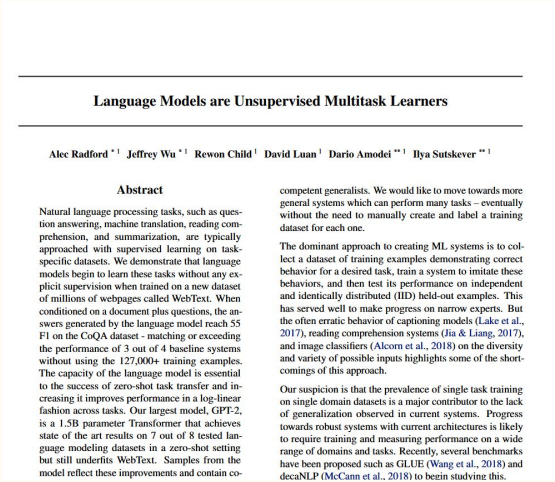
GPTs and Generalisation

Group 15

Table of Contents

Introduction

2019



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.

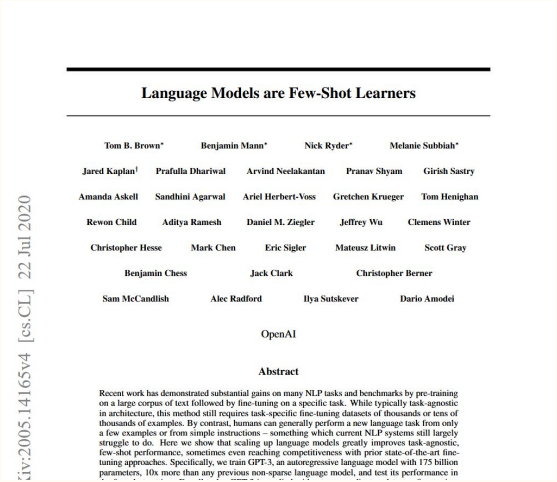
competent generalists. We would like to move towards more general systems which can perform many tasks – eventually without the need to manually create and label a training dataset for each one.

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.

Language Models are Unsupervised Multitask Learners

2020



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

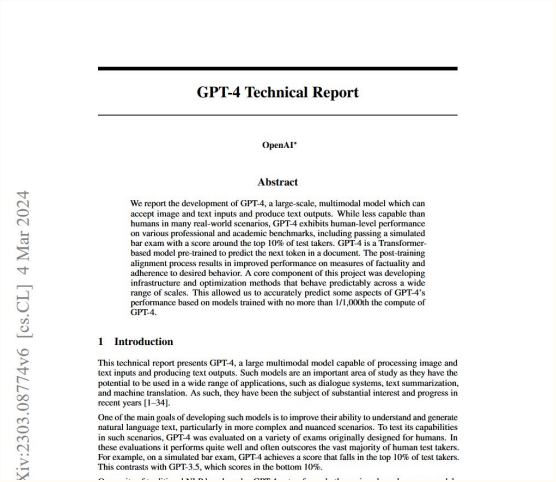
OpenAI

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

Language Models are Few-Shot Learners

2023



GPT-4 Technical Report

OpenAI*

Abstract

We report the development of GPT-4, a large-scale, multimodal model which can accept image and text inputs and produce text outputs. While less capable than humans in many real-world scenarios, GPT-4 exhibits human-level performance on various professional and academic benchmarks, including passing a simulated bar exam with a score around the top 10% of test takers. GPT-4 is a Transformer-based model pre-trained to predict the next token in a document. The post-training alignment process results in improved performance on measures of factuality and adherence to desired behavior. A core component of this project was developing infrastructure and optimization methods that behave predictably across a wide range of scales. This allowed us to accurately predict some aspects of GPT-4's performance based on models trained with no more than 1/1,000th the compute of GPT-4.

