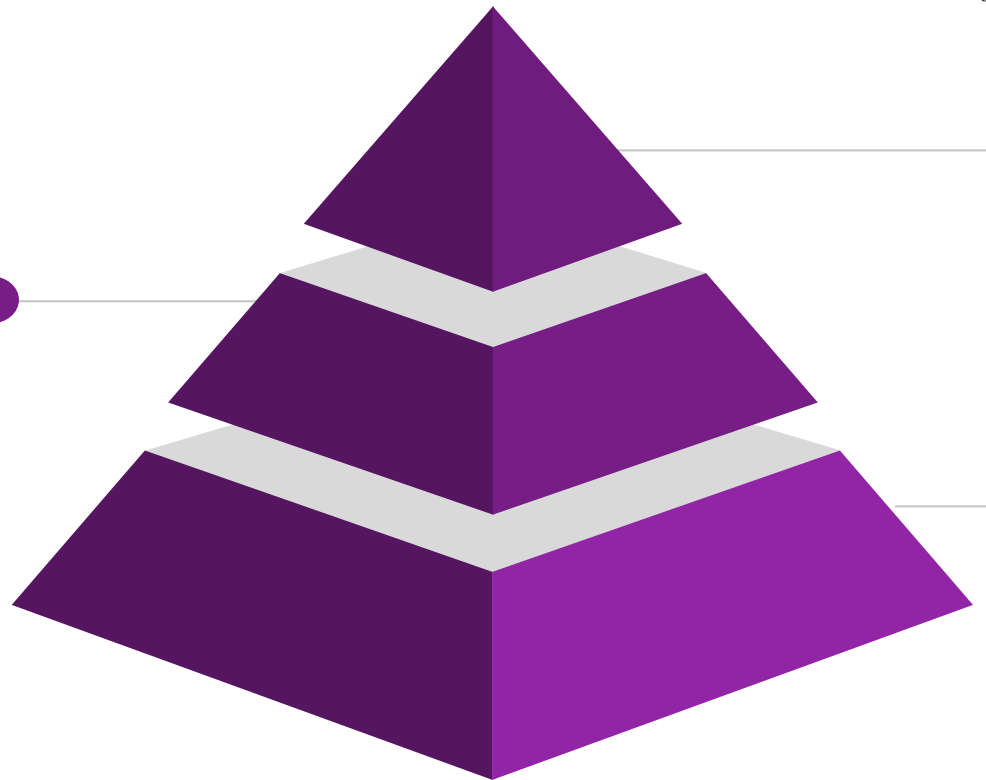


Foundation Model

On the Opportunities and Risks of Foundation Models

Rishi Bommasani* Drew A. Hudson Ehsan Adeli Russ Altman Simran Arora
Sydney von Arx Michael S. Bernstein Jeannette Bohg Antoine Bosselut Emma Brunskill
Erik Brynjolfsson Shyamal Buch Dallas Card Rodrigo Castellon Niladri Chatterji
Annie Chen Kathleen Creel Jared Quincy Davis Dorottya Demszky Chris Donahue
Moussa Doumbouya Esin Durmus Stefano Ermon John Etchemendy Kawin Ethayarajh
Li Fei-Fei Chelsea Finn Trevor Gale Lauren Gillespie Karan Goel Noah Goodman
Shelby Grossman Neel Guha Tatsunori Hashimoto Peter Henderson John Hewitt
Daniel E. Ho Jenny Hong Kyle Hsu Jing Huang Thomas Icard Saahil Jain
Dan Jurafsky Pratyusha Kalluri Siddharth Karamcheti Geoff Keeling Fereshte Khani
Omar Khattab Pang Wei Koh Mark Krass Ranjay Krishna Rohith Kuditipudi
Ananya Kumar Faisal Ladhak Mina Lee Tony Lee Jure Leskovec Isabelle Levent
Xiang Lisa Li Xuechen Li Tengyu Ma Ali Malik Christopher D. Manning
Suvir Mirchandani Eric Mitchell Zanele Munyikwa Suraj Nair Avanika Narayan
Deepak Narayanan Ben Newman Allen Nie Juan Carlos Niebles Hamed Nilforoshan
Julian Nyarko Giray Ogut Laurel Orr Isabel Papadimitriou Joon Sung Park Chris Piech
Eva Portelance Christopher Potts Aditi Raghunathan Rob Reich Hongyu Ren
Frieda Rong Yusuf Roohani Camilo Ruiz Jack Ryan Christopher Ré Dorsa Sadigh
Shiori Sagawa Keshav Santhanam Andy Shih Krishnan Srinivasan Alex Tamkin
Rohan Taori Armin W. Thomas Florian Tramèr Rose E. Wang William Wang Bohan Wu
Jiajun Wu Yuhuai Wu Sang Michael Xie Michihiro Yasunaga Jiaxuan You Matei Zaharia
Michael Zhang Tianyi Zhang Xikun Zhang Yuhui Zhang Lucia Zheng Kaitlyn Zhou
Percy Liang*[1]

Center for Research on Foundation Models (CRFM)
Stanford Institute for Human-Centered Artificial Intelligence (HAI)
Stanford University



1 **Fine-Tune Model for Multiple Downstreaming Task**

Model that can be adapted to a wide range of downstreaming task.

2 **Evaluate Model Performance**

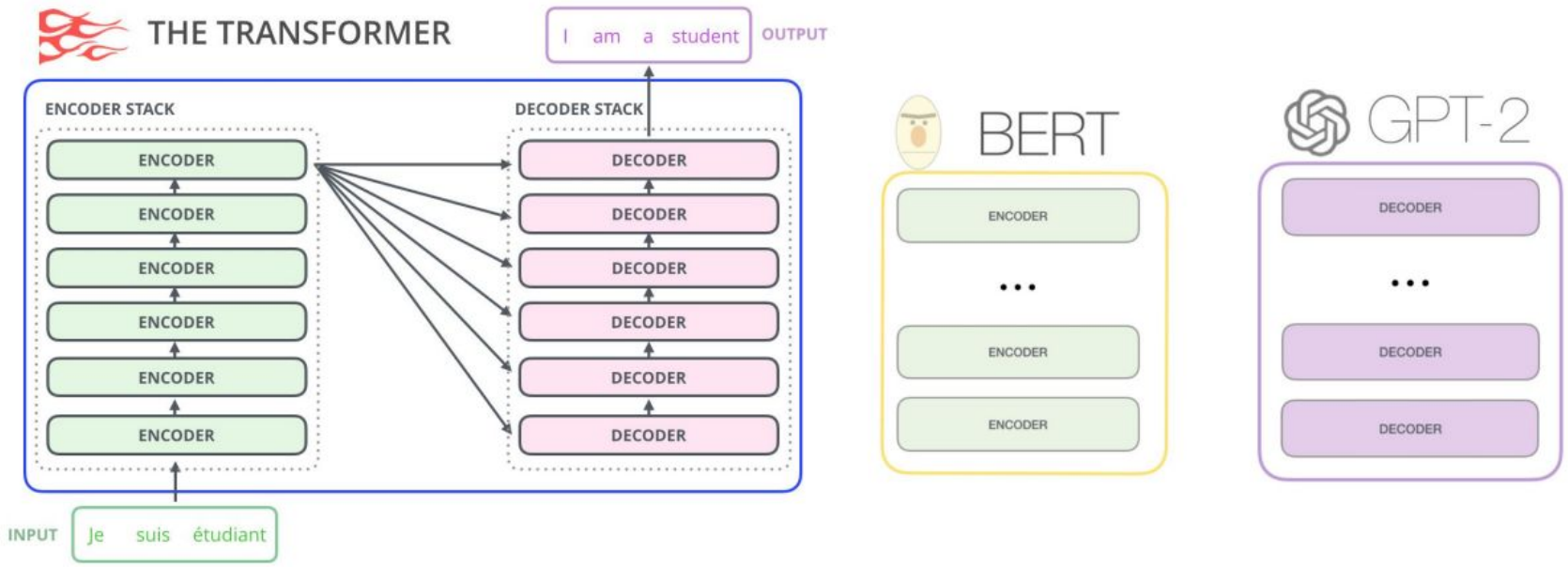
Evaluate the trained mode using Quantitative and Qualitative evaluation methods (task-dependent)

3 **Train Foundation Model One Time**

Gather the data at large scale and train the model one time.

Foundation/ Language Models: Architectures

- Encoder-only Models (BERT, RoBERTa, ELECTRA)
- Encoder-Decoder Models (T5, BART)
- Decoder-only Models (GPT-x models)



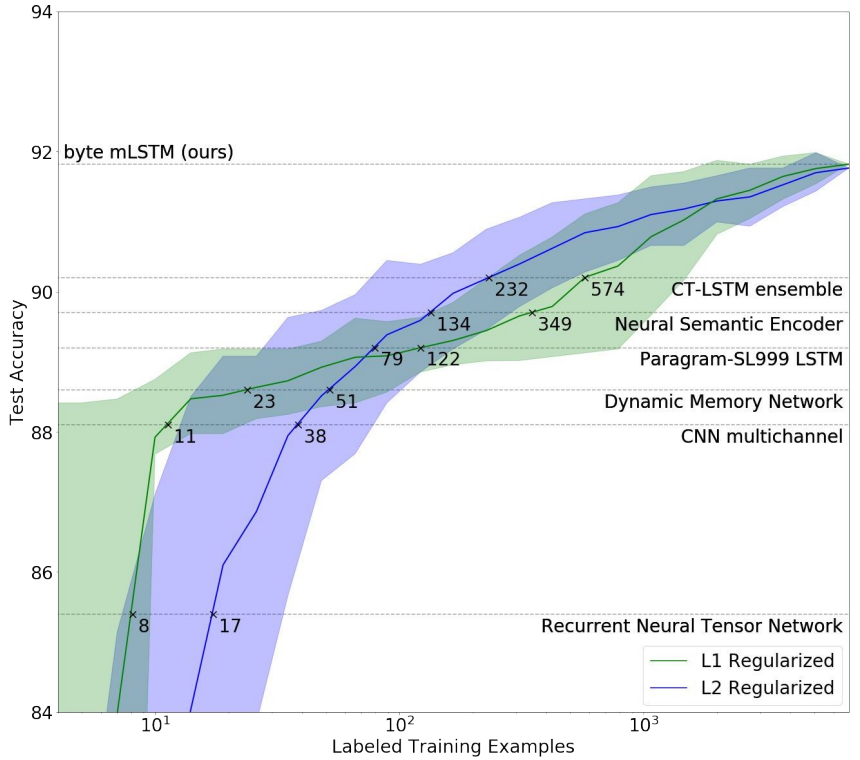
Source [(jalammar.github.io)](jalammar.github.io)

Foundational Models/ PLMs?

Let’s rewind?

A Very Brief Context: PLMs

- 2017: Introduction to Transformer with Attention Is All You Need, ELMo,Tensorflow Framework
- 2018: Open-AI GPT-1, BERT
  - Bidirectionality
  - Transformers
- 2019: Open-AI GPT-2, Baidu ERNIE, XLNET, facebook RoBERT, SpanBERT, ALBERT, Amazon Deep Composer
- 2020: GPT-3, AI21 WordTune, NVIDIA Megatron
- 2021: Google T5, Open AI DALL-E, Google T-X, LaMDA, Github Copilot
- 2022: Instruct GPT, PaLM, BLOOM,ChatGPT
- 2023: BARD, LLaMA, GPT-4



Source: https://openai.com/blog/unsupervised-sentiment-neuron/

Attention is all you Need

- Ground-breaking architecture that set SOTA on first translation and later all other NLP tasks.
- For simplicity, can just look at the encoder

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

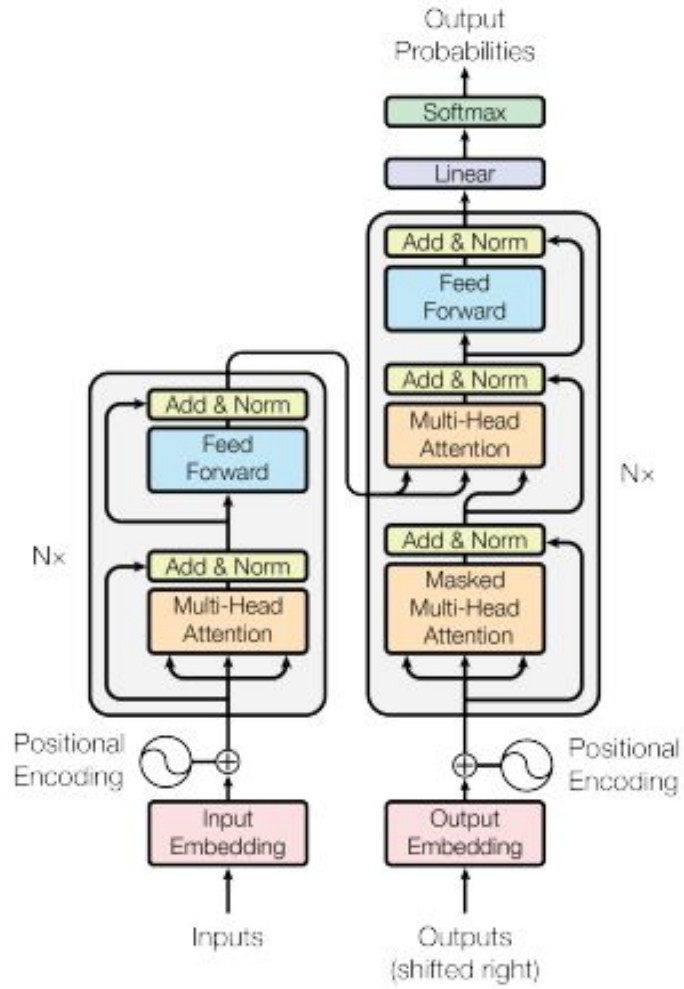
Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

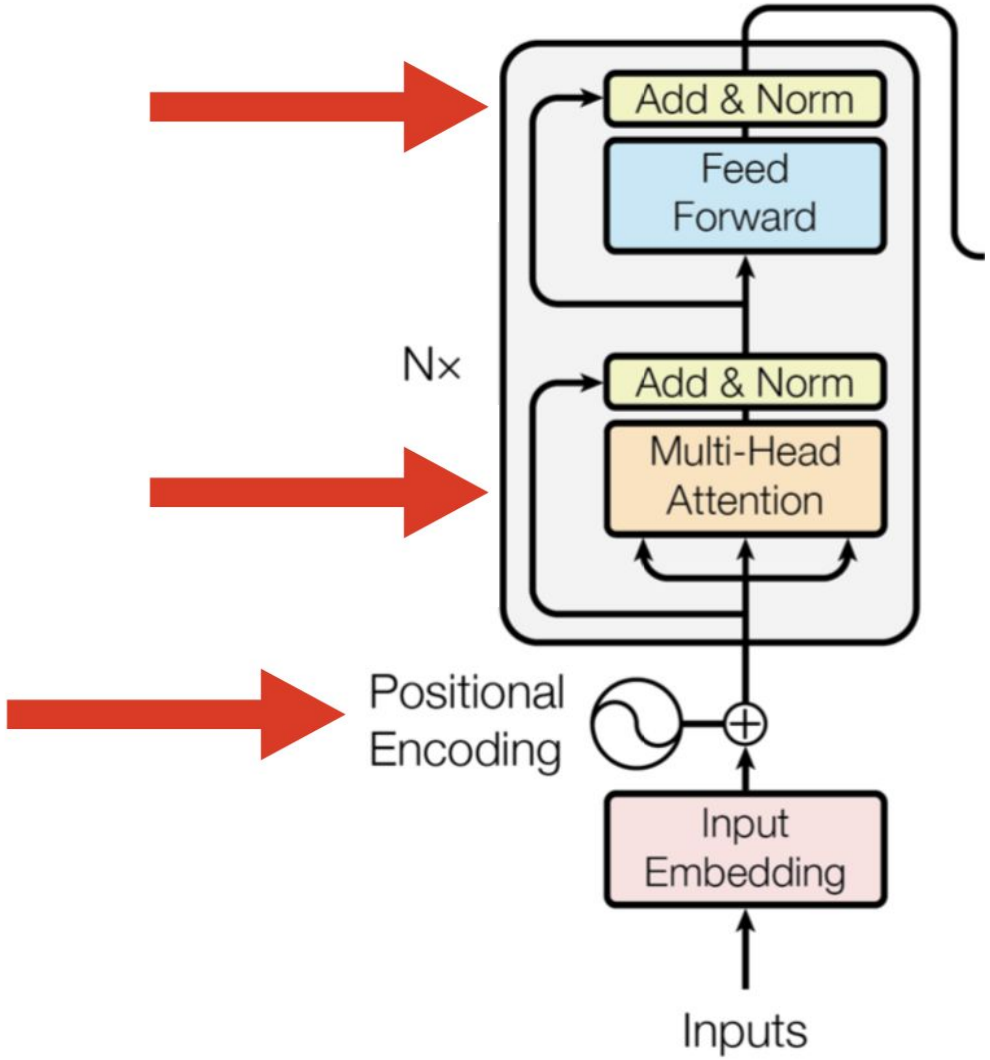


source: Vaswani et al 2017

Attention is all you Need

The Components

- Self Attention
- Positional Encoding
- Layer Normalization



Self Attention

- Input: sequence of tensors $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_t$
- Output: sequence of tensors, each one a weighted sum of the input sequence

$$\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_t$$

$$\mathbf{y}_i = \sum_j w_{ij} \mathbf{x}_j$$

- weight is just a dot product
- make it sum to 1

$$w'_{ij} = \mathbf{x}_i{}^T$$

$$w_{ij} = \frac{\exp w'_{ij}}{\sum_j \exp w'_{ij}}$$

Source *http://www.peterbloem.nl/blog/transformers*

No learned weights ⟶ Let's learn some weights!

Order of the sequence does not affect result of computation.

Multi-Head Self Attention

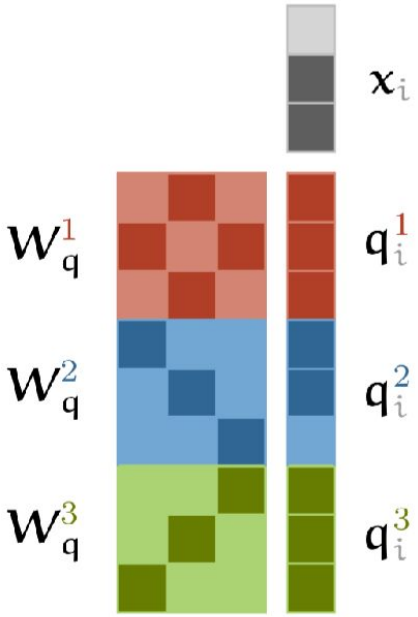
- Every input vector x_i is used in 3 ways:
  - Query
  - Key
  - Value
- We can process each input vector to fulfill the three roles with matrix multiplication
- Learning the matrices → learning attention
- Multiple “heads of attention” means different sets of W_q,W_k,W_v matrix (We learn them as different sets of weighted matrices simultaneously)
- But we implement them as a single matrix

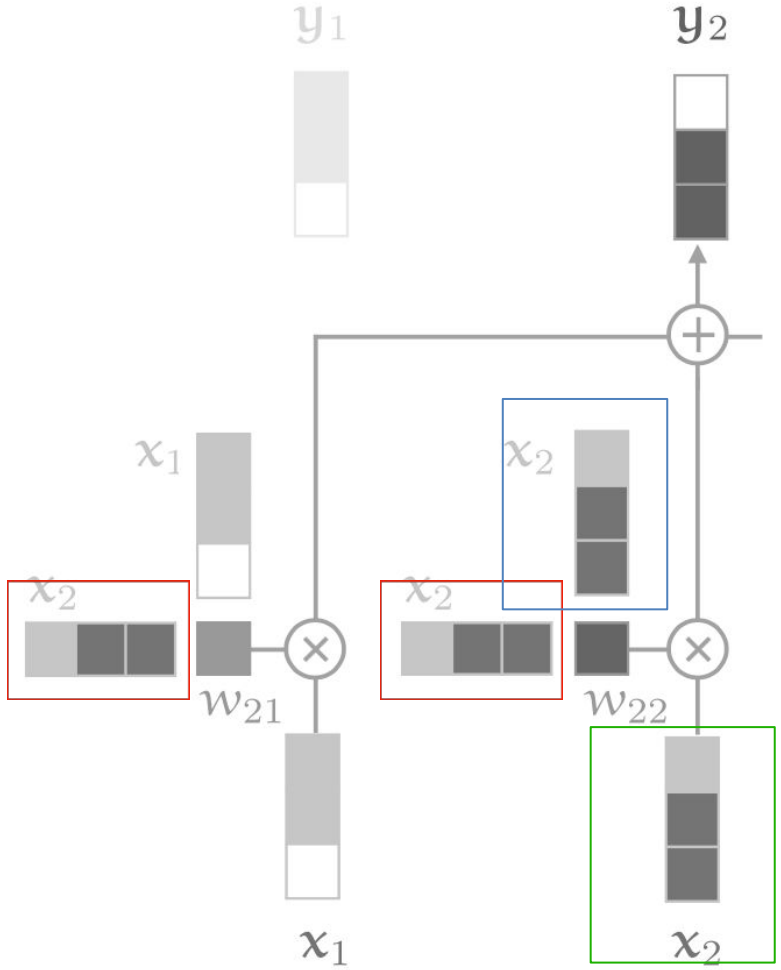
$$\mathbf{q}_i = \boldsymbol{W}_q \mathbf{x}_i \qquad \mathbf{k}_i = \boldsymbol{W}_k \mathbf{x}_i \qquad \boldsymbol{v}_i = \boldsymbol{W}_v \mathbf{x}_i$$

$$w'_{ij} = \mathbf{q}_i{}^{\mathsf{T}} \mathbf{k}_j$$

$$w_{ij} = \mathrm{softmax}(w'_{ij})$$

$$\mathbf{y}_i = \sum_j w_{ij} \boldsymbol{v}_j \,.$$





Learned query, key, value weights

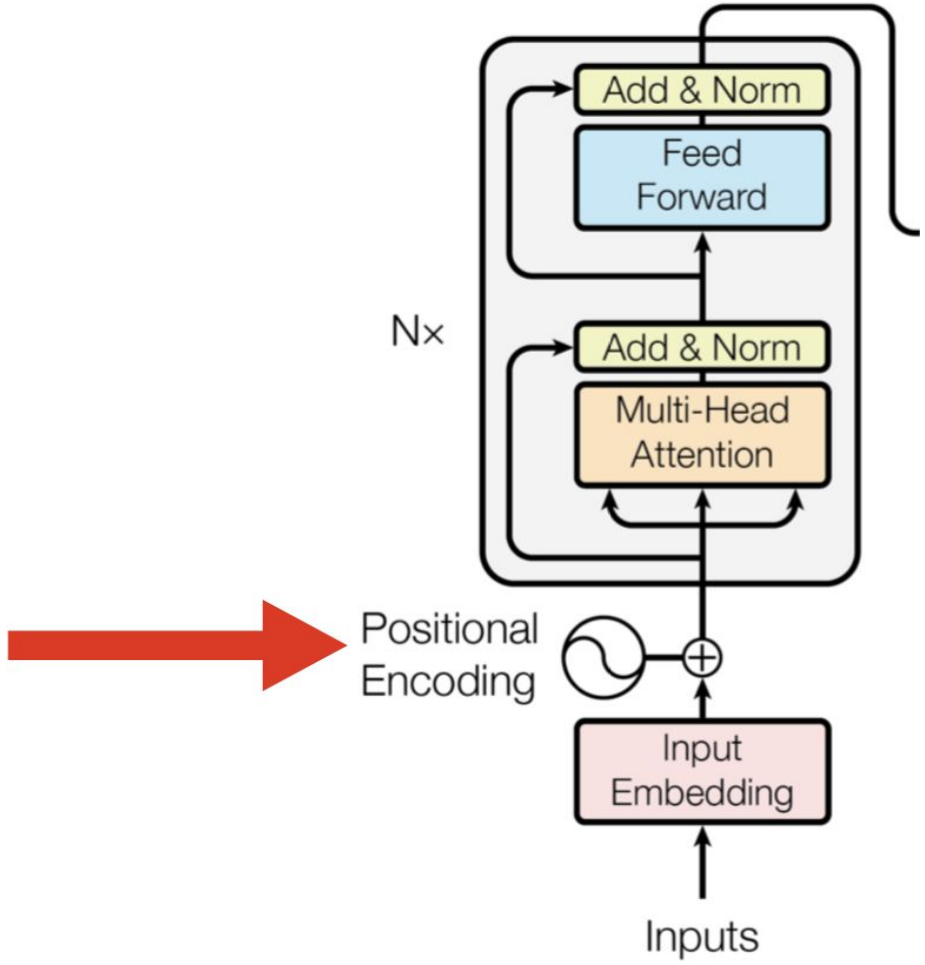
Order of the sequence does not affect result of computations ⟶

Let's encode each vector with position!

Attention is all you Need

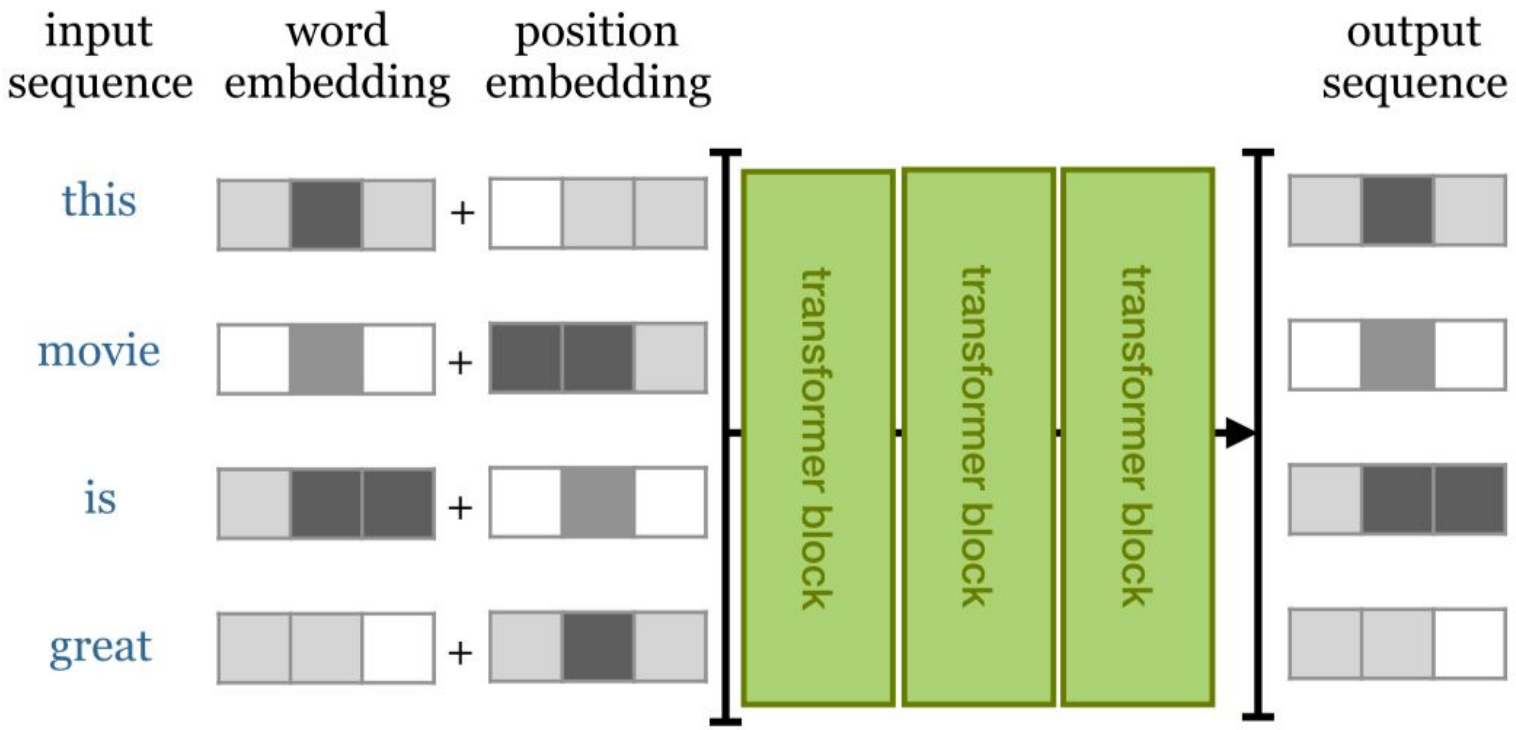
The Components:

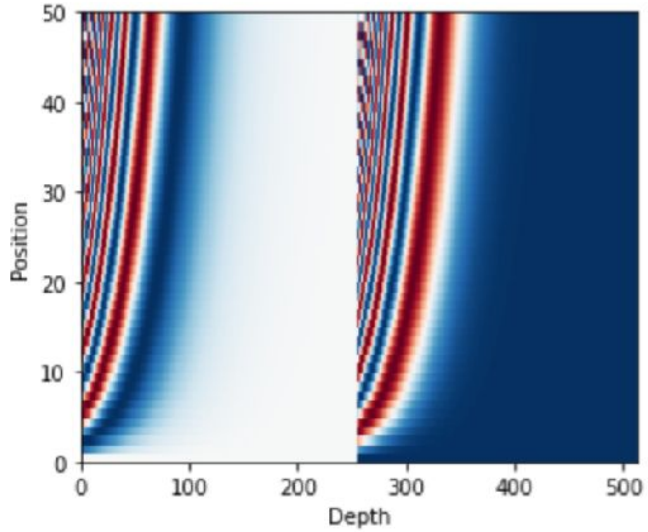
- Self attention
- Positional Encoding
- Layer Normalization



Transformers

Position Embedding: just what it sounds like!

