Training Session

on

OpenAI’s

Generative Pretrained Transformer (GPT) - 3

Ishan Jain
ishanjain.ai@gmail.com

Agenda

1. Evolution of NLP/NLG models
2. OpenAI’s GPT-3
3. GPT-3 applications
4. Conclusions & step forward

Terminologies

Open AI - A non-profit AI research organization

NLP - Natural Language Processing

NLU - Natural Language Understanding

NLG - Natural Language Generation

Language Models - A statistical language model is a probability distribution over sequences of words.

In-context learning - Learning and training of a model based on specific tasks and prompts

Parameters - A parameter is a calculation in a neural network that applies a great or lesser weighting to some aspect of the data

AGI - Artificial General Intelligence

Evolution of NLP/NLG models

No	Model	Year	Description
1	Google's BERT	2018	The self-supervised method released by Google in 2018.
2	Allen NLP's ELMo	2018	A deep contextualized word representation that models both complex and linguistic contexts.
3	ULMFit	2018	Universal Language Model Fine-tuning for Text Classification.
4	Google's XLNet	2019	A generalized autoregressive pretraining method that enables learning bidirectional contexts by maximizing the expected likelihood over all permutations of the factorization order.
5	Google's ALBERT	2019	A Lite BERT for Self-Supervised Learning of Language Representations
6	FB's RoBERTa	2019	Upgrade on BERT. Implemented in PyTorch. An optimised method for pre-training a self-supervised NLP system built on BERT's language masking strategy.
7	Microsoft's CodeBERT	2020	Google's BERT framework for NLP, has been built upon a bidirectional multi-layer neural architecture.
8	Microsoft's ZeRO-2	2020	Zero Redundancy Optimizer version 2 (ZeRO-2), a distributed deep-learning optimization algorithm.
9.	GPT-3	2020	GPT-3 is one of the most controversial pre-trained models by OpenAI.

GPT-3

The first AGI model

Natural Language Processing Model

Non-deterministic

NLP task-agnostic

Requires minimum fine-tuning

Evolution of Generative Pretrained Transformer (GPT) models

OpenAI's GPT

2018

Unsupervised

Natural Language Understanding

10X

OpenAI's GPT-2

2019

10X scaled up on parameters and training data.

1.5 billion parameters. Trained on 8 million web pages

100X

OpenAI's GPT-3

2020

175 billion parameters

A neural-network based language model.

Improving Language Understanding by Generative Pre-Training

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

Abstract

Natural language understanding comprises a wide range of diverse tasks such as textual entailment, question answering, semantic similarity assessment, and document classification. Although large unlabeled text corpora are abundant, labeled data for learning these specific tasks is scarce, making it challenging for discriminatively trained models to perform adequately. We demonstrate that large gains on these tasks can be realized by *generative pre-training* of a language model on a diverse corpus of unlabeled text, followed by *discriminative fine-tuning* on each specific task. In contrast to previous approaches, we make use of task-aware input transformations during fine-tuning to achieve effective transfer while requiring minimal changes to the model architecture. We demonstrate the effectiveness of our approach on a wide range of benchmarks for natural language understanding. Our general task-agnostic model outperforms discriminatively trained models that use architectures specifically crafted for each task, significantly improving upon the state of the art in 9 out of the 12 tasks studied. For instance, we achieve absolute improvements of 8.9% on commonsense reasoning (Stories Cloze Test), 5.7% on question answering (RACE), and 1.5% on textual entailment (MultiNLI).

1 Introduction

The ability to learn effectively from raw text is crucial to alleviating the dependence on supervised learning in natural language processing (NLP). Most deep learning methods require substantial amounts of manually labeled data, which restricts their applicability in many domains that suffer from a dearth of annotated resources [61]. In these situations, models that can leverage linguistic information from unlabeled data provide a valuable alternative to gathering more annotation, which can be time-consuming and expensive. Further, even in cases where considerable supervision is available, learning good representations in an unsupervised fashion can provide a significant performance boost. The most compelling evidence for this so far has been the extensive use of pre-trained word embeddings [10, 39, 42] to improve performance on a range of NLP tasks [8, 11, 26, 45].

