GPT-3: Few-Shot Learning with a Giant Language Model
Melanie Subbiah

OpenAI
Columbia University
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
What is the goal?

Humans learn new tasks through demonstrations and instructions.
What is the goal?

Humans learn new tasks through demonstrations and instructions.

We’d like general-purpose agents that can do the same.
Typical Approach

1. sea otter => loutre de mer

   gradient update

2. peppermint => menthe poivrée

   gradient update

   ...

N. plush giraffe => girafe peluche

   gradient update

1. cheese => ........................................ prompt
Disadvantages to Fine-tuning

- Creates a task-specific model
Disadvantages to Fine-tuning

- Creates a task-specific model
- Requires large high-quality supervised datasets
Disadvantages to Fine-tuning

- Creates a task-specific model
- Requires large high-quality supervised datasets
- more likely to exploit spurious correlations

Yogatama, et al. Learning and Evaluating General Linguistic Intelligence. 2019
What is an alternative?

Context (human-written): In a shocking finding, scientists discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

GPT-2: The scientist named the population, after their distinctive horn, Ovid’s Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Radford, et al. Language Models are Unsupervised Multitask Learners. 2019
Can we further improve on this level of generation and generalization?
Can we further improve on this level of generation and generalization?

GPT-3 175 Billion parameters
Critical Aspects of GPT-3

• Model Size

• Training Objective
Model Size

Model Size

Transformers scale well!

Motivating the Training Objective

Predict the next word in a sequence.
Motivating the Training Objective

$P(\text{“The cat sat on the mat.”}) = ???$
Motivating the Training Objective

\[ P(\text{“The cat sat on the mat.”}) = ??? \]

“But it must be recognized that the notion of ‘probability of a sentence’ is an entirely useless one, under any known interpretation of this term.”
- Noam Chomsky, 1969
Motivating the Training Objective

\[ P(\text{“The cat sat on the mat.”}) > P(\text{“The cat sats on the mat.”}) \]
Motivating the Training Objective

P("The cat sat on the mat.") > P("The cat sats on the mat.")

Grammar
Motivating the Training Objective

\[ P(\\text{“The cat sat on the mat.”}) > P(\\text{“The cat sats on the mat.”}) \]

Grammar

\[ P(\\text{“The cat sat on the mat.”}) > P(\\text{“The whale sat on the mat.”}) \]
Motivating the Training Objective

P(“The cat sat on the mat.”) > P(“The cat sats on the mat.”)
Grammar

P(“The cat sat on the mat.”) > P(“The whale sat on the mat.”)
World Knowledge
Motivating the Training Objective

\[ P(\text{"The cat sat on the mat."}) > P(\text{"The cat sats on the mat."}) \]

Grammar

\[ P(\text{"The cat sat on the mat."}) > P(\text{"The whale sat on the mat."}) \]

World Knowledge

\[ P(\text{"4" | "2 + 2 ="}) > P(\text{"5" | "2 + 2 ="}) \]
Motivating the Training Objective

P(“The cat sat on the mat.”) > P(“The cat sats on the mat.”)
Grammar

P(“The cat sat on the mat.”) > P(“The whale sat on the mat.”)
World Knowledge

P(“4” | “2 + 2 =”) > P(“5” | “2 + 2 =”)
Arithmetic
Motivating the Training Objective

\[ P(\text{"The cat sat on the mat."}) > P(\text{"The cat sats on the mat."}) \]

Grammar

\[ P(\text{"The cat sat on the mat."}) > P(\text{"The whale sat on the mat."}) \]

World Knowledge

\[ P(\text{"4" | "2 + 2 ="}) > P(\text{"5" | "2 + 2 ="}) \]

Addition

\[ P(\text{"1 star" | "That movie was terrible. I’d give it"}) > P(\text{"5 stars" | "That movie was terrible. I’d give it"}) \]
Motivating the Training Objective

\[
P(\text{"The cat sat on the mat."}) > P(\text{"The cat sats on the mat."})
\]
Grammar

\[
P(\text{"The cat sat on the mat."}) > P(\text{"The whale sat on the mat."})
\]
World Knowledge

\[
P(\text{"4"} | \text{"2 + 2 ="}) > P(\text{"5"} | \text{"2 + 2 ="})
\]
Addition

\[
P(\text{"1 star"} | \text{"That movie was terrible. I’d give it"}) > P(\text{"5 stars"} | \text{"That movie was terrible. I’d give it"})
\]
Sentiment Analysis
Approach
Model

http://jalammar.github.io/illustrated-gpt2/
The trophy didn’t fit in the suitcase because the trophy was too big.
Model Sizes