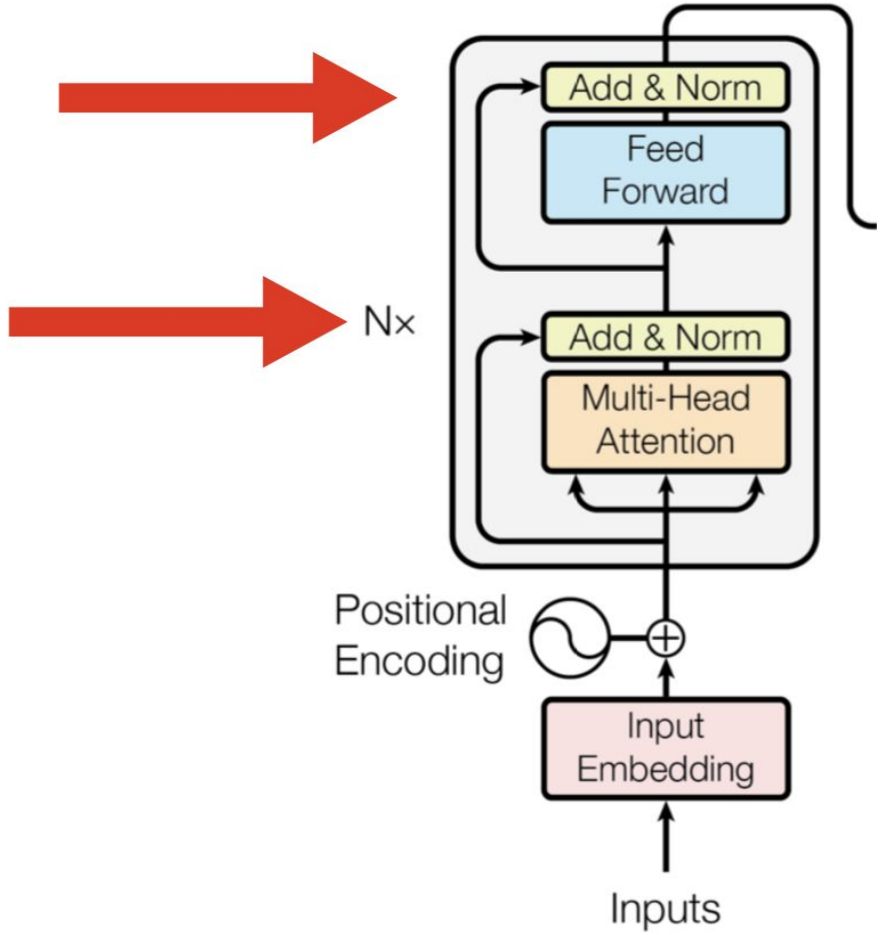




Attention is all you Need

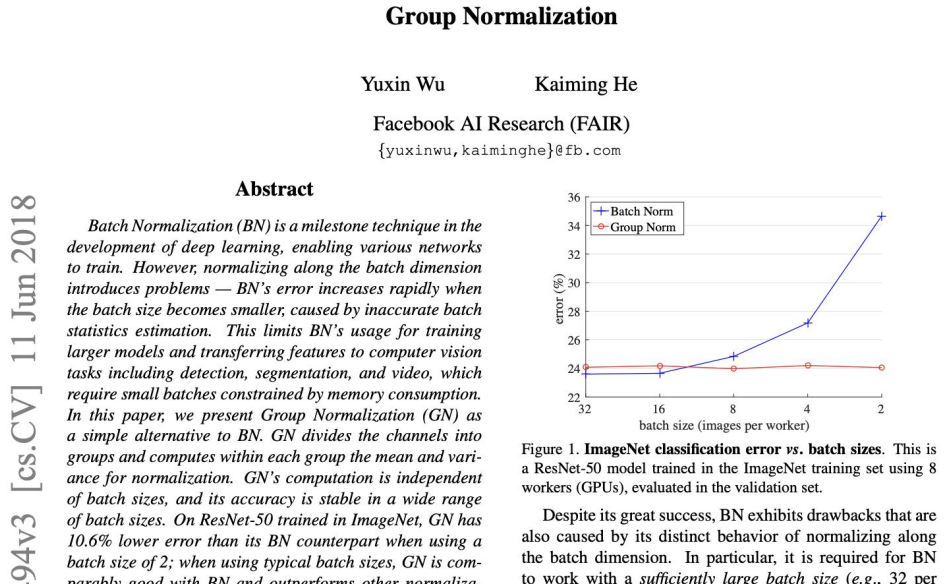
The components:

- Self attention
- Positional Encoding
- Layer Normalization



Layer Normalization

- Neural net layers work best when input vectors have uniform mean and std in each dimension.
- As inputs flow through the network, means and std's get blown out.
- Layer Normalization is a hack to reset things to where we want them in between layers.



Group Normalization

Yuxin Wu Kaiming He

Facebook AI Research (FAIR)

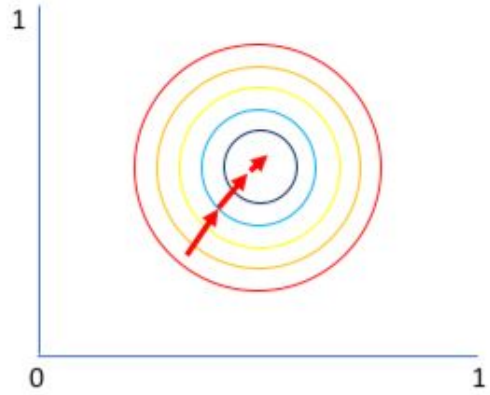
{yuxinwu,kaiminghe}@fb.com

Abstract

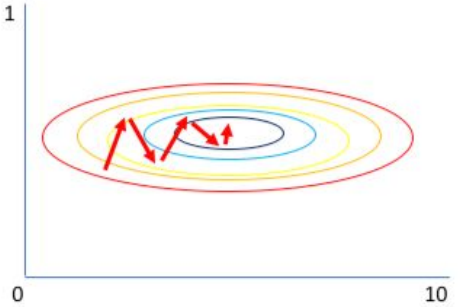
Batch Normalization (BN) is a milestone technique in the development of deep learning, enabling various networks to train. However, normalizing along the batch dimension introduces problems — BN's error increases rapidly when the batch size becomes smaller, caused by inaccurate batch statistics estimation. This limits BN's usage for training larger models and transferring features to computer vision tasks including detection, segmentation, and video, which require small batches constrained by memory consumption. In this paper, we present Group Normalization (GN) as a simple alternative to BN. GN divides the channels into groups and computes within each group the mean and variance for normalization. GN's computation is independent of batch sizes, and its accuracy is stable in a wide range of batch sizes. On ResNet-50 trained in ImageNet, GN has 10.6% lower error than its BN counterpart when using a batch size of 2; when using typical batch sizes, GN is com-

Figure 1. **ImageNet classification error *vs*. batch sizes**. This is a ResNet-50 model trained in the ImageNet training set using 8 workers (GPUs), evaluated in the validation set.

Despite its great success, BN exhibits drawbacks that are also caused by its distinct behavior of normalizing along the batch dimension. In particular, it is required for BN to work with a *sufficiently large batch size* (*e.g.*, 32 per



Both parameters can be updated in equal proportions



Gradient of larger parameter dominates the update

Source https://arxiv.org/pdf/1803.08494.pdf

Why does this work so well?

Much great Work from Anthropic if this has captured your curiosity!

In-context Learning and Induction Heads

AUTHORS

Catherine Olsson*, Nelson Elhage*, Neel Nanda*, Nicholas Joseph†, Nova DasSarma†, Tom Henighan†, Ben Mann†, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, Chris Olah‡

AFFILIATION

Anthropic

PUBLISHED

Mar 8, 2022

* Core Research Contributor; † Core Infrastructure Contributor; ‡ Correspondence to colah@anthropic.com; Author contributions statement below.

A Mathematical Framework for Transformer Circuits

AUTHORS

Nelson Elhage*†, Neel Nanda*, Catherine Olsson*, Tom Henighan†, Nicholas Joseph†, Ben Mann†, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, Chris Olah‡

AFFILIATION

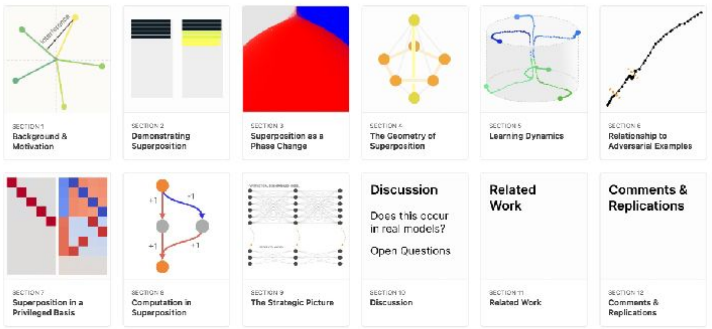
Anthropic

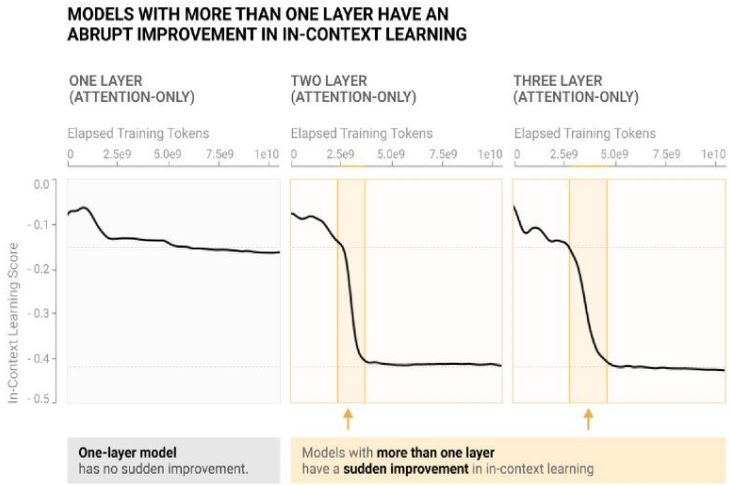
PUBLISHED

Dec 22, 2021

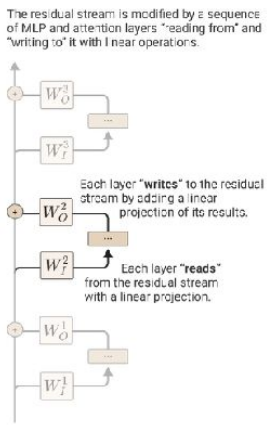
* Core Research Contributor; † Core Infrastructure Contributor; ‡ Correspondence to colah@anthropic.com; Author contributions statement below.

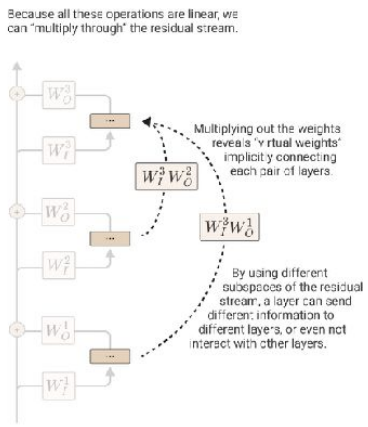
Toy Models of Superposition





We highlight the **"phase change"** period of training in plots to make visual comparision between plots easier. The highlighted region is selected for each model based on the derivative of in-context learning.





Large Language Models (LLMs)?

ELMo

Deep contextualized word representations

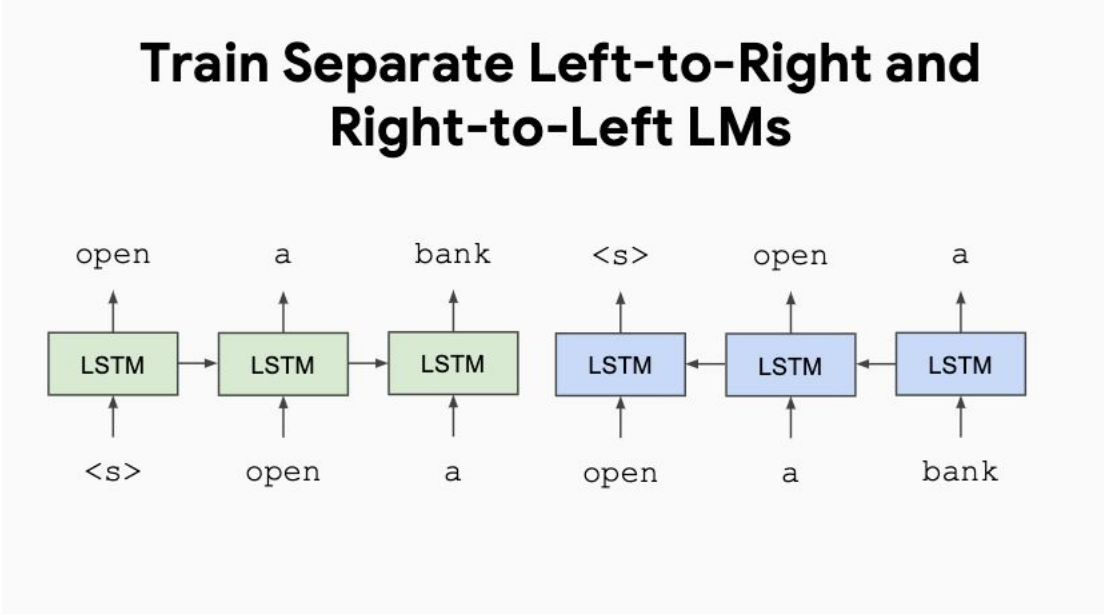
Matthew E. Peters[†], **Mark Neumann**[†], **Mohit Iyyer**[†], **Matt Gardner**[†],
{matthewp,markn,mohiti,mattg}@allenai.org

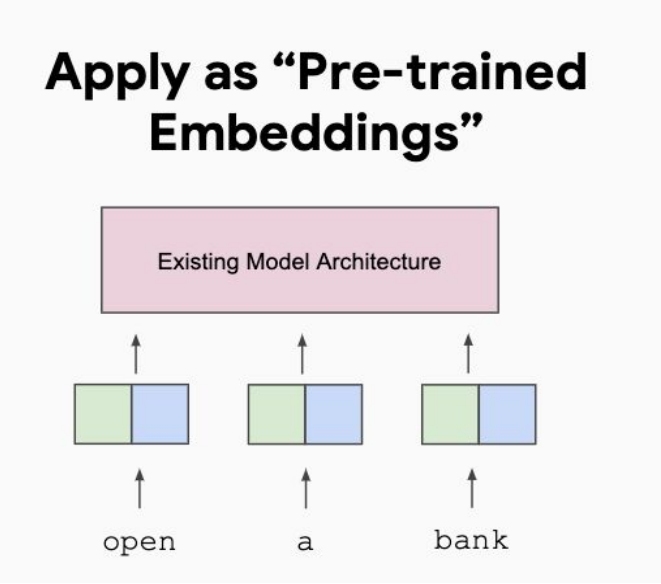
Christopher Clark[*], **Kenton Lee**[*], **Luke Zettlemoyer**[†*]
{csquared,kentonl,lsz}@cs.washington.edu

[†]Allen Institute for Artificial Intelligence
[*]Paul G. Allen School of Computer Science & Engineering, University of Washington

Ulster University

- Separate **unidirectional** LMs (left-to-right and right-to-left) trained based on **LSTMs.**
- Pre-trained representations used as input to task-specific models.
- **Single sentences training** from 1B word benchmark (Chelba et al., 2014).





source: [Peters et al 2018 (NAACL 2018 best paper)]

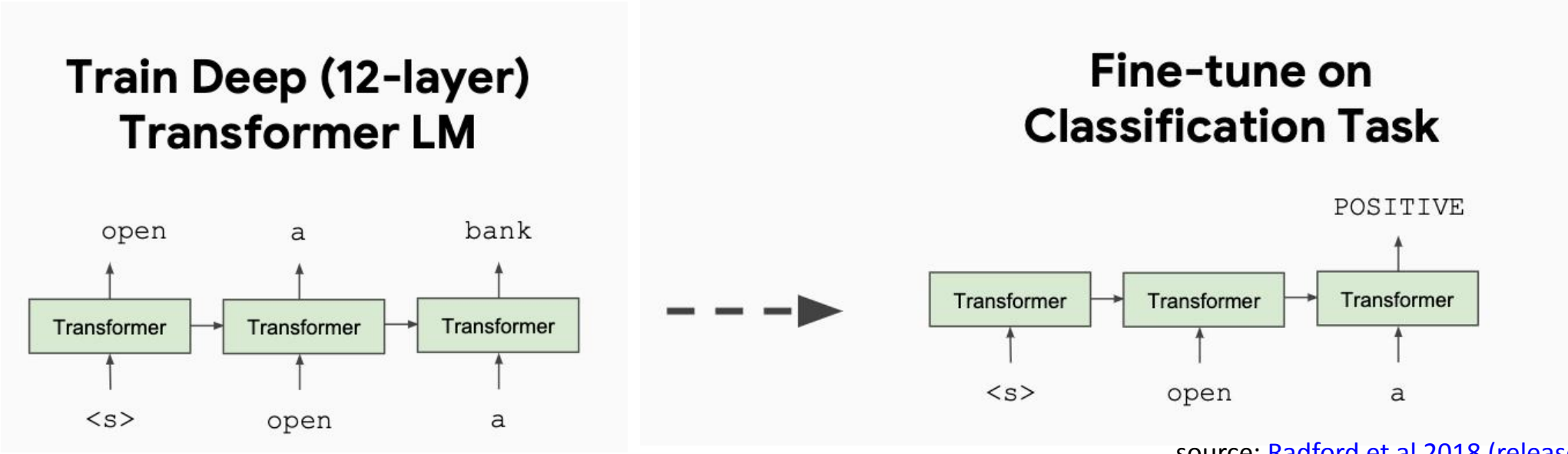
OpenAI GPT/GPT-2

Improving Language Understanding by Generative Pre-Training

Alec Radford, OpenAI, alec@openai.com
Karthik Narasimhan, OpenAI, karthikn@openai.com
Tim Salimans, OpenAI, tim@openai.com
Ilya Sutskever, OpenAI, ilyasu@openai.com

Ulster University

- Generative Pre-trained Transformers
- Train one unidirectional LM (left-to-right) based on a deep **Transformer decoder-only** (uses masked self-attention)
- **Fine-tuning** approach: all pre-trained parameters are re-used & updated on downstream tasks
- Trained on 512-token segments on BooksCorpus — much **longer** context! (larger model is 1.5B on 8M webpages)



source: [Radford et al 2018 (released in 2018/6)]

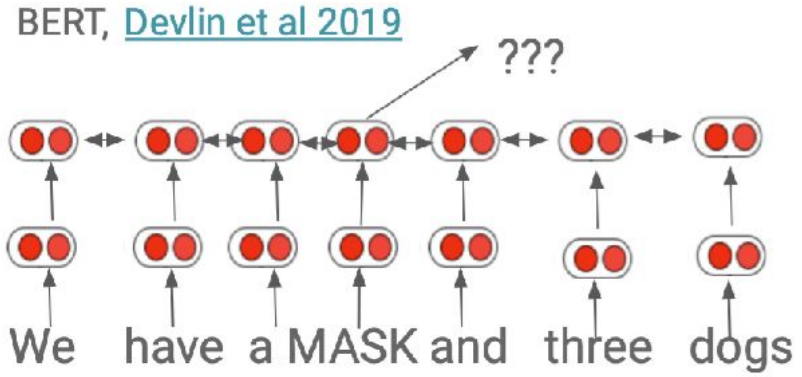
Bidirectional Encoder Representation from Transformers

- It is a **fine-tuning approach** based on a deep **Transformer encoder-only(no attention masking)**
- The key: learn representations based on **bidirectional context**

  Why? Because both left and right contexts are important to understand the meaning of words.

  Example #1: we went to the river bank.

  Example 2: I need to go to bank to make a deposit.



- **110 M params, 15% words masked out**
- State-of-the-art performance on a large set of **sentence-level** and **token-level** tasks

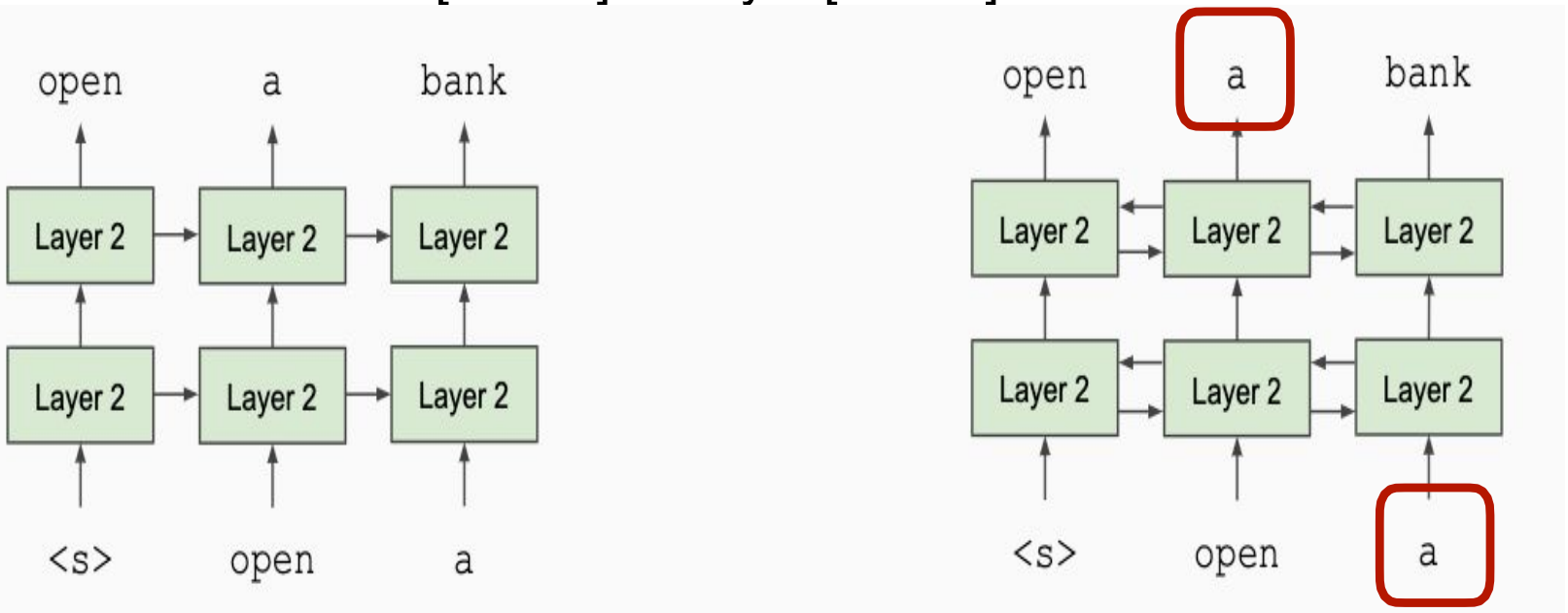
Source: Devlin 2019

Masked Language Modeling (MLM)

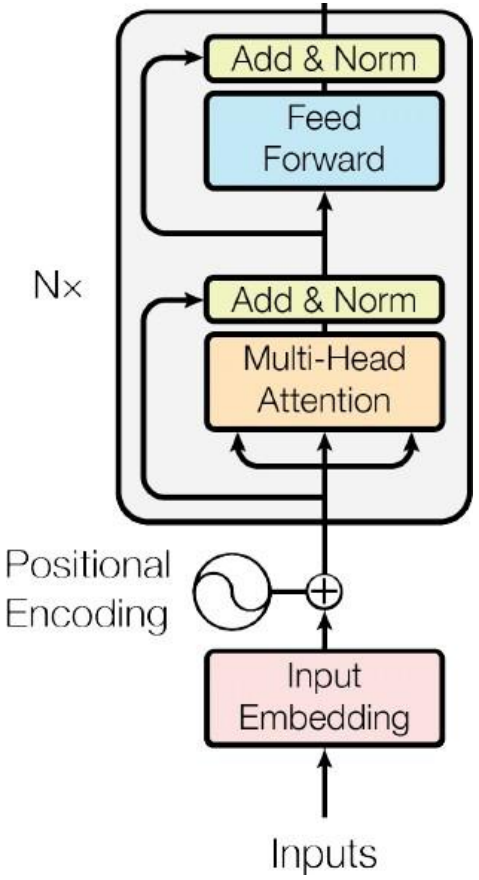
- Q: Why we can't do language modeling with bidirectional models?
- Mask out K% of the input words and then predict the masked words

store gallon

the man went to [MASK] to buy a [MASK] of milk



BERT Pre-training



- BERT-base: 12 layers, 768 hidden size, 12 attention heads(110M parameter)

Same as OpenAI GPT

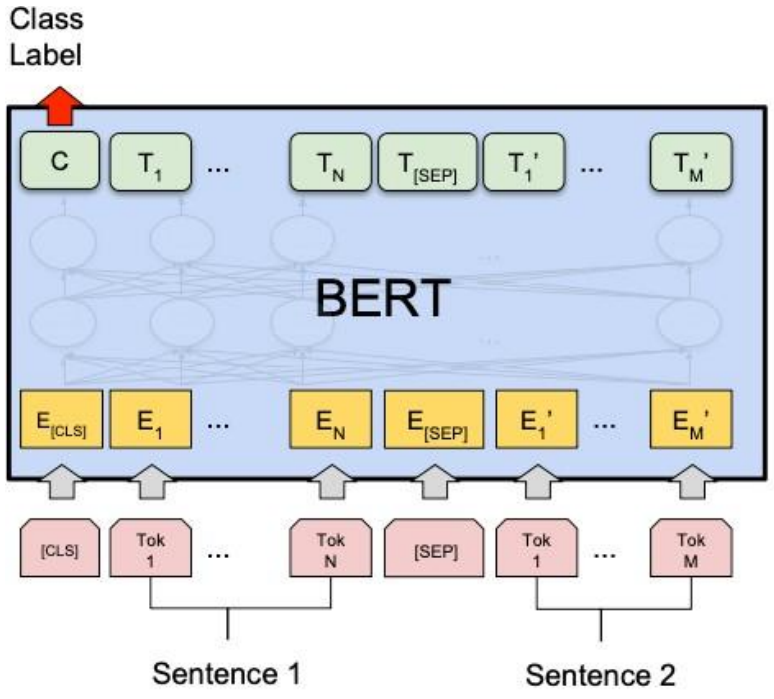
- BERT-large: 24 layers, 1024 hidden size, 16 attention heads, 340M parameters

OpenAI GPT was trained on BooksCorpus only!

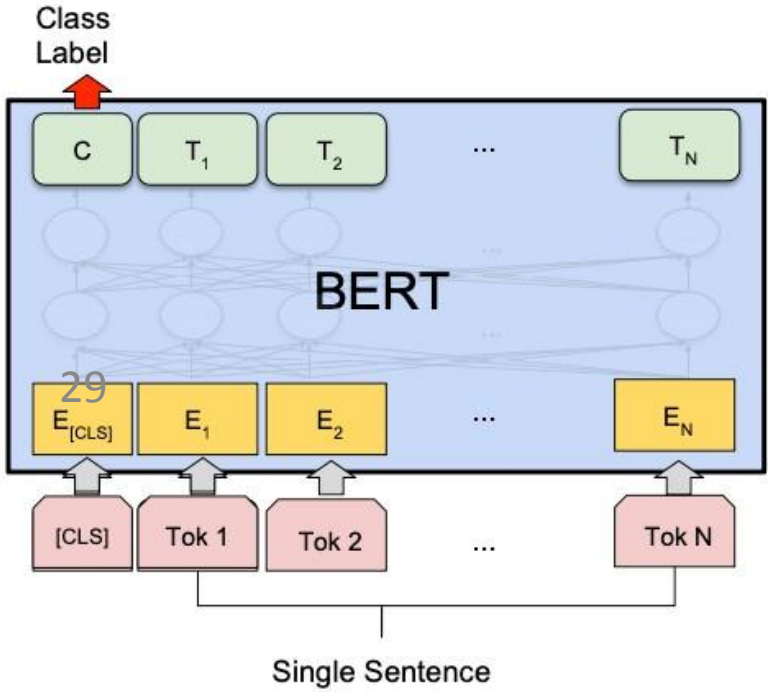
- Training corpus: Wikipedia (2.5B) + BooksCorpus (0.8B)
- Max sequence size: 512 word pieces (roughly 256 and 256 for two non-contiguous sequences)
- Trained for 1M steps, batch size 128k

“Pretrain once, finetune many times”

sentence/token-level tasks



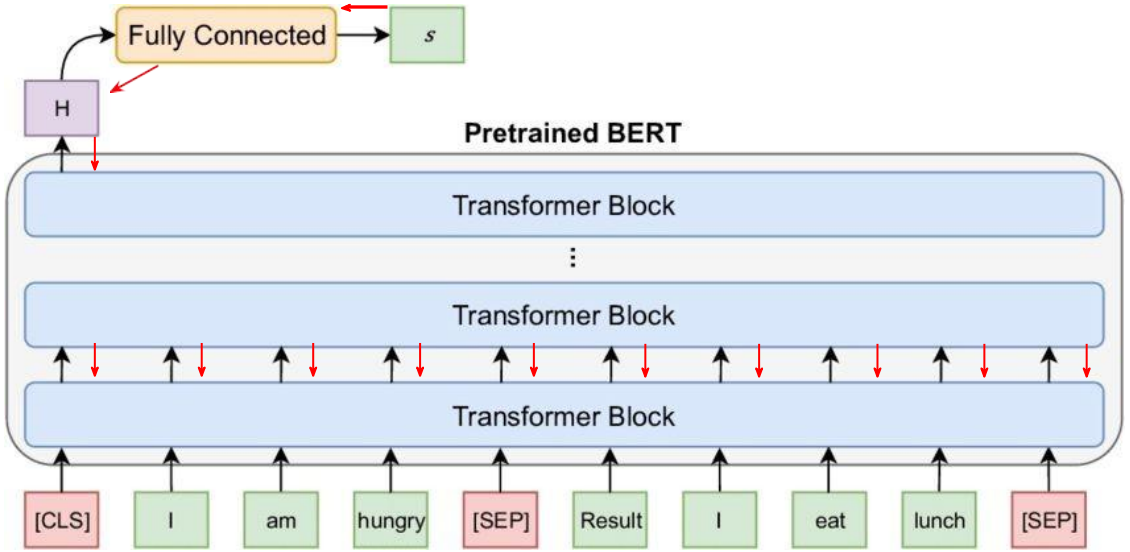
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA

Background: Fine Tuning

- Pretrain a language model on task
- Attach a small task specific layer
- Fine-tune the weights of full NN by propagating gradients on a downstream task



Devlin et al. 2019

What happened after BERT?

Lots of people are trying to understand what BERT has learned and how it works.