Leveraging more than word-level information from unlabeled text, however, is challenging for two main reasons. First, it is unclear what type of optimization objectives are most effective at learning text representations that are useful for transfer. Recent research has looked at various objectives such as language modeling [44], machine translation [38], and discourse coherence [22], with each method outperforming the others on different tasks.[1] Second, there is no consensus on the most effective way to transfer these learned representations to the target task. Existing techniques involve a combination of making task-specific changes to the model architecture [43, 44], using intricate learning schemes [21] and adding auxiliary learning objectives [50]. These uncertainties have made it difficult to develop effective semi-supervised learning approaches for language processing.

[1] https://gluebenchmark.com/leaderboard

Preprint. Work in progress.

Language Models are Unsupervised Multitask Learners

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

Abstract

Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on task-specific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. When conditioned on a document plus questions, the answers generated by the language model reach 55 F1 on the CoQA dataset - matching or exceeding the performance of 3 out of 4 baseline systems without using the 127,000+ training examples. The capacity of the language model is essential to the success of zero-shot task transfer and increasing it improves performance in a log-linear fashion across tasks. Our largest model, GPT-2, is a 1.5B parameter Transformer that achieves state of the art results on 7 out of 8 tested language modeling datasets in a zero-shot setting but still underfits WebText. Samples from the model reflect these improvements and contain coherent paragraphs of text. These findings suggest a promising path towards building language processing systems which learn to perform tasks from their naturally occurring demonstrations.

1. Introduction

Machine learning systems now excel (in expectation) at tasks they are trained for by using a combination of large datasets, high-capacity models, and supervised learning (Krizhevsky et al., 2012) (Sutskever et al., 2014) (Amodei et al., 2016). Yet these systems are brittle and sensitive to slight changes in the data distribution (Recht et al., 2018) and task specification (Kirkpatrick et al., 2017). Current systems are better characterized as narrow experts rather than competent generalists. We would like to move towards more general systems which can perform many tasks – eventually without the need to manually create and label a training dataset for each one.

The dominant approach to creating ML systems is to collect a dataset of training examples demonstrating correct behavior for a desired task, train a system to imitate these behaviors, and then test its performance on independent and identically distributed (IID) held-out examples. This has served well to make progress on narrow experts. But the often erratic behavior of captioning models (Lake et al., 2017), reading comprehension systems (Jia & Liang, 2017), and image classifiers (Alcorn et al., 2018) on the diversity and variety of possible inputs highlights some of the shortcomings of this approach.

Our suspicion is that the prevalence of single task training on single domain datasets is a major contributor to the lack of generalization observed in current systems. Progress towards robust systems with current architectures is likely to require training and measuring performance on a wide range of domains and tasks. Recently, several benchmarks have been proposed such as GLUE (Wang et al., 2018) and decaNLP (McCann et al., 2018) to begin studying this.

Multitask learning (Caruana, 1997) is a promising framework for improving general performance. However, multitask training in NLP is still nascent. Recent work reports modest performance improvements (Yogatama et al., 2019) and the two most ambitious efforts to date have trained on a total of 10 and 17 (dataset, objective) pairs respectively (McCann et al., 2018) (Bowman et al., 2018). From a meta-learning perspective, each (dataset, objective) pair is a single training example sampled from the distribution of datasets and objectives. Current ML systems need hundreds to thousands of examples to induce functions which generalize well. This suggests that multitask training many need just as many effective training pairs to realize its promise with current approaches. It will be very difficult to continue to scale the creation of datasets and the design of objectives to the degree that may be required to brute force our way there with current techniques. This motivates exploring additional setups for performing multitask learning.

The current best performing systems on language tasks

[*] Equal contribution [1] OpenAI, San Francisco, California, United States. Correspondence to: Alec Radford <alec@openai.com>.

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.

* Equal contribution
† Johns Hopkins University, OpenAI

Author contributions listed at end of paper.

What is OpenAI’s GPT-3?

- The GPT-3 model architecture itself is a transformer-based neural network that has been fed 45TB of text data.
- A language model is a model that predicts the likelihood of a sentence existing in the world.
- GPT-3 is non-deterministic, in the sense that given the same input, multiple runs of the engine will return different responses.
- GPT-3 is trained on massive datasets that covered the entire web and contained 500B tokens, humongous 175 Billion parameters, a more than 100x increase over GPT-2, which was considered state-of-the-art technology with 1.5 billion parameters.
- OpenAI’s GPT-3 is still in the experimental phase.

https://arxiv.org/abs/2005.14165

What makes GPT-3 so magical?