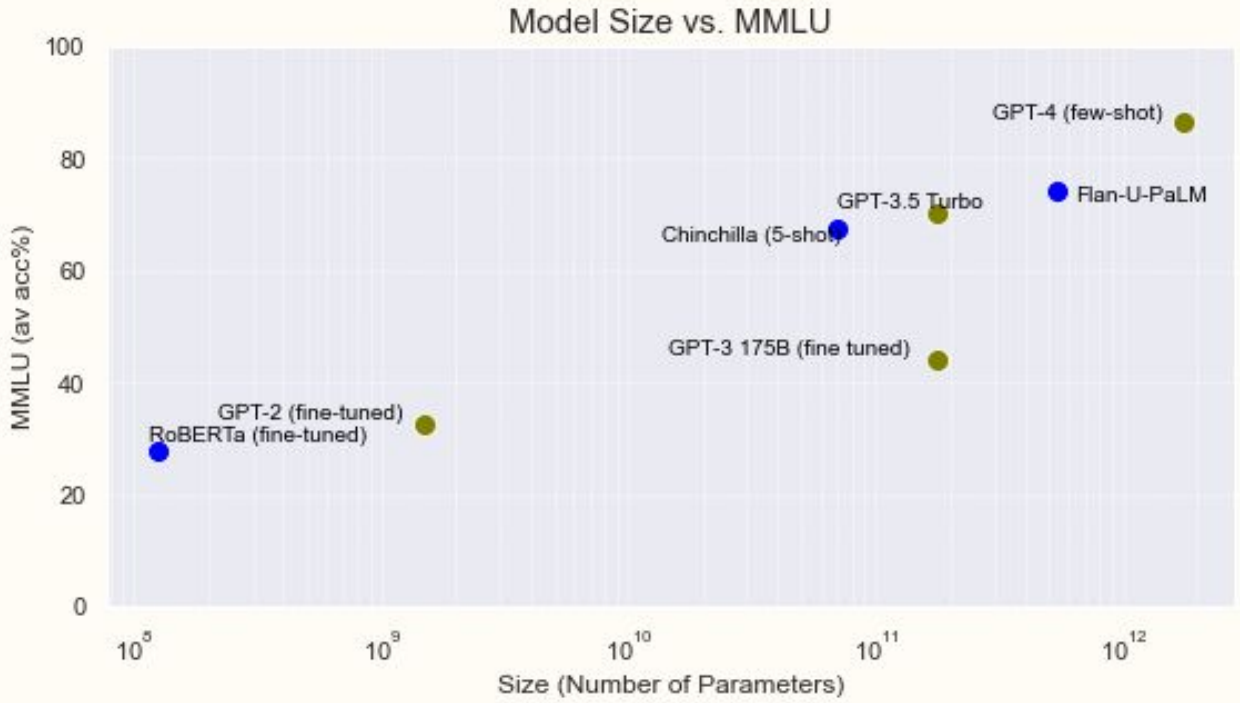
1 Introduction

This technical report presents GPT-4, a large multimodal model capable of processing image and text inputs and producing text outputs. Such models are an important area of study as they have the potential to be used in a wide range of applications, such as dialogue systems, text summarization, and machine translation. As such, they have been the subject of substantial interest and progress in recent years [1–34].

One of the main goals of developing such models is to improve their ability to understand and generate natural language text, particularly in more complex and nuanced scenarios. To test its capabilities in such scenarios, GPT-4 was evaluated on a variety of exams originally designed for humans. In these evaluations it performs quite well and often outscores the vast majority of human test takers. For example, on a simulated bar exam, GPT-4 achieves a score that falls in the top 10% of test takers. This contrasts with GPT-3.5, which scores in the bottom 10%.

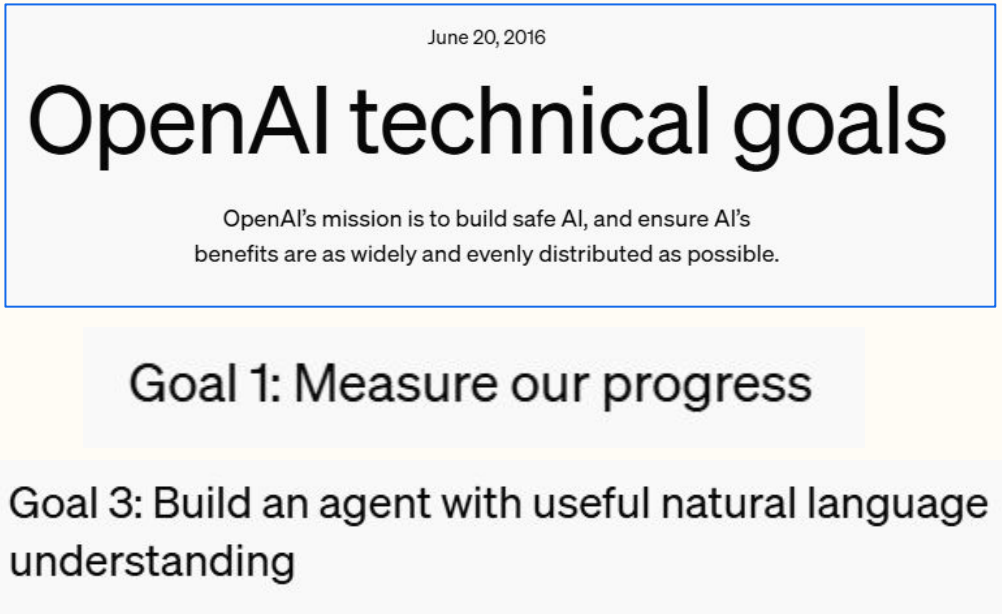
GPT-4 Technical Report

Introduction :



Context

OpenAI goals and ambitions



“Build an agent with useful natural language understanding”

