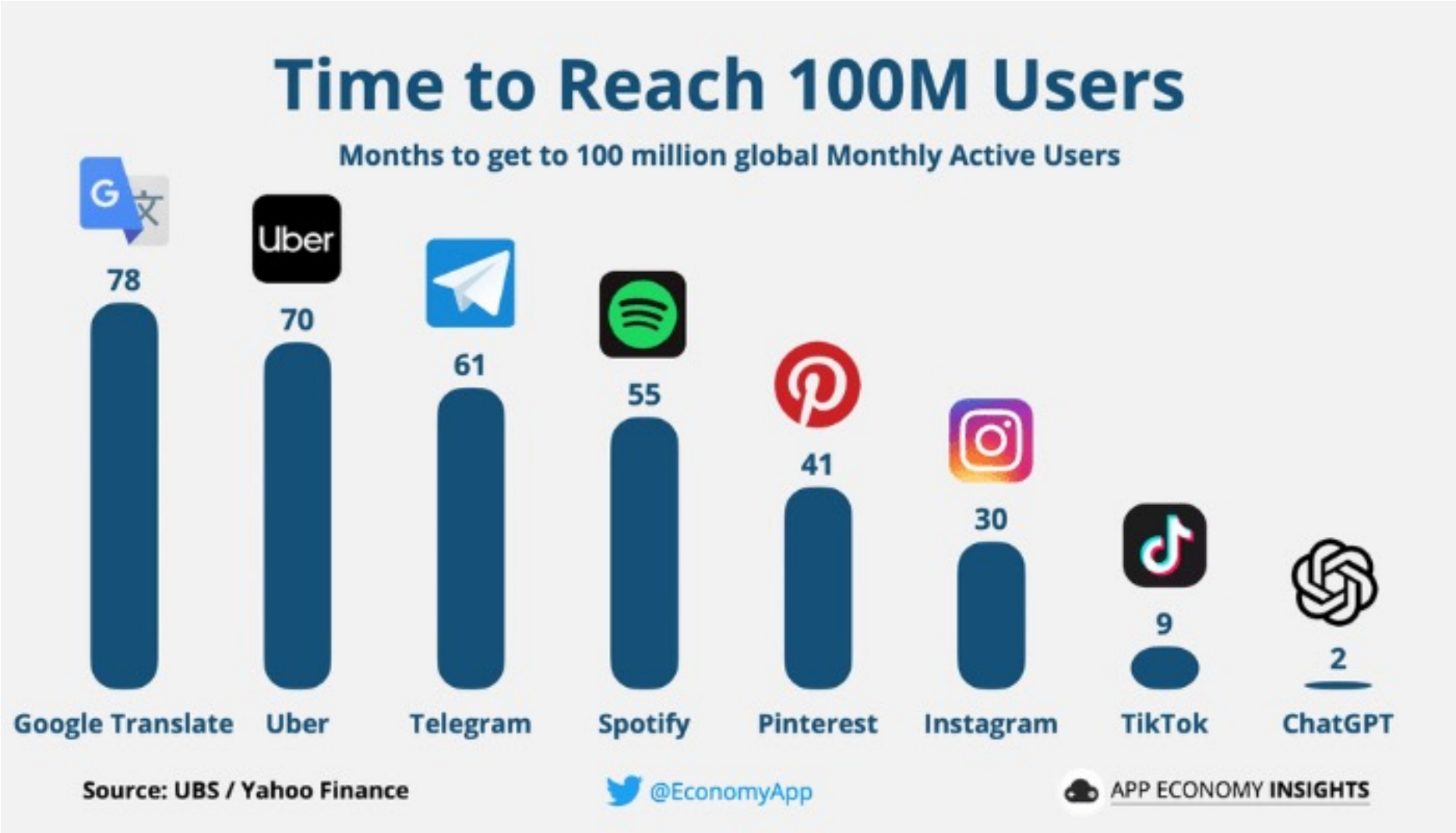
Embeddings, Transformers, RLHF

Three key ideas to understand ChatGPT

Emanuele Fabbiani
Head of AI @ xtream

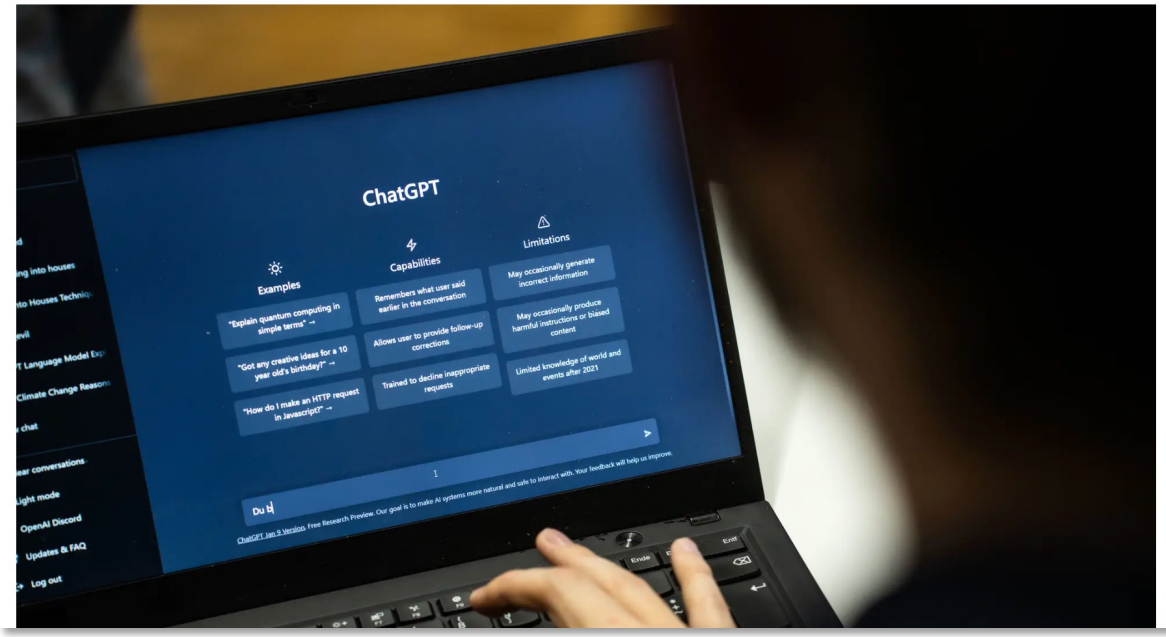

Time to Reach 100M Users
Months to get to 100 million global Monthly Active Users
78
70
61
55
41
30
9
2
Google Translate
Uber
Telegram
Spotify
Pinterest
Instagram
TikTok
ChatGPT
Source: UBS / Yahoo Finance
@EconomyApp
APP ECONOMY INSIGHTS

INSIDER

Newsletters Log in Subscribe

OpenAI's ChatGPT can actually make workers perform worse, a new study found

Aaron Mok Oct 2, 2023, 10:14 AM GMT+2

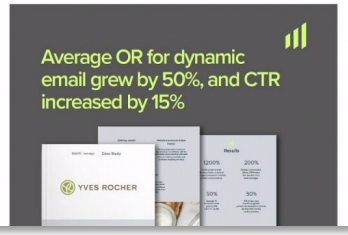


Advertisement

SALESmanago

200% better comms efficiency

Yves Rocher boosted email efficiency by 1200% with SALESmanago's dynamic messages

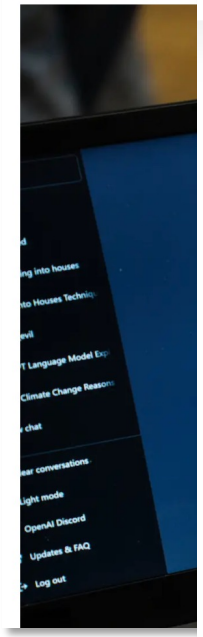


INSIDER

Newsletters Log in Subscribe

OpenAI's ChatGPT can actually make workers perform worse, a new study found

Aaron Mok Oct 2, 202



FORTUNE

Subscribe Now SIGN IN

NEWSLETTERS · CHRO DAILY

ChatGPT can boost or tank employee performance, depending on what tasks it's used for

BY PAIGE MCGLAUFLIN AND JOSEPH ABRAMS

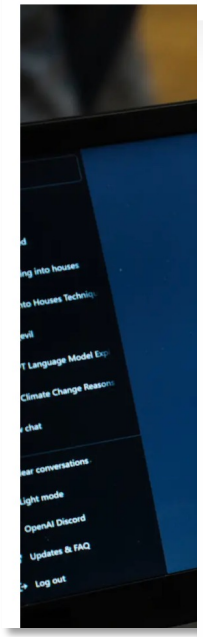
September 27, 2023 at 2:30 PM GMT+2

INSIDER

Newsletters Log in Subscribe

OpenAI's ChatGPT can actually make workers perform worse, a new study found

Aaron Mok Oct 2, 202



FORTUNE

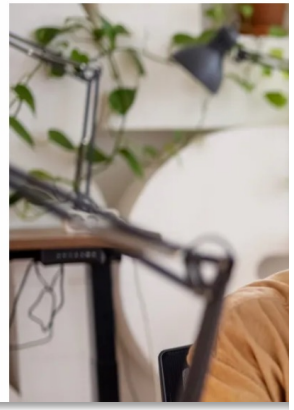
Subscribe Now SIGN IN

NEWSLETTERS · CHRO DAILY

ChatGPT can boost or tank employee performance, depending on what tasks it's used for

BY PAIGE MCGLAUFLIN AND JOSE

September 27, 2023 at 2:30 PM GMT



Consultants using AI do better, especially underperformers: study

Euan Black

Work and careers reporter

Consultants who use ChatGPT-4 can perform tasks significantly better than those who do not use the tool, according to an academic study that found below-average performers improved the most when using AI.

The Harvard Business School research found that consultants using the generative artificial intelligence tool finished 12.2 per cent more tasks,

From the original study

