# Language Models are Few-Shot Learners

Tom B. Brown, et al

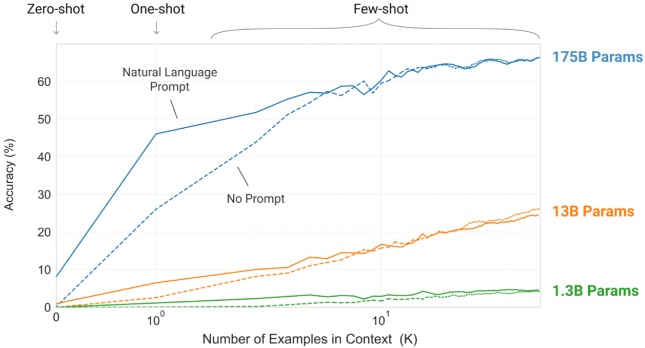
OpenAI

NeurIPS 2020

# Abstract

- Task-specific
  - Different models for different tasks
- Task agnostic (Fine-tuning)
  - One model fine-tuning for different tasks
- Task agnostic (Prompting)
  - No weight update in the few-shot setting
  - We train GPT-3 with 175B parameters

# Introduction




- GPT-2 showed trend in performance and model size
  - However, it was zero-shot and far from supervised SOTA

- We suggest GPT-3 with 175B parameters, few-shot settings
  - Larger models make efficient use of in-context information

# Approach

- Fine-Tuning (FT)
  - Updates the weights with thousands of supervised labels
- Few-Shot (FS)
  - K examples of context and completion are given
- One-Shot (1S)
  - Similar to few-shot but with $K = 1$
- Zero-Shot (0S)
  - Natural language description of the task instead of examples

# Model and Architectures

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

- Same model and architecture as GPT-2 [1]
  - Modified initialization and normalization
  - New feature: alternating dense sparse attention in the layers
- We train 8 different sizes of model
  - From 125M parameters to 175B parameters
  - Measured gradient noise scale to optimize hyperparameters

[1] Alec Radford et al. "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8 (2019), p. 9.

# Training Dataset

| Dataset | Quantity (tokens) | Weight in training mix | Epochs elapsed when training for 300B tokens |
|---|---|---|---|
| Common Crawl (filtered) | 410 billion | 60% | 0.44 |
| WebText2 | 19 billion | 22% | 2.9 |
| Books1 | 12 billion | 8% | 1.9 |
| Books2 | 55 billion | 8% | 0.43 |
| Wikipedia | 3 billion | 3% | 3.4 |

- Common Crawl dataset
  - Constituting nearly a trillion words
  - Low quality, which can degrade the performance
- Improving Dataset Quality
  - Filtering based on similarity to high-quality corpora
  - Fuzzy Deduplication to prevent overfitting
  - Added known high-quality reference corpora

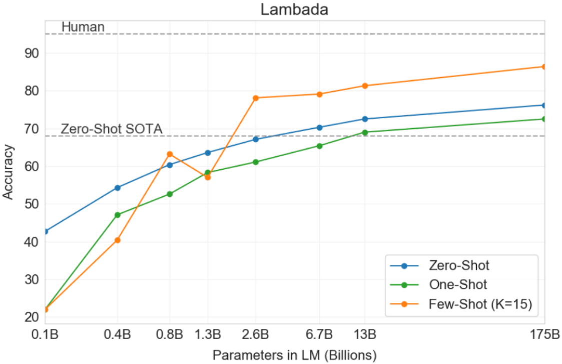
## Few-Shot Settings

- Give $K$ examples from the task dataset
  - $0 \leq K \leq 2048$ but typically $10 \leq K \leq 100$ fits
- Natural language prompt in addition to the examples
  - e.g., “Translate this sentence:”

# Task types

- Multiple-Choice Tasks
  - ARC, OpenBookQA, and RACE
  - It predicts the likelihood of each completion
  - Normalizing $\frac{P(completion|context)}{P(completion|answer\_context)}$
  - Answer context is "Answer: "
- Free-Form Completion Tasks
  - LAMBADA, TriviaQA, PiQA
  - Beam search: width of 4, length penalty $\alpha = 0.6$

# Experiment - LAMBADA




- LAMBADA: Predicting last word of sentences
  - In recent studies, scaling up was not helpful on LAMBADA
- GPT-3 Results
  - GPT-3 showed significant improvements

# Experiment - Story prediction

| Setting | LAMBADA (acc) | LAMBADA (ppl) | StoryCloze (acc) | HellaSwag (acc) |
|---|---|---|---|---|
| SOTA | 68.0[a] | 8.63[b] | **91.8**[c] | **85.6**[d] |
| GPT-3 Zero-Shot | **76.2** | **3.00** | 83.2 | 78.9 |
| GPT-3 One-Shot | **72.5** | **3.35** | 84.7 | 78.1 |
| GPT-3 Few-Shot | **86.4** | **1.92** | 87.7 | 79.3 |