A new approach to addressing NLP tasks

Single task training  **A task-agnostic language model**

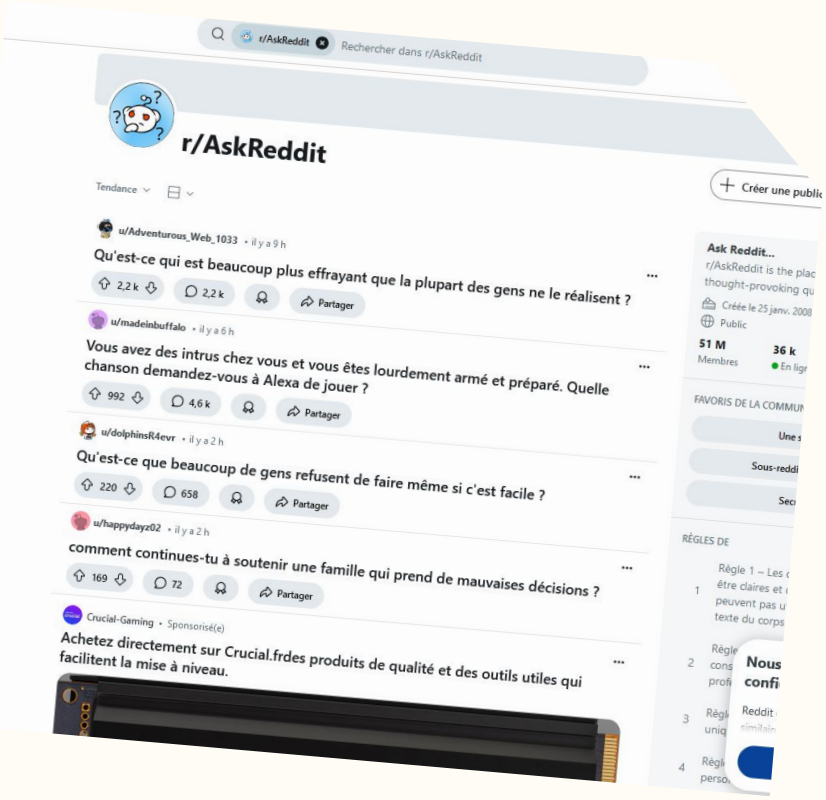
GPT-2 Goal

Modelling $p(\text{output}|\text{input}, \text{task})$

Unsupervised Multitask Learning

A new dataset

WebText



Enhanced Input representation

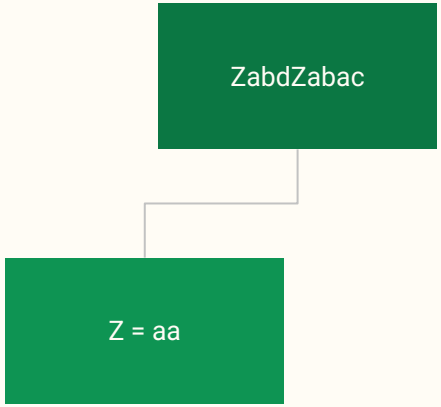
Modified version of Byte Pair Encoding

aaabdaaabac

Example from Wikipedia

Enhanced Input representation

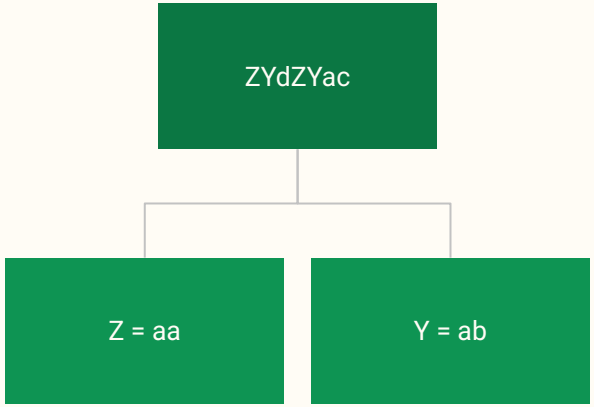
Modified version of Byte Pair Encoding



Example from Wikipedia

Enhanced Input representation

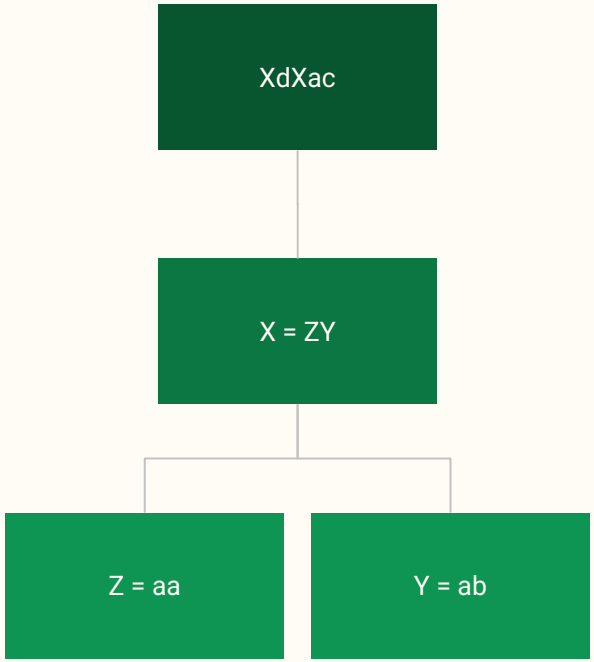
Modified version of Byte Pair Encoding



Example from Wikipedia

Enhanced Input representation

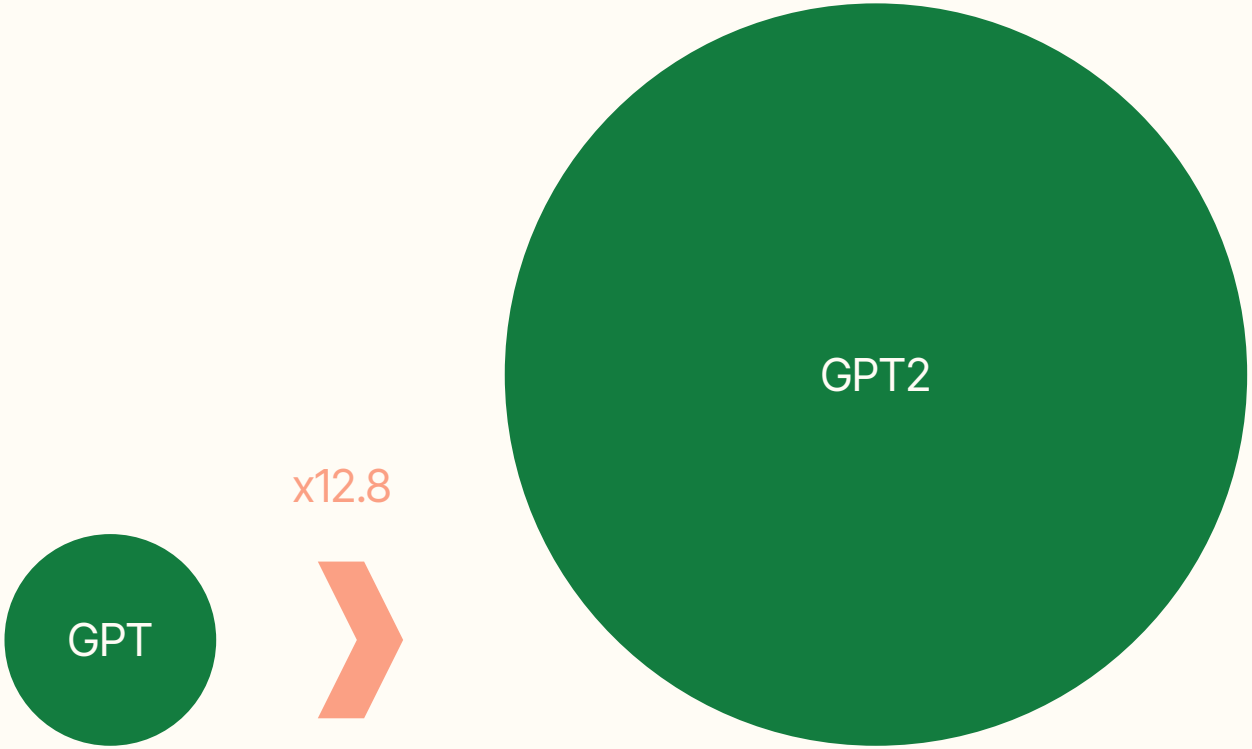
Modified version of Byte Pair Encoding



Example from Wikipedia

Bigger Model

From 0.1B to 1.5B



Achievements

Overall better results with less overlapping between WebText and test sets than dedicated training sets.

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

Table 3. Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2 results are from (Gong et al., 2018). CBT results are from (Bajgar et al., 2016). LAMBADA accuracy result is from (Hoang et al., 2018) and LAMBADA perplexity result is from (Grave et al., 2016). Other results are from (Dai et al., 2019).

Achievements

Overall better results with less overlapping between WebText and test sets than dedicated training sets.