The public release of Large Language Models (LLMs) has sparked tremendous interest in how humans will use Artificial Intelligence (AI) to accomplish a variety of tasks. In our study conducted with Boston Consulting Group, a global management consulting firm, we examine the performance implications of AI on realistic, complex, and knowledge-intensive tasks. The pre-registered experiment involved 758 consultants comprising about 7% of the individual contributor-level consultants at the company. After establishing a performance baseline on a similar task, subjects were randomly assigned to one of three conditions: no AI access, GPT-4 AI access, or GPT-4 AI access with a prompt engineering overview. We suggest that the capabilities of AI create a “jagged technological frontier” where some tasks are easily done by AI, while others, though seemingly similar in difficulty level, are outside the current capability of AI. For each one of a set of 18 realistic consulting tasks within the frontier of AI capabilities, consultants using AI were significantly more productive (they completed 12.2% more tasks on average, and completed tasks 25.1% more quickly), and produced significantly higher quality results (more than 40% higher quality compared to a control group). Consultants across the skills distribution benefited significantly from having AI augmentation, with those below the average performance threshold increasing by 43% and those above increasing by 17% compared to their own scores. For a task selected to be outside the frontier, however, consultants using AI were 19 percentage points less likely to produce correct solutions compared to those without AI. Further, our analysis shows the emergence of two distinctive patterns of successful AI use by humans along a spectrum of human-AI integration. One set of consultants acted as “Centaurs,” like the mythical halfhorse/half-human creature, dividing and delegating their solution-creation activities to the AI or to themselves. Another set of consultants acted more like “Cyborgs,” completely integrating their task flow with the AI and continually interacting with the technology.