It is really big

With 175 billion parameters, it’s the largest language model ever created

Extraordinary - Supermodel

You can ask GPT-3 to be a translator, a programmer, a poet, or a famous author, and it can do it with its user (you) providing fewer than 10 training examples.

Minimum fine-tuning

It only requires few-shot demonstrations via textual interaction with the model.

Custom language tasks without training

Task-agnostic NLP model

https://arxiv.org/abs/2005.14165

GPT-3 - a few shot learner

https://arxiv.org/abs/2005.14165

GPT-3 - a few shot learner

An overview of the original paper covering its training cost and research implications.

- GPT-3 shows that language model performance scales as a power-law of model size, dataset size, and the amount of computation.
- GPT-3 demonstrates that a language model trained on enough data can solve NLP tasks that it has never encountered. That is, GPT-3 studies the model as a general solution for many downstream jobs without fine-tuning.
- The cost of AI is increasing exponentially. Training GPT-3 would cost over $4.6M using a Tesla V100 cloud instance.
- The size of state-of-the-art (SOTA) language models is growing by at least a factor of 10 every year. This outpaces the growth of GPU memory. For NLP, the days of "embarrassingly parallel" are coming to the end; model parallelization will become indispensable.
- Although there is a clear performance gain from increasing the model capacity, it is not clear what is really going on under the hood. Especially, it remains a question of whether the model has learned to do reasoning, or simply memorizes training examples in a more intelligent way.

https://arxiv.org/abs/2005.14165

GPT-3 - a few shot learner

About GPT-3's training data

- It is a weighted mix of Common Crawl, WebText2 (a larger version of the original), two book corpora, and English Wikipedia.
- Some components (e.g. Wikipedia), were completely sampled 3+ times during training, while others like the Common Crawl, weren't even completely sampled. The authors claim that this is to help raise the overall quality of the corpus by prioritising known-good datasets.
- Altogether, the filtered/cleaned dataset is 500 billion tokens, or 700GB.
- Due to a bug, some data overlapped between the training and test sets. The paper analyzes of the impact of this leakage.

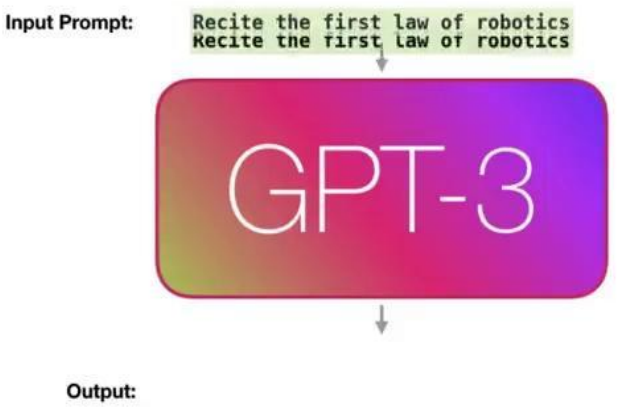
https://arxiv.org/abs/2005.14165

How does GPT-3 work?

How does GPT-3 work?

GPT- 3 language model generates text. We can optionally pass it some text as input, which influences its output.

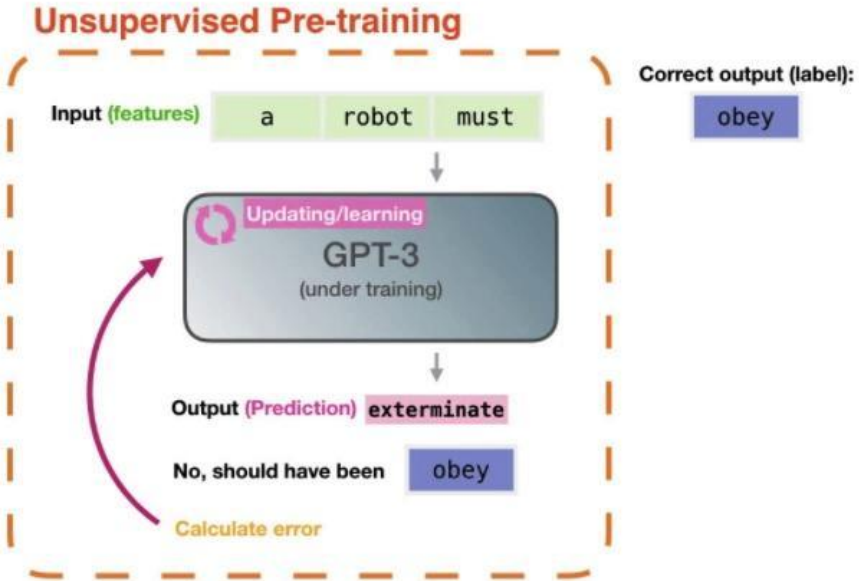
The output is generated from what the model “learned” during its training period where it scanned vast amounts of text.