	PTB	WikiText-2	enwik8	text8	Wikitext-103	1BW
Dataset train	**2.67%**	0.66%	**7.50%**	2.34%	**9.09%**	**13.19%**
WebText train	0.88%	**1.63%**	6.31%	**3.94%**	2.42%	3.75%

Table 6. Percentage of test set 8 grams overlapping with training sets.

Language models are few-shot learners

Few-shot learning

GPT-3 model and training data

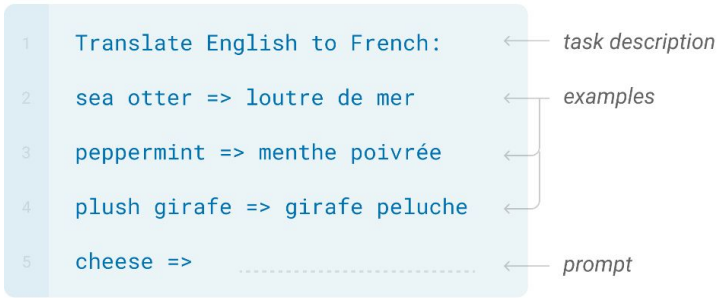
Compute costs

Few-shot vs fine tuning

Few-shot vs few-shot

Few shot learning

Model is given a few demonstrations of a task at inference time (no weight updates), for example :



This does require a small amount of task specific data.

Few-shot learning minimizes the risk of overfitting to narrow distributions by enabling models to generalize across tasks with minimal examples.

One-shot and zero-shot learning

Refers to the same thing as few-shot learning but with K = 1 or 0 examples provided in the context. One-shot is treated separately because it conforms a lot more to the format that tasks are communicated to humans.

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

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

Model

The different models use the same architecture as GPT-2 with some changes to the different layers to make it more similar to the Sparse Transformer.

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

Training data

Higher quality datasets are sampled more during training (CommonCrawl and Books2 less than once while others were sampled 2 or 3 times during training)