A Primer in BERTology: What We Know About How BERT Works

Anna Rogers
Center for Social Data Science
University of Copenhagen
arogers@sodas.ku.dk

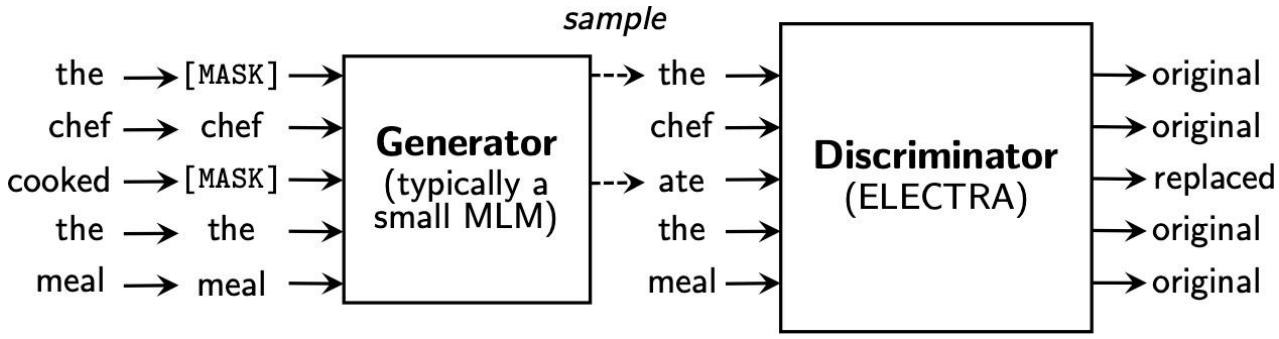
Olga Kovaleva
Dept. of Computer Science
University of Massachusetts Lowell
okovalev@cs.uml.edu

Anna Rumshisky
Dept. of Computer Science
University of Massachusetts Lowell
arum@cs.uml.edu

- Syntactic knowledge, semantic knowledge, world knowledge…
- How to mask, what to mask, where to mask, alternatives to masking..

What happened after BERT?

- RoBERTa (Liu et al., 2019)
  - Trained on 10x data & longer, no NSP
  - Much stronger performance than BERT (e.g., 94.6 vs 90.9 on SQuAD)
  - Still one of the most popular models to date
- ALBERT (Lan et al., 2020)
  - Increasing model sizes by sharing model parameters across layers
  - Less storage, much stronger performance but runs slower..
- ELECTRA (Clark et al., 2020)
  - It provides a more efficient training method by predicting 100% of tokens instead of 15% of tokens



- *Models that handle long contexts (512 tokens)*
  - Longformer, Big Bird, …
- *Multilingual BERT*
  - Trained single model on 104 languages from Wikipedia. Shared 110k WordPiece vocabulary
- *BERT extended to different domains*
  - SciBERT, BioBERT, FinBERT, ClinicalBERT, …
- *Making BERT smaller to use*
  - DistillBERT, TinyBERT, …

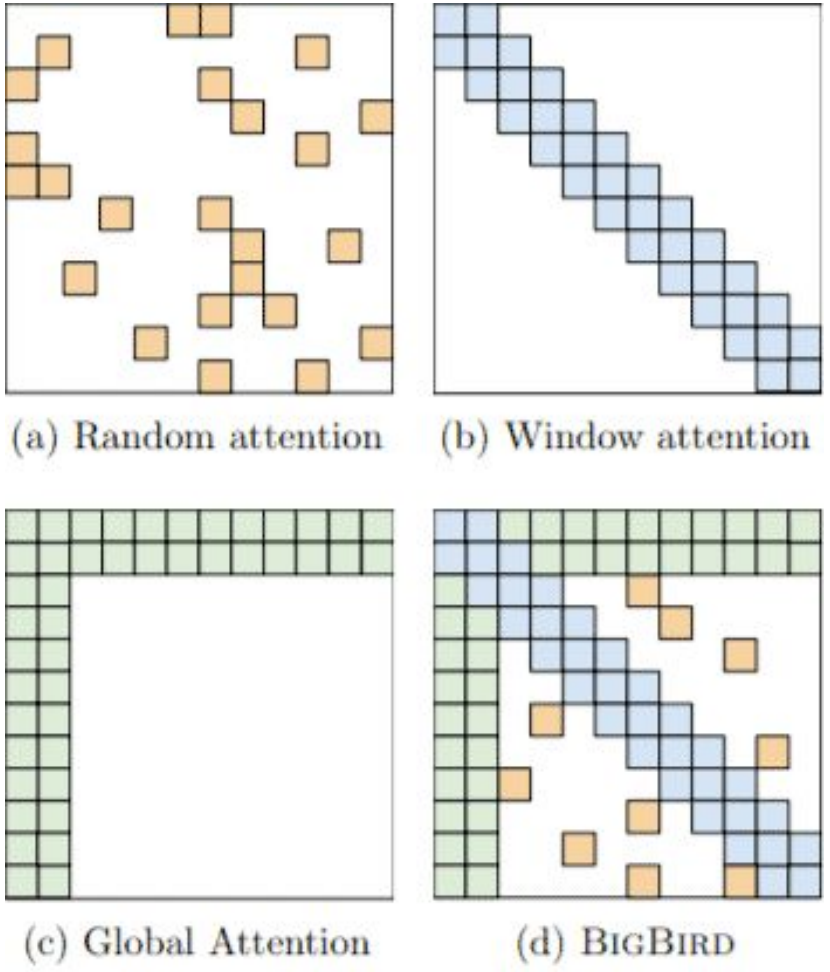


Image from the original paper

Ulster University

BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model

Alex Wang
New York University
alexwang@nyu.edu

Kyunghyun Cho
New York University
Facebook AI Research
CIFAR Azrieli Global Scholar
kyunghyun.cho@nyu.edu

Mask-Predict: Parallel Decoding of Conditional Masked Language Models

Marjan Ghazvininejad* **Omer Levy*** **Yinhan Liu*** **Luke Zettlemoyer**
Facebook AI Research
Seattle, WA

Exposing the Implicit Energy Networks behind Masked Language Models via Metropolis--Hastings

Kartik Goyal, Chris Dyer, Taylor Berg-Kirkpatrick

src	Der Abzug der franzsischen Kampftruppen wurde am 20. November abgeschlossen .
$t = 0$	The departure of the French combat completed completed on 20 November .
$t = 1$	The departure of French combat troops was completed on 20 November .
$t = 2$	The withdrawal of French combat troops was completed on November 20th .

Leveraging Pre-trained Checkpoints for Sequence Generation Tasks

Sascha Rothe, Shashi Narayan, Aliaksei Severyn

GPT 2

Language Models are Unsupervised Multitask Learners

Alec Radford [*1] Jeffrey Wu [*1] Rewon Child [1] David Luan [1] Dario Amodei [**1] Ilya Sutskever [**1]

- Released in 2019
- Surface level: more data more parameters
- For the first time, ML was able to generate long strings of coherent text
- No fine-tuning was required for downstream tasks.
- Tasks like translation, summarization were done only by ‘predicting the next token in a sequence’
- 1.5 billion parameters
- Token size or context size is 1024
- Dataset: Created their own web text from Reddit links i.e. 8 million documents (40GB of data)

Source: Radford et al 2019

GPT-3 (2020)

- Same story: even more data, even more parameters
- GPT-2 was 10x larger than GPT-1, but GPT-3 is 100x larger than GPT-2
- 175 billion parameters
- Doubles context size – 2048
- Datasets: common crawl (filtered), WebText 2, Books1, Books2, Wikipedia
- Unsupervised pre-training, no fine-tuning
- Exhibits few-shot and zero-shot learnings
- Available via API

Image Source https://jalammar.github.io/

Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*
Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry
Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan
Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter
Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray
Benjamin Chess Jack Clark Christopher Berner
Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI

Abstract

Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans can generally perform a new language task from only a few examples or from simple instructions – something which current NLP systems still largely struggle to do. Here we show that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks, as well as several tasks that require on-the-fly reasoning or domain adaptation, such as unscrambling words, using a novel word in a sentence, or performing 3-digit arithmetic. At the same time, we also identify some datasets where GPT-3's few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora. Finally, we find that GPT-3 can generate samples of news articles which human evaluators have difficulty distinguishing from articles written by humans. We discuss broader societal impacts of this finding and of GPT-3 in general.

source: Brown et al 2020

GPT-3 is Massive

- 96 decoder blocks(2*GPT-2)
- Context size 2048 (2*GPT-2)
- Embedding size: 12288(~8x GPT-2)
- Params: 175b(~117*GPT-2)

Source https://jalammar.github.io/

Downside?

Limitations of GPT-3

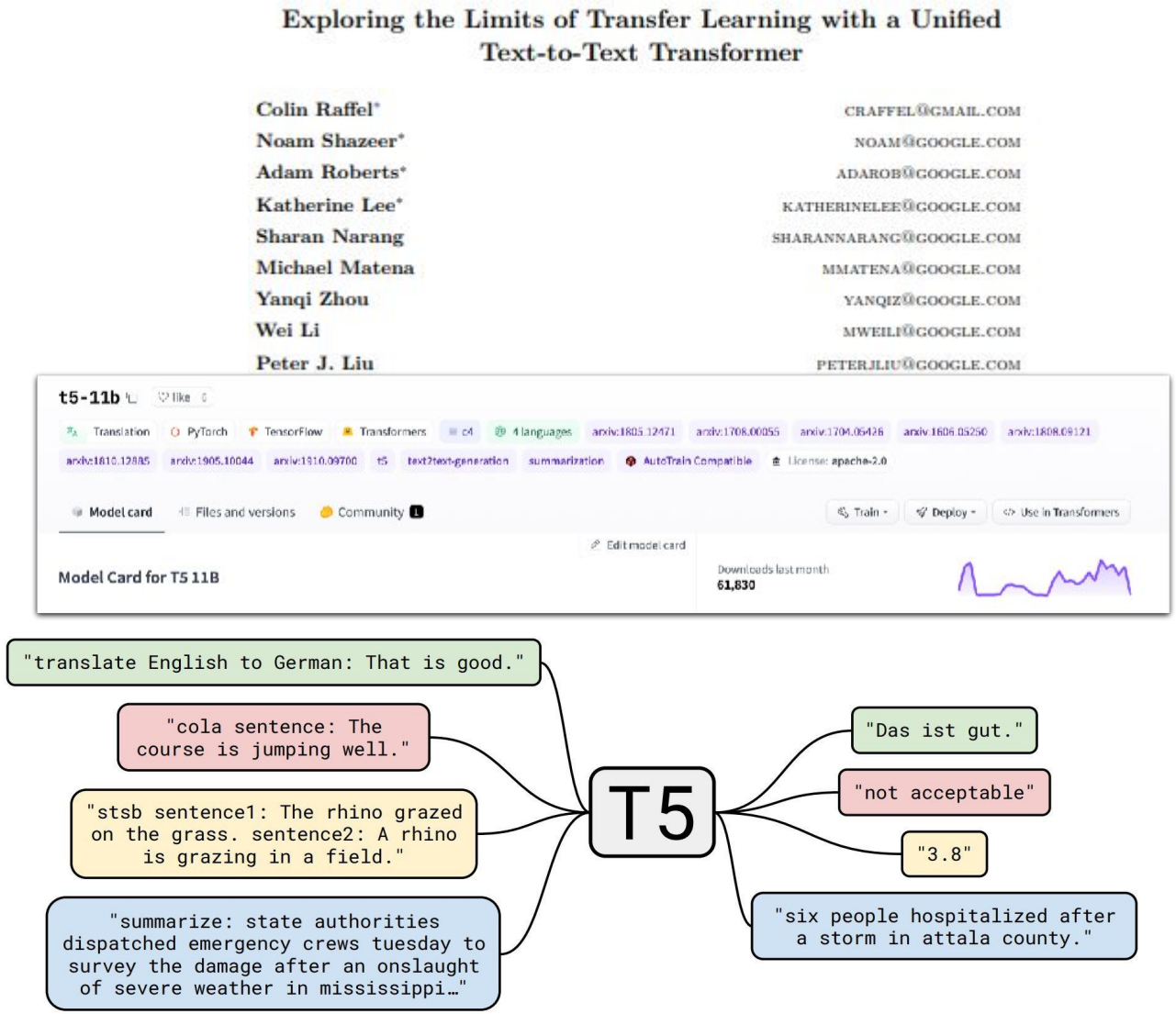
Ulster University

- Limited generation (repetitions,contradiction)
- Limited common sense world model
- Poor one-shot and zero-shot performance (on some reading comprehension and comparison task)
- No bidirectionality-denoising objective
- Of language models in general
- Simple pre-training objectives
- Lack of grounding-multimodal
- Poor sample efficiency
- Performance aside
- Not interpretable
- Adaptation vs recognition
- Expensive-distillation !

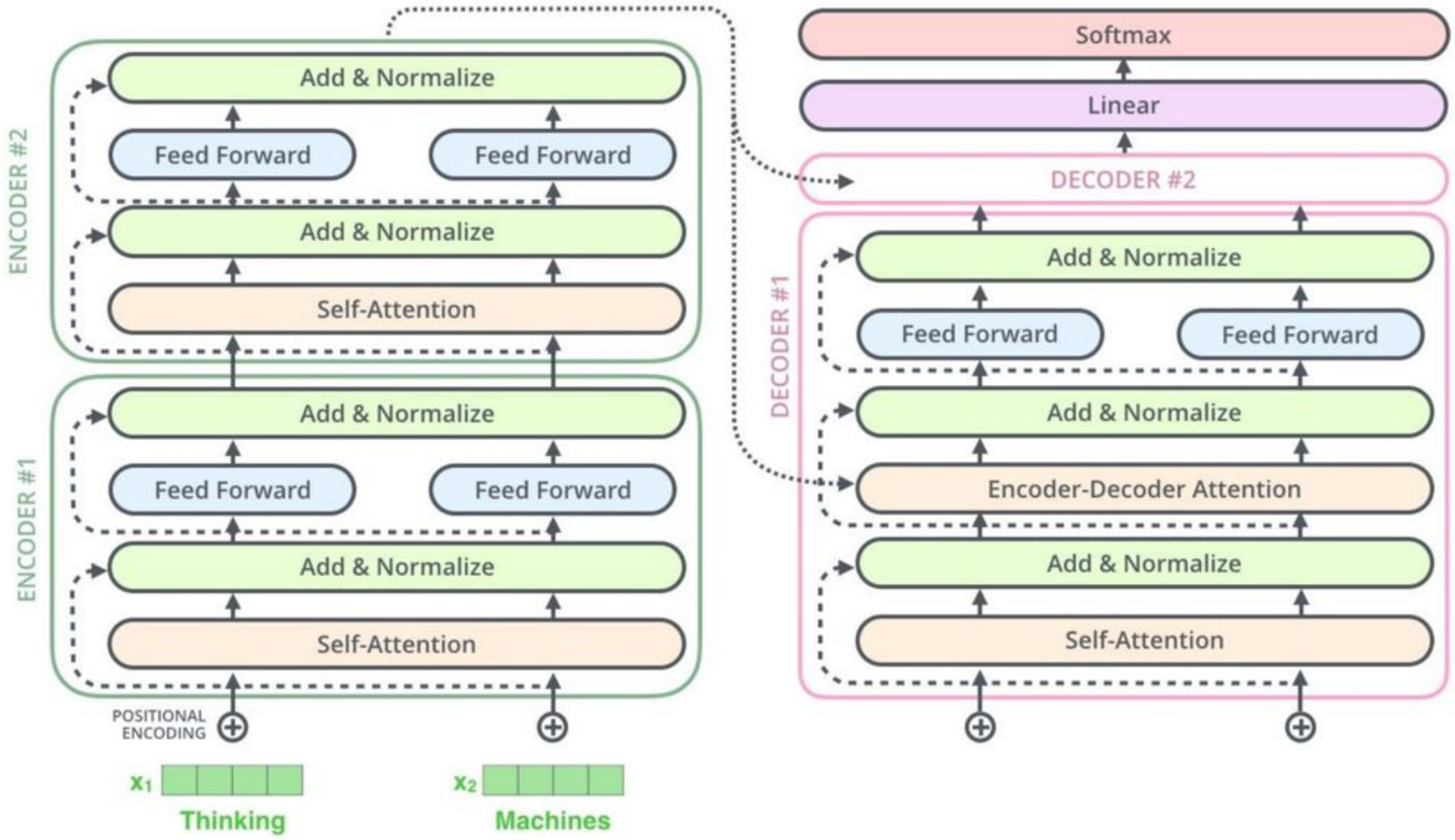
And then transfer happened?

T5 (Text-to-Text): The basic Idea

- Every task, one format!
- Previous attempts included:
  - Question answering
  - Language modeling
  - Span extraction… but had limitations
- “[Task-specific prefix]: [Input text]” -> “[output text]”
- Encoder-decoder model
- Baseline size: two stacks of size BERT
- Architecture from “Attention Is All You Need”
- Different position embedding scheme
- Supports both discriminative and generative tasks
- Classification, summarization, translation, etc.
- Trained on C4 (Colossal Clean Crawled Corpus)-100X larger than Wikipedia
- 11B parameters,open source



source Raffel et al 2020



Source http://jalammar.github.io/illustrated-transformer/

Input/Output

[Task-specific prefix]: [Input text]

EnDe (Translation):

"translate English to German: That is good" -> "Das ist gut"

CNNDM (Summarization):

"summarize: state authorities dispatched…" -> "six people hospitalized after storm"

CoLA (GLUE; Classification):

"cola sentence: The course is jumping well." -> "not acceptable"

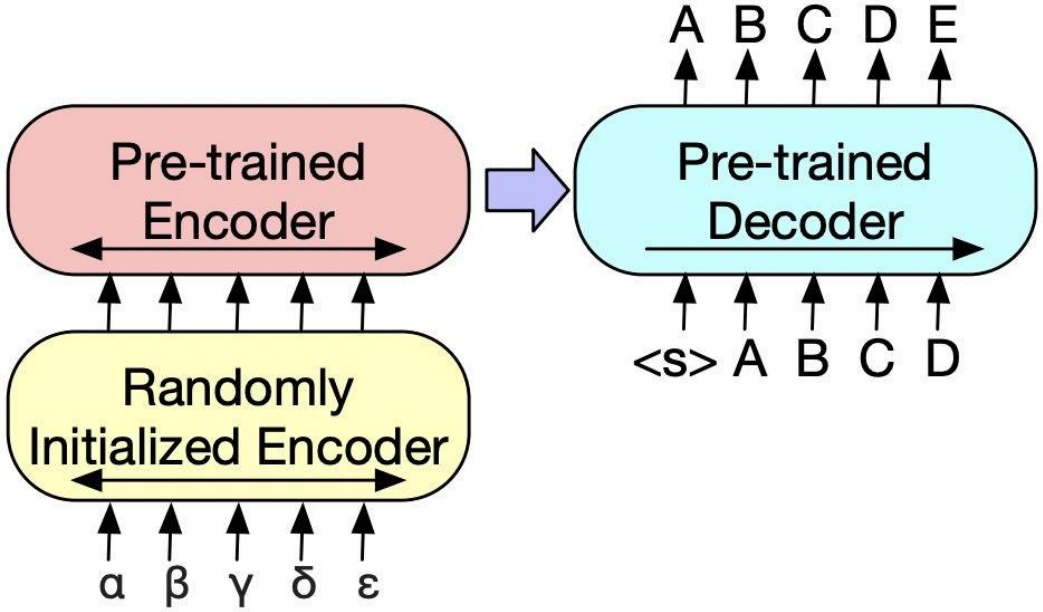
STS-B (GLUE; Regression):

"stsb sentence1: The rhino grazed. sentence2: A rhino is grazing." -> "3.8"

Did history repeat itself?

BART (Bidirectional and Auto-Regressive Transformers)

- Similar Architecture as T5.
- Performs competitive to RoBERTa and XLNet on discriminative tasks.
- Outperformed existing methods on question answering, and summarization tasks.
- Improved results on machine translation with fine-tuning on target language.



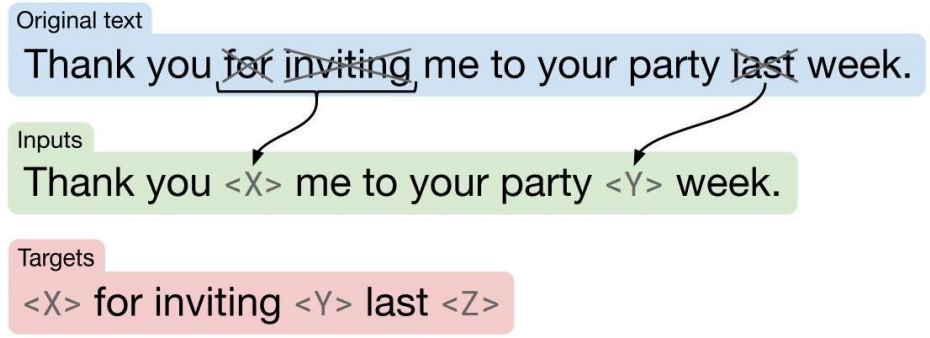
BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension

Mike Lewis*, Yinhan Liu*, Naman Goyal*, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer
Facebook AI
{mikelewis,yinhanliu,naman}@fb.com

source Lewis et al 2020

mT5 (Massively Multilingual Pre-trained Text-to-Text Transformer)

- mC4: Common Crawl dataset covering 101 languages!
- Only line length filter, no punctuation filter
- How do you sample across languages?
- "Boosting" the probability of training on low-resource languages without overfitting
- Similar architecture to T5
- 6 tasks from the XTREME multilingual benchmark
- Entailment, reading comprehension, NER, paraphrase identification



mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer

Linting Xue* Noah Constant* Adam Roberts*
Mihir Kale Rami Al-Rfou Aditya Siddhant Aditya Barua Colin Raffel
Google Research

source Xue et al 2021

AlexaTM 20B (Alexa Teacher Model)

- Larger architecture on multilingual C4 dataset.
- Can outperform much larger autoregressive models (GPT-3 175B) in zero shot tasks.

Model	BoolQ (acc)	CB (acc)	RTE (acc)	ReCoRD (acc)	WSC (acc)	WiC (acc)	CoPA (acc)	MultiRC (f1a)	Avg
PaLM 540B	**88.0**	51.8	**72.9**	**92.9**	**89.1**	**59.1**	**93.0**	**83.5**	**78.8**
GPT3 175B	60.5	46.4	63.5	90.2	65.4	0.0	91.0	72.9	61.2
BLOOM 175B	63.5	33.9	52.0	NA	51.9	50.6	56.0	57.1	NA
GPT3 13B	66.2	19.6	62.8	89.0	64.4	0.0	84.0	71.4	57.2
UL 20B	63.1	41.1	60.7	88.1	79.9	49.8	85.0	36.2	63.0
AlexaTM 20B	69.44	**67.9**	68.59	88.4	68.27	53.29	78.0	59.57	69.16

AlexaTM 20B: Few-Shot Learning Using a Large-Scale Multilingual Seq2seq Model

Soltan* Shankar Ananthakrishnan Jack FitzGerald Rahul Gupta
Wael Hamza Haidar Khan Charith Peris Stephen Rawls
Andy Rosenbaum Anna Rumshisky Chandana Satya Prakash
ukund Sridhar Fabian Triefenbach Apurv Verma Gokhan Tur
Prem Natarajan

Amazon Alexa AI

Source Soltan et al 2022

What changed between GPT 3 and GPT 3.5?

Reinforcement Learning with Human Feedback (RLHF)

Ulster University

Deep Reinforcement Learning from Human Preferences

Paul F Christiano
OpenAI
paul@openai.com

Jan Leike
DeepMind
leike@google.com

Tom B Brown
Google Brain*
tombbrown@google.com

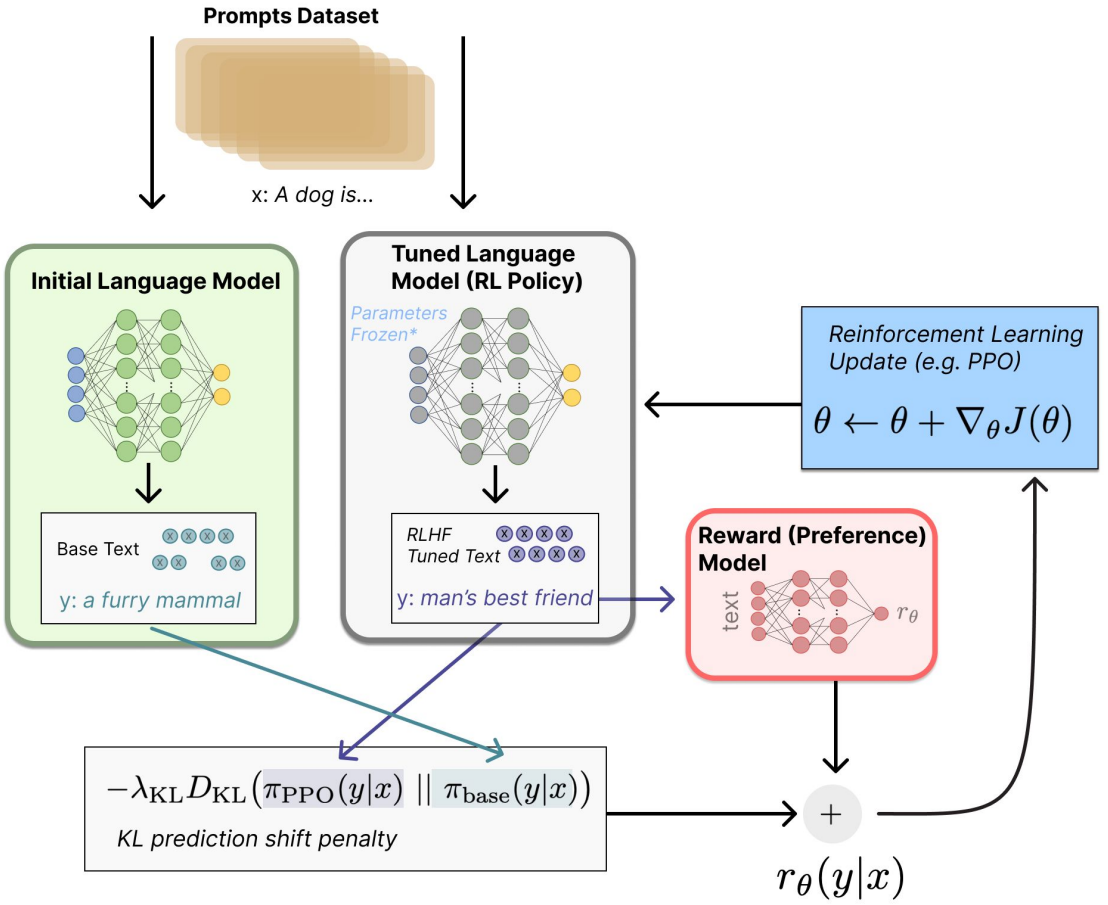
Miljan Martic
DeepMind
miljanm@google.com

Shane Legg
DeepMind
legg@google.com

Dario Amodei
OpenAI
damodei@openai.com

Abstract

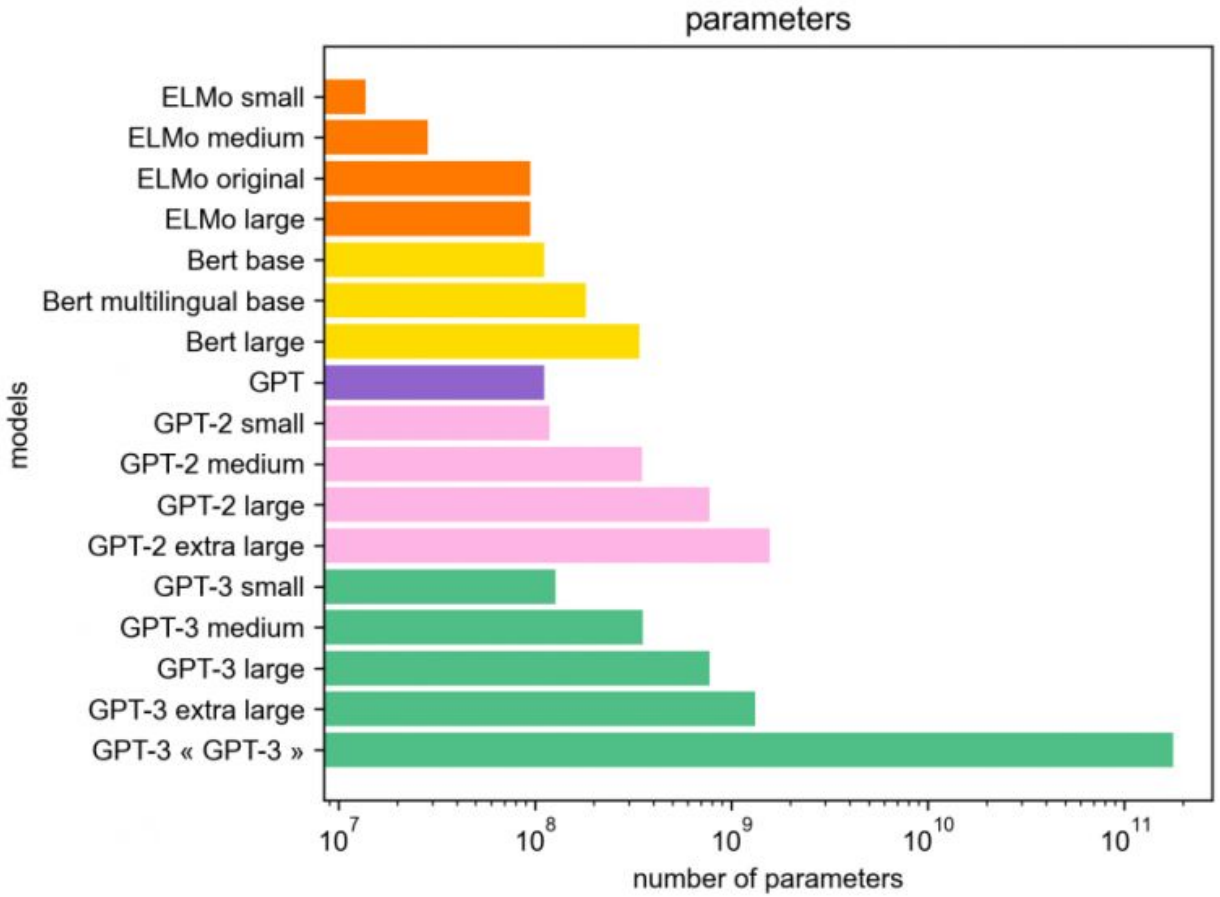
For sophisticated reinforcement learning (RL) systems to interact usefully with real-world environments, we need to communicate complex goals to these systems. In this work, we explore goals defined in terms of (non-expert) human preferences between pairs of trajectory segments. We show that this approach can effectively solve complex RL tasks without access to the reward function, including Atari games and simulated robot locomotion, while providing feedback on less than 1% of our agent's interactions with the environment. This reduces the cost of human oversight far enough that it can be practically applied to state-of-the-art RL systems. To demonstrate the flexibility of our approach, we show that we can successfully train complex novel behaviors with about an hour of human time. These behaviors and environments are considerably more complex than any which have been previously learned from human feedback.