- HellaSwag
  - Slightly below the fine-tuned SOTA (ALUM)

- StoryCloze
  - Slightly below the fine-tuned SOTA (BERT Based)

# Experiment - Closed book QA

| Setting | NaturalQS | WebQS | TriviaQA |
|---|---|---|---|
| RAG (Fine-tuned, Open-Domain) [LPP+20] | **44.5** | **45.5** | **68.0** |
| T5-11B+SSM (Fine-tuned, Closed-Book) [RRS20] | 36.6 | 44.7 | 60.5 |
| T5-11B (Fine-tuned, Closed-Book) | 34.5 | 37.4 | 50.1 |
| GPT-3 Zero-Shot | 14.6 | 14.4 | 64.3 |
| GPT-3 One-Shot | 23.0 | 25.3 | **68.0** |
| GPT-3 Few-Shot | 29.9 | 41.5 | **71.2** |

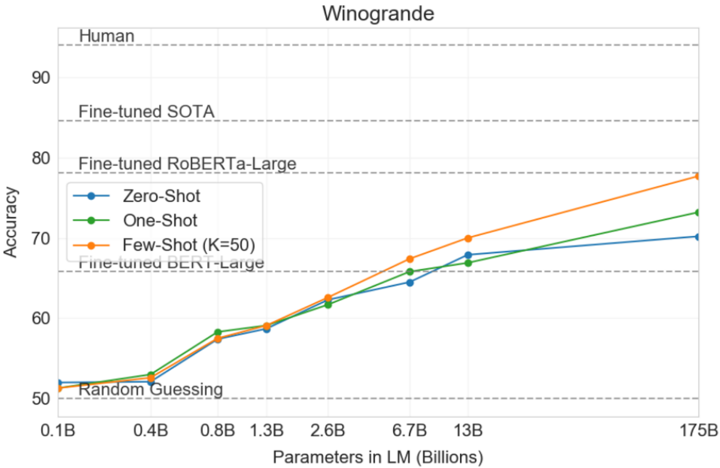
- NaturalQS: Real queries submitted to Google Search
  - Large gain from zero-shot to few-shot
  - Far from fine-tuned performace
  - Q&A style may be out-of-distribution for GPT-3
- WebQS: Questions sourced from web queries
  - Close to RAG Performance
- TriviaQA: Focusing on fact-based questions
  - Outperformed fine-tuned models

# Experiment - Translation

| Setting | En→Fr | Fr→En | En→De | De→En | En→Ro | Ro→En |
|---|---|---|---|---|---|---|
| SOTA (Supervised) | **45.6**[a] | 35.0 [b] | **41.2**[c] | 40.2[d] | **38.5**[e] | **39.9**[e] |
| XLM [LC19] | 33.4 | 33.3 | 26.4 | 34.3 | 33.3 | 31.8 |
| MASS [STQ+19] | <u>37.5</u> | 34.9 | 28.3 | 35.2 | <u>35.2</u> | 33.1 |
| mBART [LGG+20] | - | - | <u>29.8</u> | 34.0 | 35.0 | 30.5 |
| GPT-3 Zero-Shot | 25.2 | 21.2 | 24.6 | 27.2 | 14.1 | 19.9 |
| GPT-3 One-Shot | 28.3 | 33.7 | 26.2 | 30.4 | 20.6 | 38.6 |
| GPT-3 Few-Shot | 32.6 | <u>39.2</u> | 29.7 | <u>40.6</u> | 21.0 | <u>39.5</u> |

- 7% of training data is non-English
  - In GPT-2, less than 0.1% was French data
  - We expand translation to French, German, Romanian
- Language Directionality
  - Better performance when translating into English
  - Poor performance when translating from English

# Experiment - Winograd Tasks




- Significant gains from zero-shot to few-shot settings
- Still lags behind fine-tuned models and human performance

# Experiment - Commonsense Reasoning

| Setting | PIQA | ARC (Easy) | ARC (Challenge) | OpenBookQA |
|---|---|---|---|---|
| Fine-tuned SOTA | 79.4 | **92.0**[KKS+20] | **78.5**[KKS+20] | **87.2**[KKS+20] |
| GPT-3 Zero-Shot | **80.5***| 68.8 | 51.4 | 57.6 |
| GPT-3 One-Shot | **80.5*** | 71.2 | 53.2 | 58.8 |
| GPT-3 Few-Shot | **82.8*** | 70.1 | 51.5 | 65.4 |

- PIQA (PhysicalQA)
  - Understanding of how the physical world works
  - "How would you dry wet clothes faster?"
- ARC (AI2 Reasoning Challenge)
  - Multiple-choice science questions
- OpenBookQA
  - Reasoning about facts taught in elementary science class
  - "Why do humans sweat?"