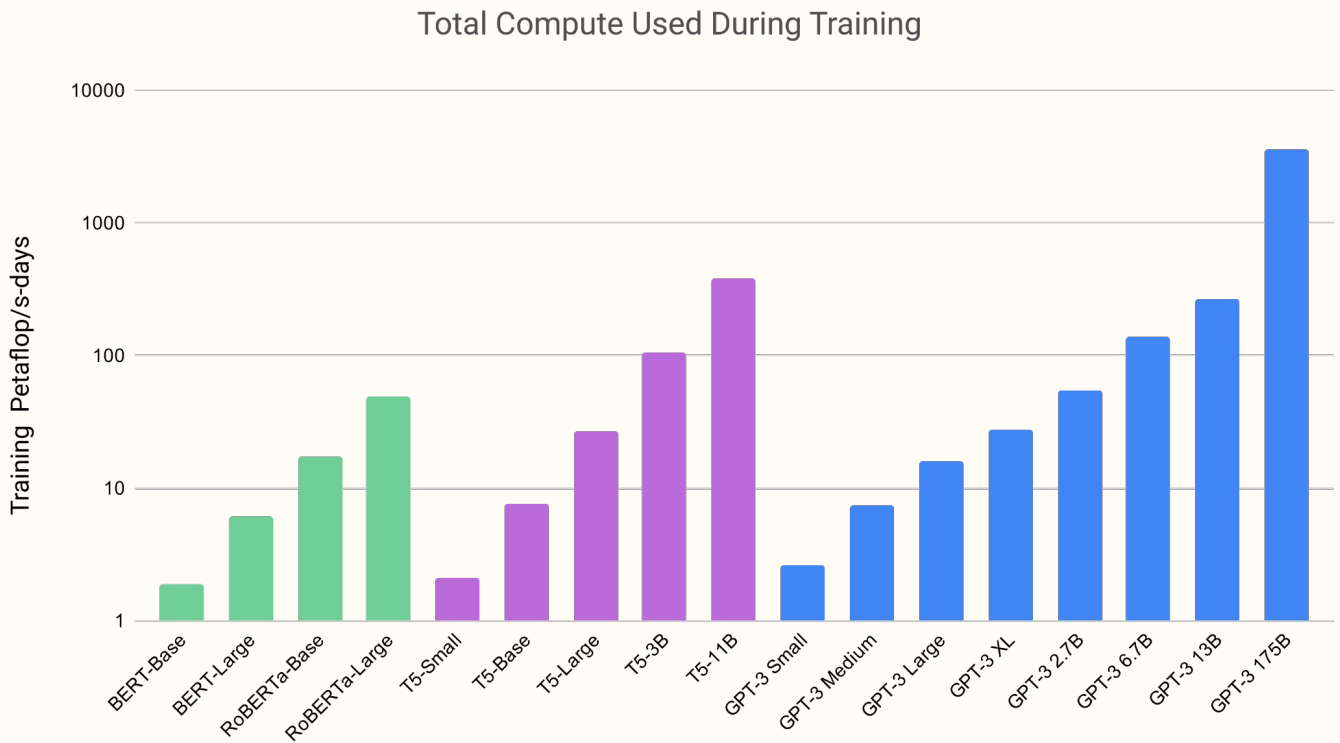
Dataset	Number of tokens	Weight in training mix
Common Crawl (filtered)	410 billion	60%
WebText2	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%

Deduplication of data within these datasets wasn't good enough which lead to a contaminated validation dataset.

Compute

Training GPT-3 (with 175B parameters) consumed thousands of petaflop/s-days of compute during pre-training, compared the tens of petaflop/s-days for a 1.5B parameter GPT-2 model.

To view the efficiency of LLMs, we shouldn't only look at the resources that go into training them but how these training costs are amortized through the lifetime of the model.


Total Compute Used During Training
Training Petaflop/s-days
10000
1000
100
10
1
BERT-Base
BERT-Large
RoBERTa-Base
RoBERTa-Large
T5-Small
T5-Base
T5-Large
T5-3B
T5-11B
GPT-3 Small
GPT-3 Medium
GPT-3 Large
GPT-3 XL
GPT-3 2.7B
GPT-3 6.7B
GPT-3 13B
GPT-3 175B

Few-shot learning leads to more efficient models by reducing the need for task specific fine tuning.

Few-shot learning vs fine tuning

GPT-3's performance is on par with a fine-tuned model on the reading comprehension task, when given K = 32 examples of said task in its prompt.

	SuperGLUE Average	BoolQ Accuracy	CB Accuracy	CB F1	COPA Accuracy	RTE Accuracy
Fine-tuned SOTA	**89.0**	**91.0**	**96.9**	**93.9**	**94.8**	**92.5**
Fine-tuned BERT-Large	69.0	77.4	83.6	75.7	70.6	71.7
GPT-3 Few-Shot	71.8	76.4	75.6	52.0	92.0	69.0