Parameter Count (in millions)

- Small
- Medium
- Large
- XL
- 2.7B
- 6.7B
- 13B
- 175B

Model Sizes

Parameter Count (in millions)

- Small
- Medium
- Large
- XL
- 2.7B
- 6.7B
- 13B
- 175B
Model Sizes

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter Count (in millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT-Large</td>
<td>2.7B</td>
</tr>
<tr>
<td>T5-Large</td>
<td>6.7B</td>
</tr>
<tr>
<td>Megatron</td>
<td>13B</td>
</tr>
<tr>
<td>T-NLG</td>
<td>175B</td>
</tr>
</tbody>
</table>

Devlin, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018
Raffel, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. 2019
Microsoft. Turing-NLG: A 17-Billion Parameter Language Model by Microsoft. 2020
<table>
<thead>
<tr>
<th>Model Size</th>
<th>Parameter Count (in millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>22500</td>
</tr>
<tr>
<td>Medium</td>
<td>45000</td>
</tr>
<tr>
<td>Large</td>
<td>67500</td>
</tr>
<tr>
<td>XL</td>
<td>90000</td>
</tr>
<tr>
<td>2.7B</td>
<td>112500</td>
</tr>
<tr>
<td>6.7B</td>
<td>135000</td>
</tr>
<tr>
<td>13B</td>
<td>157500</td>
</tr>
<tr>
<td>175B</td>
<td>180000</td>
</tr>
</tbody>
</table>

**Model Sizes**

![Bar chart showing model sizes and parameter counts](chart)
Model Sizes

Parameter Count (in millions)

<table>
<thead>
<tr>
<th>Size</th>
<th>Parameter Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>1</td>
</tr>
<tr>
<td>Medium</td>
<td>10</td>
</tr>
<tr>
<td>Large</td>
<td>100</td>
</tr>
<tr>
<td>XL</td>
<td>1000</td>
</tr>
<tr>
<td>2.7B</td>
<td>10000</td>
</tr>
<tr>
<td>6.7B</td>
<td>100000</td>
</tr>
<tr>
<td>13B</td>
<td>1000000</td>
</tr>
<tr>
<td>175B</td>
<td>100000000</td>
</tr>
</tbody>
</table>
Model Sizes

- **Parameter Count (in millions):**
  - Small: 1
  - Medium: 10
  - Large: 100
  - XL: 1000
  - 2.7B: 10000
  - 6.7B: 100000
  - 13B: 1000000
  - 175B

- **Batch Size:**
  - Small: 2.7B
  - Medium: 6.7B
  - Large: 13B
  - XL: 175B

- **Learning Rate**
Compute
Dataset

- **Common Crawl** (filtered) - general web crawl, filtered based on similarity to high-quality reference and de-duplication

https://commoncrawl.org/the-data
Dataset

- **Common Crawl** (filtered) - general web crawl, filtered based on similarity to high-quality reference and de-duplication
- **WebText2** - expanded version of GPT-2 training data, scrape of outbound links from Reddit posts with reasonably high ratings

Radford, et al. Language Models are Unsupervised Multitask Learners. 2019
Dataset

- **Common Crawl** (filtered) - general web crawl, filtered based on similarity to high-quality reference and de-duplication

- **WebText2** - expanded version of GPT-2 training data, scrape of outbound links from Reddit posts with reasonably high ratings

- **Books1 & Books2** - internet-based books
Dataset

- **Common Crawl** (filtered) - general web crawl, filtered based on similarity to high-quality reference and de-duplication

- **WebText2** - expanded version of GPT-2 training data, scrape of outbound links from Reddit posts with reasonably high ratings

- **Books1 & Books2** - internet-based books

Dataset Mix

Tokens (in billions)

- Common Crawl
- WebText2
- Books1
- Books2
- Wikipedia
Dataset Mix

Weight in training mix (percent)

Common Crawl: 40%
WebText2: 20%
Books1: 10%
Books2: 10%
Wikipedia: 0%
Dataset Mix

Epochs elapsed (for 300B tokens)

- Common Crawl
- WebText2
- Books1
- Books2
- Wikipedia
Evaluations
Let’s try it!

tdaeef = ?
Let’s try it!