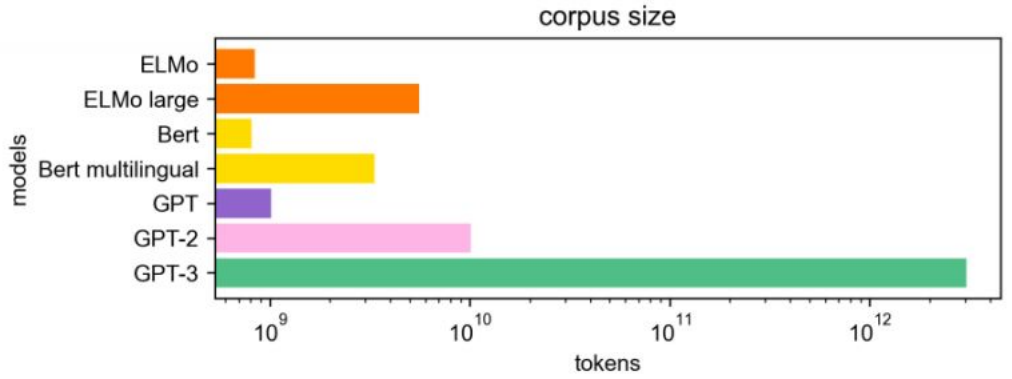


source: https://platform.openai.com/docs/models/gpt-3-5

How large are “large” LM?

How large are ”large” LM





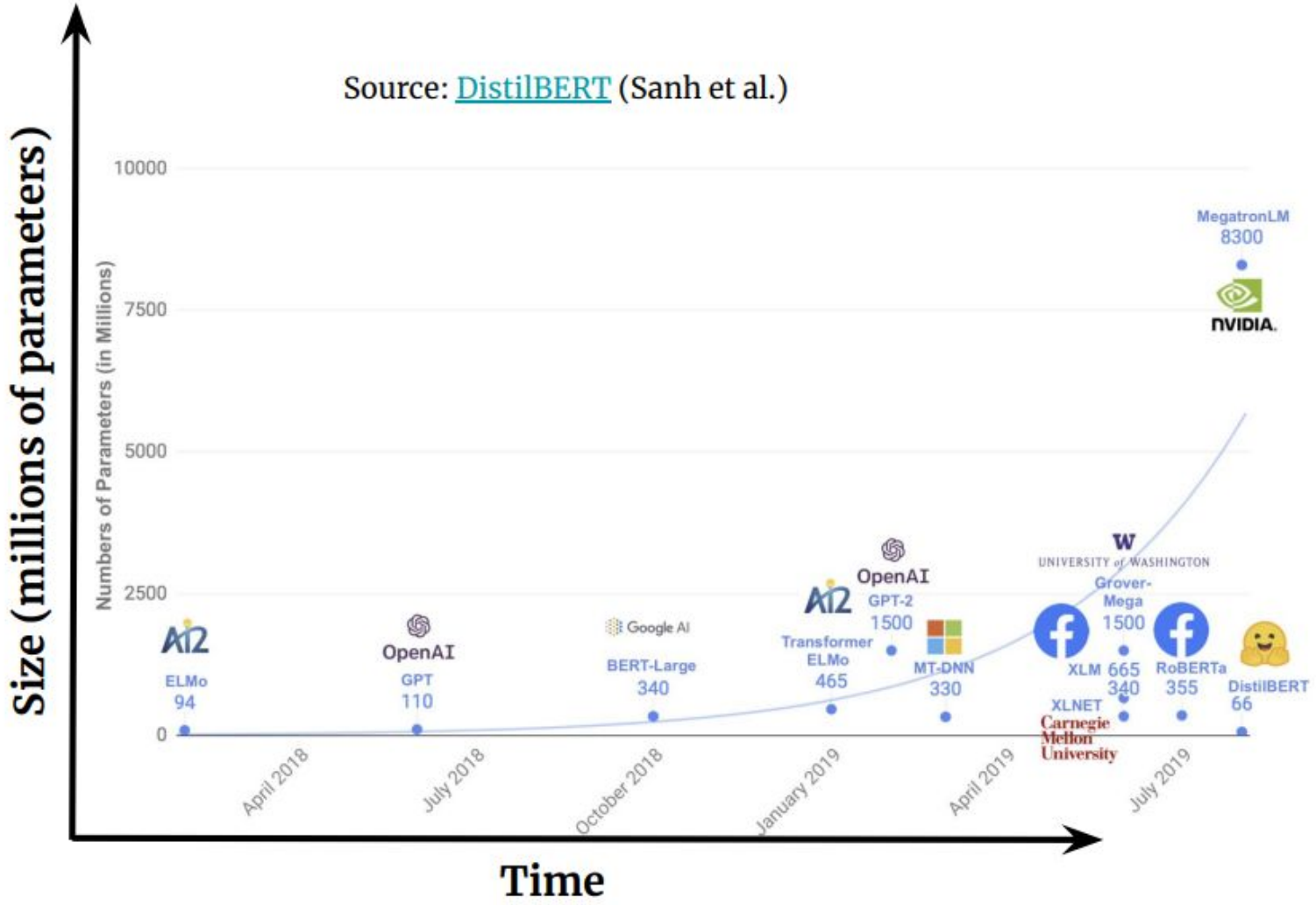
More Recent models
PaLM(540B),OPT(175B),BLOOM(176B)

Source: https://hellofuture.orange.com/en/the- gpt-3-language-model-revolution-or-evolution/

How large are ”large” LM

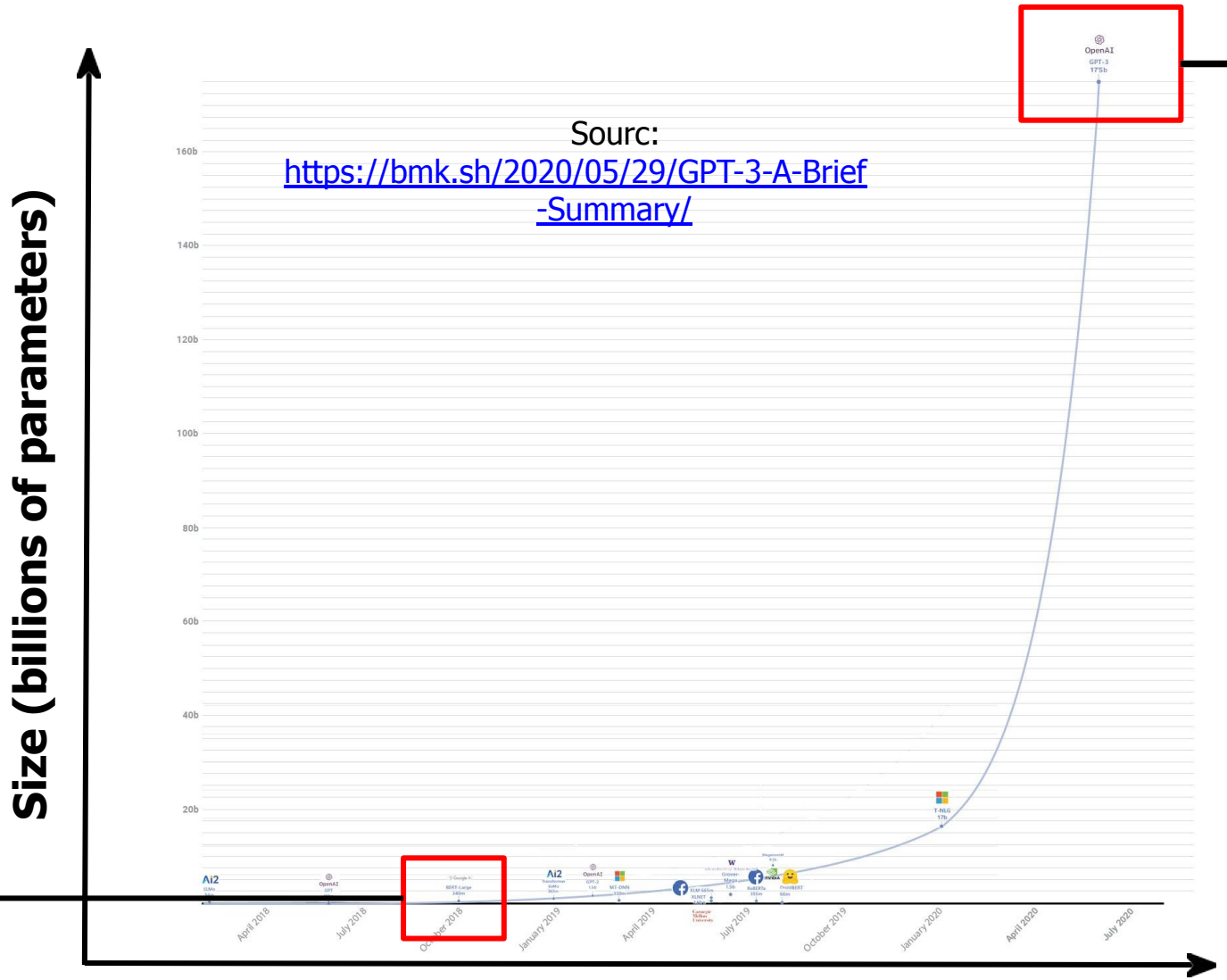
- Today, we mostly talk about two camps of models:
- Medium-sized models: BERT/RoBERTa models (100M or 300M), T5 models (220M, 770M, 3B)
- “Very” large LMs: models of 100+ billion parameters
- Larger model size larger compute, more expensive during inference
- Different sizes of LMs have different ways to adapt and use them
- Fine-tuning, zero-shot/few-shot prompting, in-context learning
- Emergent properties arise from model scale
- Trade-off between model size and corpus size

LM Landscape Pre-GPT-3



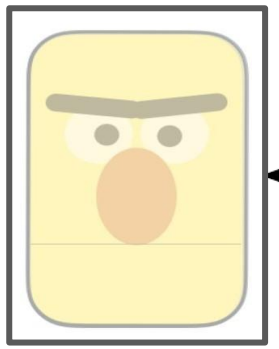
LM Landscape with GPT-3

Sourc:
https://bmk.sh/2020/05/29/GPT-3-A-Brief-Summary/



175b params!
GPT-2 was 1.5b

340m params!



Why Scale?

- Study conducted by OpenAI * Scaling Laws for Neural language models (Kaplan et al. 2020)

  A few key findings:

- Performance depends strongly on scale, weakly on model shape
- Smooth power laws ($y = ax^k$) b/w empirical performance & N - parameters, D - dataset size, C - compute
- Transfer improves with test performance
- Larger models are more sample efficient

Scaling Laws for Neural Language Models

Jared Kaplan *
Johns Hopkins University, OpenAI
jaredk@jhu.edu

Sam McCandlish*
OpenAI
sam@openai.com

Tom Henighan
OpenAI
henighan@openai.com

Tom B. Brown
OpenAI
tom@openai.com

Benjamin Chess
OpenAI
bchess@openai.com

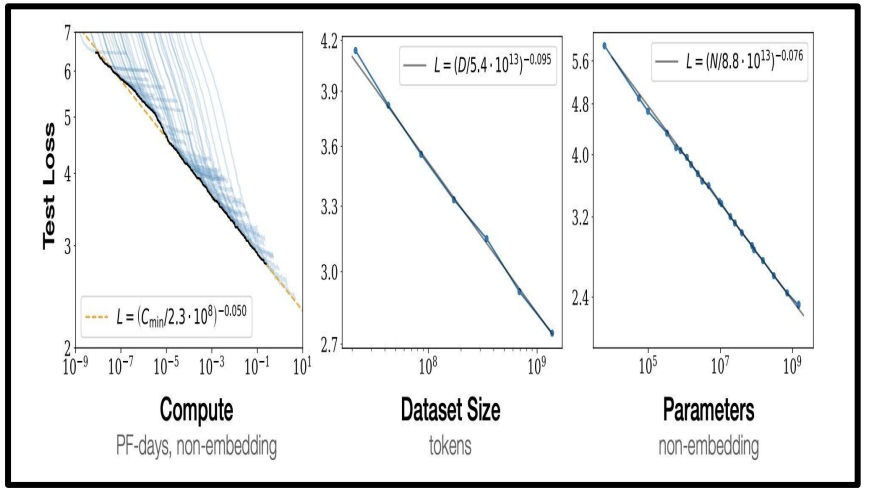
Rewon Child
OpenAI
rewon@openai.com

Scott Gray
OpenAI
scott@openai.com

Alec Radford
OpenAI
alec@openai.com

Jeffrey Wu
OpenAI
jeffwu@openai.com

Dario Amodei
OpenAI
damodei@openai.com



Pre-trained LLMs

- Instruct GPT-3 (ada 350M, babbage 1.3B, curie 6.7B, and davinci 175B) (Ouyang et al., 2022) (text-davinci)
- PaLM (8B, 62B, 540B) (Chowdhery et al., 2022)
- LaMDA (422M, 2B, 8B, 68B, 137B) (Thoppilan et al., 2022)
- GPT-3 (ada 350M, babbage 1.3B, curie 6.7B, davinci 175B)
- GPT-2 (1.5B)
- GPT-Neo (2.7B) (Yang et al. 2022), GPT-J (6B), T0 (11B) (Sanh et al., 2022), OPT (13B) (Zhang et al., 2022)

Ulster University

bigger models → better performance

*This may be true, but is increasing model size the most **efficient** way of improving performance?*

Do the largest model always give the best performance today?

Do largest models always give the best performance today?

- Chinchilla outperforms Gopher on 51 task
- Achieves similar performance on 2 task
- Underperforms on 4 task namely college mathematics, econometrics, moral scenarios, formal logic)

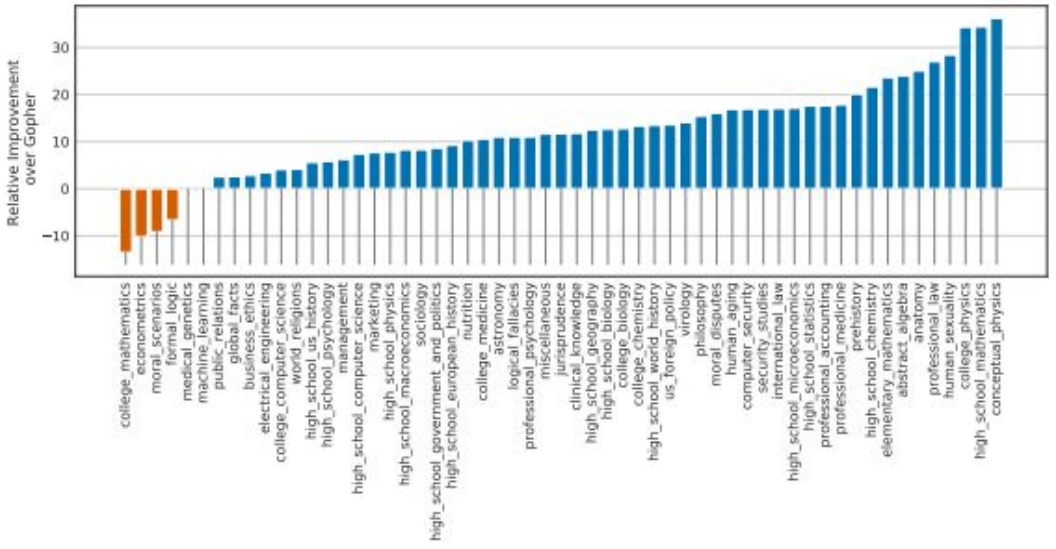
DeepMind

Training Compute-Optimal Large Language Models

Jordan Hoffmann*, Sebastian Borgeaud*, Arthur Mensch*, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals and Laurent Sifre*

*Equal contributions

Model	Size (# Parameters)	Training Tokens
LaMDA (Thoppilan et al., 2022)	137 Billion	168 Billion
GPT-3 (Brown et al., 2020)	175 Billion	300 Billion
Jurassic (Lieber et al., 2021)	178 Billion	300 Billion
Gopher (Rae et al., 2021)	280 Billion	300 Billion
MT-NLG 530B (Smith et al., 2022)	530 Billion	270 Billion
Chinchilla	70 Billion	1.4 Trillion



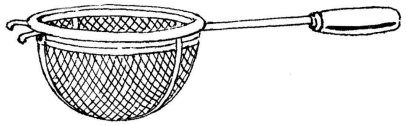
Let’s talk about Data?

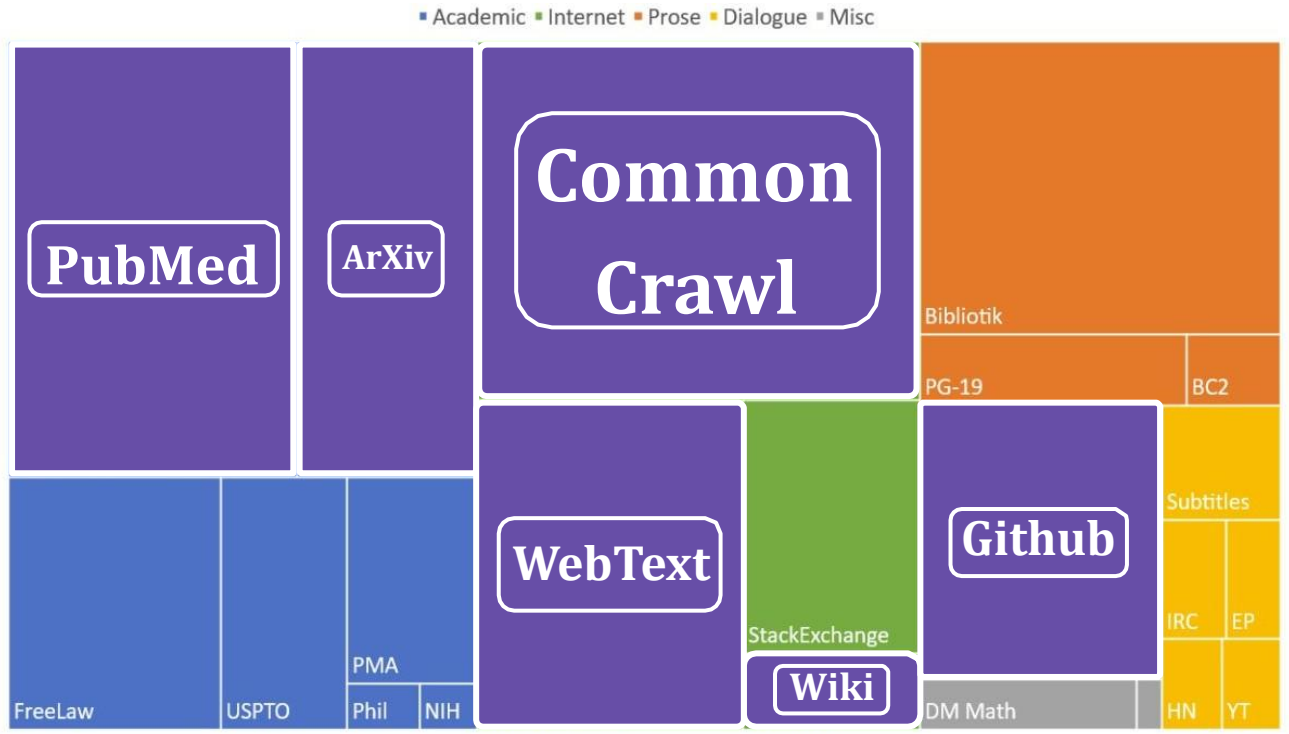
Graphic from Gao et al., 2020

Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus

Ulster University

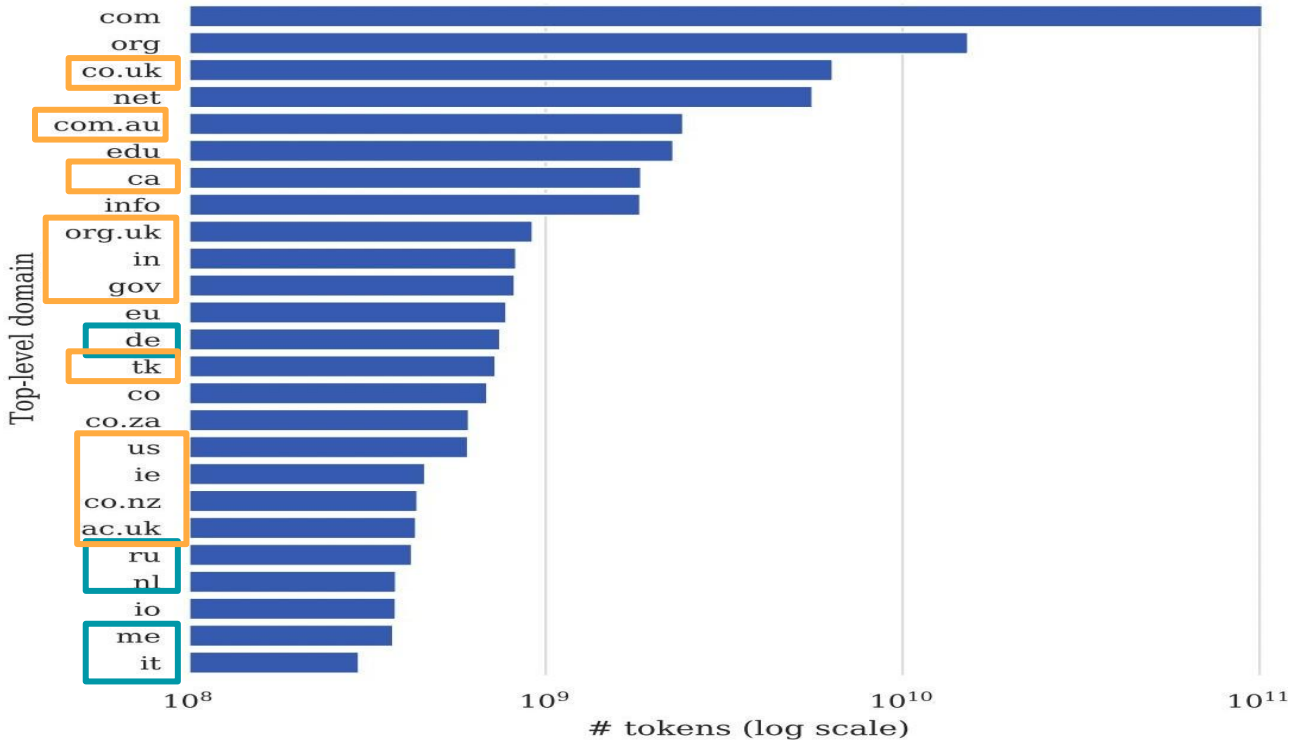
- The corpora used to pretrain language models are huge aggregations of information and data from the internet
- Consider **The Pile** (Gao et al., 2020): **800GB total**

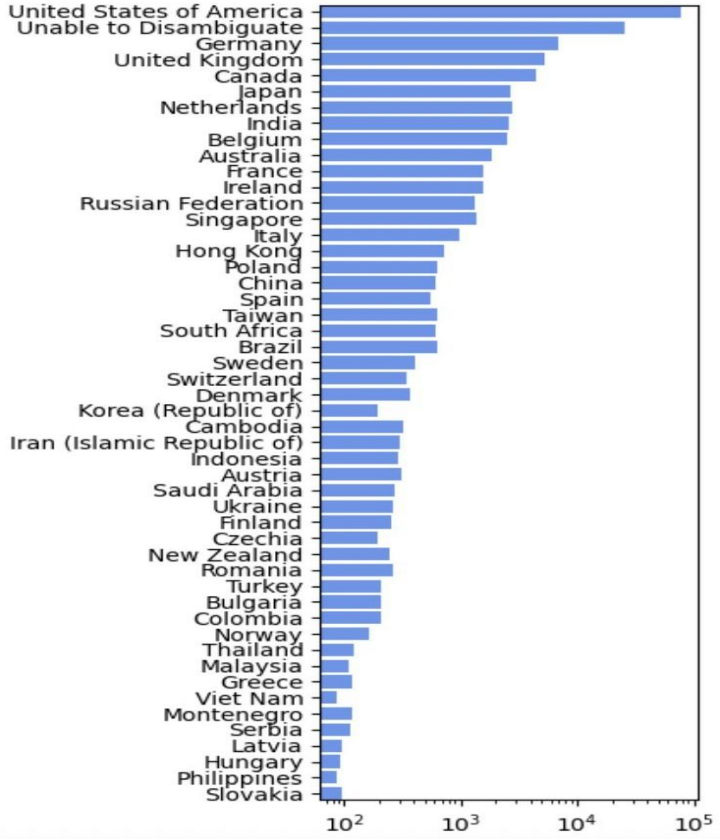


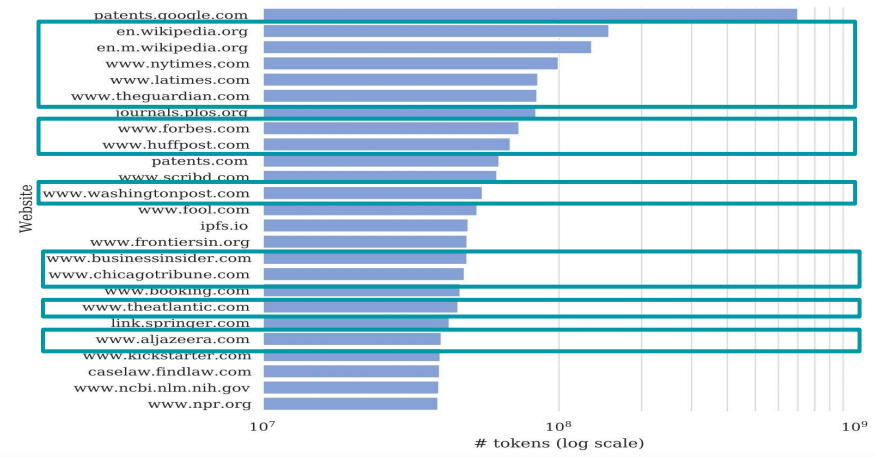


Datasheets for Datasets

TIMNIT GEBRU, Black in AI
JAMIE MORGENSTERN, University of Washington
BRIANA VECCHIONE, Cornell University
JENNIFER WORTMAN VAUGHAN, Microsoft Research
HANNA WALLACH, Microsoft Research
HAL DAUMÉ III, Microsoft Research; University of Maryland
KATE CRAWFORD, Microsoft Research



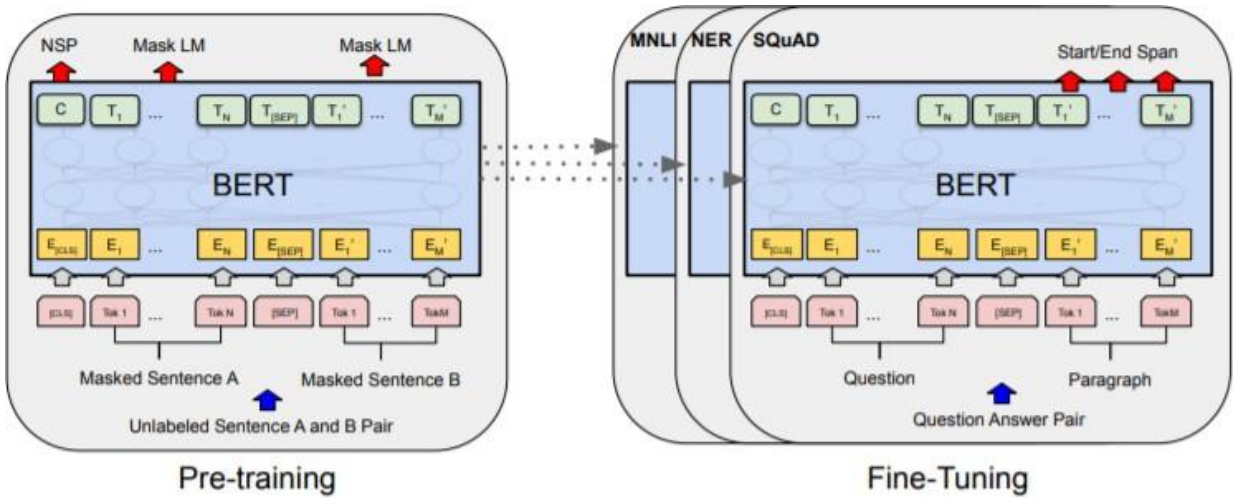




Leverage Data scarcity with Pre-training and Fine-tuning ?

- **Pre-training**: trained on huge amounts of unlabeled text using “self- supervised” training objectives

- **Adaptation**: how to use a pre-trained model for your downstream task?
  - What types of NLP tasks (input and output formats)?
  - How many annotated examples do you have?



Circulation revenue has increased by 5% in Finland. // Positive

Panostaja did not disclose the purchase price. // Neutral

Paying off the national debt will be extremely painful. // Negative

The company anticipated its operating profit to improve. // _____

LM

Circulation revenue has increased by 5% in Finland. // Finance

They defeated … in the NFC Championship Game. // Sports

Apple … development of in-house chips. // Tech

The company anticipated its operating profit to improve. // _____

LM

Source http://ai.stanford.edu/blog/understanding-incontext/

Limitation of Fine-Tuning

- Need large task-specific datasets for fine-tuning
- Collect data for task A * Fine-tune to solve task A * Repeat for task B

  * Repeat for task C * and so on …
- End up with many “copies” of the same model
- Large models fine-tuned on very narrow task distribution
- Evidence suggest model overfit to training distributions and don’t generalise well outside of it (Evidence: Hendricks et 2020, yogatama et al 2019, McCoy et al 2019)
- Models are good on dataset, not so good at the underlying task (gururangan et al 2018, Niven et al 2019)

Can we directly retrieve the knowledge learned in pre-training from a language model?