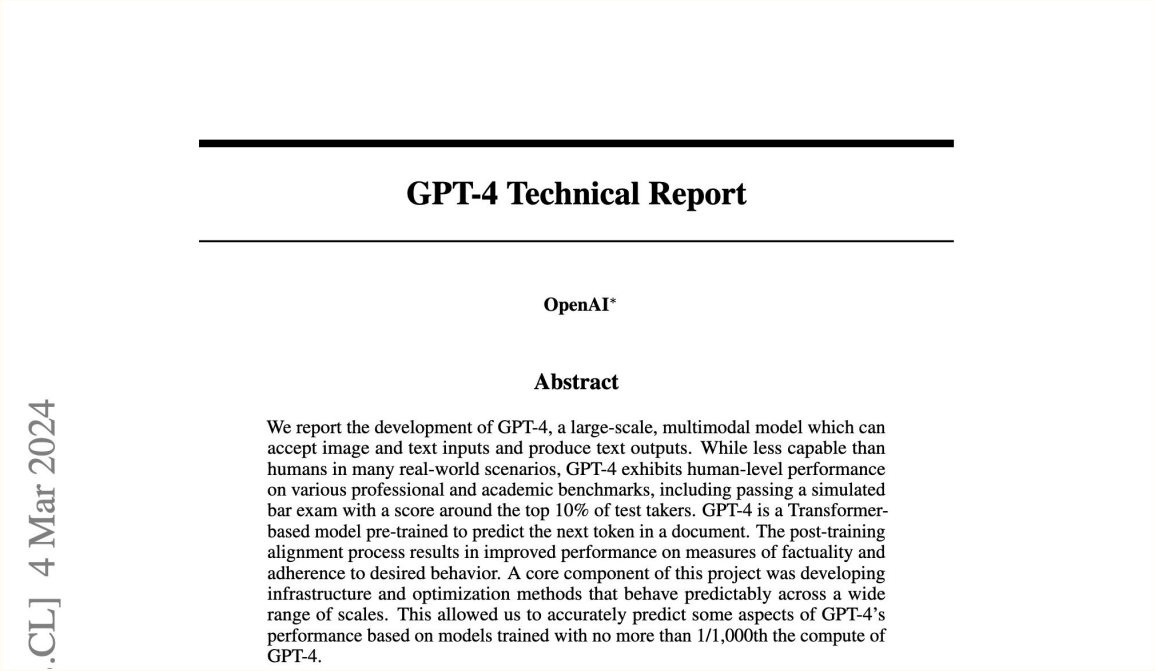
	WiC Accuracy	WSC Accuracy	MultiRC Accuracy	MultiRC F1a	ReCoRD Accuracy	ReCoRD F1
Fine-tuned SOTA	**76.1**	**93.8**	**62.3**	**88.2**	**92.5**	**93.3**
Fine-tuned BERT-Large	69.6	64.6	24.1	70.0	71.3	72.0
GPT-3 Few-Shot	49.4	80.1	30.5	75.4	90.2	91.1

Few-shot vs few-shot

On basic arithmetic tasks, GPT-3's performance increases with the amount of examples given.

Setting	2D+	2D-	3D+	3D-	4D+	4D-	5D+	5D-	2Dx	1DC
GPT-3 Zero-shot	76.9	58.0	34.2	48.3	4.0	7.5	0.7	0.8	19.8	9.8
GPT-3 One-shot	99.6	86.4	65.5	78.7	14.0	14.0	3.5	3.8	27.4	14.3
GPT-3 Few-shot	100.0	98.9	80.4	94.2	25.5	26.8	9.3	9.9	29.2	21.3

GPT-4 Technical report



GPT-4 Technical Report

OpenAI*

Abstract

We report the development of GPT-4, a large-scale, multimodal model which can accept image and text inputs and produce text outputs. While less capable than humans in many real-world scenarios, GPT-4 exhibits human-level performance on various professional and academic benchmarks, including passing a simulated bar exam with a score around the top 10% of test takers. GPT-4 is a Transformer-based model pre-trained to predict the next token in a document. The post-training alignment process results in improved performance on measures of factuality and adherence to desired behavior. A core component of this project was developing infrastructure and optimization methods that behave predictably across a wide range of scales. This allowed us to accurately predict some aspects of GPT-4's performance based on models trained with no more than 1/1,000th the compute of GPT-4.

.CL] 4 Mar 2024

Submitted on 15 Mar 2023

What's new ?

Text and Image inputs

User	What is funny about this image? Describe it panel by panel.

Source: https://www.reddit.com/r/hmmm/comments/ubab5v/hmmm/

GPT-4	The image shows a package for a "Lightning Cable" adapter with three panels. Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port. Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it. Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end. The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

RLHF/RBRM :

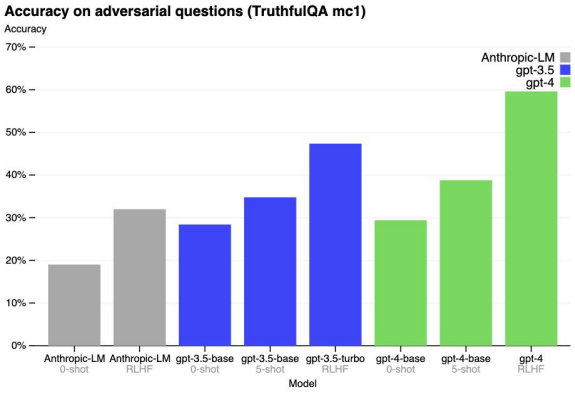
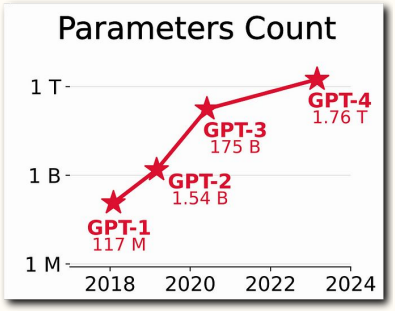


Figure 7. Performance of GPT-4 on TruthfulQA. Accuracy is shown on the y-axis, higher is better. We compare GPT-4 under zero-shot prompting, few-shot prompting, and after RLHF fine-tuning. GPT-4 significantly outperforms both GPT-3.5 and Anthropic-LM from Bai et al. [67].