HBR, MIT, BCG, Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality

Embeddings

seal

AI models only understand numbers

«You shall know a word by the company it keeps»

1957

2013

Efficient Estimation of Word Representations in Vector Space

Tomas Mikolov
Google Inc., Mountain View, CA
tmikolov@google.com

Kai Chen
Google Inc., Mountain View, CA
kaichen@google.com

Greg Corrado
Google Inc., Mountain View, CA
gcorrado@google.com

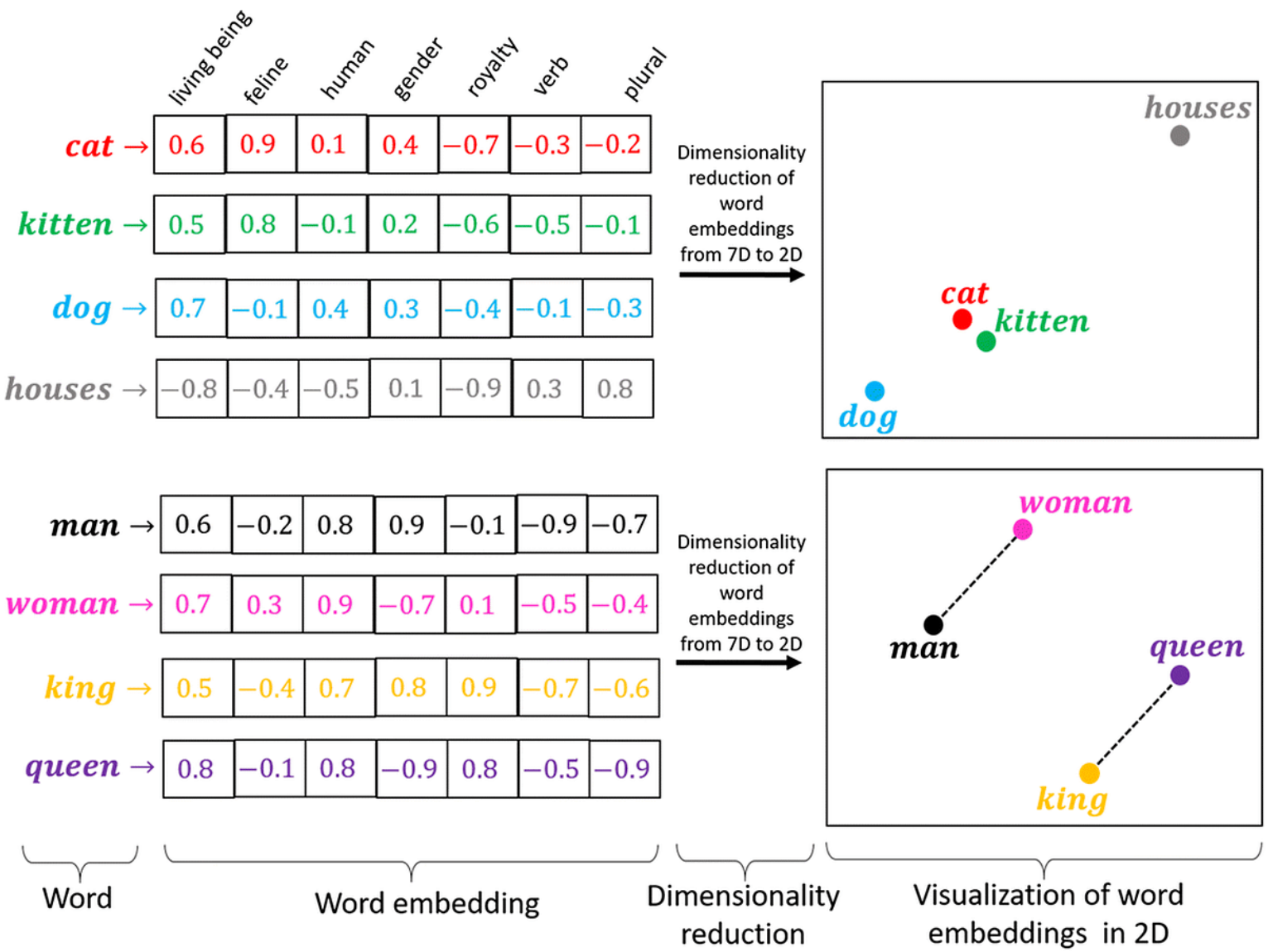
Jeffrey Dean
Google Inc., Mountain View, CA
jeff@google.com

781v3 [cs.CL] 7 Sep 2013

Abstract

We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words data set. Furthermore, we show that these vectors provide state-of-the-art performance on our test set for measuring syntactic and semantic word similarities.

https://arxiv.org/pdf/1301.3781.pdf



https://medium.com/@hari4om/word-embedding-d816f643140

The seal put her seal of approval.