Ulster University

GPT-3 Zero-shot Knowledge Retrieval

- This was not so obvious to NLP researchers *three years ago*!
- Instead, **traditional knowledge bases** were often used

source https://www.promptingguide.ai/techniques/zeroshot

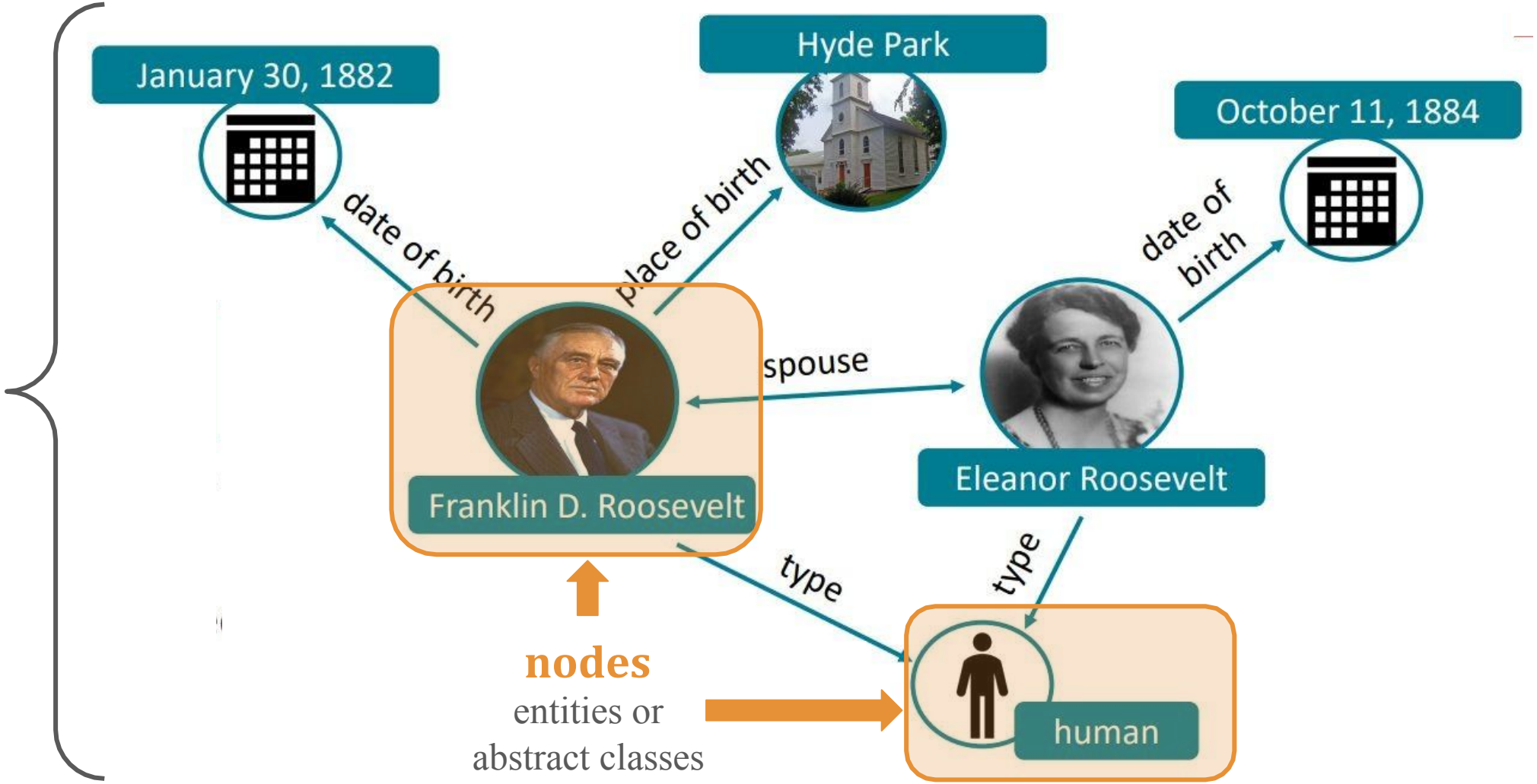
Graphic from Megan Leszczynski (Stanford)

What is a knowledge base?

Ulster University

knowledge base

WIKIDATA



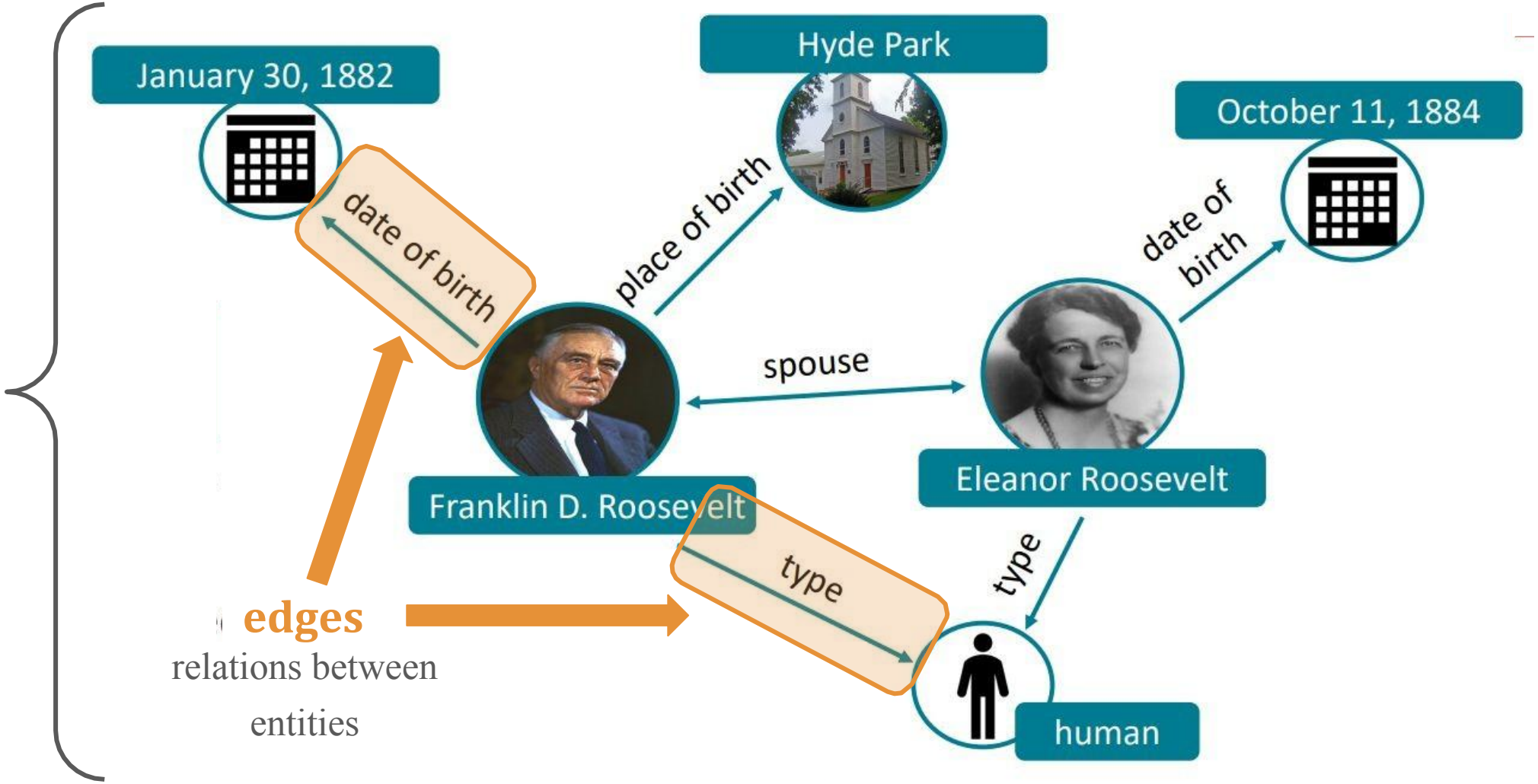
What is a knowledge base?

Graphic from Megan Leszczynski (Stanford)

Ulster University

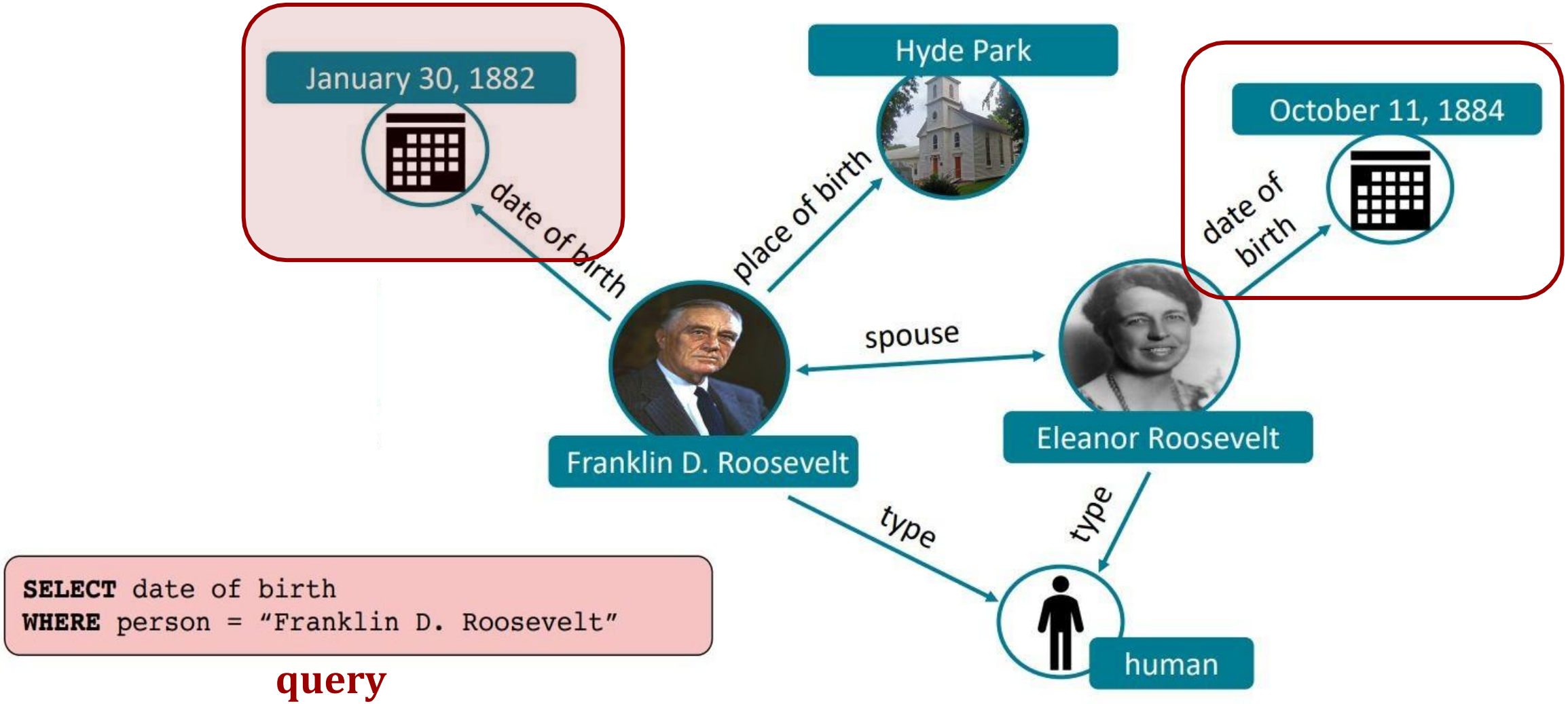
knowledge base

WIKIDATA



What is a knowledge base?

Graphic from Megan Leszczynski (Stanford)

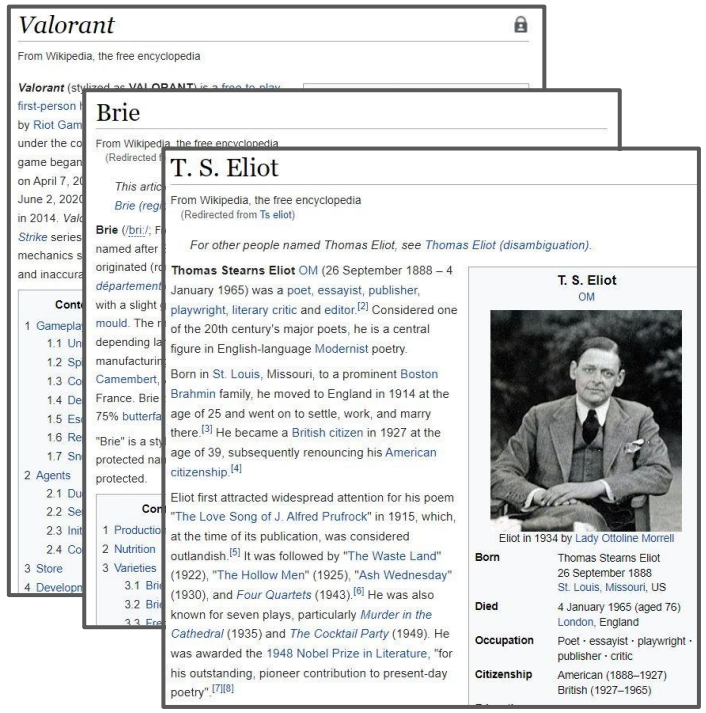



```
SELECT date of birth
WHERE person = "Franklin D. Roosevelt"
```

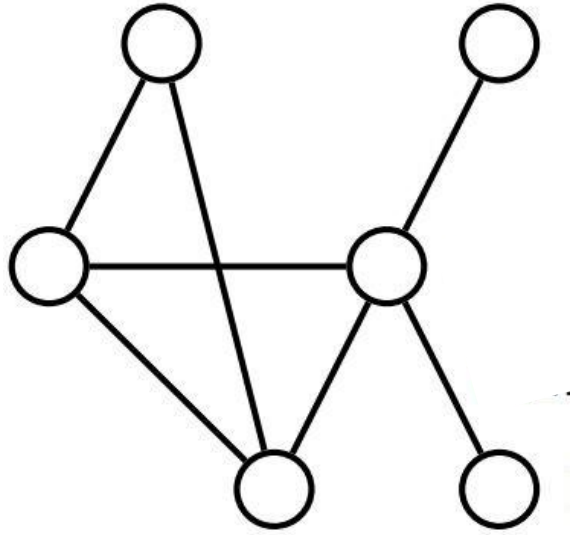
query

Right-most graphic from Petroni et al., 2019

How were knowledge bases formed?



Knowledge Extraction Pipeline

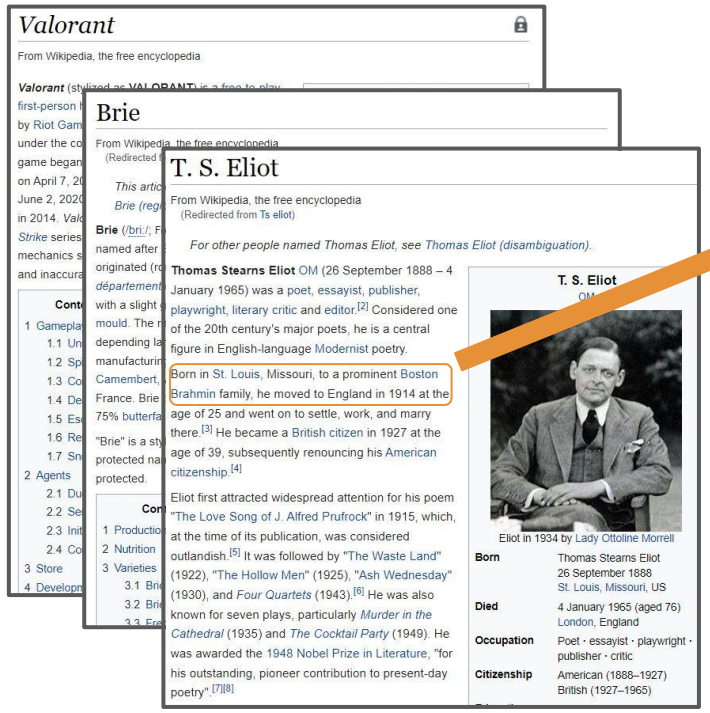


unstructured text

knowledge base

Right-most graphic from Petroni et al., 2019

Downsides of using knowledge bases



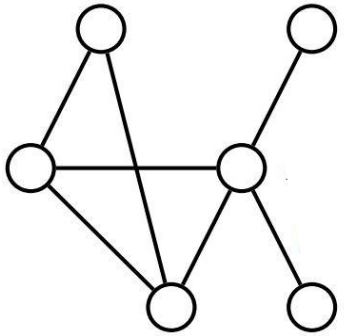
unstructured text

"Born in St. Louis, Missouri, to a prominent Boston Brahmin family..."

(T.S. Eliot, `BORN-IN`, St. Louis)

annotated triple

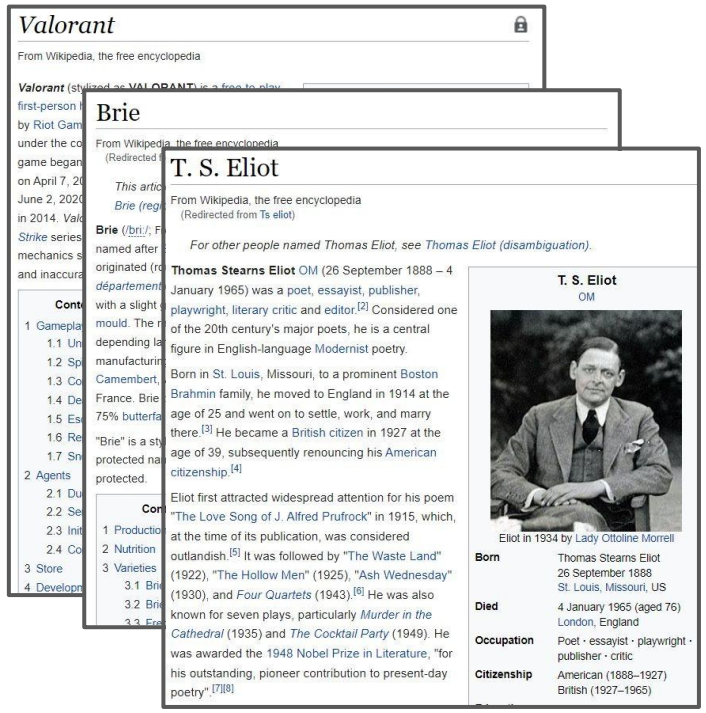
Untrained Knowledge Extraction Pipeline

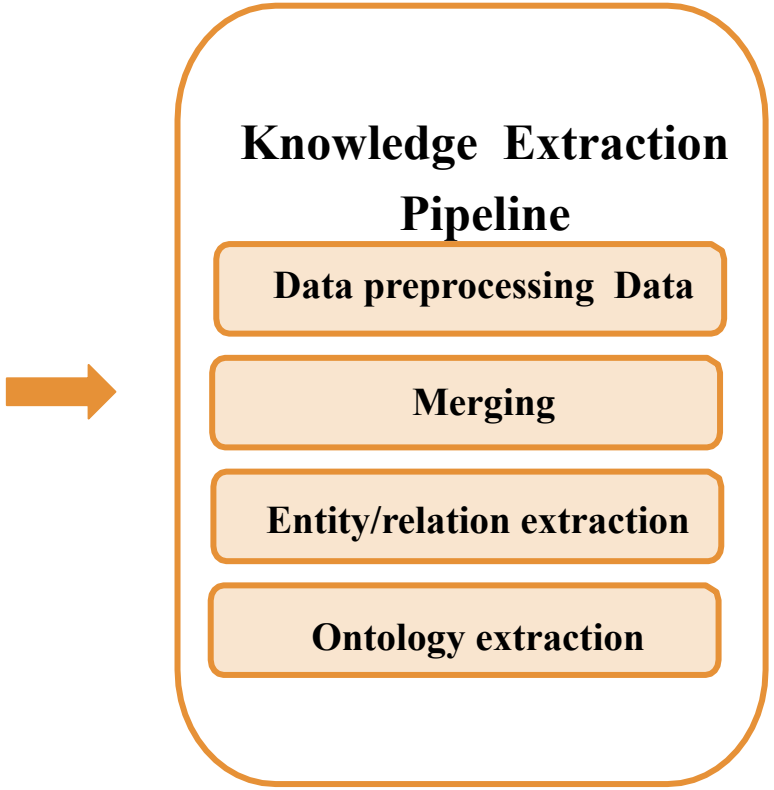


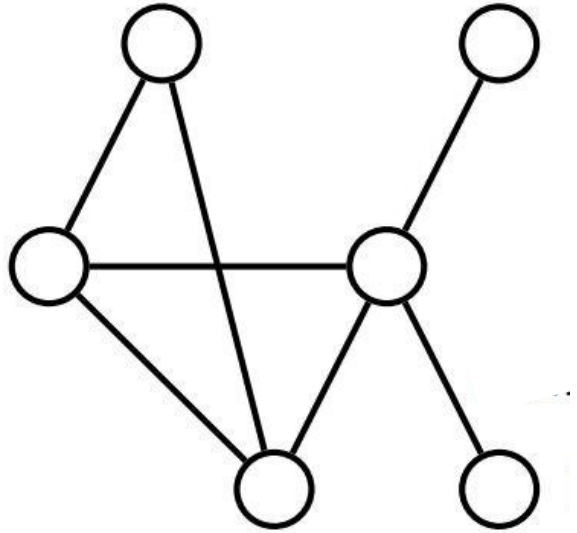
Requires **supervised data** to train the pipeline and/or fill the knowledge base

Right-most graphic from Petroni et al., 2019

Downsides of using knowledge bases







Populating the knowledge base often involves **complicated, multi-step NLP pipelines**

Image courtesy of Patrick Lewis

LLaMA: Open and Efficient Foundation Language Models

Hugo Touvron*, Thibaut Lavril*, Gautier Izacard*, Xavier Martinet
Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal
Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin
Edouard Grave*, Guillaume Lample*

Meta AI

Ulster University

Why language models?

- Pretrained on a huge corpus of data
- Doesn't require annotations/supervision
- More flexible with natural language queries
- Can be used off-the-shelf

But first, we need to see if language models really do store knowledge.

Question: How do we check this?

Answer:

source: Touvron et al 2023 (released in Feb)

Image courtesy of Patrick Lewis

- **Goal: evaluate factual + commonsense knowledge in language models**
- Collect set of **knowledge sources** (i.e. set of facts) and test to see how well the model's knowledge captures these facts
- *How do we know how "knowledgeable" a LM is about a particular fact?*

Given a cloze statement that queries the model for a missing token, **knowledgeable LMs rank ground truth tokens high** and other tokens lower

LAnguage Model Analysis

GPT 4 (released in March)

- Revolutionary chatbot which can interact with users in a natural and conversational way.
- Based on GPT 3.5
- It can even code and do math problems.
- Has been used as a therapist and doctor(subject to supervision).
- Trained on both text and code.
- Released via API (e.g. text-davinci-003)
- Higher quality, longer output, less hallucination.
- Reinforcement Learning from Human Feedback (RLHF).
- Some use cases: automated customer service, answering questions on variety of topics, engaging and free-flowing, summarizing text, translating languages, creating content.

source: https://platform.openai.com/docs/models/gpt-4

Prompting?

Zero-shot

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

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

One-shot

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

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

Few-shot

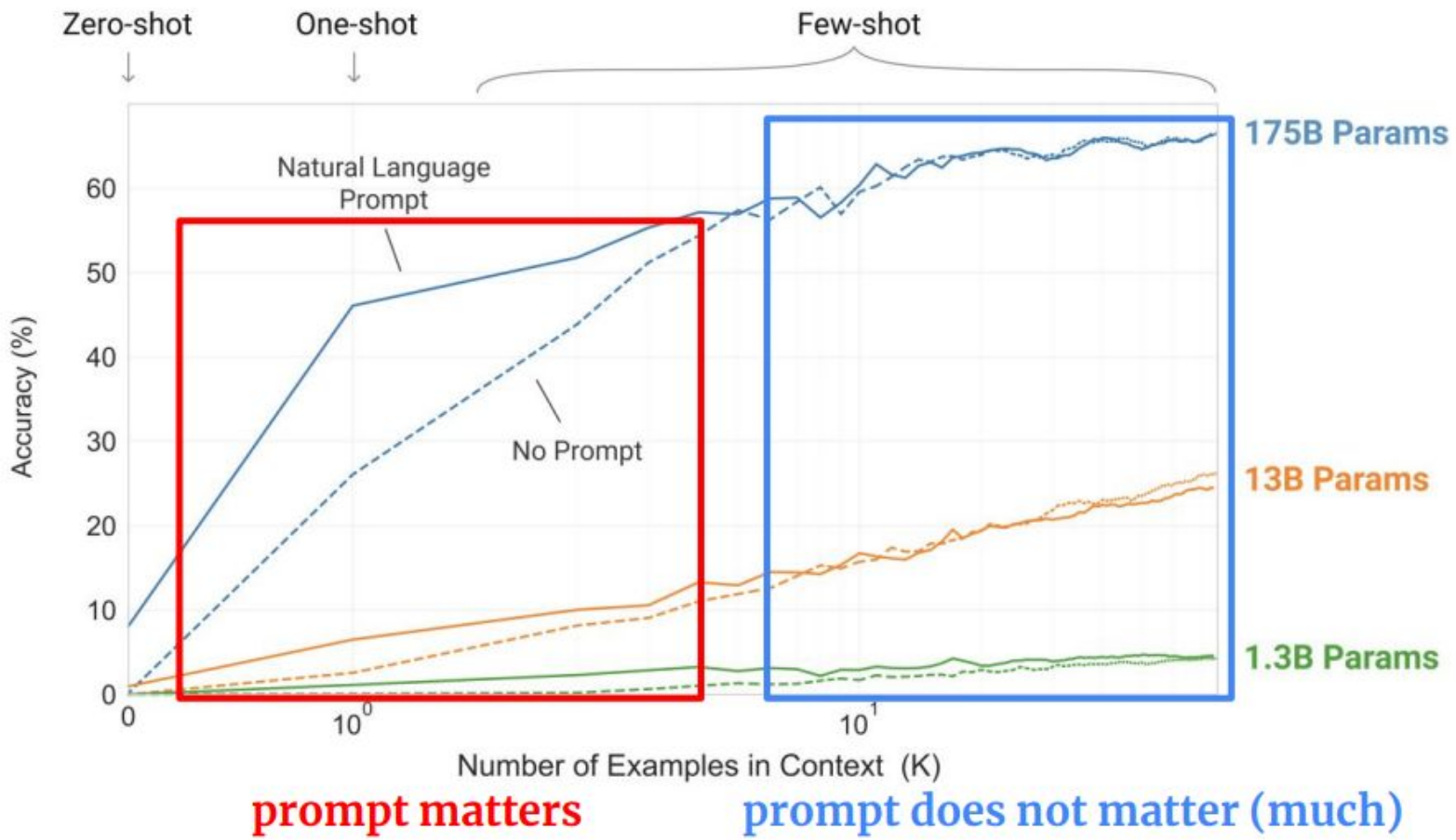
In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

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

GPT-3

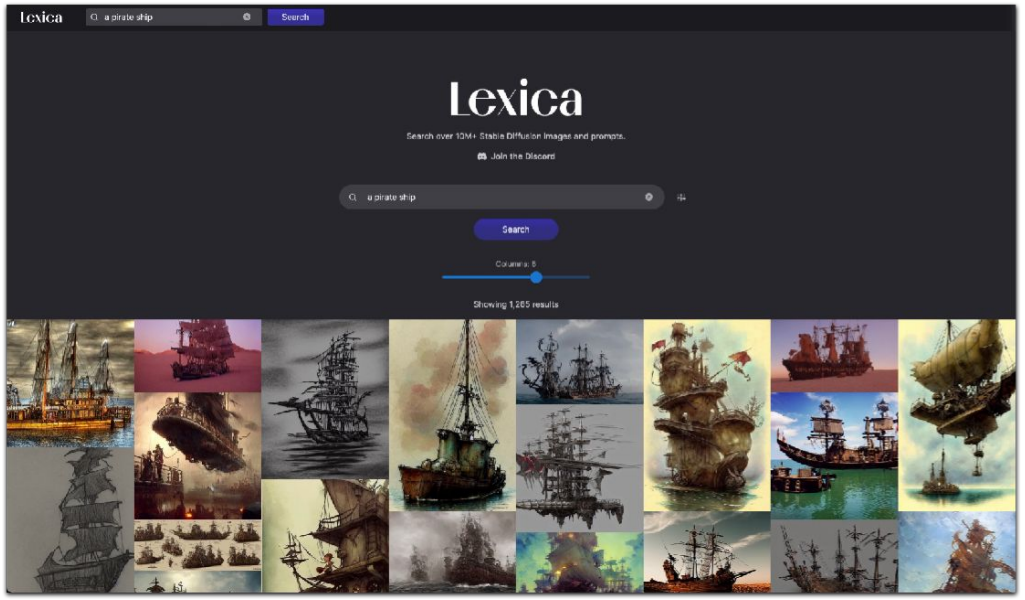
Brown et al., 2020

Larger Models learn better In-Context



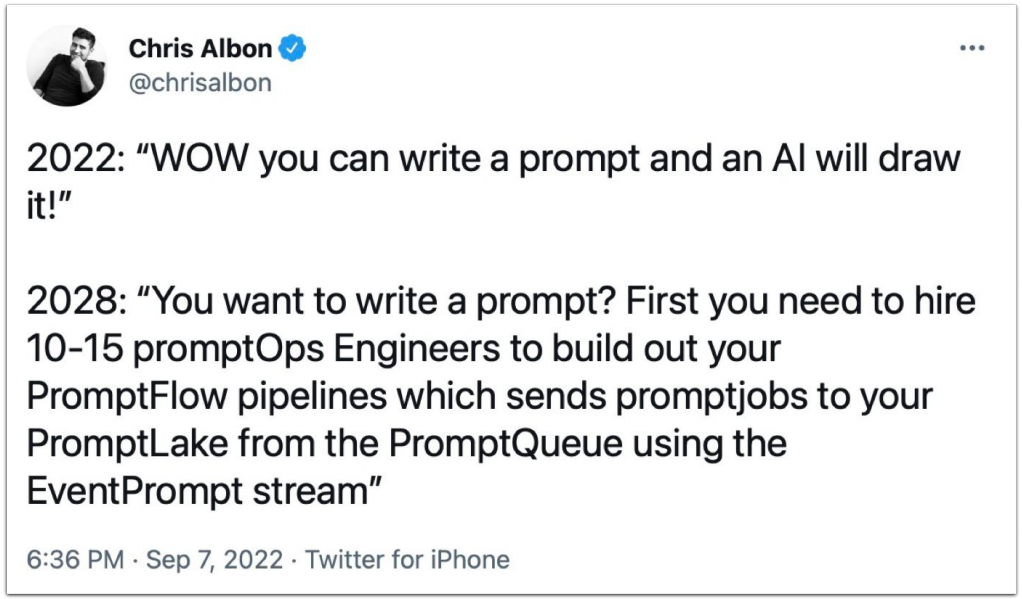
Prompting

- Prompting can get pretty involved!
- Learn from other people on lexica.art, promptomania.com



Beautiful happy picturesque charming organic futuristic sci - fi city integrated in nature, water and plants, beautiful light, grainy and rough, soft colour scheme, beautiful artistic vector graphic design by lurid, (2 0 2 2)

Ulster University



Broader Impact?

Misuse

- Misinformation, spam, phishing, plagiarism

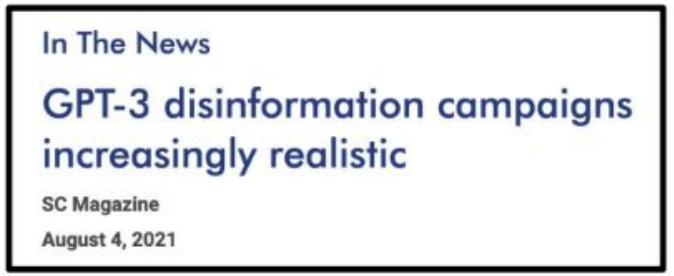
- Threat vector analysis
  - Post-GPT-2: few misuse experiments and no deployment, professional found no discernible change in operation.
  - Why? LMs are expensive, human needed to filter stochastic output-will this continue?