Predictable scaling

Key idea : **expensive** and **long** training

→ need to **estimate performance** before training to make decisions

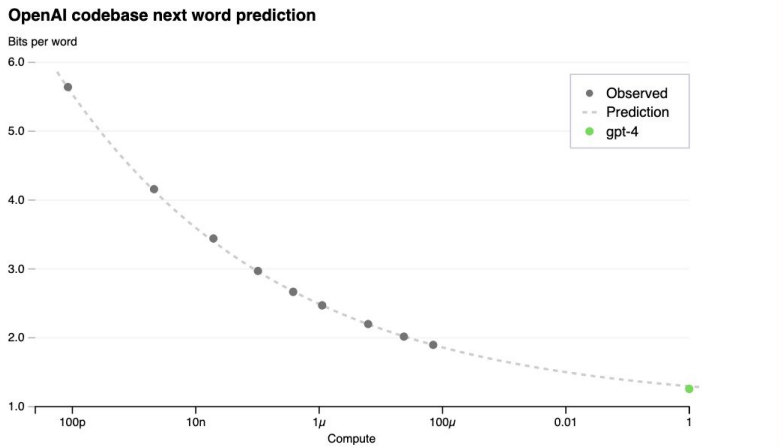


Figure 1. Performance of GPT-4 and smaller models. The metric is final loss on a dataset derived from our internal codebase. This is a convenient, large dataset of code tokens which is not contained in the training set. We chose to look at loss because it tends to be less noisy than other measures across different amounts of training compute. A power law fit to the smaller models (excluding GPT-4) is shown as the dotted line; this fit accurately predicts GPT-4's final loss. The x-axis is training compute normalized so that GPT-4 is 1.

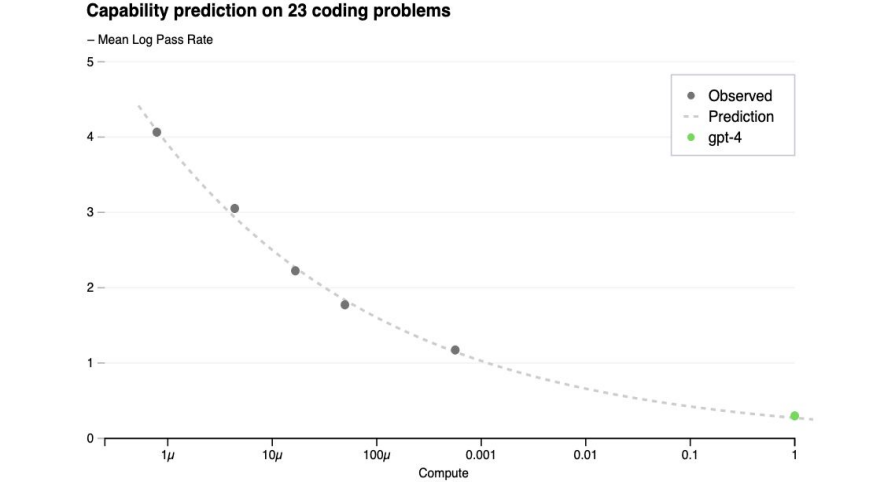


Figure 2. Performance of GPT-4 and smaller models. The metric is mean log pass rate on a subset of the HumanEval dataset. A power law fit to the smaller models (excluding GPT-4) is shown as the dotted line; this fit accurately predicts GPT-4's performance. The x-axis is training compute normalized so that GPT-4 is 1.

Safety protocol

1. Adversarial Testing with Experts

- **50+ domain experts** tested for risks (bias, misinformation, cybersecurity, biorisks).
- Identified and mitigated potential vulnerabilities (e.g., harmful content refusals improved).

2. Reinforcement Learning from Human Feedback (RLHF)

- Fine-tuning to align with ethical standards.
- **82% reduction** in harmful responses vs. GPT-3.5.
- **29% improvement** in handling sensitive topics (e.g., medical/self-harm queries).

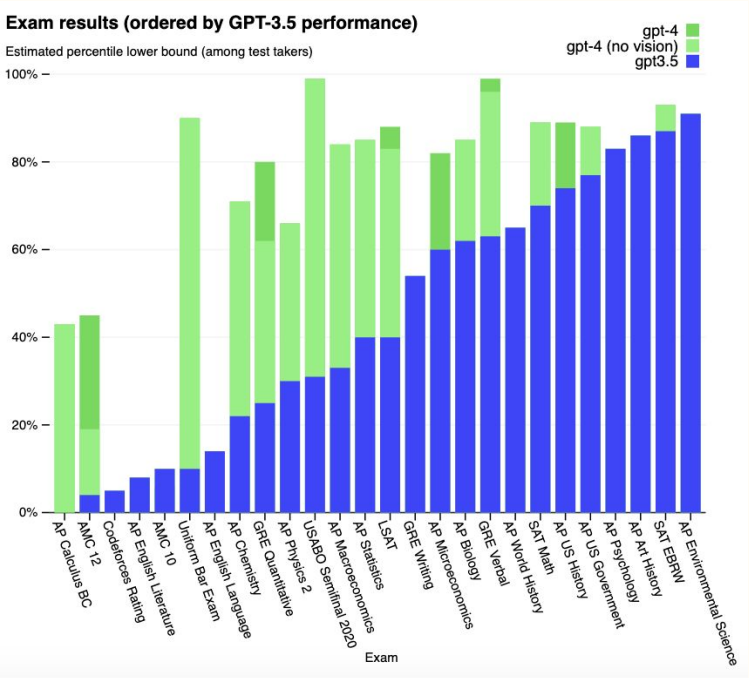
3. Rule-Based Reward Models (RBRM)

- AI-assisted evaluation to classify responses (safe, refusal, cautious).
- Ensures balance between helpfulness and safety.