2018

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

04805v2 [cs.CL] 24 May 2019

Abstract

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

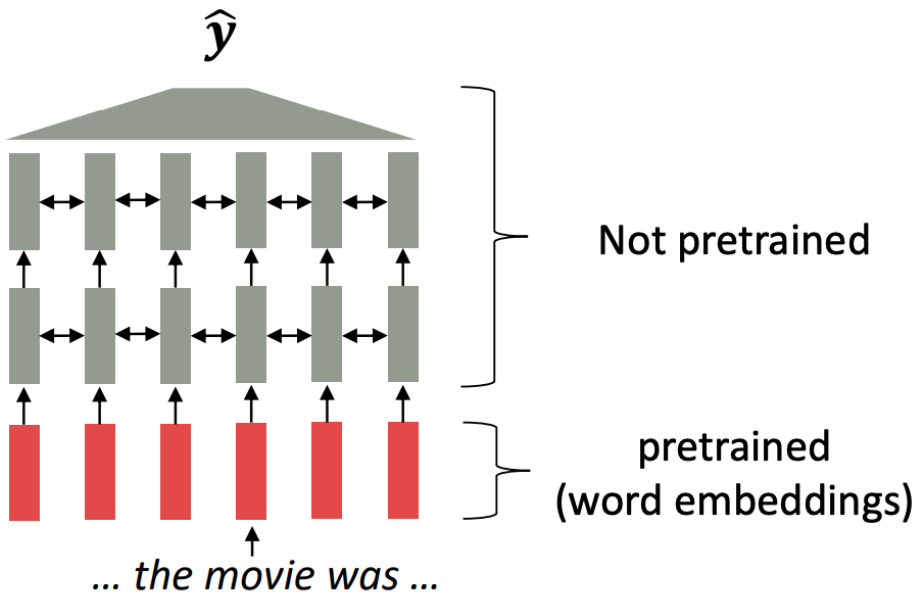
BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answer-

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pre-trained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

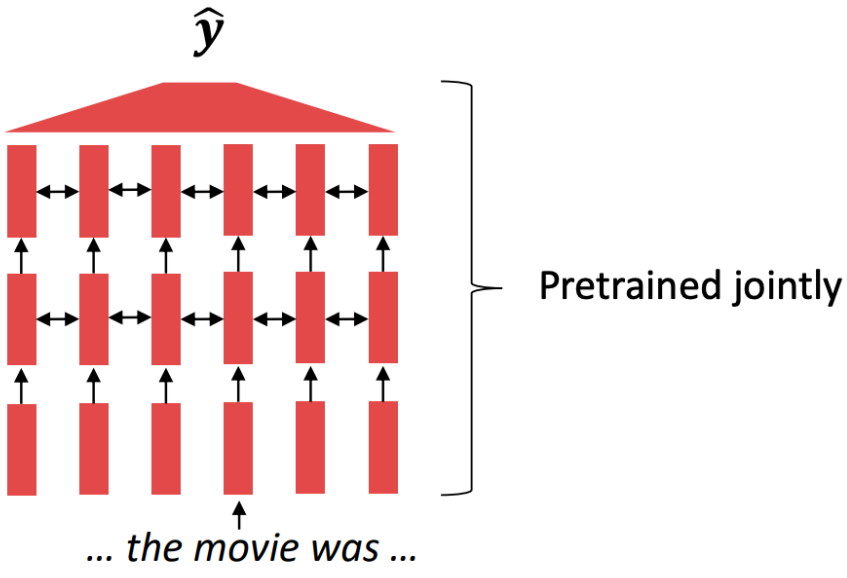
We argue that current techniques restrict the power of the pre-trained representations, especially for the fine-tuning approaches. The major limitation is that standard language models are unidirectional, and this limits the choice of architectures that can be used during pre-training. For

https://arxiv.org/pdf/1810.04805.pdf

Before 2018



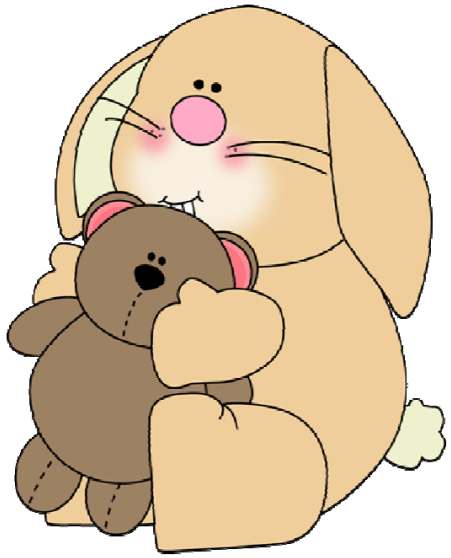
After 2018



https://web.stanford.edu/class/cs224n/

Read and fill in the missing words from the box.

bunny,	**only,**	**parents,**	**years,**	**teddy bear,**
school,	**have got,**	**garden,**	**~~name~~,**	**birthday,** **nice**



My _____*name*_____ is Lisa. I'm seven _____ old.

I go to _____ , my class is 2.b. I live in a house with my

_____ . I haven't got any brothers or sisters. I'm an

_____ child. My dad is really _____ and

my mum is great. We _____ a cat. It's name is Riki.

He likes to play in the _____ very much.

I have got a lot of toys, but I'd like to have a big _____

and a small _____ for my _____ .

____ / 10

How much does GPT-4 cost?

Written by Joshua J.
Updated over a week ago

We are excited to announce GPT-4 has a new pricing model, in which we have reduced the price of the prompt tokens.