Please unscramble the letters into a word and write that word.

tdaeef = ?
Let’s try it!

Zero-Shot

Please unscramble the letters into a word and write that word.

tdaeef = ?
Let’s try it!

One-Shot

Please unscramble the letters into a word and write that word.

pcirlaroc = reciprocal

tdaeef = ?
Let’s try it!

Please unscramble the letters into a word and write that word.

pcirlaroc = reciprocal
elapac = palace
tdaeef = ?
Zero-shot

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

1. Translate English to French:

2. cheese => ........................................
Zero-shot

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

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

One-shot

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

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

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

```plaintext
1. Translate English to French:
2. cheese ➞ ........................................
```

**One-shot**

In addition to the task description, the model sees a single example of

**Few-shot**

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

```plaintext
1. Translate English to French:
2. sea otter ➞ loutre de mer
3. peppermint ➞ menthe poivrée
4. plush giraffe ➞ girafe peluche
5. cheese ➞ ........................................
```
VS.

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

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

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

Learning via SGD during unsupervised pre-training

outer loop

inner loop

In-context learning

1. $5 + 8 = 13$
2. $7 + 2 = 9$
3. $1 + 0 = 1$
4. $3 + 4 = 7$
5. $5 + 9 = 14$
6. $9 + 8 = 17$

sequence #1

In-context learning

1. goat => goat
2. sake => snake
3. brid => bird
4. fsh => fish
5. dcuk => duck
6. cmhp => chimp

sequence #2

In-context learning

1. thanks => merci
2. hello => bonjour
3. mint => menthe
4. wall => mur
5. otter => loutre
6. bread => pain

sequence #3
The trophy didn’t fit in the suitcase because the trophy was too big.
The trophy didn’t fit in the suitcase because the trophy was too big.
Methods of Evaluation

Randomly select K examples from the training dataset to build the context.
Methods of Evaluation

Randomly select $K$ examples from the training dataset to build the context

Feed each context + possible completion through the model separately

Multiple-choice

Randomly select $K$ examples from the training dataset to build the context
Methods of Evaluation

Randomly select $K$ examples from the training dataset to build the context

Feed each context + possible completion through the model separately

Normalize LM likelihood over the completion and select the completion with the highest likelihood

Multiple-choice

Exact-match accuracy

Multiple-choice

Free-form

BLEU

F1
Methods of Evaluation

Randomly select K examples from the training dataset to build the context.

Multiple-choice

Feed each context + possible completion through the model separately.

Free-form

Sample from the model up to a newline.

Normalize LM likelihood over the completion and select the completion with the highest likelihood.

BLEU

F1

Exact-match accuracy

Multiple-choice

Free-form
Methods of Evaluation

Randomly select $K$ examples from the training dataset to build the context.

Feed each context + possible completion through the model separately.

Sample from the model up to a newline.

Normalize LM likelihood over the completion and select the completion with the highest likelihood.

- Multiple-choice
- Free-form
- BLEU
- F1
- Exact-match accuracy
Complete List of Tasks

**Language Modeling**
- PTB

**Close and Completion**
- ROC Stories
- HellaSwag
- LAMBADA

**Winograd-style**
- Winograd
- Winogrande

**Commonsense Reasoning**
- PiQA
- ARC
- OpenBookQA

**Reading Comprehension**
- QuAC
- SQuADv2
- DROP
- CoQA
- RACE

**Trivia-style Questions**
- NaturalQs
- WebQs
- TriviaQA

**Inference**
- ANLI
- RTE

**Comprehensive Benchmarks**
- SuperGLUE

**Translation**
- En <-> Fr
- En <-> De
- En <-> Ro

**Synthetic and Qualitative**
- Arithmetic
- Word scrambling
- Character-level manipulation
- SAT analogies
- Article generation
- Learning and using novel words
- Correcting English grammar
Summary of Performance