https://jalammar.github.io/how-gpt3-works-visualizations-animations/

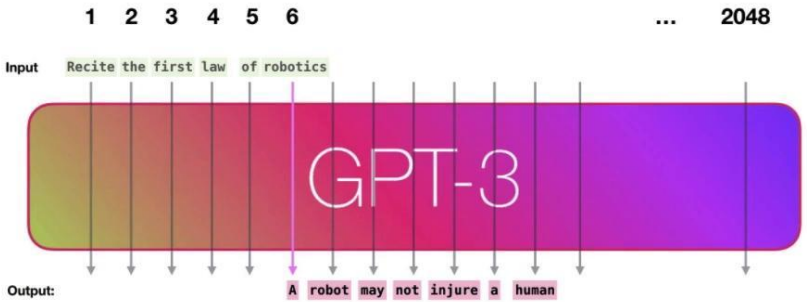
How does GPT-3 work?

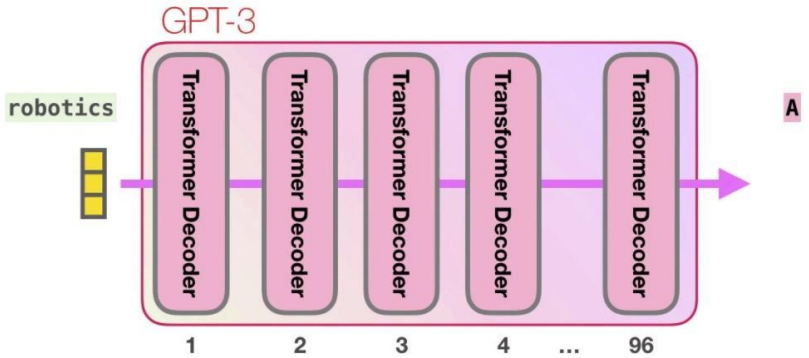
- Training is the process of exposing the model to lots of text.
- **All the experiments are from that one trained model.**
- It was estimated to cost 355 GPU years and cost $5m - $12m.



https://jalammar.github.io/how-gpt3-works-visualizations-animations/

How does GPT-3 work?