For our models with 8k context lengths (e.g. `gpt-4` and `gpt-4-0314`), the price is:

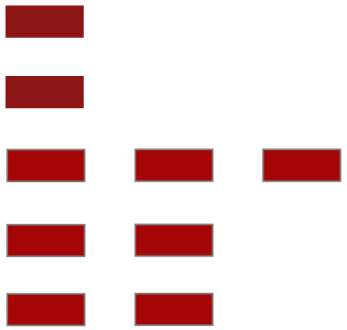
- $0.03/1k prompt tokens
- $0.06/1k sampled tokens

For our models with 32k context lengths (e.g. `gpt-4-32k` and `gpt-4-32k-0314`), the price is:

- $0.06/1k prompt tokens
- $0.12/1k sampled tokens

	word		vocab mapping
Common words	hat	→	hat
	learn	→	learn
Variations	taaaaasty	→	taa## aaa## sty
misspellings	laern	→	la## ern##
novel items	Transformerify	→	Transformer## ify

embedding

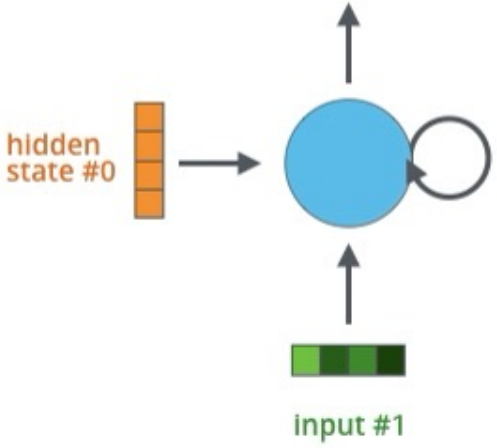


https://web.stanford.edu/class/cs224n/

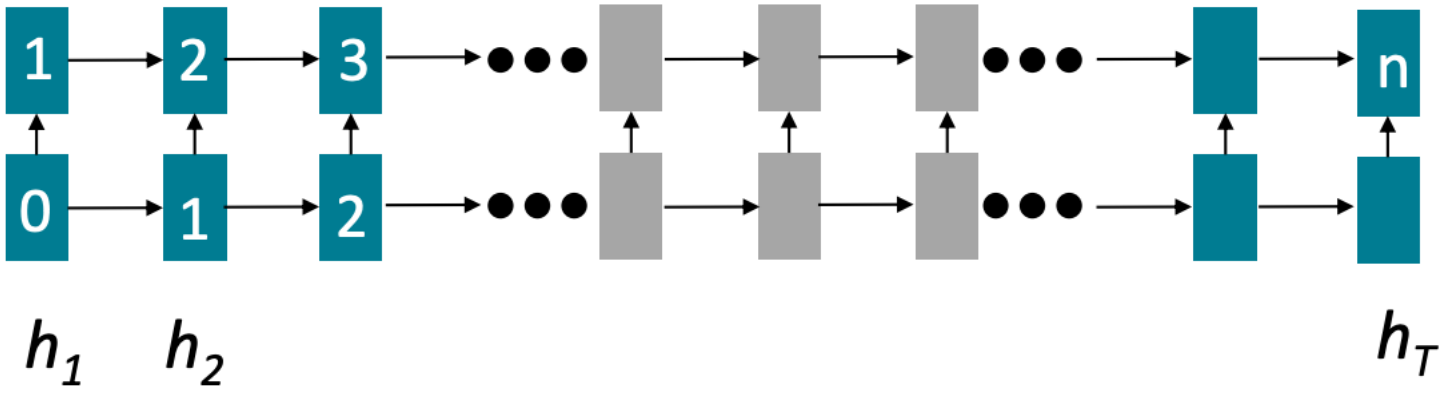
Transformers

Recurrent Neural Network

Time step #1:
An RNN takes two input vectors:



https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/



Inherently sequential

Could not "remember" context

2017

Attention Is All You Need

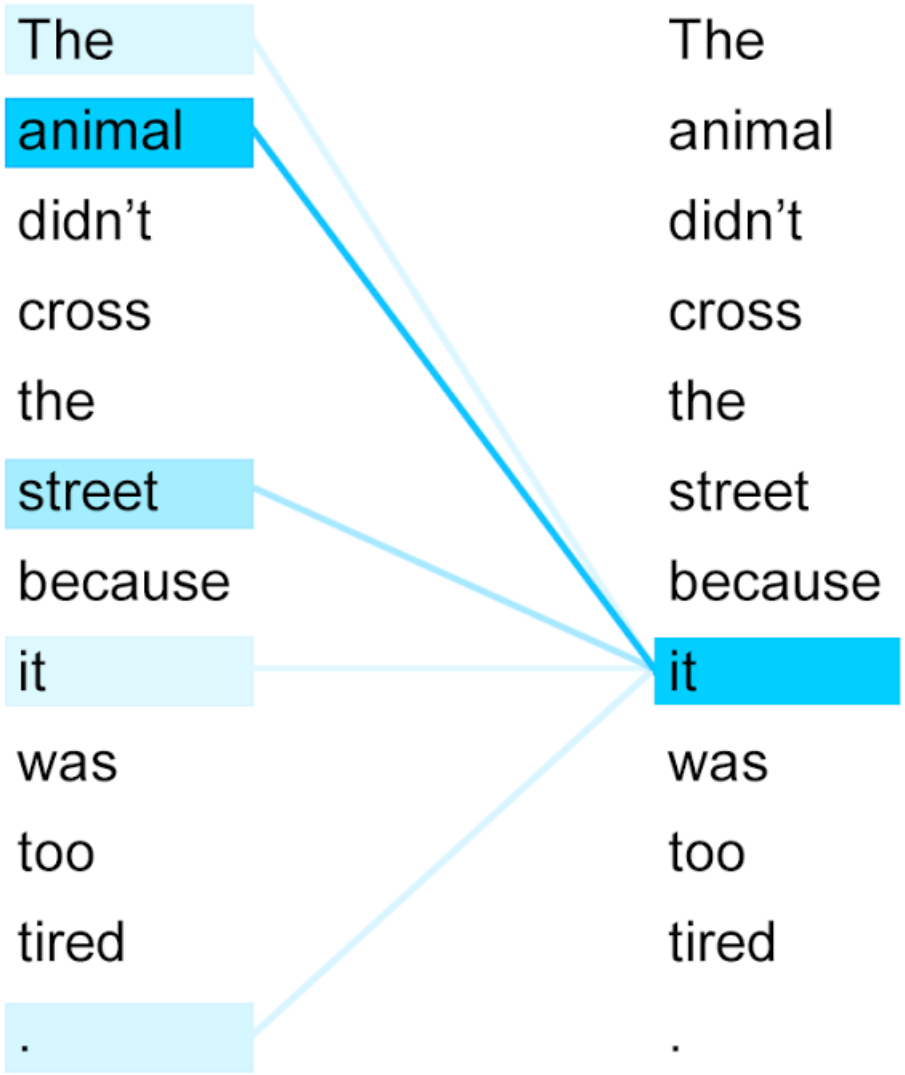
Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