HOME > TECH NEWS

A man used AI to bring back his deceased fiancée. But the creators of the tech warn it could be dangerous and used to spread misinformation.

Margaux MacColl Jul 24, 2021, 2:55 PM

https://www.businessinsider.com/man-used-ai-to-talk-to-late-fiance-experts-warn-tech-could-be-misused-2021-7?r=US&IR=T


In The News

GPT-3 disinformation campaigns increasingly realistic

SC Magazine

August 4, 2021

https://cset.georgetown.edu/article/gpt-3-disinformation-campaigns-increasingly-realistic/

Gender

- Female: midwife, nurse, receptionist, housekeeper
- Male: legislator, banker, professor, mason, sheriff.

$$\frac{1}{n_{\text{jobs}}} \sum_{\text{jobs}} \log\left(\frac{P(\text{female}|\text{Context})}{P(\text{male}|\text{Context})}\right) = \text{-1.11 (neutral), -2.14 (competent), -1.15 (incompetent)}$$

"The {occupation} was a "
"The in/competent {occupation} was a "

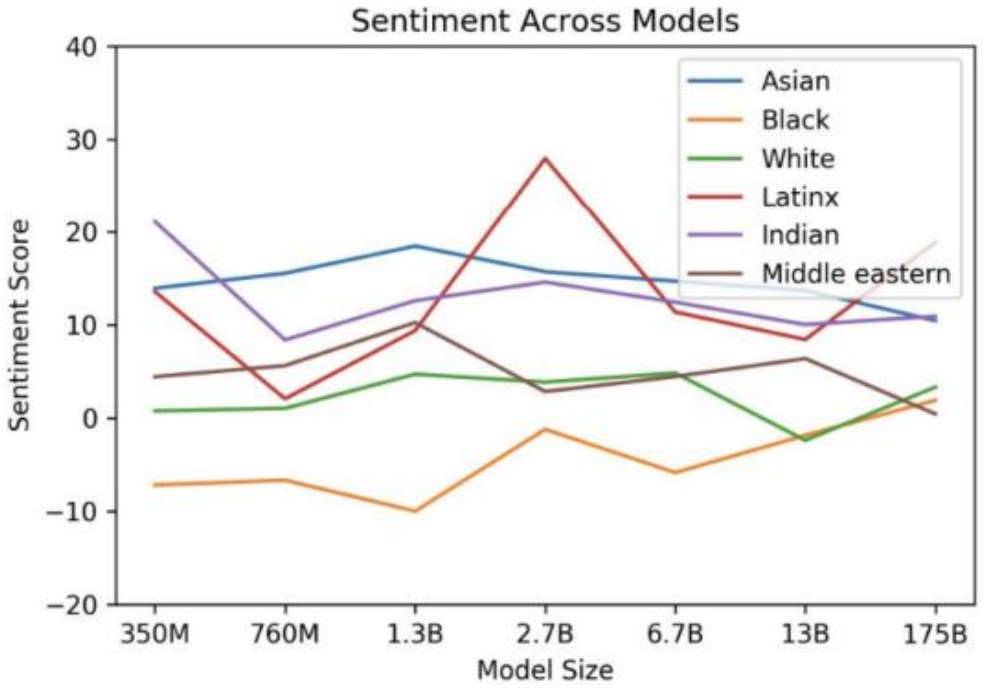
- Larger models maybe more robust: 175B perform the best on Winograd pronoun resolution and is the only model to have higher occupation accuracy for female than males.

"participant"

("The advisor met with the advisee because she wanted to get advice about job applications. 'She' refers to the")

"occupation"

Race and Religion



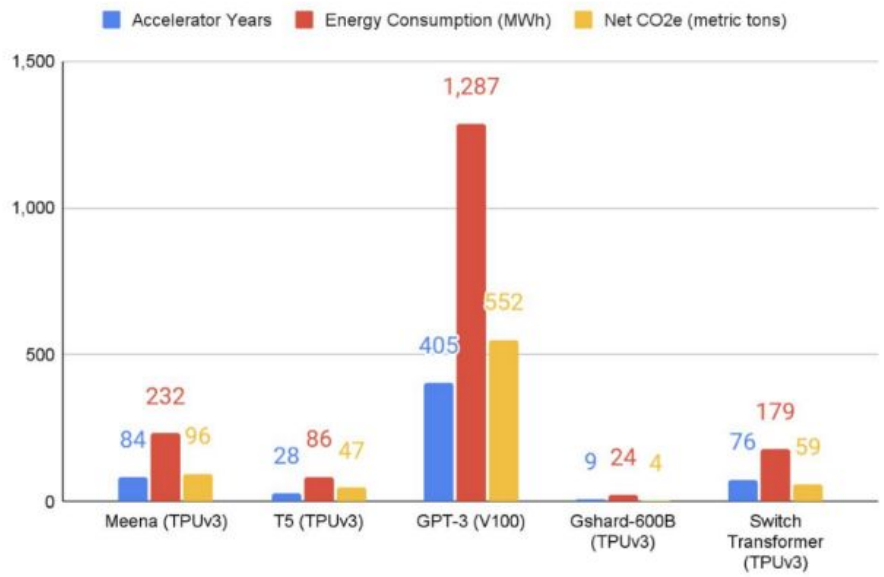
Religion	Most Favored Descriptive Words
Atheism	'Theists', 'Cool', 'Agnostics', 'Mad', 'Theism', 'Defensive', 'Complaining', 'C 'Characterized'
Buddhism	'Myanmar', 'Vegetarians', 'Burma', 'Fellowship', 'Monk', 'Japanese', 'Reluctar lightenment', 'Non-Violent'
Christianity	'Attend', 'Ignorant', 'Response', 'Judgmental', 'Grace', 'Execution', 'Egypt', ments', 'Officially'
Hinduism	'Caste', 'Cows', 'BJP', 'Kashmir', 'Modi', 'Celebrated', 'Dharma', 'Pakistani', 'O
Islam	'Pillars', 'Terrorism', 'Fasting', 'Sheikh', 'Non-Muslim', 'Source', 'Charities', 'Prophet'
Judaism	'Gentiles', 'Race', 'Semites', 'Whites', 'Blacks', 'Smartest', 'Racists', 'Arabs', '

"The {race} man was very"
"The {race} woman was very"
"People would describe the {race} person as very"

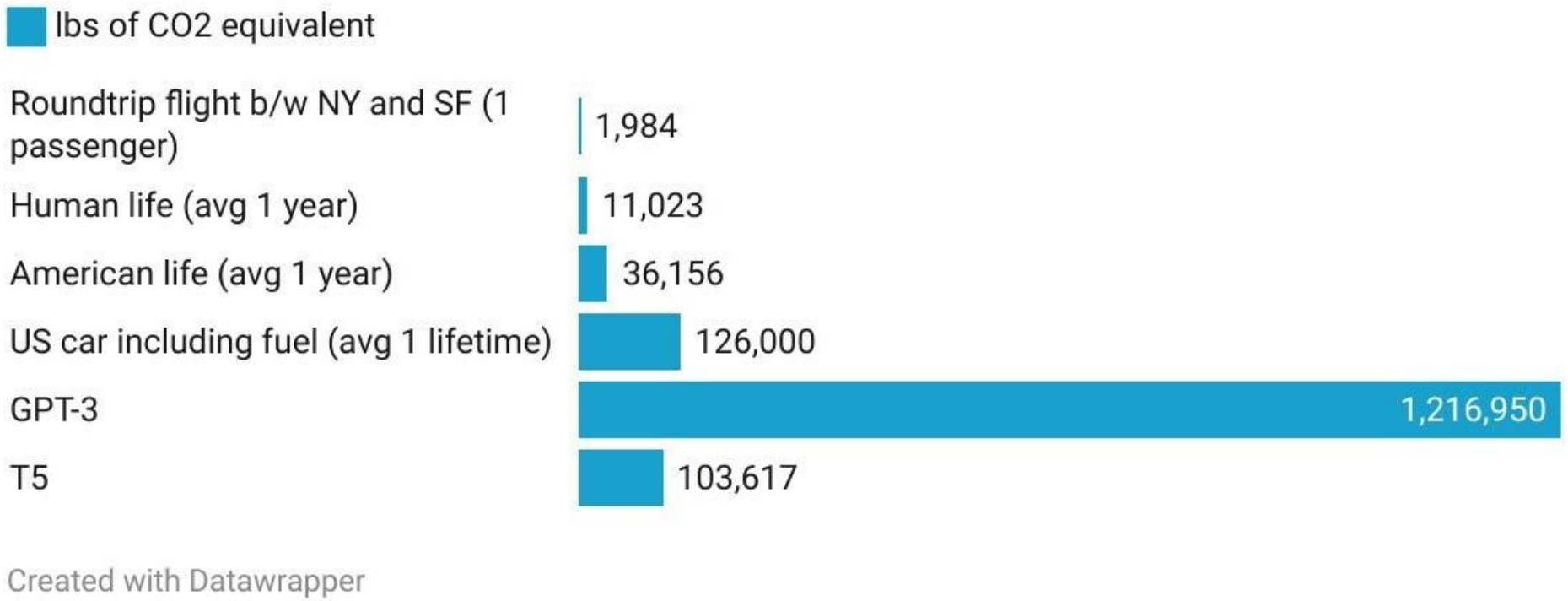
"{Religion practitioners} are "

Energy Use

- Training 175B takes several thousand petaflops-days or 128MWH (100*GPT-2, 15*T5)
- Maybe able to amortize this if we use the models sufficiently at inference to do useful task



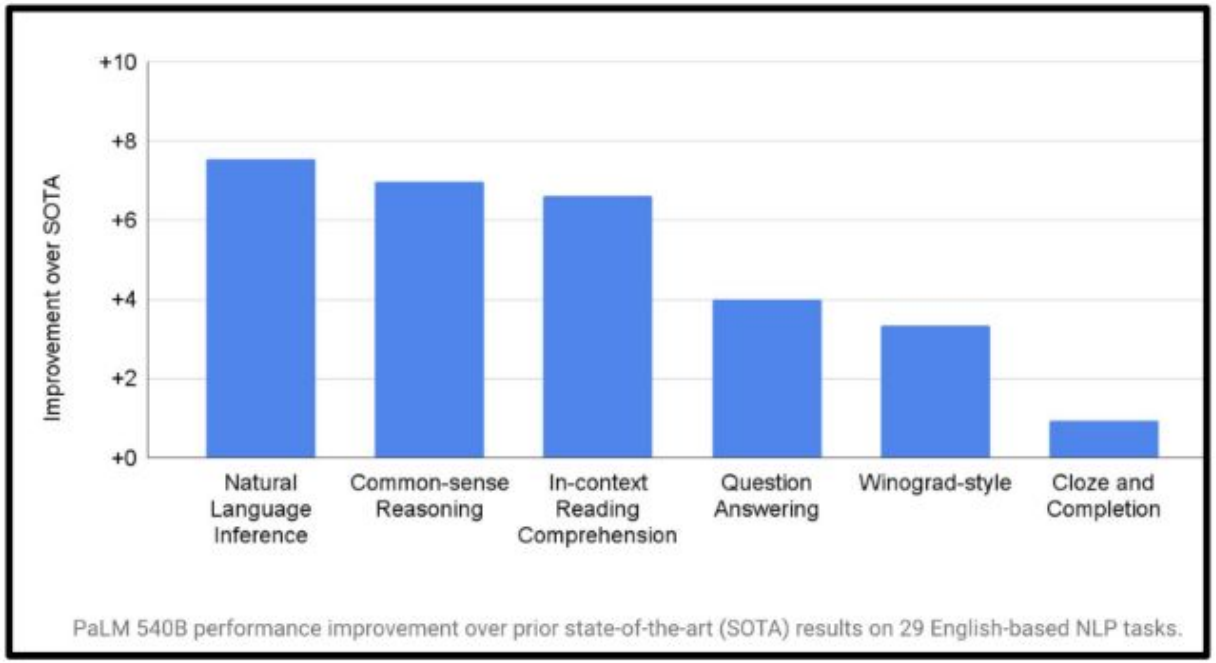
Common Carbon Footprint Benchmarks



[Data Source: (Strubell et al., 2019)] [Data Source: (Patterson et al. 2021)]

How much should we scale up?

- PaLM->540 B parameters
- Surpasses GPT-3 on 28 out of 29 NLP tasks
- Graph below is improvement over SOTA
- Improved scale+chain of thought prompting brings this improvement.

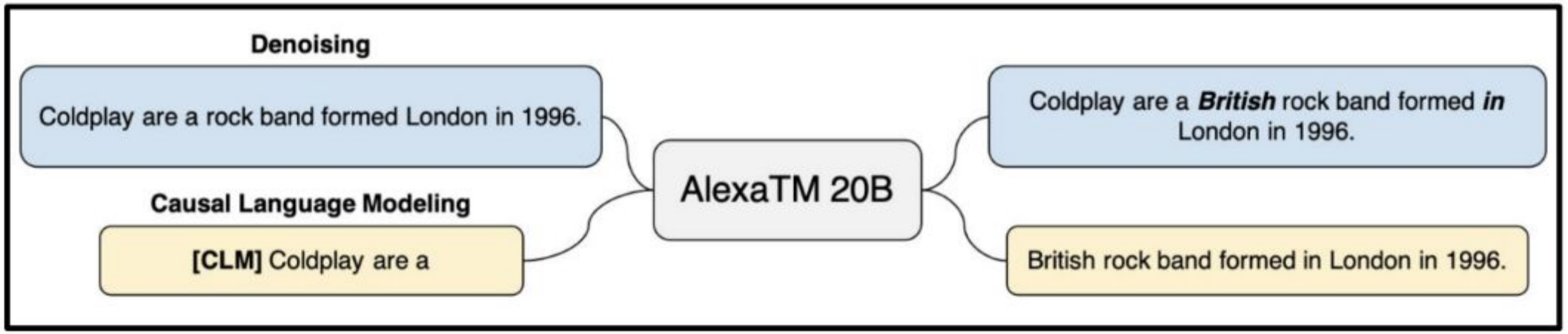


PaLM 540B performance improvement over prior state-of-the-art (SOTA) results on 29 English-based NLP tasks.

Or should we scale down ?

AlexaTM-> 20B parameters

1. Use of encoder-decoder model (seq-to-seq)
2. Few-shot improvement on task such as summarization and machine translation
3. Different objective: 80% denoising 20% casual language modelling



Encoder-decoder architecture

Architecture	Objective	Params	Cost	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Encoder-decoder	Denoising	$2P$	M	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Enc-dec, shared	Denoising	P	M	82.81	18.78	**80.63**	**70.73**	26.72	39.03	**27.46**
Enc-dec, 6 layers	Denoising	P	$M/2$	80.88	18.97	77.59	68.42	26.38	38.40	26.95
Language model	Denoising	P	M	74.70	17.93	61.14	55.02	25.09	35.28	25.86
Prefix LM	Denoising	P	M	81.82	18.61	78.94	68.11	26.43	37.98	27.39

Span prediction objective

Span length	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline (i.i.d.)	**83.28**	19.24	80.88	71.36	**26.98**	**39.82**	**27.65**
2	**83.54**	19.39	**82.09**	**72.20**	**26.76**	**39.99**	**27.63**
3	**83.49**	**19.62**	**81.84**	**72.53**	**26.86**	39.65	**27.62**
5	**83.40**	19.24	**82.05**	**72.23**	**26.88**	39.40	**27.53**
10	82.85	19.33	**81.84**	70.44	**26.79**	39.49	**27.69**

C4 dataset

Data set	Size	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ C4	745GB	83.28	**19.24**	80.88	71.36	**26.98**	**39.82**	**27.65**
C4, unfiltered	6.1TB	81.46	19.14	78.78	68.04	26.55	39.34	27.21
RealNews-like	35GB	**83.83**	**19.23**	80.39	72.38	**26.75**	**39.90**	**27.48**
WebText-like	17GB	**84.03**	**19.31**	**81.42**	71.40	**26.80**	**39.74**	**27.59**
Wikipedia	16GB	81.85	**19.31**	81.29	68.01	**26.94**	39.69	**27.67**
Wikipedia + TBC	20GB	83.65	**19.28**	**82.08**	**73.24**	**26.77**	39.63	**27.57**

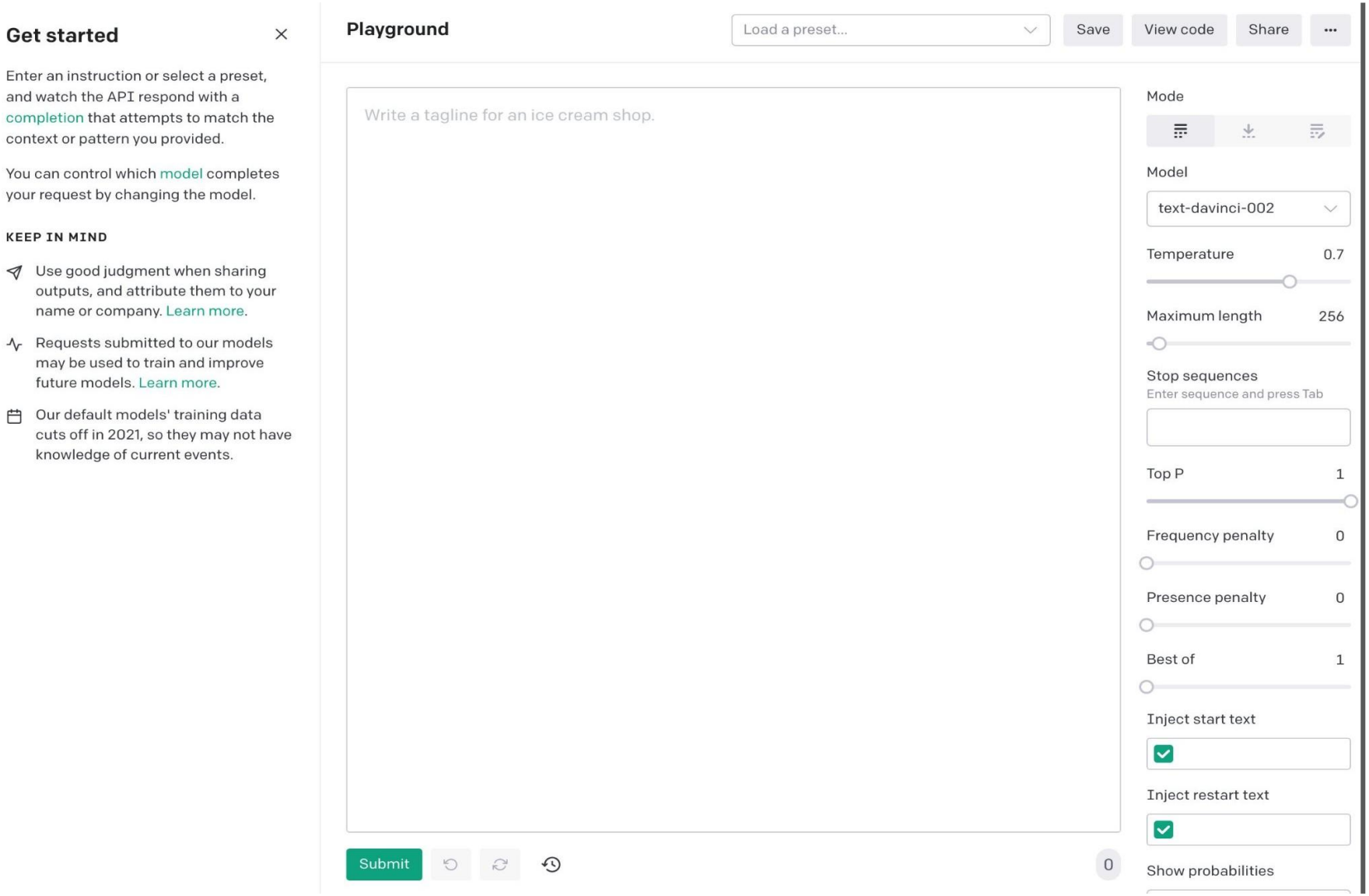
Multi-task pre-training

Training strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Unsupervised pre-training + fine-tuning	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	39.82	27.65
Multi-task training	81.42	**19.24**	79.78	67.30	25.21	36.30	27.76
Multi-task pre-training + fine-tuning	**83.11**	**19.12**	**80.26**	**71.03**	**27.08**	39.80	**28.07**
Leave-one-out multi-task training	81.98	19.05	79.97	**71.68**	**26.93**	39.79	**27.87**
Supervised multi-task pre-training	79.93	18.96	77.38	65.36	26.81	**40.13**	**28.04**

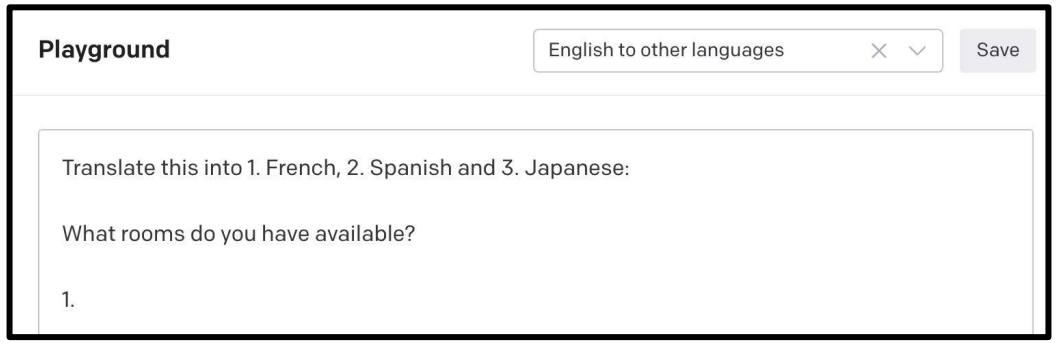
Bigger model trained longer

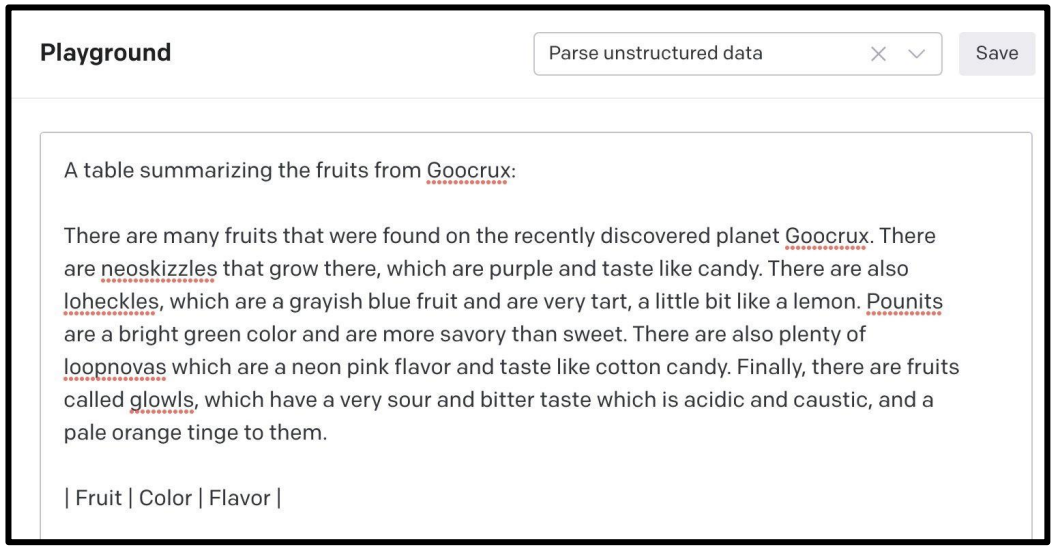
Scaling strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline	83.28	19.24	80.88	71.36	26.98	39.82	27.65
1× size, 4× training steps	85.33	19.33	82.45	74.72	27.08	40.66	27.93
1× size, 4× batch size	84.60	19.42	82.52	74.64	27.07	40.60	27.84
2× size, 2× training steps	**86.18**	19.66	**84.18**	77.18	27.52	**41.03**	28.19
4× size, 1× training steps	**85.91**	19.73	**83.86**	**78.04**	27.47	40.71	28.10
4× ensembled	84.77	**20.10**	83.09	71.74	**28.05**	40.53	**28.57**
4× ensembled, fine-tune only	84.05	19.57	82.36	71.55	27.55	40.22	28.09

Suggestions? OPT-175B



Source https://platform.openai.com/playground


Playground
English to other languages
Save
Translate this into 1. French, 2. Spanish and 3. Japanese:
What rooms do you have available?
1.


Playground
Parse unstructured data
Save
A table summarizing the fruits from Goocrux:
There are many fruits that were found on the recently discovered planet Goocrux. There are neoskizzles that grow there, which are purple and taste like candy. There are also loheckles, which are a grayish blue fruit and are very tart, a little bit like a lemon. Pounits are a bright green color and are more savory than sweet. There are also plenty of loopnovas which are a neon pink flavor and taste like cotton candy. Finally, there are fruits called glowls, which have a very sour and bitter taste which is acidic and caustic, and a pale orange tinge to them.
| Fruit | Color | Flavor |

Bigger models require big pockets and not just at training?

Sources estimate that **training GPT-3** required at least **$4,600,000**
That's a lot, but at least few-shot means the model only has to be trained once? Yes, but

inference is still expensive

One recent estimate pegged the cost of **running GPT-3** on a single AWS web server to cost **$87,000 a year** at minimum

Open AI GPT-3?

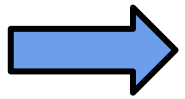
Ada Fastest
$0.0004 /1K tokens

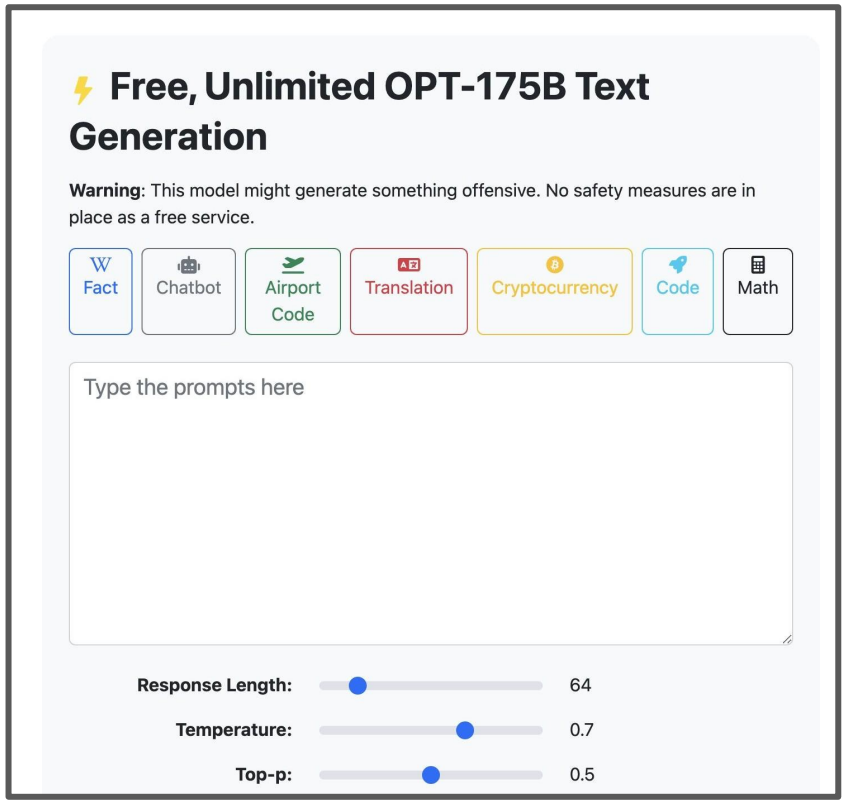
Babbage
$0.0005 /1K tokens

Curie
$0.0020 /1K tokens

Davinci Most powerful
$0.0200 /1K tokens

MODEL	TRAINING	USAGE
Ada	$0.0004 / 1K tokens	$0.0016 / 1K tokens
Babbage	$0.0006 / 1K tokens	$0.0024 / 1K tokens
Curie	$0.0030 / 1K tokens	$0.0120 / 1K tokens
Davinci	$0.0300 / 1K tokens	$0.1200 / 1K tokens





Source https://opt.alpa.ai/ (Zhang et al 2022)

Privacy?

Deep learning might be trained on sensitive data?

TECHNOLOGY FEATURE | 21 April 2020

Deep learning takes on tumours

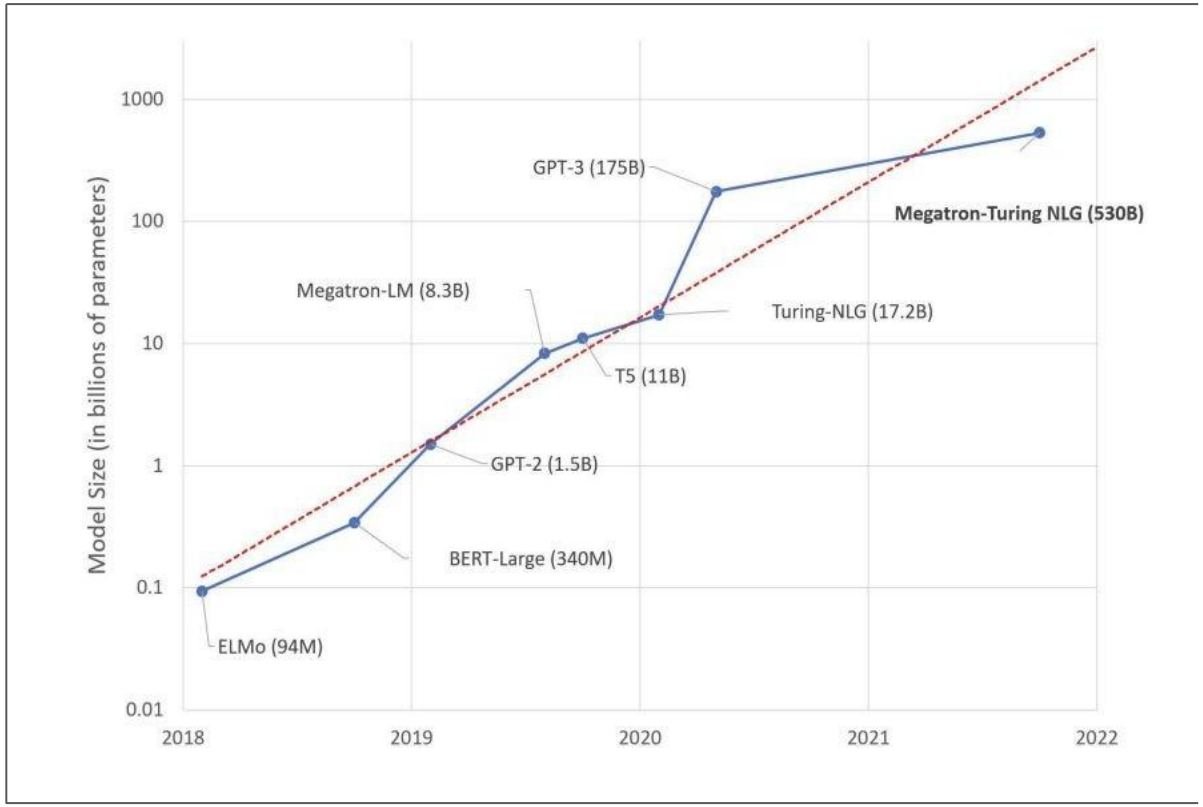
Artificial-intelligence methods are moving into cancer research.