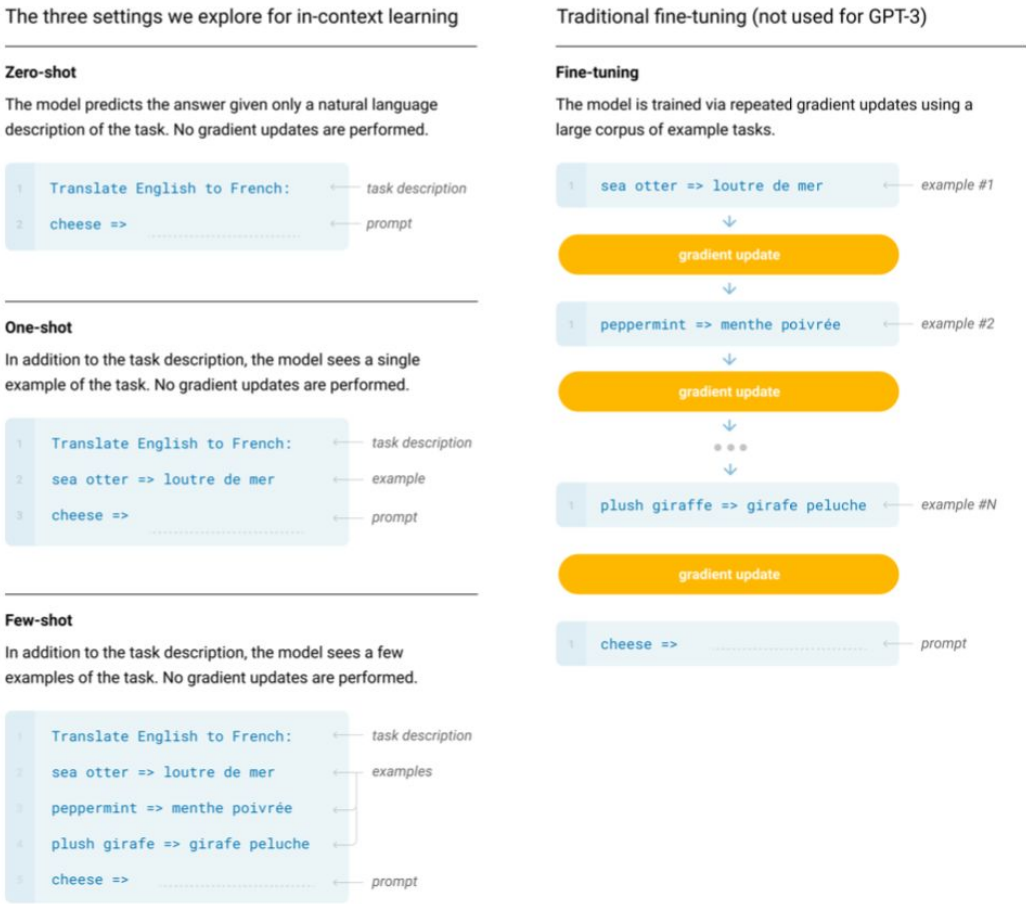


- GPT3 actually generates output one token at a time (let's assume a token is a word).
- The important calculations of the GPT3 occur inside its stack of 96 transformer decoder layers. This is the "depth" in "deep learning".
- Each of these layers has its own 1.8B parameter to make its calculations. That is where the "magic" happens.
- GPT3 is 2048 tokens wide. That is its "context window". That means it has 2048 tracks along which tokens are processed.

How does GPT-3 work?

The model is evaluated in three different settings:

- **Few-shot learning**, when the model is given a few demonstrations of the task (typically, 10 to 100) at inference time but with no weight updates allowed.
- **One-shot learning**, when only one demonstration is allowed, together with a natural language description of the task.
- **Zero-shot learning**, when no demonstrations are allowed and the model has access only to a natural language description of the task.



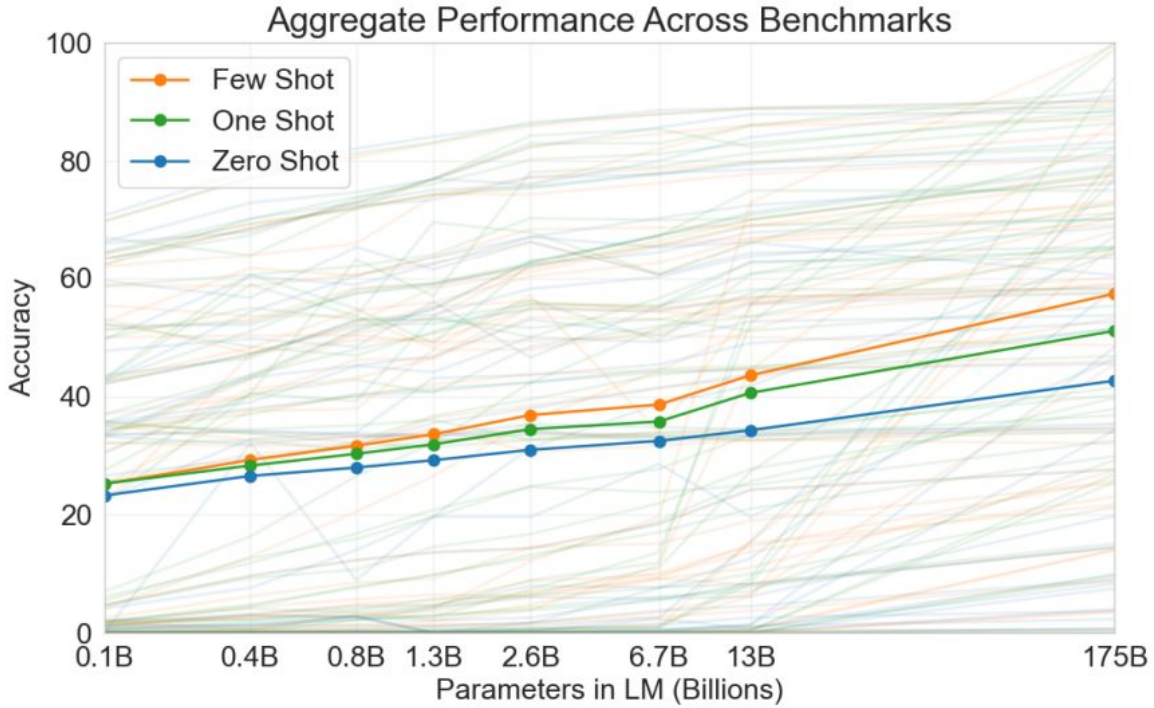
Zero-shot, one-shot and few-shot, contrasted with traditional fine-tuning.