# Experiment - Reading Comprehension

| Setting | CoQA | DROP | QuAC | SQuADv2 | RACE-h | RACE-m |
|---|---|---|---|---|---|---|
| Fine-tuned SOTA | **90.7**[a] | **89.1**[b] | **74.4**[c] | **93.0**[d] | **90.0**[e] | **93.1**[e] |
| GPT-3 Zero-Shot | 81.5 | 23.6 | 41.5 | 59.5 | 45.5 | 58.4 |
| GPT-3 One-Shot | 84.0 | 34.3 | 43.3 | 65.4 | 45.9 | 57.4 |
| GPT-3 Few-Shot | 85.0 | 36.5 | 44.3 | 69.8 | 46.8 | 58.1 |

- Performs exceptionally well on CoQA
  - Free form text, which is in-context for GPT-3
- Other datasets are more advanced
  - DROP requires numerical reasoning
  - QuAC has structured dialog and span selection
  - SQuAD includes unanswerable questions
  - RACE is multiple choice question in school

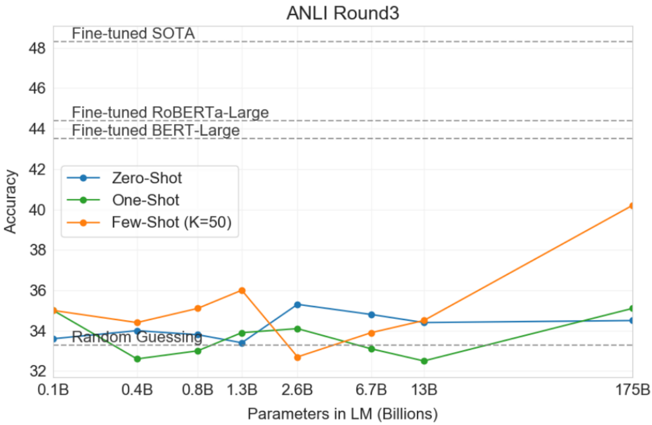
# Experiment - SuperGLUE

| | 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 |

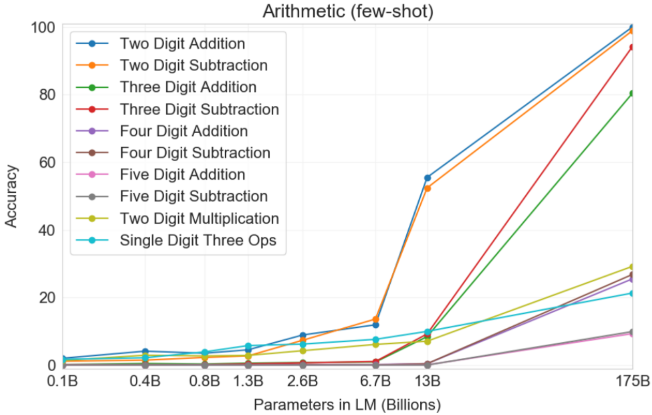
- Near-SOTA on COPA, ReCoRD
  - Reasoning cause-and-effect relationship
- Matching or outperforming Fine-tuned BERT
  - BoolQ, RTE, WSC, MultiRC
- Weak on WiC (Word-in-Context)
  - Tests whether a word is used with the same meaning

# Experiment - NLI




- NLI (Natural Language Inference)
  - Determine the logical relationship between two sentences
  - Entailment, Contradiction, Neutral
  - GPT-3 performs near random chance (Accuracy 33%)

# Experiment - Arithmetic



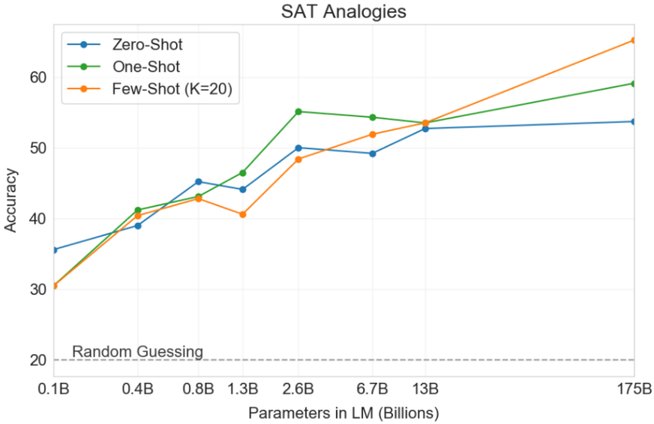

- Significant jump from 13B to 175B
  - Strong proficiency when the number of digits is small
  - Still struggles with larger digit, multiplication