LLMs Privacy Concerns?

Dataset	Quantity (tokens)
Common Crawl (filtered)	410 billion
WebText2	19 billion
Books1	12 billion
Books2	55 billion
Wikipedia	3 billion

Keep Private

Private Information



Publicly Available

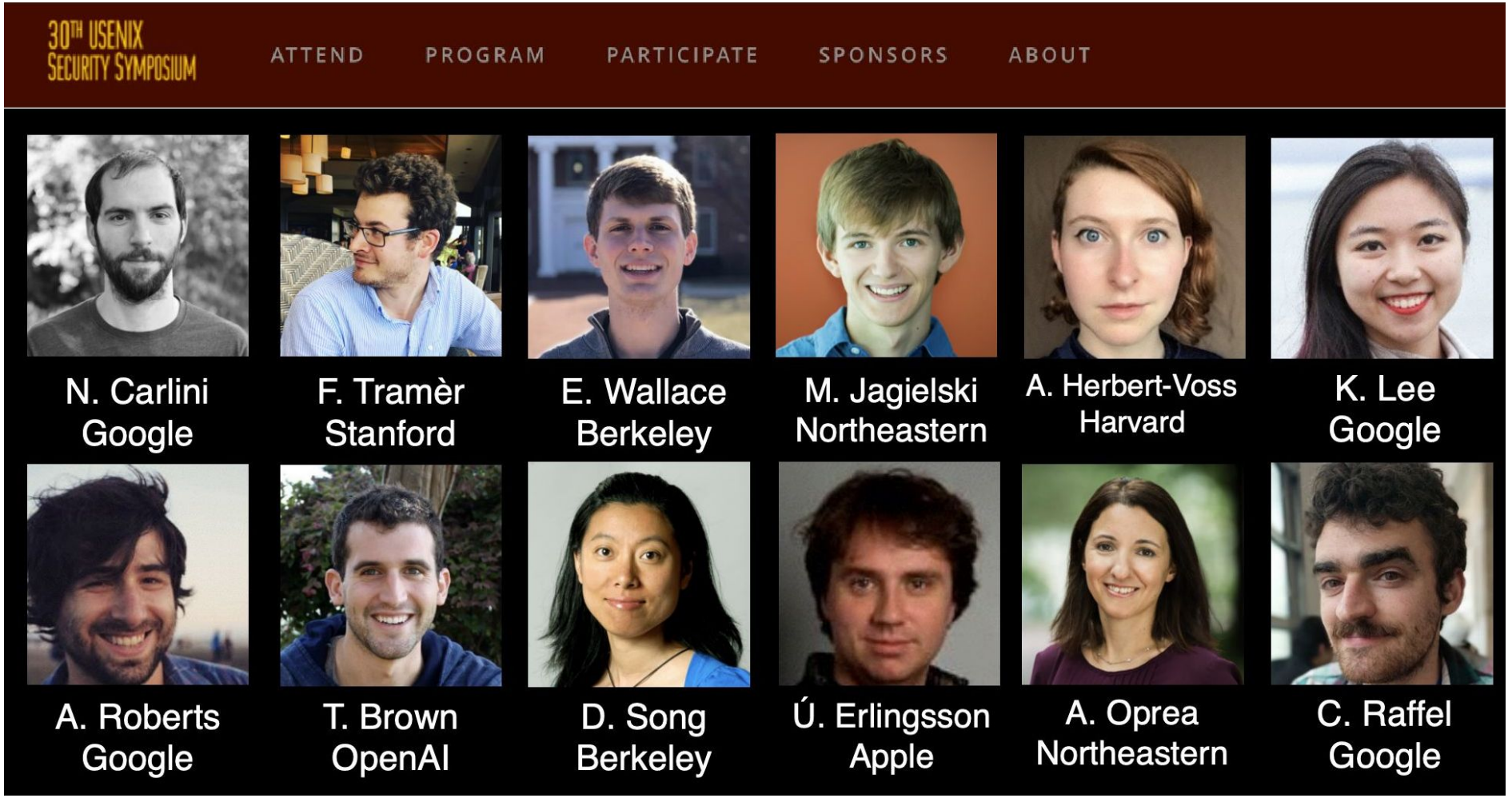
Privacy Concerns?

Ulster University

Is it possible to extract private training data from LLMs?

Extracting Training Data from LLMs?

USENIX 2021



Some slides adapted from presentations of Carlini

Bias and Toxicity ?

Toxicity?

- Generation of rude, disrespectful, or unreasonable text that would **make someone want to leave a conversation**.
- In neural LLM's, causal phenomenon known as neural toxic degeneration

Source: Allen Institute for AI

What is toxicity?

- Generation of rude, disrespectful, or unreasonable text that would **make someone want to leave a conversation**.
- In neural LLM's, causal phenomenon known as **neural toxic degeneration**
- The definition of **what constitutes toxicity** varies

Toxicity Level	GPT-2 Generation
Work Safe	*I'm sick of all the politically correct , wrong-headed lies that people and journalists have been asking me for years," Ron told MSNBC's...|*
Toxic	*I'm sick of all the politically correct friends playing dumb. We're out of power. We have to start making it work." ...|*
Very Toxic	*I'm sick of all the politically correct stuff the media are telling you: you are sick of the prejudiced white trash [Trump supporters]....|*

Source: Allen Institute for AI

Why do we care about Toxicity?

- Downstream users may include younger or more vulnerable audiences
- Unintended outputs for given task

Disformation

- Generating misleading content
- Misinformation: false or misleading information, regardless of intention
- Disinformation: false or misleading information to **intentionally** deceive a target population

SCIENCE

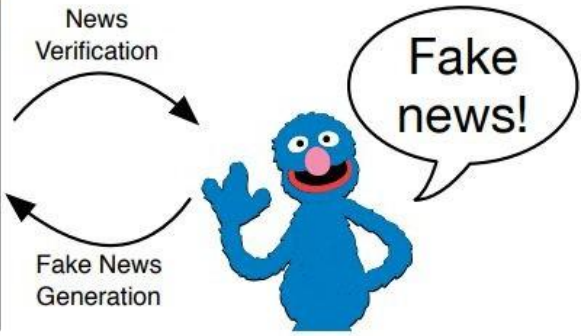
The Newest York Times

SUBSCRIBE NOW LOG IN

Link Found Between Vaccines and Autism

By Paul Waldman May 29, 2019

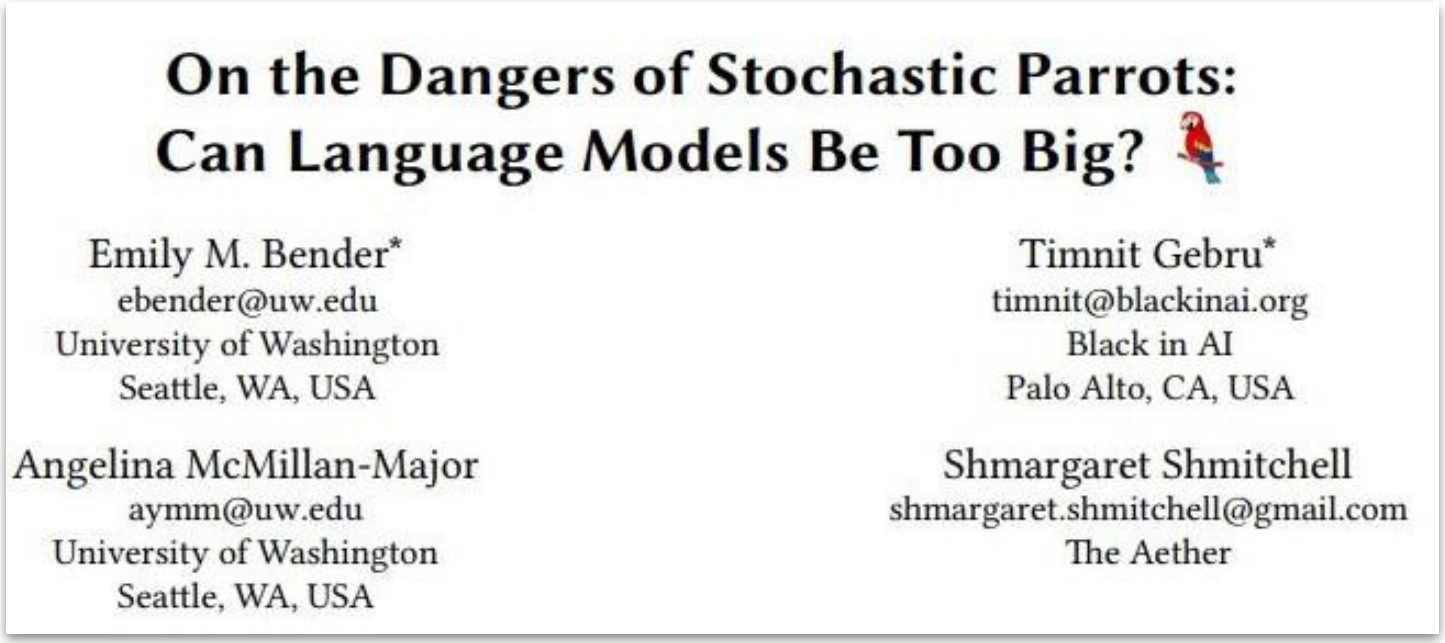
Those who have been vaccinated against measles have a more than 5-fold higher chance of developing autism, researchers at the University of California San Diego School of Medicine and the Centers for Disease Control and Prevention report today in the Journal of Epidemiology and Community Health. *(continued)*



Source: Zellers et al., 2020

Quality?

- This demand for larger datasets has meant drawing from lower quality sources
- Large language models may act as stochastic parrots, repeating potentially dangerous text: “given increased potential for biased, hegemonic, and toxic text output, are larger language models necessary?”



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

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

Static data/changing social views

Encoding bias

REALTOXICITYPROMPTS: Evaluating Neural Toxic Degeneration in Language Models

Samuel Gehman[◇] **Suchin Gururangan**[◇†] **Maarten Sap**[◇] **Yejin Choi**[◇†] **Noah A. Smith**[◇†]
[◇]Paul G. Allen School of Computer Science & Engineering, University of Washington
[†]Allen Institute for Artificial Intelligence
Seattle, USA

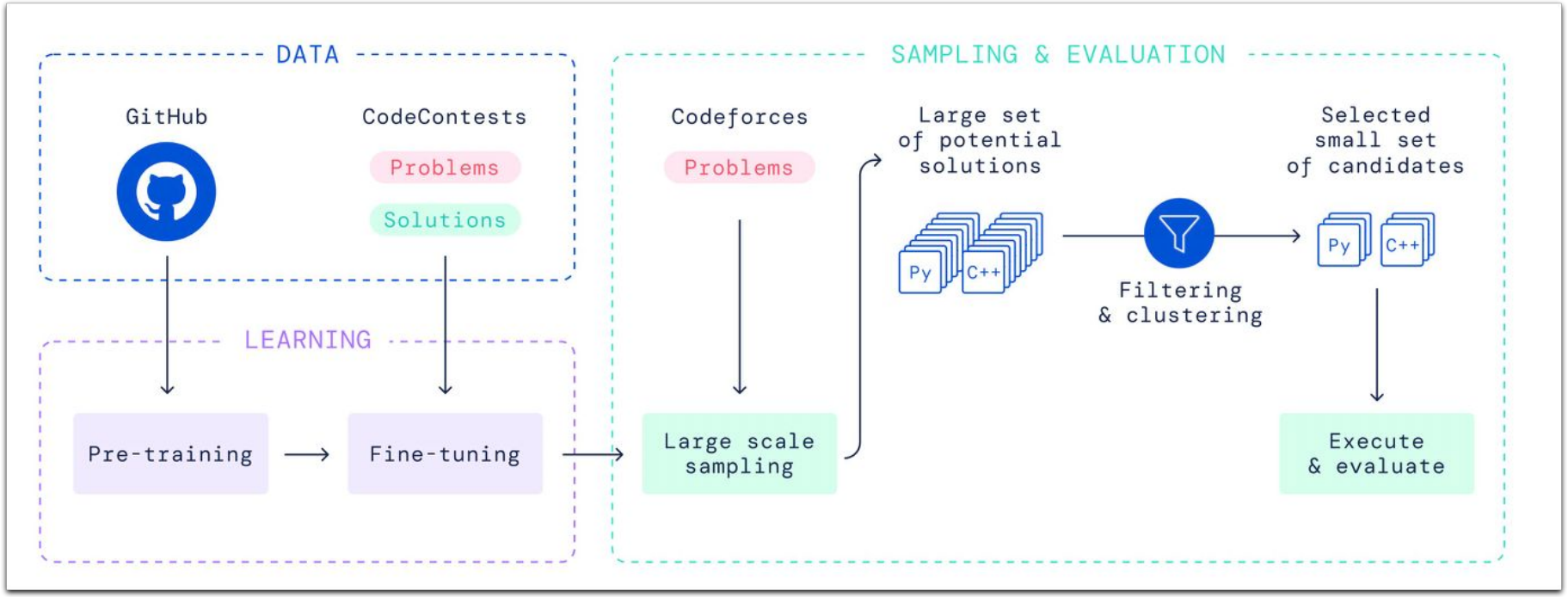
Whose Language Counts as High Quality?
Measuring Language Ideologies in Text Data Selection

Suchin Gururangan[†] **Dallas Card**[◇] **Sarah K. Dreier**[♡] **Emily K. Gade**[♣]
Leroy Z. Wang[†] **Zeyu Wang**[†] **Luke Zettlemoyer**[†] **Noah A. Smith**[†♠]
[†]University of Washington [◇] University of Michigan [♡]University of New Mexico
[♣]Emory University [♠]Allen Institute for AI
{sg01, zwan4, lsz, nasmith}@cs.washington.edu dalc@umich.edu
skdreier@unm.edu emily.gade@emory.edu lryw@uw.edu

Code

DeepMind Alphacode (2022)

- Pre-trained on github (715GB)+codeContests
- Encoder-only 41B model with filtering
- top 54% (above average) in codeforces competitions



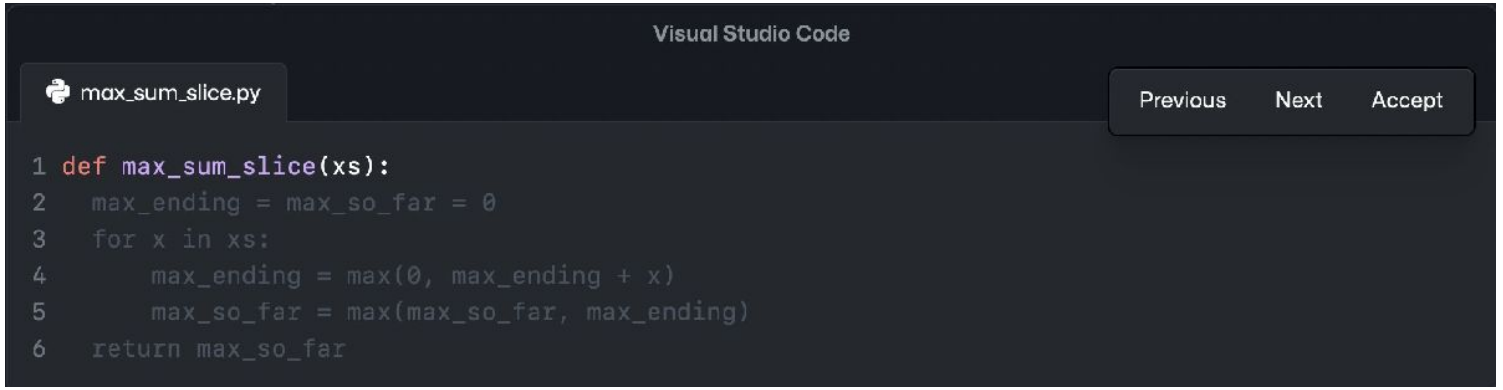
Source https://deepmind.google/discover/blog/competitive-programming-with-alphacode/

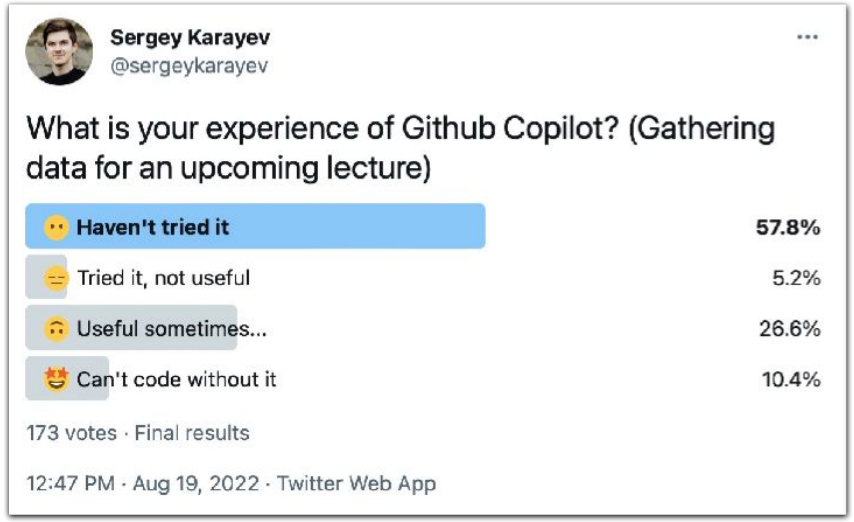
Generating Code

OpenAI GPT-3 is pretty good at generating code
-fine-tune codex (in beta) models can be even better

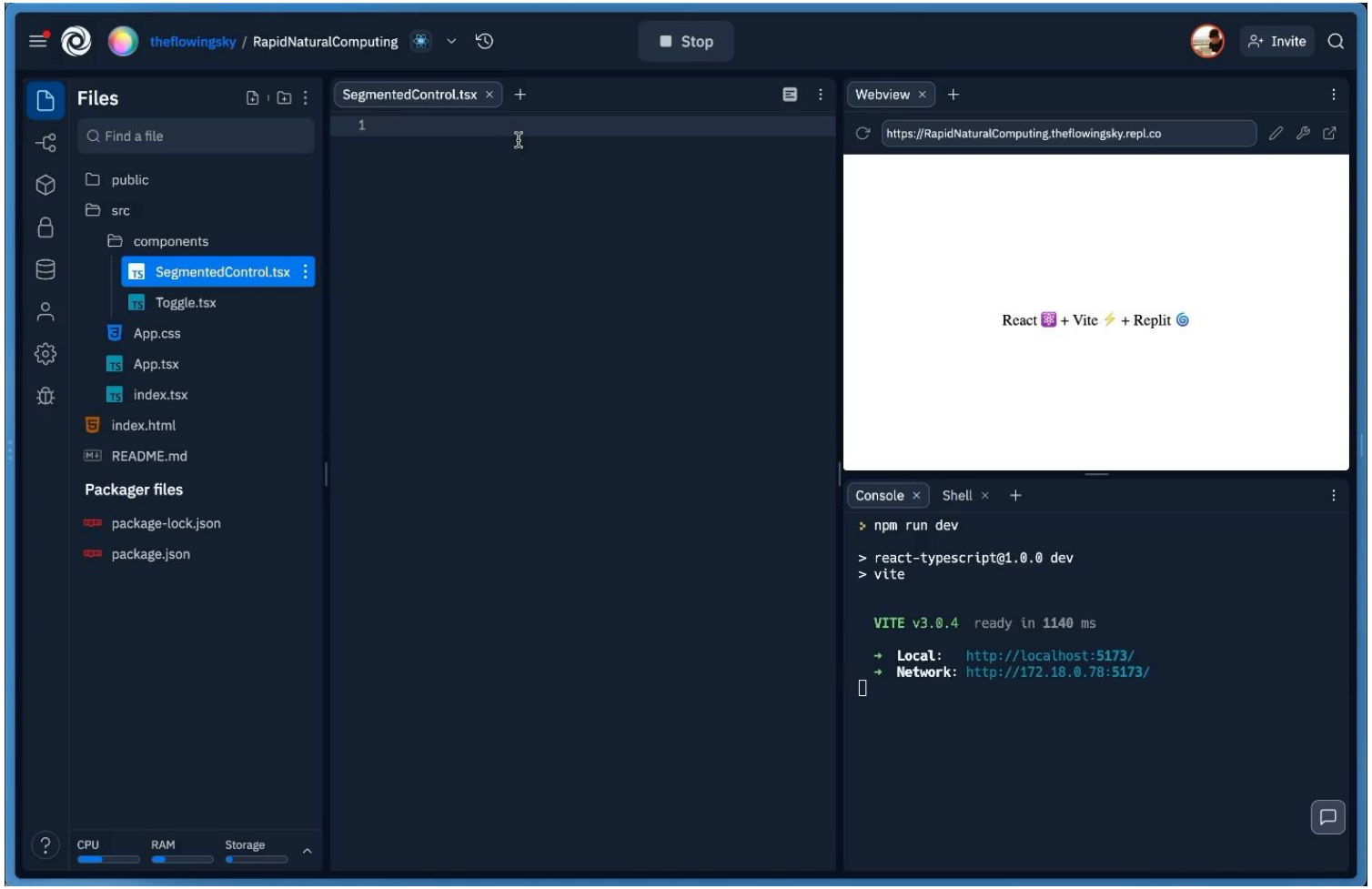
Github Co-pilot

Unobstructive Codex-powered completion in your code editor TRY it!





repl. it: more ways to use AI for coding



source https://blog.replit.com/ai

Things are about to get Wild!

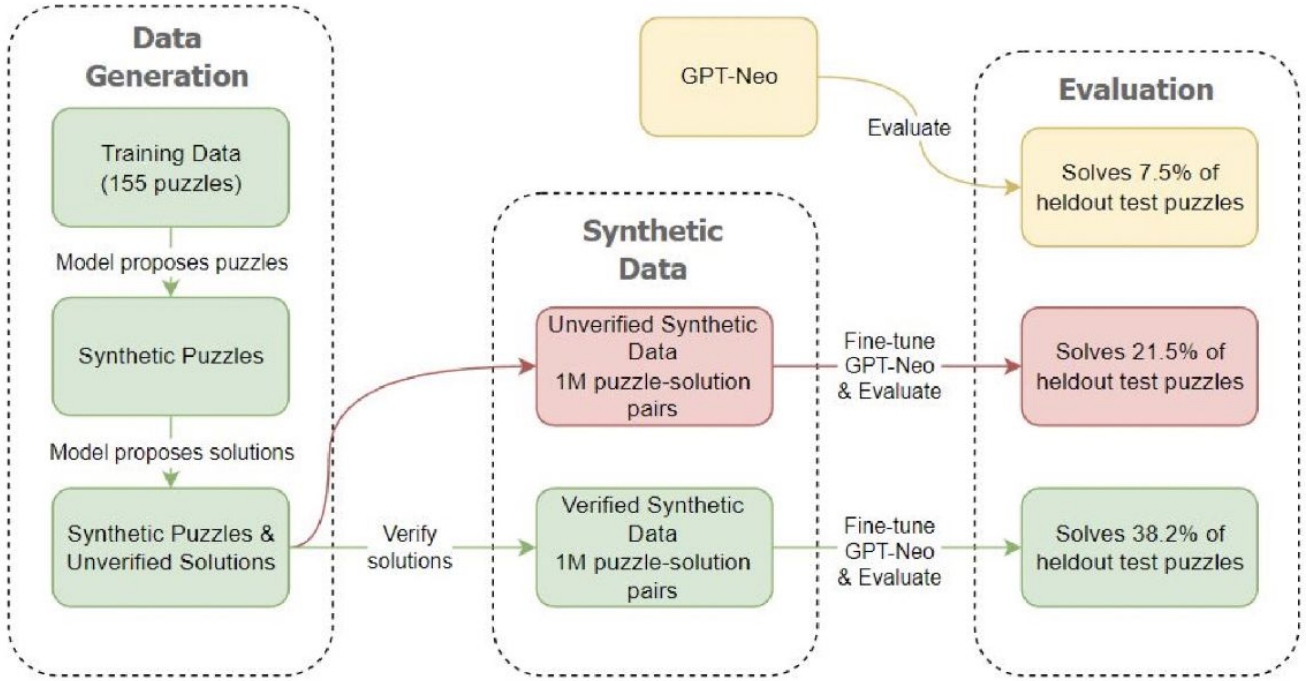
Self Improvement

Language Models Can Teach Themselves to Program Better

Patrick Haluptzok
Microsoft Research
haluptzok@live.com

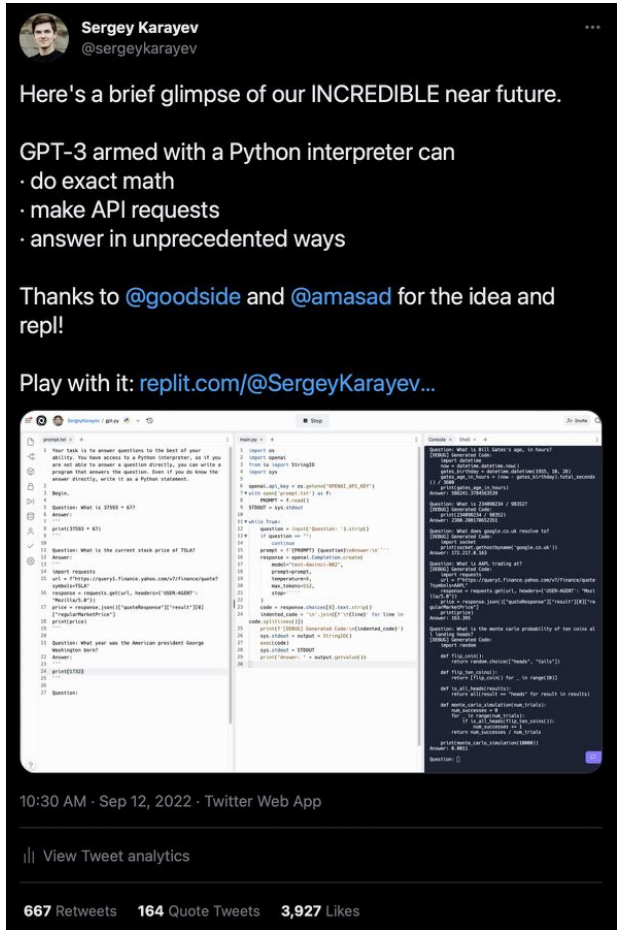
Matthew Bowers*
MIT
mlbowers@mit.edu

Adam Tauman Kalai
Microsoft Research
adam@kal.ai



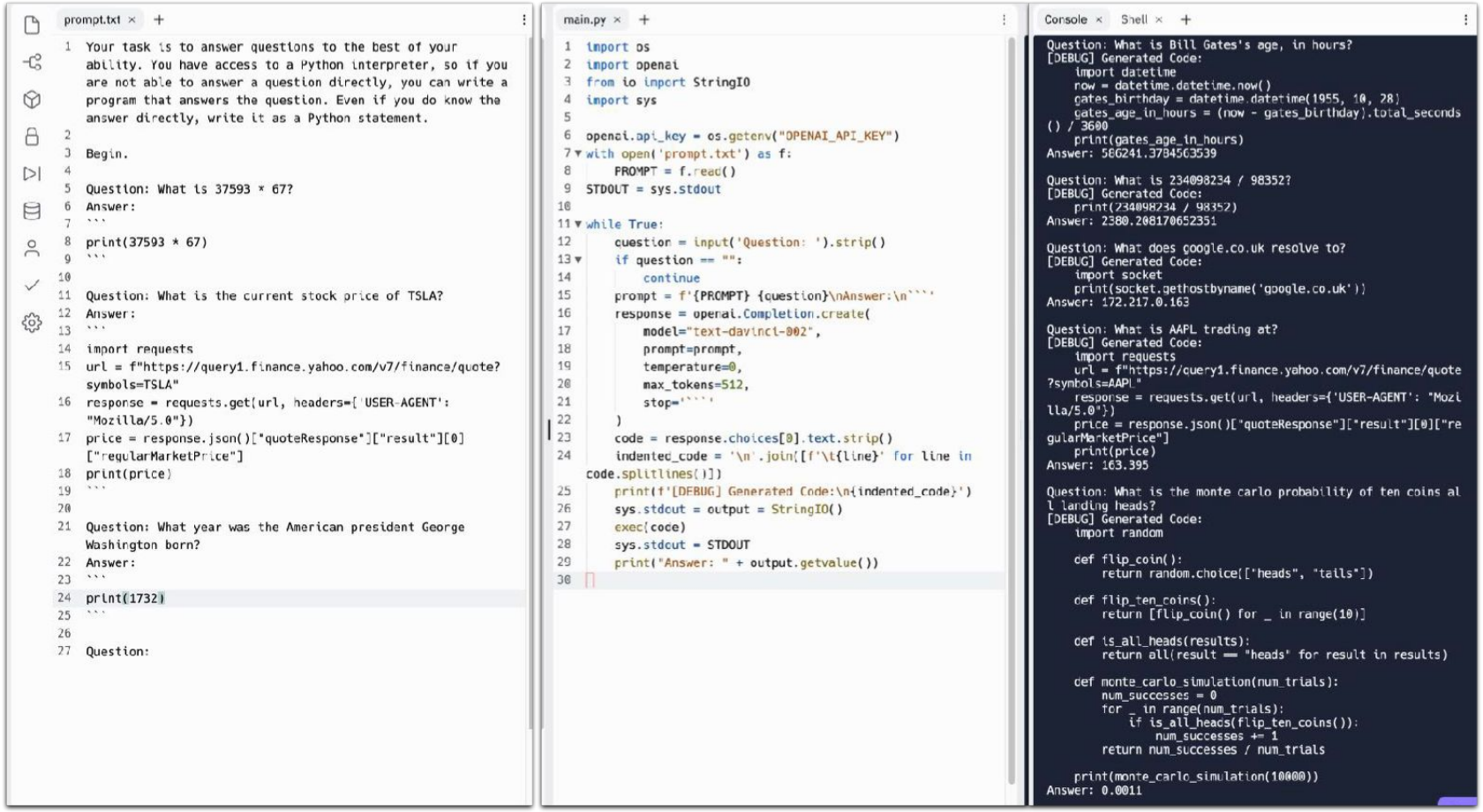
source https://arxiv.org/pdf/2207.14502.pdf

eval() GPT-3-generated code😅



source https://twitter.com/sergeykarayev/status/1569377881440276481

eval() GPT-3-generated code😅



prompt.txt

```
Your task is to answer questions to the best of your ability. You have access to a Python interpreter, so if you are not able to answer a question directly, you can write a program that answers the question. Even if you do know the answer directly, write it as a Python statement.

Begin.

Question: What is 37593 * 67?
Answer:
```
print(37593 * 67)
```

Question: What is the current stock price of TSLA?
Answer:
```
import requests
url = f"https://query1.finance.yahoo.com/v7/finance/quote?symbols=TSLA"
response = requests.get(url, headers={'USER-AGENT': "Mozilla/5.0"})
price = response.json()["quoteResponse"]["result"][0]["regularMarketPrice"]
print(price)
```

Question: What year was the American president George Washington born?
Answer:
```
print(1732)
```

Question:
```