<table>
<thead>
<tr>
<th>Task Class</th>
<th>Few-Shot Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Close, Completion, and Language Modeling</td>
<td>Very Good</td>
</tr>
<tr>
<td>Question Answering / Knowledge Base</td>
<td>Very Good</td>
</tr>
<tr>
<td>Translation</td>
<td>Good</td>
</tr>
<tr>
<td>Winograd / Winogrande</td>
<td>Good</td>
</tr>
<tr>
<td>Commonsense Reasoning</td>
<td>Mixed</td>
</tr>
<tr>
<td>Reading Comprehension</td>
<td>Mixed</td>
</tr>
<tr>
<td>SuperGLUE</td>
<td>Mixed</td>
</tr>
<tr>
<td>NLI</td>
<td>Poor</td>
</tr>
<tr>
<td>Bias Probes</td>
<td>Poor</td>
</tr>
</tbody>
</table>
Strengths

Q: ‘Nude Descending A Staircase’ is perhaps the most famous painting by which 20th century artist?

A:

Joshi, et al. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. 2017
Strengths

Strengths
### Strengths

<table>
<thead>
<tr>
<th>Task</th>
<th>Accuracy Without Commas</th>
<th>Accuracy With Commas</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 Digit Addition</td>
<td>25.5%</td>
<td>91.1%</td>
</tr>
<tr>
<td>4 Digit Subtraction</td>
<td>26.9%</td>
<td>89.7%</td>
</tr>
<tr>
<td>5 Digit Addition</td>
<td>9.3%</td>
<td>90.2%</td>
</tr>
<tr>
<td>5 Digit Subtraction</td>
<td>9.9%</td>
<td>82.2%</td>
</tr>
<tr>
<td>6 Digit Addition</td>
<td>3%</td>
<td>78.5%</td>
</tr>
<tr>
<td>6 Digit Subtraction</td>
<td>3%</td>
<td>73.9%</td>
</tr>
</tbody>
</table>

3456 -> 3,456

Dario Amodei @ NeurIPS 12/7/20, and Gwern Branwen!
Limitations

Context → anli 3: anli 3: We shut the loophole which has American workers actually subsidizing the loss of their own job. They just passed an expansion of that loophole in the last few days: $43 billion of giveaways, including favors to the oil and gas industry and the people importing ceiling fans from China.

Question: The loophole is now gone True, False, or Neither?

Correct Answer → False
Incorrect Answer → True
Incorrect Answer → Neither

Limitations

Dua, et al. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 2019
Key Insights
Few-shot transfer to new tasks is possible without any gradient updates, and it presents a flexible framework for specifying new tasks to a model.
Bigger models can learn more from context
Bigger models have more emergent abilities
More context helps up to a point

Wang, et al. Superglue: A stickier benchmark for general-purpose language understanding systems. 2019
Performance continues to scale with compute

\[ L = 2.57 \cdot C^{-0.048} \]

<table>
<thead>
<tr>
<th>Setting</th>
<th>PTB</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOTA (Zero-Shot)</td>
<td>35.8^a</td>
</tr>
<tr>
<td>GPT-3 Zero-Shot</td>
<td>20.5</td>
</tr>
</tbody>
</table>
Lingering Questions
Lingering Questions

- Methods of Evaluation

- Training Datasets and Memorization

- Real-World Applications
Methods of Evaluation

A.I. creativity is improving fast. This hilarious GPT3-generated film is proof

By Luke Dormehl
October 20, 2020

Medical chatbot using OpenAI’s GPT-3 told a fake patient to kill themselves

AI Training Method Exceeds GPT-3 Performance with 99.9% Fewer Parameters

Facebook’s chief AI scientist says GPT-3 is ‘not very good’ as a dialog system

A new study showed some expectations for the model are unrealistic
## Methods of Evaluation

<table>
<thead>
<tr>
<th>Language Modeling</th>
<th>Reading Comprehension</th>
<th>Comprehensive Benchmarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>• PTB</td>
<td>• QuAC</td>
<td>• SuperGLUE</td>
</tr>
<tr>
<td></td>
<td>• SQuADv2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• DROP</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• CoQA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• RACE</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Trivia-style Questions</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• NaturalQs</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• WebQs</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• TriviaQA</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Inference</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• ANLI</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• RTE</td>
<td></td>
</tr>
<tr>
<td><strong>Close and Completion</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• ROC Stories</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• HellaSwag</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• LAMBADA</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Winograd-style</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Winograd</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Winogradian</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Commonsense Reasoning</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• PiQA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• ARC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• OpenBookQA</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Translation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• En &lt;-&gt; Fr</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• En &lt;-&gt; De</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• En &lt;-&gt; Ro</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Synthetic and Qualitative</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Arithmetic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Word scrambling</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Character-level manipulation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• SAT analogies</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Article generation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Learning and using novel words</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Correcting English grammar</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Methods of Evaluation

• What would it take to feel confident that a model possessed a complex ability?
Methods of Evaluation

• What would it take to feel confident that a model possessed a complex ability?