v:1706.03762v5 [cs.CL] 6 Dec 2017

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with

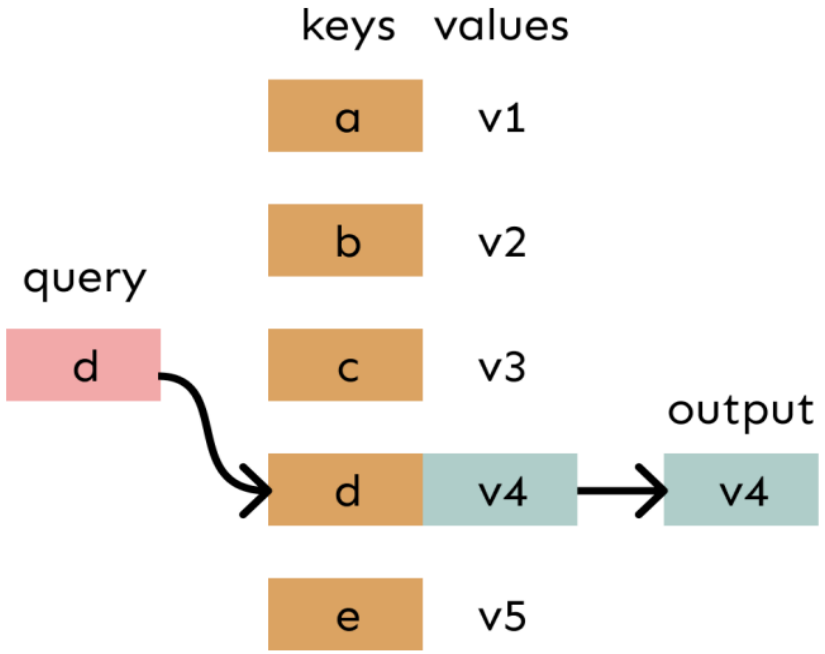
https://arxiv.org/pdf/1706.03762.pdf



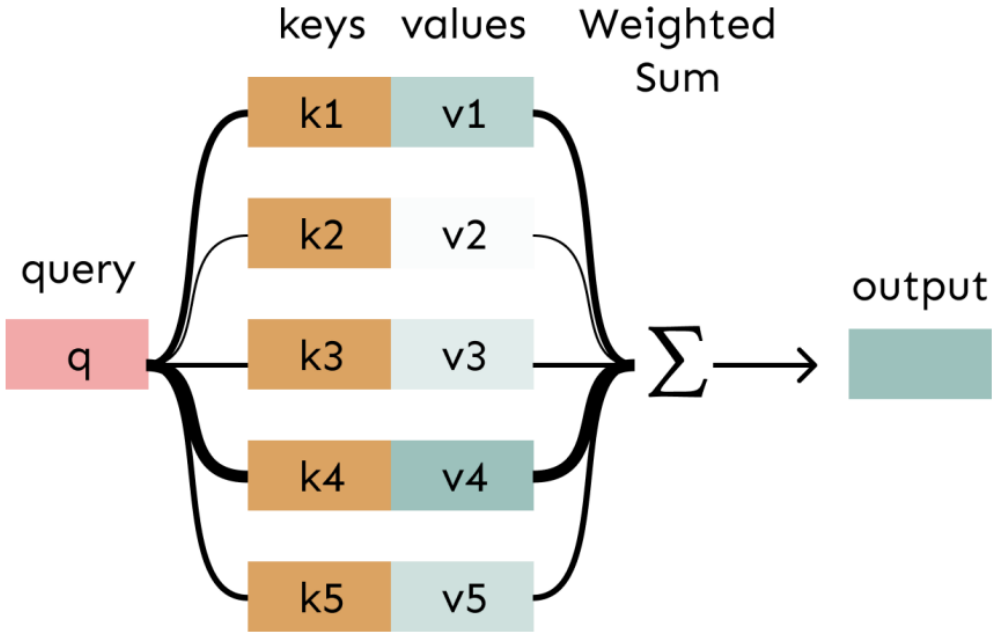
https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