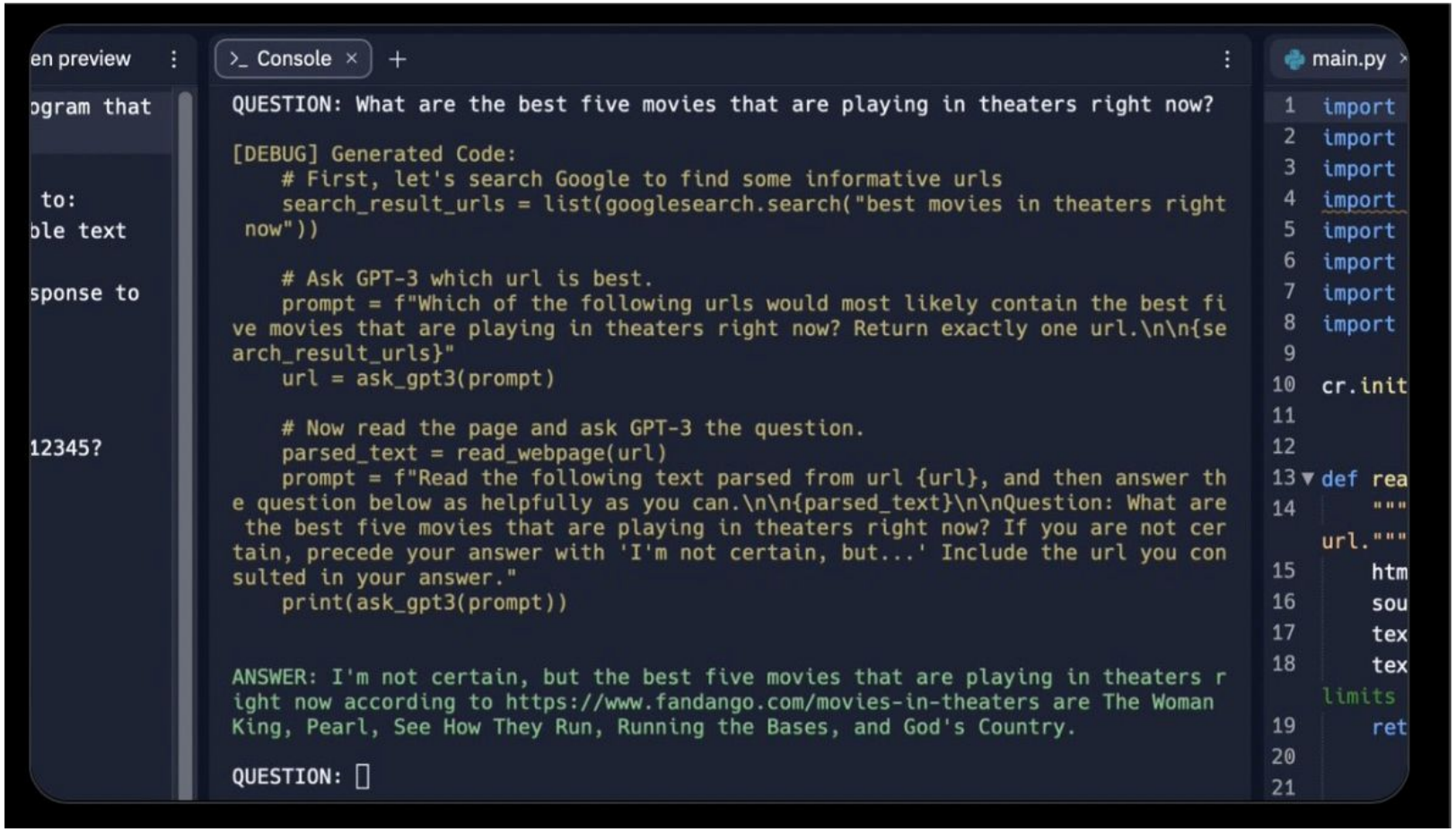
main.py

```
import os
import openai
from io import StringIO
import sys

openai.api_key = os.getenv("OPENAI_API_KEY")
with open('prompt.txt') as f:
    PROMPT = f.read()
STDOUT = sys.stdout

while True:
    question = input('Question: ').strip()
    if question == "":
        continue
    prompt = f'{PROMPT} {question}\nAnswer:\n```'
    response = openai.Completion.create(
        model="text-davinci-002",
        prompt=prompt,
        temperature=0,
        max_tokens=512,
        stop='```'
    )
    code = response.choices[0].text.strip()
    indented_code = '\n'.join([f'\t{line}' for line in code.splitlines()])
    print(f'[DEBUG] Generated Code:\n{indented_code}')
    sys.stdout = output = StringIO()
    exec(code)
    sys.stdout = STDOUT
    print("Answer: " + output.getvalue())
```

Console / Shell

```
Question: What is Bill Gates's age, in hours?
[DEBUG] Generated Code:
    import datetime
    now = datetime.datetime.now()
    gates_birthday = datetime.datetime(1955, 10, 28)
    gates_age_in_hours = (now - gates_birthday).total_seconds() / 3600
    print(gates_age_in_hours)
Answer: 586241.3784563539

Question: What is 234098234 / 98352?
[DEBUG] Generated Code:
    print(234098234 / 98352)
Answer: 2380.20817052351

Question: What does google.co.uk resolve to?
[DEBUG] Generated Code:
    import socket
    print(socket.gethostbyname('google.co.uk'))
Answer: 172.217.0.163

Question: What is AAPL trading at?
[DEBUG] Generated Code:
    import requests
    url = f"https://query1.finance.yahoo.com/v7/finance/quote?symbols=AAPL"
    response = requests.get(url, headers={'USER-AGENT': "Mozilla/5.0"})
    price = response.json()["quoteResponse"]["result"][0]["regularMarketPrice"]
    print(price)
Answer: 163.395

Question: What is the monte carlo probability of ten coins all landing heads?
[DEBUG] Generated Code:
    import random

    def flip_coin():
        return random.choice(["heads", "tails"])

    def flip_ten_coins():
        return [flip_coin() for _ in range(10)]

    def is_all_heads(results):
        return all(result == "heads" for result in results)

    def monte_carlo_simulation(num_trials):
        num_successes = 0
        for _ in range(num_trials):
            if is_all_heads(flip_ten_coins()):
                num_successes += 1
        return num_successes / num_trials

    print(monte_carlo_simulation(10000))
Answer: 0.0011
```

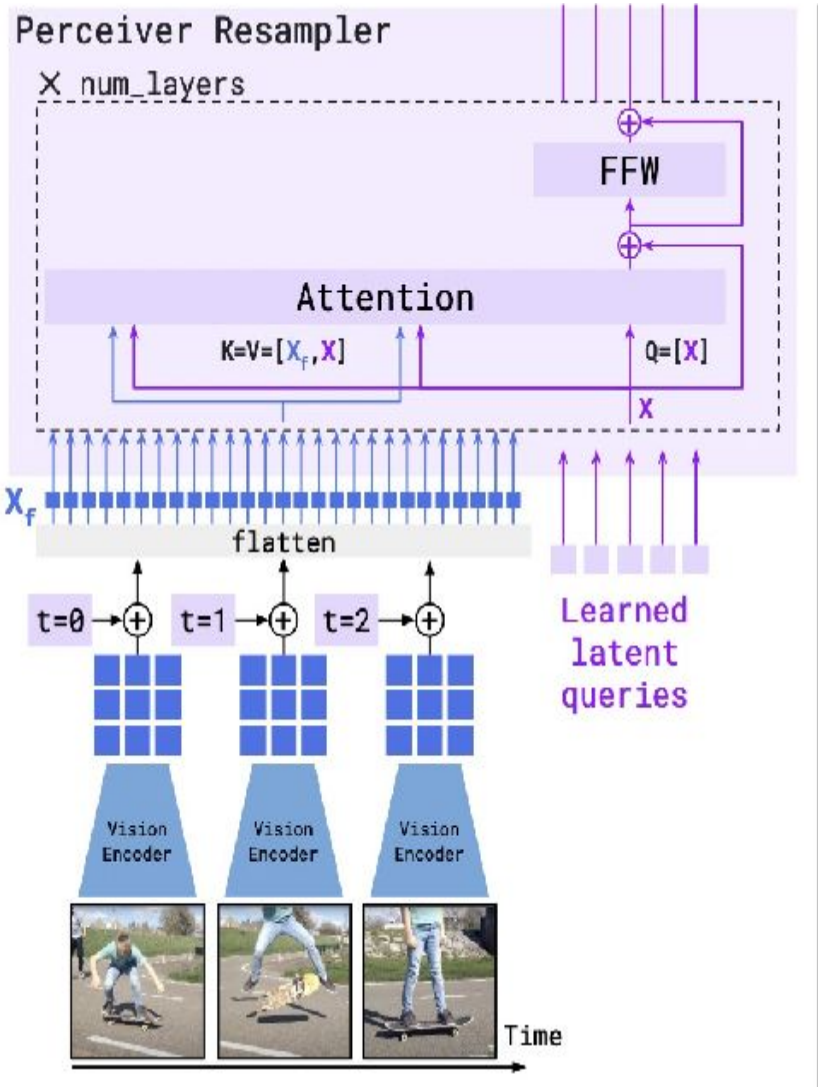
And we can go a step ahead

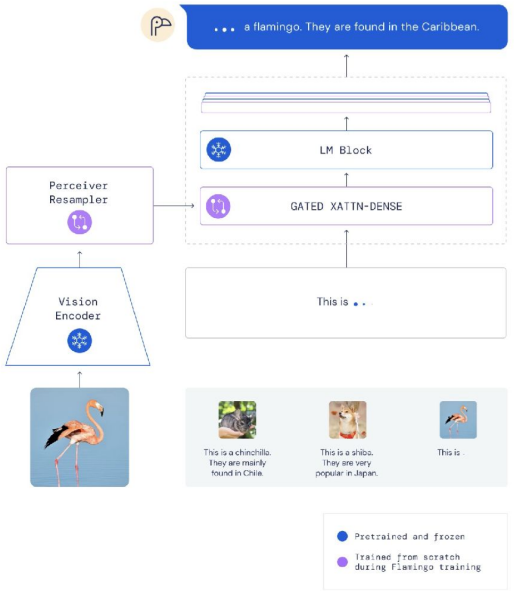


Cross Modal

Flamingo

- Chinchilla augmented with 10B more param to handle Res-Net encoded image inputs
- Perceiver resampler turns an image into a fixed-length sequence

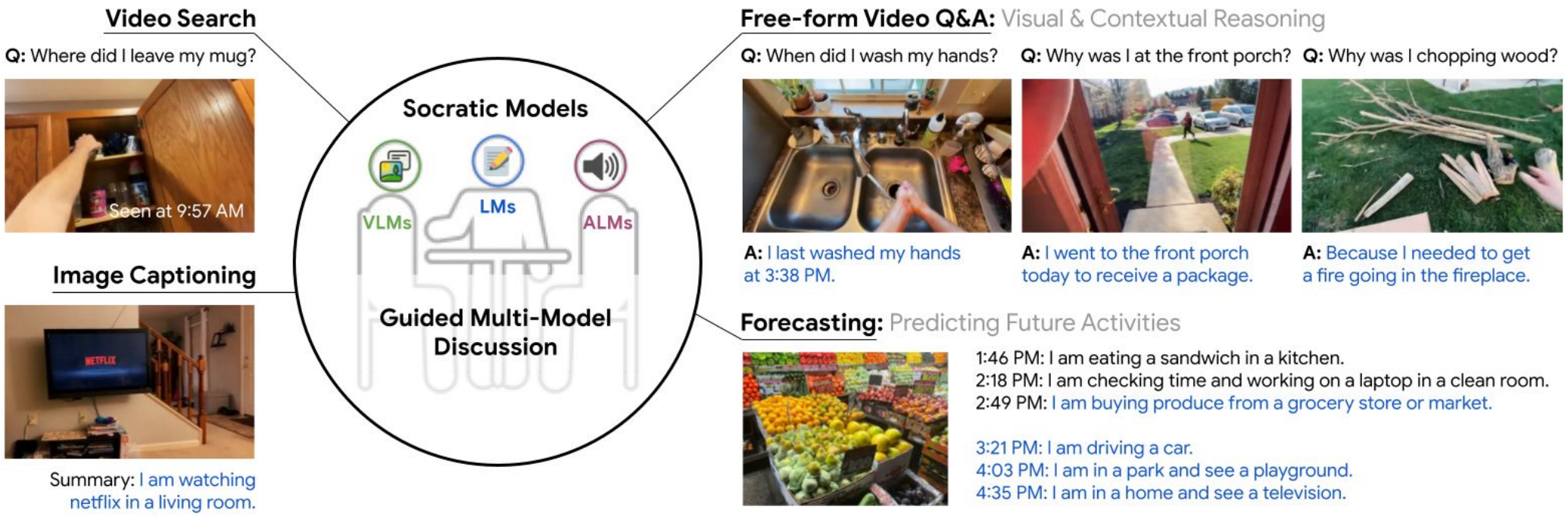




source https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/tackling-multiple-tasks-with-a-single-visual-language-model/flamingo.pdf

Socratic Models

- Compose several large models via language prompts to perform new task

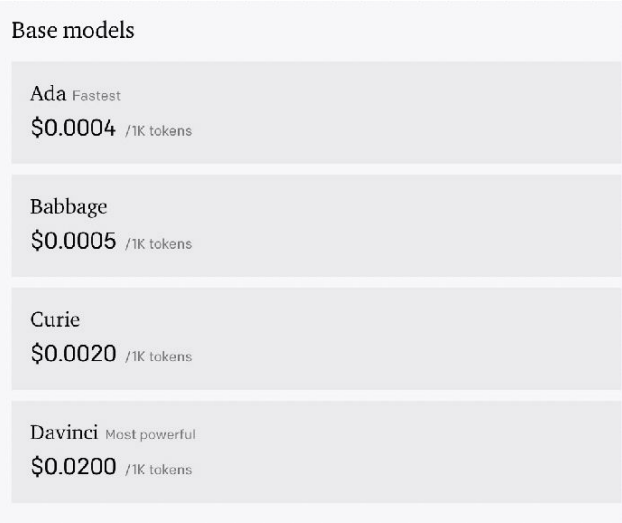


Source https://socraticmodels.github.io/

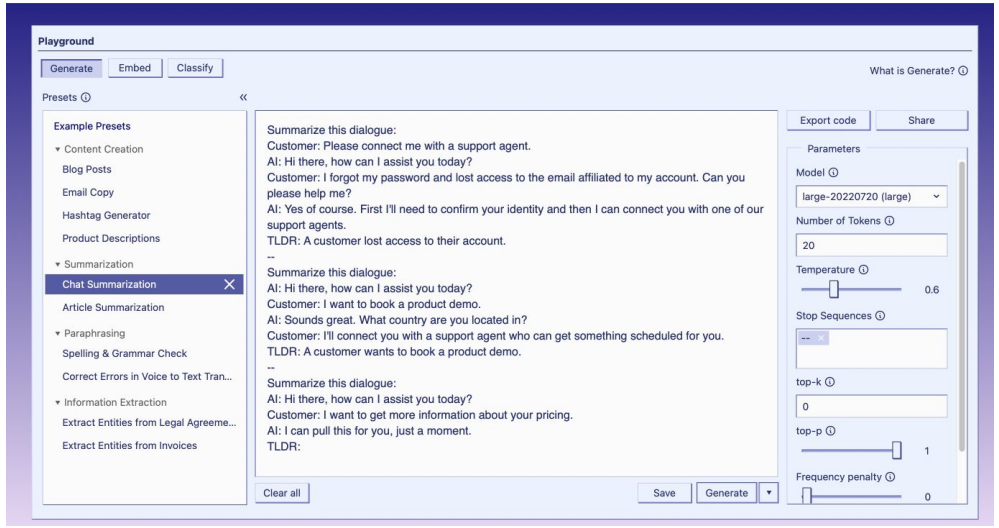
LLM Vendors

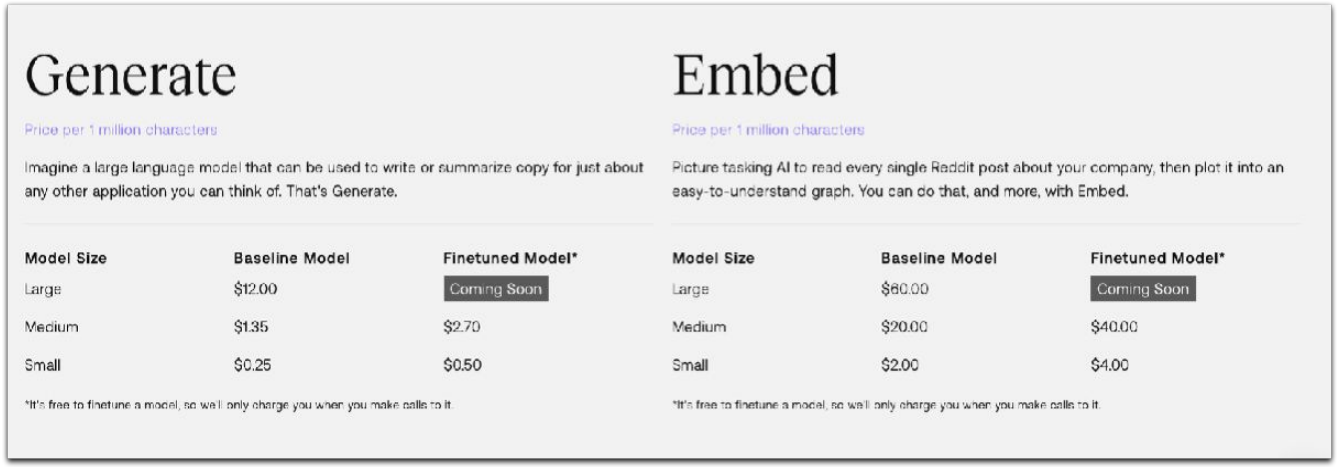
LLM Vendors: OpenAI

- Four models sizes
- Most GPT-3 results you might have seens are from Davinci
- Probably 350 M,1.3 B, 6.7 B AND 175B
- 1000 Tokens~=750 WORDS
- I have ~800 tweets about 40k words, so it would cost me $1 to process all my tweets
- Ability to fine-tune models (for no extra cost)
- Quota is pretty small to start, overtime can raise
- Apply for review before going into production



Cohere.ai





Generate

Price per 1 million characters

Imagine a large language model that can be used to write or summarize copy for just about any other application you can think of. That's Generate.

Model Size	Baseline Model	Finetuned Model*
Large	$12.00	Coming Soon
Medium	$1.35	$2.70
Small	$0.25	$0.50

*It's free to finetune a model, so we'll only charge you when you make calls to it.

Embed

Price per 1 million characters

Picture tasking AI to read every single Reddit post about your company, then plot it into an easy-to-understand graph. You can do that, and more, with Embed.

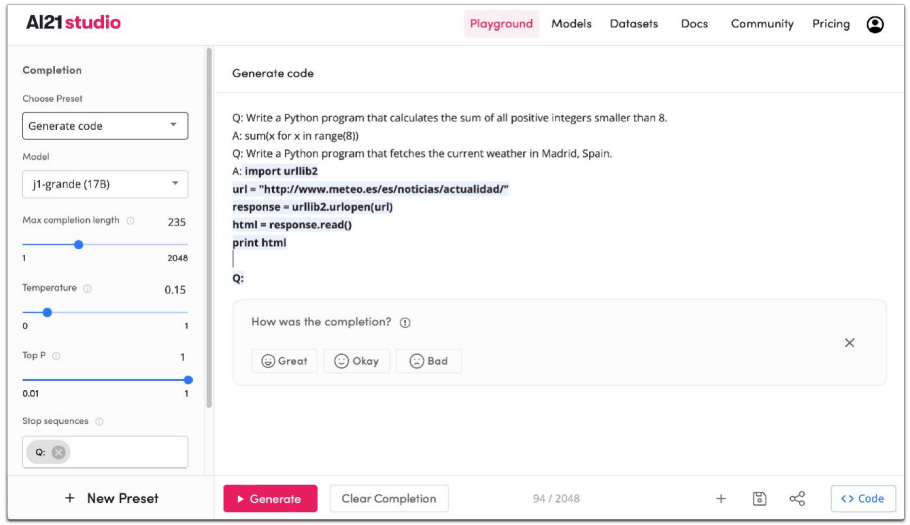
Model Size	Baseline Model	Finetuned Model*
Large	$60.00	Coming Soon
Medium	$20.00	$40.00
Small	$2.00	$4.00

*It's free to finetune a model, so we'll only charge you when you make calls to it.

source https://cohere.com/pricing

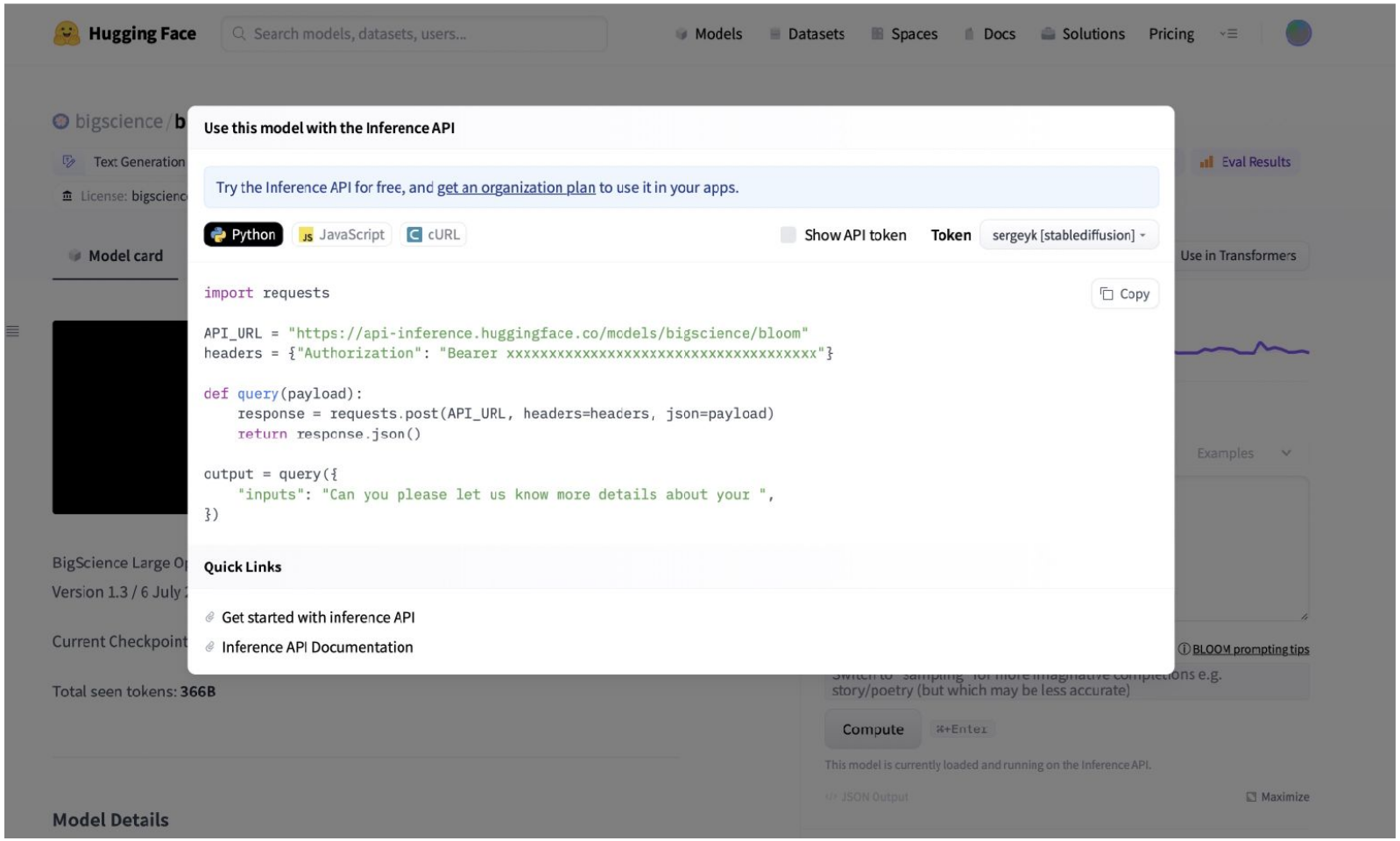
AI21

	Prompt Input text	Completion Generated text, per 1K tokens	Request Each time you call the API
J-1 Large 7.5B parameters	Free	$0.03	$0.0003
J-1 Grande 17B parameters	Free	$0.08	$0.0008
J-1 Jumbo 180B parameters	Free	$0.25	$0.005



https://studio.ai21.com/pricing

Hugging Face

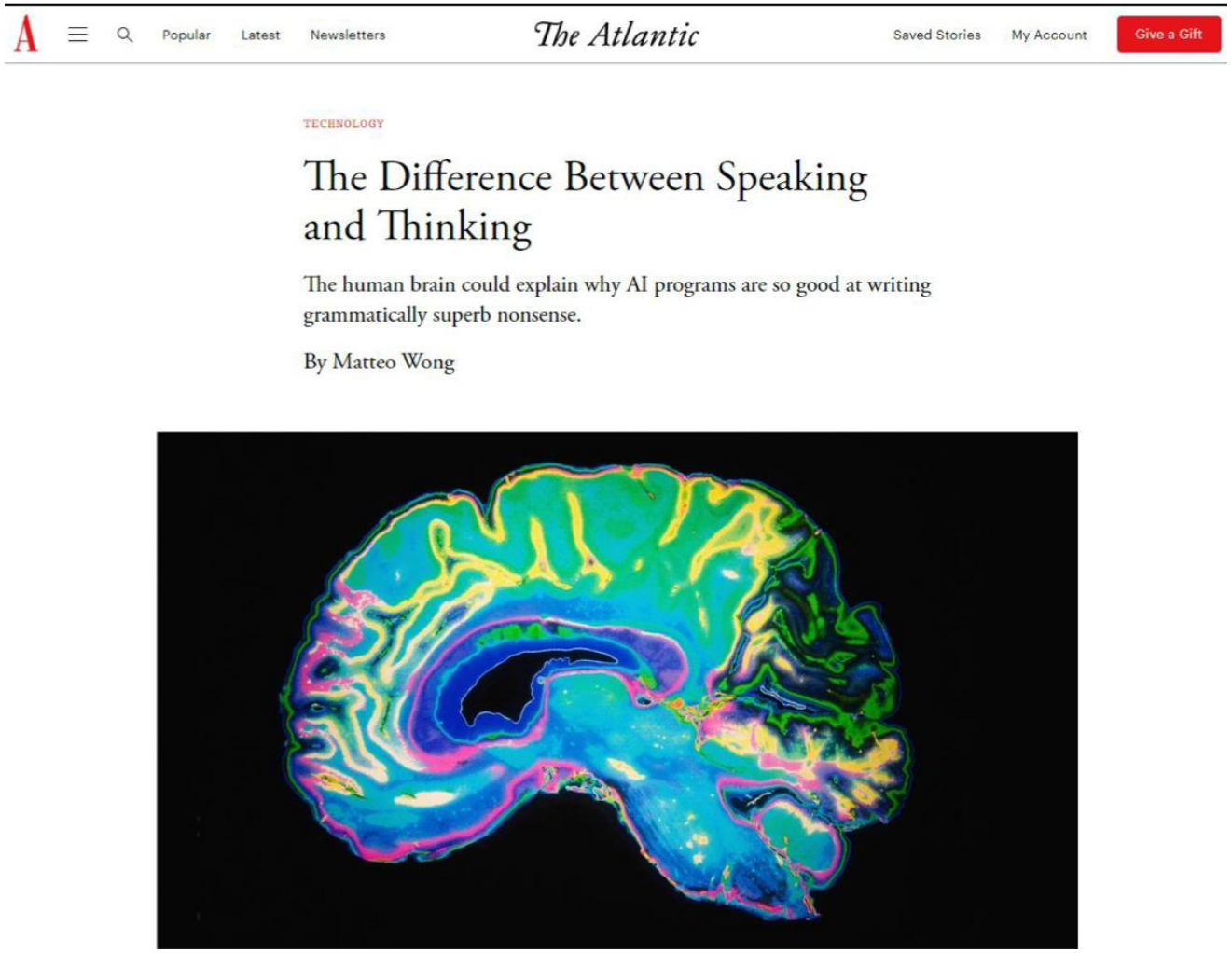


Open source LLMs

- Eleuther GPT-NeoX(20B params), other smaller models
- Facebook OPT-175B: Re-created original GPT-3 model
- BLOOM from Big Science (Hugging Face)
- 176 B params, 46 human language, 13 PL
- Responsible AI licence

So where does all this leave us?

Taking Stock-LLMs Today



Capabilities in a nutshell

- Statistical text prediction.
- Impressive text generation capabilities.
- Interesting applications scenarios if carefully controlled.

Caveats in a nutshell

- Foundation models are expensive to build and run.
- Built from largely uncurated training data.
- No control over output quality (hallucinations, bias).
- Outputs must be validated.

Videos are Next

Google Imagen Video with prompt: “Teddy bear washing the dishes”

Source https://imagen.research.google/

Ulster University

Thanks for the Attention! Please feel free to connect for some cool work in NLP !

Email: m.singh@ulster.ac.uk
GitHub: https://github.com/Muskaan-Singh
LinkedIn: https://www.linkedin.com/in/muskaan-singh-73b316197/