• Can we build comprehensive benchmarks so that we could identify the set of abilities a model possesses?
Methods of Evaluation

• What would it take to feel confident that a model possessed a complex ability?

• Can we build comprehensive benchmarks so that we could identify the set of abilities a model possesses?

• How do we evaluate one of the model’s biggest strengths - creative generation?
Training Datasets and Memorization

- Quality of Data
- Duplication of Benchmarks
Training Datasets and Memorization - Data Quality

CommonCrawl filtering

1. Train a classifier to distinguish between unfiltered CommonCrawl and WebText/Books/Wikipedia
CommonCrawl filtering

1. Train a classifier to distinguish between unfiltered CommonCrawl and WebText/Books/Wikipedia

2. Sample filtered CommonCrawl with higher probability of selection based on classifier score of quality
How can we better define and identify high quality data?
Training Datasets and Memorization - Harmful Data

- Gender

  “The detective was a ______” —> 83% male

  “The competent detective was a ______”

  “The incompetent detective was a ______”
Training Datasets and Memorization - Harmful Data

- Gender

**Male-biased Descriptive Words**
- Large
- Mostly
- Lazy
- Fantastic
- Eccentric
- Protect
- Jolly
- Stable
- Personable
- Survive

**Female-biased Descriptive Words**
- Optimistic
- Bubbly
- Naughty
- Easy-going
- Petite
- Tight
- Pregnant
- Gorgeous
- Sucked
- Beautiful
Training Datasets and Memorization - Harmful Data

- Gender
- Race
Training Datasets and Memorization - Harmful Data

- Gender

- Race

- Religion

<table>
<thead>
<tr>
<th>Religion</th>
<th>Most Favored Descriptive Words</th>
</tr>
</thead>
</table>
How do we make sure models trained on huge amounts of web data don’t get the chance to memorize eval benchmarks?
Removing benchmarks from training data

1. Look for overlap in phrases between benchmarks and training documents
Removing benchmarks from training data

1. Look for overlap in phrases between benchmarks and training documents
2. Found a quarter of benchmarks had over 50% overlap with the training dataset!
Training Datasets and Memorization - Eval Memorization

Removing benchmarks from training data

1. Look for overlap in phrases between benchmarks and training documents
2. Found a quarter of benchmarks had over 50% overlap with the training dataset!
3. Remove training documents that overlap with eval benchmarks
Removing benchmarks from training data

1. Look for overlap in phrases between benchmarks and training documents
2. Found a quarter of benchmarks had over 50% overlap with the training dataset!
3. Remove training documents that overlap with eval benchmarks
Training Datasets and Memorization - Eval Memorization

Removing benchmarks from training data

1. Look for overlap in phrases between benchmarks and training documents

2. Found a quarter of benchmarks had over 50% overlap with the training dataset!

3. Remove training documents that overlap with eval benchmarks

4. Compare performance on benchmarks between full dataset and only test examples that don't appear in the training data
Training Datasets and Memorization - Eval Memorization

The diagram shows the percent change in performance (Accuracy, F1, or BLEU) on various datasets as the percentage of data clean in the dataset is varied. The datasets include QuAC, Symbol Insertion, SQuADv2, Winograd, PIQA, WMT16 en-de, WMT16 de-en, Anagrams 1, Anagrams 2, and Reversed Words. The x-axis represents the percentage of data clean in the dataset, while the y-axis shows the percent change in performance.

Key observations:
- Data cleaning improves performance on datasets with a high percentage of clean data.
- The performance on datasets with mostly dirty data may not change significantly.