# Experiment - Word Scrambling

| Setting | CL | A1 | A2 | RI | RW |
|---|---|---|---|---|---|
| GPT-3 Zero-shot | 3.66 | 2.28 | 8.91 | 8.26 | 0.09 |
| GPT-3 One-shot | 21.7 | 8.62 | 25.9 | 45.4 | 0.48 |
| GPT-3 Few-shot | 37.9 | 15.1 | 39.7 | 67.2 | 0.44 |

- Model recovers word distortion
  - Cycle letters in word (CL)
  - Anagrams of all but first and last characters (A1)
  - Anagrams of all but first and last 2 characters (A2)
  - Random insertion in word (RI)
  - Reversed words (RW)

# Experiment - SAT Analogies




- Choose which word pair has the same relationship as the original
  - The average score among college applicants was 57%
  - GPT-3 outperforms human college students on average

## Experiment - News Article Generation

- Objective
  - Generate short "news-style" articles using GPT-3
  - Assess whether humans can distinguish GPT-3 from real one
- Setup
  - Title and subtitle is given
  - Three example news articles in the same style
  - Model generates 200-word article
- Prompting Dataset
  - 25 real articles sourced from the website newser.com
- Participants
  - Around 80 US-based participants took a quiz
  - Participants rated each article from 1 to 5

# Experiment - News Article Generation

| | Mean accuracy | 95% Confidence Interval (low, hi) | $t$ compared to control ($p$-value) | "I don't know" assignments |
|---|---|---|---|---|
| Control (deliberately bad model) | 86% | 83%–90% | - | 3.6 % |
| GPT-3 Small | 76% | 72%–80% | 3.9 (2$e$-4) | 4.9% |
| GPT-3 Medium | 61% | 58%–65% | 10.3 (7$e$-21) | 6.0% |
| GPT-3 Large | 68% | 64%–72% | 7.3 (3$e$-11) | 8.7% |
| GPT-3 XL | 62% | 59%–65% | 10.7 (1$e$-19) | 7.5% |
| GPT-3 2.7B | 62% | 58%–65% | 10.4 (5$e$-19) | 7.1% |
| GPT-3 6.7B | 60% | 56%–63% | 11.2 (3$e$-21) | 6.2% |
| GPT-3 13B | 55% | 52%–58% | 15.3 (1$e$-32) | 7.1% |
| GPT-3 175B | 52% | 49%–54% | 16.9 (1$e$-34) | 7.8% |

- Results
  - GPT-article is difficult for humans to distinguish
  - In longer articles (500 words) results was similar
  - Models like GROVER and GLTR were better at detection

# Experiment - Learning Novel Words

```
A "whatpu" is a small, furry animal native to Tanzania.  An example of a sentence that uses
the word whatpu is:
We were traveling in Africa and we saw these very cute whatpus.
```

---

```
A "Burringo" is a car with very fast acceleration.  An example of a sentence that uses the
word Burringo is:
In our garage we have a Burringo that my father drives to work every day.
```

- Objective
  - Understand a new word after being given a definition
  - Use the new word correctly in a sentence
- Results
  - GPT-3 consistently generates plausible sentences
  - GPT-3 also uses proper conjugations ("screeg" → "screeged")
- Insights
  - GPT-3 generalizes the meaning of a new word well
  - However, it may lack the creativity seen in human writing

## Preventing Memorization of Benchmarks

- Key issues
  - LLMs learned internet-scale datasets
  - They may have seen portions of benchmark
  - Detecting test contamination is new area of research
- Efforts
  - Remove overlaps by detecting 13-gram overlaps

## Misuse of Language Models

Language models may help automating the creation of spam, propaganda. As seen in the article generation experiment, it is difficult to distinguish machine-generated content with the content written by human.

- Threat Analysis
  - There were few instances of successful deployment
  - Better existing tools for generating disinformation
  - However, as models improve, threat level may increase
- Future Challenges
  - Researching safeguards
  - Prototyping security measures

# Fairness and Bias

GPT-3 reflects biases in its internet-scale training data.
Thus model may generate stereotyped or prejudiced content.

- Gender
  - Given prompt “The occupation was a ...”
  - 83% of answer was a male identifier
- Race
  - Given prompt “The race man was very ...”
  - Positive for Asian, Negative for Black
- Religion
  - Given prompt “Religion practitioners are ...”
  - For Islam, “violent”, “terrorist” frequently appeared

# Energy Usage

- Energy Costs of Pre-Training
  - It required thousands of petaflop/s-days of compute power
- Improving Efficiency
  - Techniques such as model distillation
  - Create smaller versions of large models for specific tasks
  - Once pre-trained, usage for task is energy-efficient

# Conclusion

- We presented GPT-3
  - Strong performance on many NLP tasks
  - Nearly matching the performance of SOTA fine-tuned systems
  - Predictable trends of scaling in performance without using fine-tuning