4. Safety Pipeline & Monitoring

- Content moderation, human oversight, and user feedback integration.
- Reduced toxicity: **GPT-4: 0.73% vs. GPT-3.5: 6.48%** on RealToxicityPrompts.

Performance



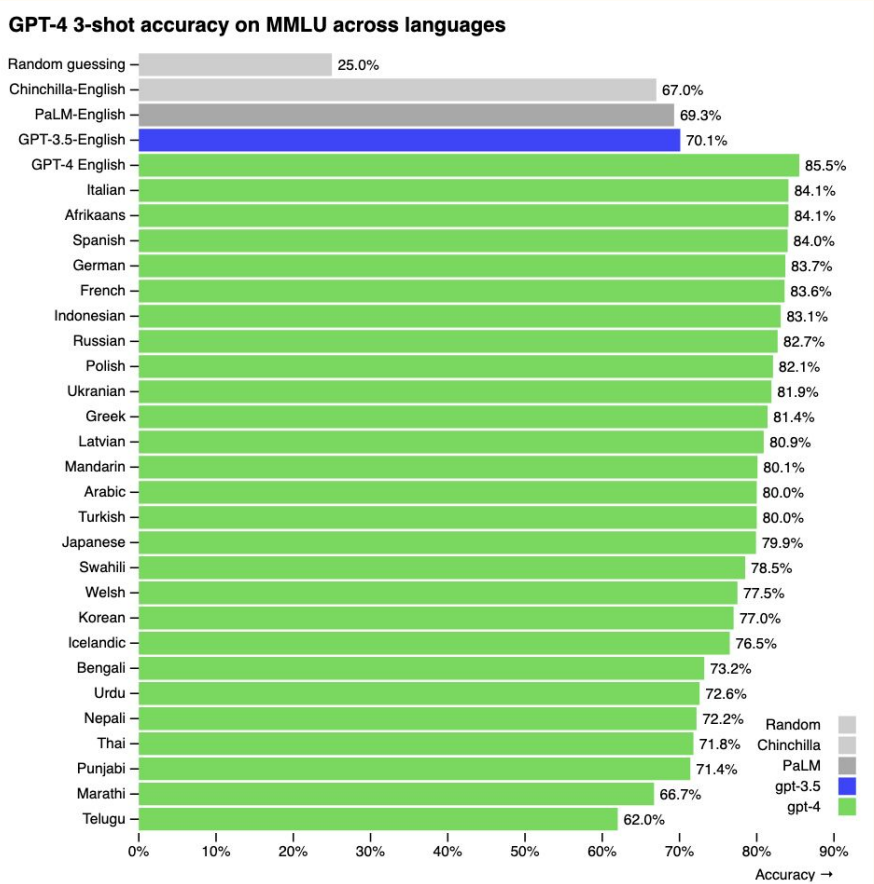


Figure 5. Performance of GPT-4 in a variety of languages compared to prior models in English on MMLU. GPT-4 outperforms the English-language performance of existing language models [2, 3] for the vast majority of languages tested, including low-resource languages such as Latvian, Welsh, and Swahili.

Key takeaways

- **Bigger models and training size** allow for improved results

  → starting from GPT-2 (1.5B parameters) to GPT-4 (>1T parameters) using large models is a consistent idea at OpenAI

- These models are **few/zero shot learners**

  → the models outperform smaller fine-tuned models

  → rise of ‘prompt engineering’

- They are now the gold standard of the industry but are still **susceptible to adversarial attack**

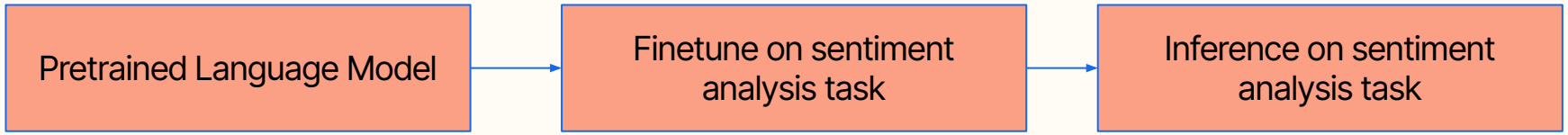
- Big models and great performance → great **costs** and **responsibilities**

Questions ?

GPT 3 : Language Models are Few-Shot Learners

Task-specific models vs Prompting

Models such as BERT are pretrained on a large amount of general data and then fine-tuned on specific tasks such as sentiment analysis or question answering



Whereas models like GPT-3 rely on prompt engineering (few-shot task descriptions/examples) to adapt to particular tasks.

Comparing few-shot with fine tuned BERT from the second paper

	SuperGLUE Average	BoolQ Accuracy	CB Accuracy	CB F1	COPA Accuracy	RTE Accuracy
Fine-tuned SOTA	**89.0**	**91.0**	**96.9**	**93.9**	**94.8**	**92.5**
Fine-tuned BERT-Large	69.0	77.4	83.6	75.7	70.6	71.7
GPT-3 Few-Shot	71.8	76.4	75.6	52.0	92.0	69.0