22
10
73
15
41
9
39
24
36
3
84
12
OL

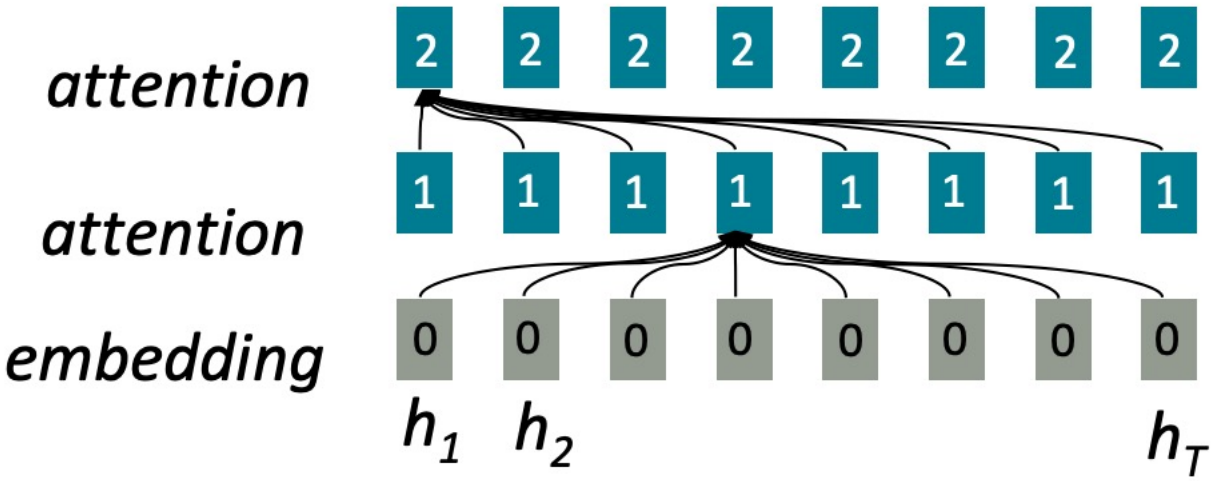
In a standard **lookup table**, we have a set of keys mapping to values. The query matches one of the keys, allowing to retrieve the value.



In **attention**, the query has a "match score" with each key. The values are retrieved and averaged, using the "match scores" as weights.