https://arxiv.org/abs/2005.14165

Performance of GPT-3

Performance of GPT-3

- Unlike its predecessors, GPT-3 can infer a task from one or a few example: this is a massive step towards generalization.
- GPT-3 can be “tuned” by providing instructions in plain English, whereas its precessors require task-specific tuning.
- Increasing model size improves performance across almost all tasks; in contrast, fine-tuning limits performance gains to one task and risks catastrophic forgetting and overfitting.



https://arxiv.org/abs/2005.14165

Performance of GPT-3

Key Achievements:

- The GPT-3 model without fine-tuning achieves promising results on a number of NLP tasks, and even occasionally surpasses state-of-the-art models that were fine-tuned for that specific task:
    - On the **CoQA** benchmark, 81.5 F1 in the zero-shot setting, 84.0 F1 in the one-shot setting, and 85.0 F1 in the few-shot setting, compared to the 90.7 F1 score achieved by fine-tuned SOTA.
    - On the **TriviaQA** benchmark, 64.3% accuracy in the zero-shot setting, 68.0% in the one-shot setting, and 71.2% in the few-shot setting, surpassing the state of the art (68%) by 3.2%.
    - On the **LAMBADA** dataset, 76.2 % accuracy in the zero-shot setting, 72.5% in the one-shot setting, and 86.4% in the few-shot setting, surpassing the state of the art (68%) by 18%.
- **The news articles generated by the 175B-parameter GPT-3 model are hard to distinguish from real ones,** according to human evaluations (with accuracy barely above the chance level at ~52%).

https://arxiv.org/abs/2005.14165

GPT-3 Applications

GPT-3 as a Search Engine

The OpenAI' GPT-3 API allows searching over documents based on the natural-language meaning of queries rather than keyword matching.

Casetext automates litigation tasks to help attorneys

Casetext Demo

Search Engine Demo

Conversation & Chat

The API can enable fast, complex and consistent natural language discussions. With a brief prompt, the API generates dialogues spanning a range of topics, from space travel to history.

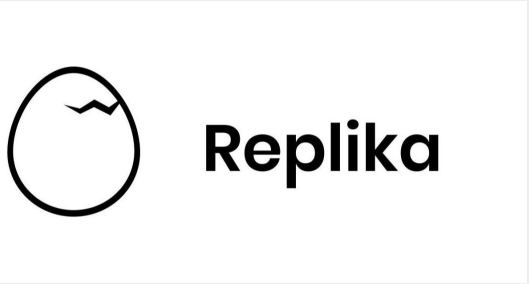
AI Channels is a social network for people and artificial intelligence agents.

AI Channels Demo

Content Comprehension & Generation

The API can be used to build tools to help individuals consume content more efficiently.

The API can generate complex and consistent natural language, and enables use cases like creative writing.



Replika, an AI companion

Replica Demo

The API can transform dense text into simplified summaries.

Summarization Demo

Meme Generation Demo

Software Engineer
&
Data Scientist

Context-aware code suggestions.

Translate natural language to unix commands.

Generate ML codes.

Words to Websites.

After fine-tuning with code from thousands of Open Source GitHub repositories, the API completes code based on function names and comments.

Code Completion Demo

Natural Language Shell Demo

No Code AI Demo

Website Template Demo

Personalization

The API can be used to build tools to help individuals consume content more efficiently.

Art of Problem Solving (AoPS) is helping to effectively prepare the next generation of STEM professionals through engaging online instruction, at a time when the traditional nature of in-person education is being challenged.

AoPS Demo

Conclusions

- GPT-3 can write.
- GPT-3 can have a conversation.
- GPT-3 can self reflect.
- GPT-3 is amazing. It’s scary. It’s exhilarating. It’s the biggest thing since bitcoin. It’s the future.