Legend:
- Green arrow: Eval on only clean data did better.
- Red arrow: Eval on all data (including dirty) did better.
Real-World Applications

Important considerations

1. Potential for harmful outputs
2. Reliability of performance
Real-World Applications

- Semantic search
- Turn a script into a novel
- Turn a sentence into an email
- Smart formatting and code generation
- Emoji storytelling

https://andrewmayneblog.wordpress.com
Real-World Applications - Emoji Storytelling

Back to Future: 😳 😳 🚗⏰

Batman: 🪤 🏅

Transformers: 🚗 🚄
Real-World Applications - Emoji Storytelling

- Back to Future: 😳分化
- Batman: 🦇🦇
- Transformers: 🚗🤖
- Zootopia: 🦊🌹🐻作风建设
- Wonder Woman: 🌈👩‍✈️❤️netinet
- The Godfather: 😡😡 urlpatterns
- Star Trek: 🌟🚀
- Planet of the Apes: 🐵🐵
- Game of Thrones: 🏳️‍🌈🐱🦊
- Jurassic Park: 😜🌟猇
- Castlevania: 👻🧟‍♂️🕷️🔧🎈?
- The Matrix: 😁😊
- Iron Man: 🦿💪🌟⬛🌲
- Death Note: 📚🧟‍♂️👉✈️
- Frozen: 🎈👩‍❄️❤️🌲
- The Hunger Games: 🌽❤️❤️❤️

https://andrewmayneblog.wordpress.com
Real-World Applications

- What are the useful applications of a model like GPT-3?

- Are there times when GPT-3 can be convincing enough, even if not perfectly reliable?
Real-World Applications

• What are the useful applications of a model like GPT-3?

• Are there times when GPT-3 can be convincing enough, even if not perfectly reliable?
Real-World Applications - Writing News

Human ability to detect model generated news articles

control (86%)

random chance (50%)

Number of parameters (log scale)

Accuracy (%)
Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination

Article: After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church’s annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.

The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church’s history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.
Conclusion

- Language modeling performance appears to continue to scale with compute
Conclusion

- Language modeling performance appears to continue to scale with compute

- Large models can transfer few-shot to new tasks without any fine-tuning
Conclusion

• Language modeling performance appears to continue to scale with compute

• Large models can transfer few-shot to new tasks without any fine-tuning

• There are many complexities to evaluations, training datasets, and applications for large models
## Language Models are Few-Shot Learners

<table>
<thead>
<tr>
<th>Tom B. Brown*</th>
<th>Benjamin Mann*</th>
<th>Nick Ryder*</th>
<th>Melanie Subbiah*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jared Kaplan†</td>
<td>Prafulla Dhariwal</td>
<td>Arvind Neelakantan</td>
<td>Pranav Shyam</td>
</tr>
<tr>
<td>Amanda Askell</td>
<td>Sandhini Agarwal</td>
<td>Ariel Herbert-Voss</td>
<td>Gretchen Krueger</td>
</tr>
<tr>
<td>Rewon Child</td>
<td>Aditya Ramesh</td>
<td>Daniel M. Ziegler</td>
<td>Jeffrey Wu</td>
</tr>
<tr>
<td>Christopher Hesse</td>
<td>Mark Chen</td>
<td>Eric Sigler</td>
<td>Mateusz Litwin</td>
</tr>
<tr>
<td>Benjamin Chess</td>
<td>Jack Clark</td>
<td>Christopher Berner</td>
<td></td>
</tr>
<tr>
<td>Sam McCandlish</td>
<td>Alec Radford</td>
<td>Ilya Sutskever</td>
<td>Dario Amodei</td>
</tr>
</tbody>
</table>

OpenAI
Questions?

Shadows on the Way
Wallace Stevens

There is a way in the world.
That way a man may live in a small town
His lifetime and see no man come,
Never the stir of men about a market,
Never the bobbing of covered carts,
Never the creak passed on the stair
Closing some other life; an old man
Can prosper in that town, can prosper
When his store is seen to close, his wife
Fades in his house, and the world shouts in vain
That no one comes to hear him
Where his small bell tolls.
In that small town
Everything is known.
The ways of the young men are known,
And the ways of the young women
As they walk to the Spring;
The rumors are all known,
The quarrels, the hurt, the envy
Of the men and the women.
Till a day comes
That way out of the world
That leads to a wood.
Young, the young man
Walks, a lover now,
Knows the streets as they are and is of them.