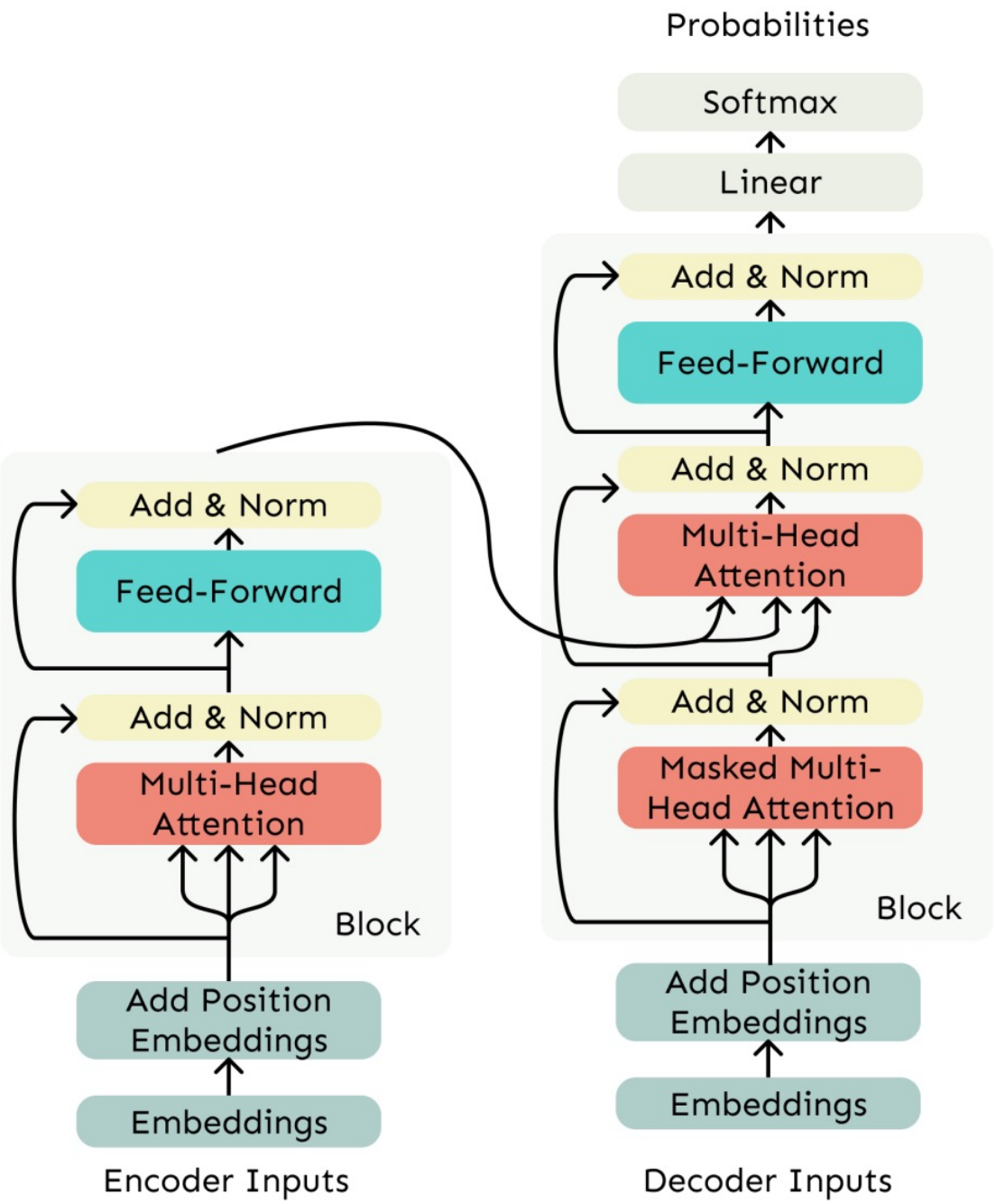
https://web.stanford.edu/class/cs224n/



Completely parallelizable

Chooses which context to "remember"

https://web.stanford.edu/class/cs224n/



https://web.stanford.edu/class/cs224n/

Reinforcement Learning with Human Feedback

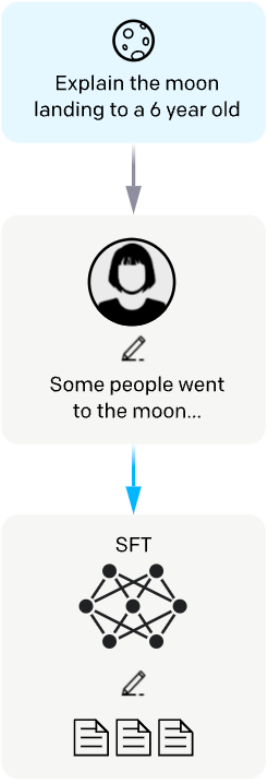
Step 1

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.



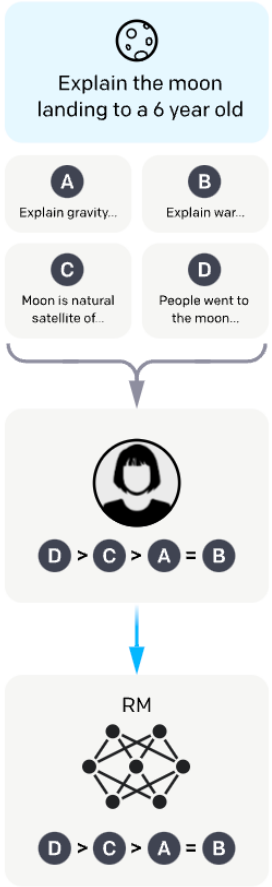
Step 2

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.



Step 3

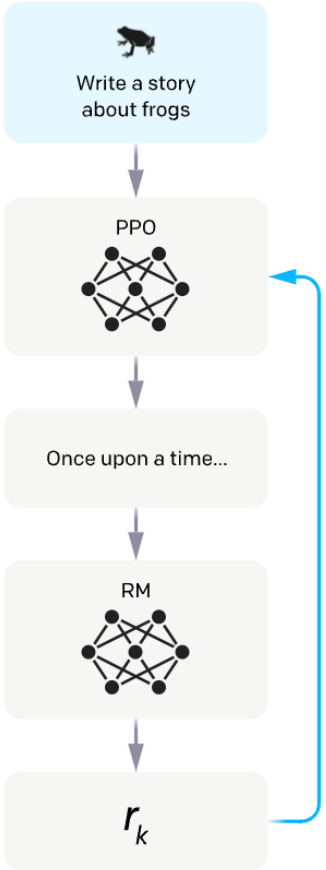
Optimize a policy against the reward model using reinforcement learning.

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.



Building GPT-3

GPT Assistant training pipeline

Stage	Pretraining	Supervised Finetuning	Reward Modeling	Reinforcement Learning
Dataset	**Raw internet** text trillions of words low-quality, large quantity	**Demonstrations** Ideal Assistant responses, ~10-100K (prompt, response) written by contractors low quantity, high quality	**Comparisons** 100K –1M comparisons written by contractors low quantity, high quality	**Prompts** ~10K-100K prompts written by contractors low quantity, high quality
Algorithm	**Language modeling** predict the next token	**Language modeling** predict the next token	**Binary classification** predict rewards consistent w preferences	**Reinforcement Learning** generate tokens that maximize the reward
		init from	init from	init from SFT use RM
Model	**Base model**	**SFT model**	**RM model**	**RL model**
Notes	1000s of GPUs months of training ex: GPT, LLaMA, PaLM can deploy this model	1-100 GPUs days of training ex: Vicuna-13B can deploy this model	1-100 GPUs days of training	1-100 GPUs days of training ex: ChatGPT, Claude can deploy this model

fiverr.
Gabby @ggerbus
Freelance AI Copywriter
AI TOOK MY JOB
TO THE NEXT LEVEL

Interesting Resources

- OpenAI, OpenAI Cookbook
- Dair AI, Prompt Engineering Guide
- Microsoft, State of GPT
- Mistral AI, Retrieval vs. Fine-tuning
- HBR, MIT, BCG, Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality
- G. Kasparov, Don't Fear Intelligent Machines, Work with Them

- Follow me on LinkedIn ☺

Ego Slide

Emanuele Fabbiani

Head of AI at xtream
PhD from UniPV

Speaker and track organizer at AMLD Lausanne, ODSC, PyCon, eRUM, PyMI, Pisa.dev.

Course creator at UniPV, Almo Collegio Borromeo, Treccani.

Lecturer in Italy, Poland, and Switzerland.